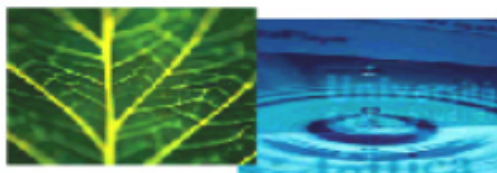


PhD Dissertation



**International Doctorate School in Information and
Communication Technologies**

DISI - University of Trento

**A LEXI-ONTOLOGICAL RESOURCE FOR
CONSUMER HEALTHCARE
The Italian Consumer Medical Vocabulary**

Elena Cardillo

Advisor:

Dr. Luciano Serafini

Fondazione Bruno Kessler

Co-Advisor:

Dr. Stefano Forti

Fondazione Bruno Kessler

April 2011

To my family, and all the people who trusted in me during my Ph.D

Abstract

In the era of Consumer Health Informatics, healthcare consumers and patients play an active role because they increasingly explore health related information sources on their own, and they become more responsible for their personal healthcare, trying to find information on the web, consulting decision-support healthcare systems, trying to interpret clinical notes or test results provided by their physician, or filling in parts of their own Personal Health Record (PHR). In spite of the advances in Healthcare Informatics for answering consumer needs, it is still difficult for laypersons who do not have a good level of healthcare literacy to find, understand, and act on health information, due to the communication gap which still persists between consumer and professional language (in terms of lexicon, semantics, and explanation). Significant effort has been devoted to promote access to and the integration of medical information, and many standard terminologies have been developed for this aim, some of which have been formalized into ontologies. Many of these terminological resources are used in healthcare information systems, but one of the most important problems is that these types of terminologies have been developed according to the physicians' perspective, and thus cannot provide sufficient support when integrated into consumer-oriented applications, such as Electronic Health Records, Personal Health Records, etc. This highlights the need for intermediate consumer-understandable terminologies or ontologies being integrated with more technical ones in order to support communication between patient-applications and those designed for experts. The aim of this thesis is

to develop a lexical-ontological resource for consumer-oriented healthcare applications which is based on the construction of a Consumer-oriented Medical Vocabulary for Italian, able to reflect the different ways consumers and patients express and think about health topics, helping to bridge the vocabulary gap, integrated with standard medical terminologies/ontologies used by professionals in the general practice for representing the process of care, by means of Semantic Web technologies, in order to have a coherent semantic medical resource useful both for professionals and for consumers. The feasibility of this consumer-oriented resource and of the Integration Framework has been tested by its application to an Italian Personal Health Record in order to help consumers and patients in the process of querying healthcare information, and easily describe their problems, complaints and clinical history.

Keywords

Consumer-oriented Healthcare Vocabularies, Medical Terminology Integration, Knowledge Engineering, Semantic Web Technologies.

Acknowledgements

I want to start this long list of acknowledgement saying endlessly thank you to my scientific advisor and inspirer Luciano Serafini. For introducing me in the world of research, for motivating me every step of our long activity together, for your scientific guidelines, suggestions for improving my thesis, for your support and for teaching me that research is not only studying but also relationships with people, exposing and expressing our works, thoughts and believes without worries, and having funny, to you, Luciano, "Grazie di cuore per tutto", hoping that we can keep on collaborating together in FBK for research activities involving Knowledge Organization and Representation...I think about you as a point of reference. I thank members of the thesis defence committee, Giovanni Adamo, Annette ten Tejie, Robert Vander Stichele, for their the time dedicated in reading my thesis, for their availability, for the important contributions provided during the thesis' defence and for the really accurate and detailed review of the thesis, proving me constructive feedbacks. In particular, I would like to thank you Olivier Bodenreider for his hospitality and our productive collaboration during my visit to the National Library of Medicine (NLM - NHI, Bethesda) in spring 2010, where he supported me as my supervisor in the activity of integrating consumer-oriented terminologies with professional UMLS vocabularies using Semantic Web Technologies. A deep thank you goes to Stefano Forti, my second advisor, who allowed me to be part also of the e-Health research unit in FBK funding my Ph.D during these three years work, and also to include my research into the TreC project, and to integrate my Italian Consumer-oriented

Medical Vocabulary with the Italian PHR, still for his availability, patience, encouragement and personal support, and finally for offering me to keep on working with him and his staff in FBK. I gratefully acknowledge all of my friends and colleagues in Trento for the opened and friendly environment in which it was really a pleasure to work and live in. In particular I would mention Andrei, for his support, both technical and personal one, for all the interesting activities we carried out together, breaks taking about everything, his suggestions and encouragement and for helping me in the practical tasks of my thesis. The same deep thank you goes to Elena, my Ph.D sister :-), Rachele, Manuela and Charles, not only colleagues but friends, for their constant support, even in this case, scientific and personal...they allowed me to feel at home even in a place so far from my native Town, my habits and my family. Thank you for being truly close and helpful and for the funny moments together. A smiley thank to my e-Health colleagues (Flavio, Emiliano, Roberta, Marco, Francesco, Barbara, Luca, Claudio ed Enrico, for being so helpful, for their patience and for the really friendly environment they provide everyday to me. In the same way I would thank my friends and colleagues of the university of Calabria, Maria, Ivana, Teresa, and Antonella, who keeping on collaborating with me during this Ph.D and, even if physically far from me, truly close and helpful every time I needed, and a special thank goes to the prof. Roberto Guarasci, my Master thesis' advisor and my constant scientific guide, who convinced me in applying for this Ph.D. And finally, the most of my gratitude goes to God and to my family, my father, my sisters and my mother from the Sky, whose love and support invaluablely helped me in completing this thesis work and in living far from their embraces.

Contents

I	PRELIMINARIES	1
1	Introduction	3
1.1	The Context	3
1.2	The Problem	5
1.3	The Solution	6
1.4	Outline	7
2	State of the Art	9
2.1	Medical Terminologies	10
2.2	Consumer-oriented Vocabularies	14
2.3	Semantic Web and Ontologies	18
2.3.1	Ontologies	18
2.3.2	Semantic Web Technologies	21
2.3.3	Publishing data on Semantic Web	23
2.4	The Semantic Web for Healthcare	25
2.4.1	Healthcare Ontologies	25
2.4.2	Semantic-based Integration	27
2.4.3	Publishing Data for Healthcare	28
2.4.4	Applications in Healthcare	30
3	The Problem Statement and the Approach	33
3.1	The Linguistic Gap	33

3.2	The Approach: an overview	37
II	DEVELOPMENT	43
4	The Italian Consumer Medical Vocabulary	45
4.1	Theoretical background	46
4.1.1	Terminology and Lexicology	46
4.1.2	Generating Vocabularies	49
4.1.3	Knowledge Acquisition	51
4.2	Knowledge Acquisition Process	53
4.2.1	Wiki based Acquisition	55
4.2.2	Nurse-assisted Acquisition	58
4.2.3	Focus Group Acquisition	61
4.2.4	Web-based medical consultations	62
4.3	Term Extraction	65
4.4	Clinical Review	74
4.5	Familiarity Degree	76
4.6	Finding ICMV correspondences in ICPC2	77
5	Formalizing Medical Terminologies in Semantic Web Languages	85
5.1	Encoding resources in OWL	86
5.1.1	ICPC2 and ICD10 Background	87
5.1.2	ICPC2, ICD10 and their clinical mapping in OWL	93
5.2	RDF Encoding	106
5.2.1	Materials	108
5.2.2	RDF N-Triples Encoding	114
5.2.3	Encoding ICMV in RDF Schema	117
5.3	Results and Evaluation	119

6	Integrating ICMV with Standard Medical Terminologies	123
6.1	Mapping to ICPC2	124
6.2	Mapping to UMLS	126
6.2.1	The UMLS Enrichment Process	127
6.2.2	Querying the repository	128
6.2.3	Quality Assurance of the Mappings	135
6.2.4	Results and Evaluation	136
6.3	Manual vs. Automatic Mapping	146
III	APPLICATIONS	147
7	Experimental use of ICMV	149
7.1	Applying the ICMV	150
7.2	ICMV for PHRs	155
7.2.1	TreC - Cartella Clinica del Cittadino	156
7.2.2	Integrating ICMV in TreC	161
7.2.3	Results and Services for TreC	165
8	Conclusion	171
8.1	Summary and Contributions	171
8.1.1	Comparison with Similar Approaches	179
8.1.2	Current and Potential Impacts	181
8.1.3	Future Work	183
	Bibliography	185

List of Tables

2.1	Medical Terminologies and Classification Systems	11
2.2	Mapping example between ICPC2 and ICD10	13
2.3	Disease representation as ontological facts in SNOMED CT . . .	26
4.1	Demographic characteristics of consumers	54
4.2	Corpus Summary	65
4.3	List of the 20 most frequent terms extracted with T2K from the three datasets	68
4.4	Fragments of taxonomical chains	69
4.5	Clusters of related terms	70
4.6	Summary of T2K Term Extraction on the three datasets	70
4.7	Example of KX output list of keyphrases	73
4.8	Wiki term collection	79
4.9	Nurse-assisted term collection	80
4.10	Focus Group/Games with Elderly Person	80
4.11	Mappings between “Medicitalia” lay terms and ICPC2 concepts	81
4.12	Results overview	82
5.1	The ICD10 chapter headings	90
5.2	The structure of ICPC	92
5.3	One-to-one mappings between ICPC and ICD10	93
5.4	Breaking open of ICD10 rubrics and mapping to one ICPC rubric	94
5.5	Mapping of a group of ICD10 classes to one ICPC rubric	94

5.6	ICPC2 Chapters names in ICPC2 Ontology	98
5.7	Parts description in LOINC for “Pain in back”	114
5.8	ICMV properties in RDF	119
5.9	Metrics for ICPC2, ICD10 and Mapping ontologies	120
6.1	Example of manual mapping between ICMV and ICPC2	125
6.2	Query result for the ICPC2 concept Headache	132
6.3	Statistical summary of the mappings among resources	141
6.4	Example of mappings between ICMV terms and UMLS Italian concepts	144
6.5	Comparing the mapping approaches	146
7.1	ICMV Wiki content statistics	154
7.2	ICMV Wiki most visited pages	154
7.3	Statistics of terminological tables integrated into TreC widgets .	166

List of Figures

2.1	The Semantic Web Stack	21
3.1	Typical scenario for the use of a Consumer-oriented health vocabulary	36
3.2	Approach Overview	39
4.1	Wiki page example created by users to express the symptom Absence of voice	56
4.2	Nurse-based acquisition table extracted from the Triage records summary	60
4.3	XML view of a medical consultation about “Arthro-RMI” in Medicitalia	63
4.4	Sample marked-up document (Medicitalia-C4d2) from two an- notators using CLarK	64
4.5	T2K at work	66
5.1	Example of taxonomy in ICD10 for the concept “Neoplasm of pancreas”.	91
5.2	OWL encoding of the ICPC2 class A71 “Measles” in Protégé . .	100
5.3	OWL encoding of ICD10 class C25.0 “Head of pancreas” in Protégé	103
5.4	Extract of ICPC2-ICD10 mapping ontology in Protégé	105
5.5	RDF Triple format	107

5.6	Example of the use of UMLS identifiers for the concept Atrial Fibrillation	111
5.7	MeSH record for the concept Heartburn - D006356	113
5.8	SNOMED CT N-triples for the concept 25064002 - Headache	115
5.9	MeSH N-triples for Descriptor D006261 - Headache	115
5.10	ICD10 N-triples for the concept N01 - Headache	116
5.11	MeSH N-triples for the concept R51 - Headache	116
5.12	LOINC N-triples for the concept 55466-7, Influenza virus A and and for the part “Headache” LP74908-2	117
5.13	Excel view of ICMV	118
5.14	Extract of the RDF/S triples for the ICMV lay term “Cuore in gola”	119
6.1	SNOMEDCT N-triples enriched with UMLS CUI and LUIs corresponding to “Headache”	128
6.2	Example of Transitive Closure constructs output for some ICPC2 concepts	134
6.3	ICPC2 mappings to the other resources	137
6.4	Overlap among resources	138
6.5	Examples of multiple mappings between ICPC2 and other terminologies	139
6.6	Exemple of extended synonyms for ICPC2 U04 <i>Incontinence Urine</i> by mapping to SNOMED CT	140
6.7	Example of shared parent relation between ICPC2 and SNOMED CT via UMLS Parent CUI	142
6.8	Mapping to UMLS Italian Resources	143
7.1	The ICMV Wiki main page	151
7.2	ICMV categories shown in alphabetical order in the wiki	152
7.3	Representation of the term “Ematoma” in the ICMV Wiki	153

7.4	TreC Architecture	158
7.5	TreC widgets in the user interface	161
7.6	Searching and adding Allergies or Intolerances in TreC	167
7.7	Searching and adding Vaccinations in TreC	168
7.8	Searching and adding Reasons for Encounter in TreC	169
7.9	Searching and adding Diseases in TreC	170

Part I

PRELIMINARIES

Chapter 1

Introduction

1.1 The Context

In the era of Consumer Health Informatics, “healthcare consumers”¹ and patients play a more proactive role than in the past as they increasingly explore health-related information sources on their own, and they become more responsible for their personal healthcare. In fact, there has been an increase in searching information on the web, consulting decision-support healthcare information systems, and using patient-oriented healthcare systems which allow consumers to directly access clinical notes or test results and to fill in their Personal Health Record. Furthermore, significant effort has been devoted to promoting access to and the integration of medical information, and many standard terminologies have been developed for this goal, some of which, thanks to the advent of the Semantic Web, have been formalized into ontologies. In this context, the terminological aspect is of great importance together with that concerning Medical Information Integration. This importance can be understood by taking a look at the huge quantity of medical lemmas hosted in hundreds of vocabularies. These are “technical terms” included in Medical Classification Systems or terminolo-

¹Consumers are members of the public who seek information for themselves about medical and health issues to support decision making in healthcare. This group is known as healthcare consumers because their behaviour and needs are similar to those of people trying to learn about “traditional” products and services.

gies mainly used by physicians, and adopted by healthcare information systems to allow physicians to encode symptoms, diagnoses and diseases during their clinical activity. However, we can recognize that a linguistic and semantic discrepancy still exists between specialized medical terminology used by healthcare professionals, and the “lay” medical terminology used by patients and consumers in general. A lay medical terminology, often called “Consumer Health Vocabulary” (CHV) in the Healthcare literature, is a collection of forms used in health-oriented communication for a particular task or need by a substantial percentage of consumers from a specific discourse group, which includes for each form a correspondence to professional concepts. This type of vocabulary can have three possible bridging roles between consumers and health applications or information systems:

- Information Retrieval, because CHV would facilitate automated mapping of consumer-entered queries to technical terms, producing better search results;
- Medical Records, since medical records and test results are becoming available to patients, they frequently contain jargon, so a CHV could supplement these terms with consumer-understandable names to help patients interpret these terms;
- Health Care Applications, where patients may enter consumer expressions such as “sudden hair loss” or “shortness of breath”, receiving help via the integrated CHV, which would facilitate automated mapping of these expressions to technical concepts (in this case “alopecia” and “dyspnea”) enabling subsequent analysis and response.

1.2 The Problem

As mentioned above, in spite of the advances in healthcare informatics for answering consumer needs, it is still difficult for laypersons who do not have a good level of healthcare literacy to interact with health information systems due to the linguistic gap which still exists between consumer and professional language (in terms of lexicon, semantics, and explanation). This gap consists in the fact that non-professionals (above all lay persons) are often unable to access or comprehend information in specialized domains such as the medical one because of technical terminology. This discrepancy has become even more evident ever since consumers have begun to actively access and manage healthcare information on the web. In fact, during the interaction with healthcare applications on the web, without the intermediation of physicians, consumers can rely only on their knowledge and experience, and this can often generate a wrong inference of the meaning of a term, or the mis-association of a term with its context. The problem is emphasized by the fact that, existing terminologies and classification systems have been developed according to a physicians' perspective, so they cannot provide sufficient support when integrated into consumer-oriented applications, such as Personal Health Records, etc. This highlights the necessity of integrating consumer-understandable terminologies or ontologies with more technical ones in order to support communication between applications designed for patients and those designed for experts. Moreover, medical knowledge integration in healthcare information systems is facilitated by the use of Semantic Web technologies (e.g. formal representation of medical terminologies, medical ontology mappings, and healthcare ontology repositories), helping consumers during their access to healthcare information and improving the exchange of their personal clinical data. But even if significant effort has been devoted to the creation of these medical resources, used above all to help physicians in filling out Electronic Health Records, there is little work based on

the use of lay (consumer-oriented) medical terminology, and in addition most existing studies have been done only for English, so there is no support for Italian or a multilingual perspective.

1.3 Thesis Objective and Approach

To bridge the linguistic gap, the present thesis proposes the construction of a consumer-oriented medical vocabulary for Italian integrated with standard medical terminologies/ontologies through Semantic Web technologies, in order to have a coherent semantic medical resource useful both for professionals and for consumers. This resource could be used in healthcare systems, such as Personal Health Records, to help consumers during the process of querying and accessing healthcare information, so as to bridge the communication gap. In the present research activity the problem spans over the following dimensions: 1. Terminology Acquisition and Vocabulary Generation, which have been addressed by the creation of a Consumer-oriented Medical Vocabulary for Italian (ICMV), containing common medical expressions and terms used by Italians (principally older people, hospitalized patients, chronic patients, etc.); 2. A formal representation of ICMV and standard medical terminologies using Semantic Web technologies; 3. Creation of an Integration Framework for medical information, where all the formalized resources taken into account are semantically integrated with the help of formal mappings; and 4. Evaluation of the feasibility of this medical resource through its application to a new Personal Health Record, developed at the Fondazione Bruno Kessler for the Province of Trento.

The present work has three important innovative aspects. The first is the creation of the ICMV, a consumer-oriented vocabulary for the healthcare domain not yet present in the Italian context. A second important aspect is the use of Semantic Web technologies, with the consequent advantages for integrating the

ICMV with the specialized medical terminologies used by healthcare providers, in a context of Italian Personal Health Record which requires the exchange of data between patients and physicians. The final innovative aspect is the Cross-Domain characterization of this thesis, i.e. the use of a hybrid methodology which associates Semantic Web technologies with Healthcare Informatics technologies on the one hand, and Linguistics and Terminology issues on the other. We expect that treating linguistic and terminological aspects together with the new technologies in Healthcare Informatics will improve consumer-oriented applications.

1.4 Outline

Besides the current introduction to the thesis work (Chapter 1), the remainder of the thesis is organized in the following four main parts:

Part I: Preliminaries :

- Chapter 2 gives an overview of the state of the art in the field of Medical Terminologies and Classification Systems, Consumer Health Vocabularies and in Semantic Web technologies for Healthcare;
- Chapter 3 studies in more detail the problem statement and limitations of the State of the Art and provides an overview of the approach;

Part II: Development :

- Chapter 4 describes in detail the first step of our approach, that is the creation of the Italian Consumer-oriented Medical Vocabulary (ICMV), focusing on the three principal steps: Medical Terminology/Knowledge acquisition; Information Extraction, Normalization and Candidate Terms Detection; and finally a clinical review where lay candidate terms are manually mapped to a standard medical terminology (ICPC2);

- Chapter 5 is about the formalization of medical terminologies, in particular ICPC2, ICD10, using the Semantic Web languages OWL, and other terminologies from UMLS Metathesaurus and our Italian Consumer-oriented Medical Vocabulary in RDF, for improving their integration;
- Chapter 6 describes the integration process where the ICMV has been mapped to the standard medical terminologies under investigation. Here we compared the manual mapping performed by healthcare professionals to the automatic one performed using semantic web technologies;

Part III: Applications :

- Chapter 7 reports an experimental use case for the application of our integration framework and in particular the ICMV;

Chapter 8 concludes the thesis with a summary of the work presented in the thesis, its comparisons to closest related approaches, and outlines some future directions.

Chapter 2

State of the Art

The chapter overviews the background material of the thesis starting from the studies and improvements on the topic of Medical Terminology, than showing the existing frameworks for creating consumer-oriented medical vocabularies, Ontologies and finally Semantic web application in the Healthcare domain, recalling in particular the problem of formalization and integration of medical terminologies. The selected presentation scheme of frameworks highlights critical questions in using standard medical terminologies, the need for new lay terminologies to be integrated with these medical terminologies and above all the advantages in using Semantic Web technologies and languages to overcome the interoperability problem and to provide reasoning and knowledge services on semantic-based healthcare information systems which use formalized and integrate medical terminologies.

The chapter is organized as follows: Section 2.1 gives an overview of the most used medical terminologies and classification systems, explaining also the needs for their clinical mappings and highlighting the problems of their integration and of their use in consumer-oriented healthcare systems. Section 2.2 overviews the few frameworks present in literature which address the linguistic gap between specialized and lay medical terminologies with the creation of consumer-oriented medical vocabularies. Section 2.3 provides some theoretical

background about ontologies and Semantic Web technologies and finally Section 2.4 describes the recent efforts in their application to the healthcare domain.

2.1 Medical Terminologies and Medical Classification Systems

Over the last two decades research on Medical Terminologies and Classification Systems has become a popular topic and much work has been undertaken to develop systems which could be efficiently used by physicians during their patient's health care visits. These standardization efforts have established a number of classification systems as well as conversion mappings between them. Medical Classification Systems are not the same of Medical terminologies, in fact this second concept refers to "pre-established hierarchies of terms used to constrain selections made by users in annotating large document corpora" (Ceusters *et. al*, 2003 [31]). According to (Rector, 1999 [96]), they "concern the meaning, expression, and use of concepts in statements in the medical record or other clinical information system". Medical Classification Systems, on the other hand, provide an essential instrument for unambiguously labelling clinical concepts in processes and services in healthcare (communication of diseases and their causes, statistical analysis for epidemiological studies, etc.) and for improving the accessibility and elaboration of medical content in clinical information systems. Even though medical classifications and terminologies are distinct in their definition, or more properly, one is included within the other, in Healthcare Informatics they are used synonymously (more details on the definitions and use of clinical classifications and terminologies can be found in Chute, 2000 [34]).

Existing medical terminologies vary in their coverage and completeness, and are differentiated on the basis of their purpose: diagnostic, procedural, pharmaceutical, and topographical. Systems designed to cover clinical information

have tended to cover a relatively narrow subset of healthcare, such as nursing procedures, problem lists, etc. Furthermore, some systems that concentrate on coding fine-grained primary clinical data have been proprietary, custom-built, limited, or difficult for clinicians to use.

In recent years significant effort has been spent by the United States National Library of Medicine in its UMLS Metathesaurus project¹, and the older MeSH²; by the UK National Health Service and its Center for Coding and Classification, by SNOMED CT³ and the GALEN Project⁴ of the European Community. Also the WHO (World Health Organization) and WONCA (The World Organization of Family Doctors) developed their representative standard classification systems, respectively the International Classification of Diseases(so far at his 10th revision)⁵ and the International Classification of Primary Care(at his second edition)⁶. Terminology standards have also been a major effort for CEN/TC251⁷ and ISO/TC215 [3]. Table 2.1 summarizes the most used terminologies and classification systems, along with their purposes:

Name	Organization	Goal
ICPC2	WONCA	Codify primary care reason for encounters
ICD10	WHO	Codify diagnoses for recording morbidity and mortality statistics
MeSH Thesaurus	US-NLM	Indexing articles
UMLS	US-NLM	Retrieve and integrate relevant health info
SNOMED	IHTSDO	Index the entire medical vocabulary
LOINC	REGENSTRIEF INST.	Provide universal codes for laboratory and other clinical observations

Table 2.1: Medical Terminologies and Classification Systems

The dream of a unique Terminology.

In this scenario one could think that the ideal would be the creation of a sin-

¹<http://www.nlm.nih.gov/research/umls/>

²<http://www.ncbi.nlm.nih.gov/mesh>

³<http://www.ihtsdo.org/>

⁴<http://www.opengalen.org/>

⁵<http://www.who.int/classifications/icd/en/>

⁶<http://www.globalfamilydoctor.com/wicc/sensi.html>

⁷<https://login.cen.eu/sso/Authn/UserPassword>

gle unified medical terminology, which would improve the automated flow of clinical information. However, such a terminology does not exist and, even if modern medicine had a firm scientific basis on which to facilitate terminological standardization, other components, such as traditional nomenclatures, some structures of physical and natural sciences specialities and sub-domains, and areas of uncertainty or gaps of knowledge, preclude this dream. It is not clear that having a unique terminology is the best option as multiple terminologies allow for a more flexible architecture and organization. They, in fact, are created and tailored to serve different needs and tasks requiring possibly conflicting levels of granularity or perspectives, from mortality and morbidity statistics to billing to electronic patient records, and in addition, as (Cabr e, 1999 [23]) pointed out “terminology is not an end in itself, but addresses social needs and attempts to optimize communication among specialists and professionals...”.

Need for mapping medical terminologies.

In this context, the need to establish unambiguous mappings between different coding systems to guarantee their interoperability was raised. The purpose of mappings medical terminologies is to provide a link between one terminology and another in order to: (i) use data which are collected for a specific aim also for other purposes; (ii) retain the value of data when migrating to newer database formats and schemas; and (iii) avoid entering data multiple times with the risk of driving up cost and errors. Mappings between medical terminologies or classification systems are defined for a given purpose, representing the agreement reached between medical specialists expressed in the form of correspondence tables, as shown in Table 2.2, and they are created with a specific purpose in mind, and so they must be refined for particular use cases and users in diverse settings (Foley, 2006 [44]):

In the presence of all these medical terminologies, even providing clinical mappings between them, interoperability is still a significant problem, because

Concept in ICPC2	Concept in ICD10
A71 (Measles)	B05 (Measles)
N01 (Headache)	R51 (Headache) G44.3 (Chronic post-traumatic headache) G44.8 (Other specified headache syndromes)

Table 2.2: Mapping example between ICPC2 and ICD10

content, structure, completeness, level of detail, cross-mapping, taxonomy, and definitions vary between existing vocabularies. Many established medical coding systems lack a precise semantic underpinning (the recent emergence of description logic encoded medical terminologies - particularly SNOMED CT - aims to address this problem). Furthermore, there is no clear formal reference establishing the precise meaning of mappings between the coding systems, and the interpretation of mappings cannot be done coherently — different groups of mappings can hold different implicit semantics. This seems obvious since each terminology or classification system is created for a precise purpose, and, as asserted above, mapping has to be based on a clinical consent. Hence, a general standard for this is hard to achieve.

A number of researchers have spent a good deal of effort on computerized terminology mapping, also trying to face the more difficult task of formalization of terminologies and their mappings. Despite some critical considerations on the use of new techniques for creating, formalizing and integrating medical terminologies (See Ceusters *et al.*, 2003 [31]), they have been found practically useful in different application scenarios. In fact, the medical community needs medical terminologies and widely uses them. In terminology research there has been a general trend to make the definitions of medical concepts more explicit, both to disambiguate concepts and to facilitate automated terminology merging efforts.

On the one hand, a large number of the proposed algorithms and heuristics for discovering mapping between medical classifications are based on the

extensive use of the Metathesaurus UMLS as a Knowledge Resource for semantic mapping of concepts belonging to different classifications and terminologies. The notable works in this direction are the ones proposed by (Fung and Bodenreider, 2005 [45], Elink and Brown, 2002 [41]) on mapping between SNOMED RT (Systematized Nomenclature of Medicine-Reference Terminology) and ICD9-CM (International Classification of Diseases version 9 - Clinical Modifications); (Wang *et al.*, 2006 [120]) on mapping ICPC2-Plus to SNOMED CT, among others.

On the other hand, a number of formally grounded mapping approaches have been proposed. For example, Dolin *et al.*, 1998 [39] applied description logics to a semantically-based mapping, evaluating a LALR (lexically assign, logically refine) strategy for merging overlapping healthcare terminologies. De Keizer and Abu-Hanna, 2000 [38] provided a representation formalism based on Entity Relationship Diagrams (ERD) and First Order Logic (FOL), as part of a framework for representing the structure of 5 terminological systems: ICD, SNOMED, NHS clinical terms, UMLS, and GALEN. Finally Lomax and McCray, 2004 [73] described their experiences with mapping the Gene Ontology (GO) to UMLS.

2.2 Consumer-oriented Medical Vocabularies

The problem of having many terminologies continues to plague health professionals and their information systems. Moreover, all the existing medical terminologies and classification systems are designed by and for professionals. This lack of lay language vocabulary continues to hinder consumers and in particular laypersons, who are the most damaged by the increased communication gap. Since consumers have information needs to support personal healthcare decision-making - apart from mediators and professionals in their professional capacities - they need a controlled medical terminology or vocabulary to bridge

this communication gap.

Consumers generally use a subset of the vocabulary from Language for General Purposes (LGP) to describe medical concepts, including both Language for Specialized Purposes (LSP) terms that have crossed-over to the mainstream and folk classification. However, LGP terms borrowed from LSP are often less precise, leading to possible misunderstanding. To respond to consumer needs for supporting them in medical language understanding and in personal healthcare decision-making, over the last several years, many researchers have worked at the creation of lexical resources that reflect the way consumers/patients express and think about health topics. One of the largest initiatives in this direction is the Consumer Health Vocabulary Initiative⁸, promoted by Q. Zeng and her colleagues at Brigham and Women’s Hospital, Harvard Medical School, who developed the Open Access Collaborative Consumer Health Vocabulary (OAC CHV) for English. It includes lay medical terms and synonyms connected to their corresponding technical concepts in the UMLS Metathesaurus, and can be used for extending research on informatics-based tools to help consumers in health information seeking, retrieval, and understanding (for instance, “nose - bleed” mapped to “epistaxis”, “heart attack” mapped to “myocardial infarction”, etc.) (Zeng *et al.*, 2007a [124]). Two years ago the OAC CHV was submitted to the National Library of Medicine for incorporation into the UMLS. In developing this kind of vocabulary Zeng *et al.* combined corpus-based text analysis with a human review approach, including the identification of consumer forms, that means compound terms and expressions used by consumers and laypersons for “standard” health-related concepts. Also Tse and Soergel, 2004 [109] tried to create a Consumer Medical Vocabulary identifying consumer medical terms and expressions used by lay people and health mediators, associating a Mediator Medical Vocabulary with this vocabulary, and mapped them to a Professional Medical Vocabulary, mainly stressing the mapping pro-

⁸<http://www.consumerhealthvocab.org>

cedure. As a matter of fact, they provided an additional mapping of reviewed terms to UMLS to find synonyms and quasi-synonyms, and extended this functionality with the creation of a medical Interpretative Layer, to mediate between lay and professional perspectives, at all levels. All these studies examined large numbers of consumer utterances (i.e., hundreds of thousands of tokens) and consistently found that between 20% and 50% of consumer health expressions were not represented by professional health vocabularies. A further subset of these unrepresented expressions (i.e., hundreds to thousands) underwent human review to be acquired and validated. An overview of all these studies can be found in Keselman *et al.*, 2008 [68].

During the last three years some of these researchers started to collaborate with the CHVs Initiative trying to develop a unique framework which could answer the multiple consumer needs not yet resolved, and a common term identification method for the creation of a CHV. However, many obstacles have to be overcome to achieve such an aim: rapidity in changing uses and variability in general language expressions, both culturally and temporally; variable length of lay health expressions; and reliance on local and personal context for meaning, contrary to terminology where, ideally, terms are unambiguous in meaning within a domain. Furthermore, a subset of these unrepresented expressions underwent human review. In most of these cases automatic term extraction was performed from written texts, such as healthcare consumer queries on medical web sites, postings and medical publications. Another important observation is that there is no real proof of application in healthcare informative systems of the lexical resources that came out of the CHVs Initiative. Only in Kim *et al.*, 2007 [70] do we find an attempt to face syntactic and semantic issues in an effort to improve PHR readability, using the CHV to map content in Electronic Health Records (EHR) and Personal Health Records (PHR); and in Zeng *et al.*, 2007b [122] which designed and implemented a prototype text translator to make EHR and PHR more comprehensible to consumers and patients. The

translator identifies difficult terms, replaces them with easier synonyms, and generates and inserts explanatory text for them (they used UMLS and their CHV as sources of vocabulary knowledge). On the other hand, Rosembloom *et al.*, 2006 [100] developed a clinical interface terminology, a systematic collection of healthcare-related phrases (terms) to support clinicians' entries of patient-related information into computer programs such as clinical "note capture" and decision support tools, facilitating display of computer-stored patient information to clinician-users as simple human-readable texts. These kinds of interface terminologies have been used for problem list entry, clinical documentation in electronic health record systems, text generation and care provider order entry with decision support.

To address the critical need of vocabulary support for consumer health applications, for-profit organizations have also developed consumer health vocabulary products, such as Apelon, and WellMed Inc.'s Consumer Health Terminology (CHT) Thesaurus, which is based on the SNOMED nomenclature (Marshall, 2000 [76]) and contains more than 20,000 patient-friendly, culture- and dialect-specific, self-care, patient compliant, and health risk-oriented terms. Concerning multilingual consumer-oriented health vocabularies, we can mention only the initiative of the European Commission's Multilingual Glossary of Popular and Technical Medical Terms [40] in nine European languages⁹, but it is a limited medical vocabulary for medicinal product package inserts accessible to consumers. In fact, it consists of a list of 1,400 technical terms frequently encountered in inserts, with corresponding consumer terms in all the languages of the EC. Greater overlap between technical and lay terms was observed for the Romance languages and Greek than for the Germanic languages (except English) and some technical terms had no lay equivalent.

⁹<http://users.ugent.be/?rvdstich/eugloss/information.html>

2.3 Ontologies and the Semantic Web: an overview

In this section we give some theoretical background on Semantic Web and Ontologies, before walking through the state of the art in using these technologies for the healthcare domain.

The Semantic Web (henceforth SW) is a project that intends to create a universal medium for information exchange by putting documents with computer-processable meaning (semantics) on the World Wide Web. It is aimed to make web pages understandable by computers, so that they can search websites and perform actions in a standardized way. Proposed by Tim Berners-Lee in 2001 [16], this idea of the web provides a common framework for the integration, sharing and reuse of data from multiple sources on the web which overcomes the limits of the inconsistent and disconnected “syntactic web”. This type of framework is possible, even if hard to apply to the whole Web, by adding Semantics (Greek *semantikos*, giving signs, significant, symptomatic, from *sema*, sign) to the web resources, which refers to the aspects of meanings that are expressed in a language, code, or other form of representation. In other words, semantics refers to the meanings assigned to symbols and sets of symbols in a language. Essentially, SW is about two things. It is about common formats for integration and the combination of data drawn from diverse sources, and about language for recording how the data relates to real world objects. This allows a person, or a machine, to start off in one database, and then move through a large set of databases which are connected not by wires but by being about the same thing¹⁰.

2.3.1 Ontologies

At the foundation of the Semantic Web there are the ontologies, which have origins as philosophical concepts. In fact, Ontology is a branch of philosophy

¹⁰<http://www.w3.org/2001/sw/>

that deals with the nature and the organization of reality, and aims to find out what entities and type of entities exist and how things should be classified.

Having been borrowed by the Artificial Intelligence community, the term ontology gained new definitions and found a broad spectrum of applications in various branches of computer science (Guarino, 1998 [51]). In AI, an ontology is considered to be an engineering artefact: it is constituted by a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary. Thus, paraphrasing Gruber's definition of an ontology, "An explicit specification of a conceptualization" (Gruber, 1993 [50]), it describes a formal specification of a certain domain. In order to build such explicit conceptualization, ontologies expressed in a SW language (as we will see later the OWL language) utilize a set of constructs for describing the world in terms of classes, properties, and individuals, as well as provide a set of constructs for expressing complex descriptions in terms of relations between classes, cardinality, equality, richer typing of properties, characteristics of properties, and enumerated classes. Consequently, an ontology consists of a set of definitions of classes, properties, and individuals, together with a set of axioms expressing the relations between classes and properties, and a set of facts about particular individuals. Ontologies define a common vocabulary (for a specific domain) and a shared understanding. We can have different ontologies according to the level of formalism used for representing the domain of interest: light-weight ontologies are ontologies which represent only the hierarchical level of concepts and relations in a domain, in practical terms "taxonomies" (e.g. the Amazon catalog), and heavy-weight ontologies, which are lightweight ontology enriched with axioms used to fix the semantic interpretation of concepts and relations .

Ontologies allow sharing knowledge between people, agents, and software; they enable reuse of domain knowledge, make domain assumptions explicit, and allow access to and evolution of legacy data. Furthermore, they enable

the separation of domain knowledge from operational knowledge and the re-use of domain and operational knowledge separately (e.g., configuration based on constraints), and can manage combinatorial explosion and enable automated reasoning.

The importance of an ontology as means of structuring knowledge has been recognized in areas of knowledge representation, natural language processing, knowledge management, multi-agent systems, database integration, cooperation of distributed applications, web services, and others.

During the last ten years a significant amount of academical research has been directed at developing a theoretical and practical basis of ontology technology. Among other achievements, the most notable developments have been the world wide web consortium standardization of expressive *representational languages* for publishing ontologies on the web (Gomez-Perez and Gorcho, 2002 [48], Bechhofer *et al.*, 2004 [14]), proving the appropriateness of a Description Logics formalism as an *underpinning theory* for performing the formal analysis of ontologies (Antoniou and van Harmelen, 2004, Horrocks *et al.*, 2003, and Horrocks, 2005 [8, 61, 59]), and finally the development of effective *reasoning algorithms* (Horrocks, 1997, Horrocks *et al.*, 2000a and 2000b, Tessaris, 2001, and Horrocks and Sattler, 2005 [58, 63, 64, 114, 62]) and *practical implementations* of automatic inference systems (Horrocks and Patel-Schneider, 1998, Haarslev and Moller, 2001, and Cuenca Grau *et al.*, 2004 [60, 52, 49]) facilitating the automated processing and use of ontologies.

Ontology technology has found many practical applications in a range of scenarios where a *single ontology* is considered – all the tools are off the shelf, and many improvements have been reached for building, mapping, and integrating *multiple ontologies* (Plessers *et al.*, 2007, Kuipers, 2010, Maedche *et al.*, 2002, and Euzenat and Shvaiko, 2007 [94, 71, 75, 43], and many others).

2.3.2 Semantic Web Technologies

In order to reach its aims, and together with the use of ontologies, the SW perspective proposes the combined use of many technologies, as summarized by the following figure, most of them recommended standards of the W3C consortium.

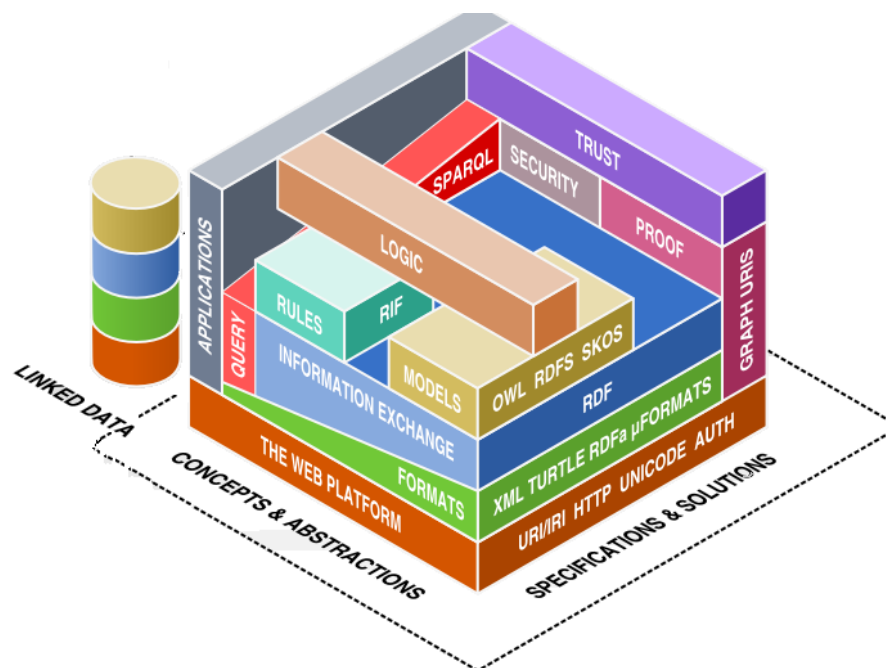


Figure 2.1: The Semantic Web Stack

The SW stack, presented by Tim Berners-Lee in 2003, is based on the Web, and adding structured formats to web documents, arrives at a standardized information exchange, which is the Key of the SW, and finally to query this information, allowing fine-grained data access. Among SW technologies shown in the figure as pieces of the stack, most applications use only a subset of them (XML, RDF, RDFS, OWL, SPARQL, and few others). Here we mention the most important and innovative SW technologies:

- **Uniform Resource Identifiers (URIs)**, which identify the name of a resource on the Internet. The URI identification enables interaction with

representations of the resource over the World Wide Web using specific protocols;

- **Extensible Markup Language (XML)**¹¹, extended by XML Schema, is a markup language, derived from SGML (Generalized Markup Language), composed of a set of rules for encoding documents in machine-readable form, and it is currently considered the standard format for structuring documents;
- **Resource Description Framework (RDF)**¹², the standard model for data interchange on the Web; prescribed framework for representing resources in a common format. It describes information in the form of subject-predicate-object triples, enabling information to be represented in the form of a graph. Using this simple model, structured and semi-structured data can be mixed, exposed, and shared across different applications. Just like XML, RDF has an extension, RDF Schema (RDFS), developed to facilitate content description.
- **Web Ontology Language (OWL)**¹³, is a semantic markup language for publishing and sharing ontologies on the world wide web, and together with RDFS is considered a more expressive way for describing things in the world and how they are related using classes and properties. OWL is particularly designed for use by applications that need to process the content of information, and to facilitate greater machine interpretability of Web content than what is supported by XML, RDF, and RDF Schema (RDF-S), by providing additional vocabulary along with a formal semantics. In fact, it allows constraints on properties, equivalence and disjointness of classes; union, intersection and complement of classes, and finally to characterize properties.

¹¹<http://www.w3.org/XML/>

¹²<http://www.w3.org/RDF/>

¹³<http://www.w3.org/TR/owl-features/>

- **Simple Knowledge Organization System (SKOS)**¹⁴, is a model for expressing knowledge organization systems in a machine-understandable way, within the framework of the Semantic Web. The SKOS Core Vocabulary is an RDF application. Using RDF allows data to be linked and merged with other RDF data by Semantic Web applications. SKOS Core provides a model for expressing the basic structure and content of concept schemes, including thesauri, classification schemes, subject heading lists, taxonomies, terminologies, and other types of controlled vocabulary used for representing semantic Knowledge Organization Systems.
- **SPARQL query language**¹⁵, a query language for pattern matching against RDF graphs, with a syntax resembling to SQL, but which is more powerful, enabling queries spanning multiple disparate (local or remote) data sources containing heterogeneous semi-structured data. It allows for pulling values from structured and semi-structured data, exploring data by querying unknown relationships, performing complex joins of disparate databases into a single one, and transforming RDF data from one vocabulary to another (Hitzler *et al.*, 2009 [56]).

Applying SW technologies can provide many advantages. In particular it is possible to automate operations, for example, from completing all that we need for travel to updating our personal records. The SW then can be defined as a web of information on the Internet and Intranet that contains characteristics of annotation which enables accessing of the precise information that one needs.

2.3.3 Publishing data on the Semantic Web: Linking Data

The trend towards publishing data on the Web is gaining momentum, particularly spurred by the Linking Open Data (LOD) project. This project is promoted

¹⁴<http://www.w3.org/2004/02/skos/>

¹⁵<http://www.w3.org/TR/rdf-sparql-query/>

by the W3C SWEO Linking Open Data community with the aim of extending the Web by publishing various open data sets as RDF on the Web and by setting RDF links between data items from different data sources¹⁶. and several government initiatives publishing public sector data.

The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web. These best practices have been adopted by an increasing number of data providers over the last three years, leading to the creation of a global data space containing billions of assertions - the Web of Data. Linked Data builds upon standard Web and Semantic Web technologies, such as HTTP, URIs, RDF/OWL and SPARQL - 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, and Heath and Bizer, 2011 [17, 53]). Linked Data provides a method for publishing data on the Semantic Web that encourages reuse, reduces redundancy, maximizes its real and potential inter-connectedness, and finally enables network effects to add value to data. We can see many examples of publishing data on the web as Linked Data. The most interesting example is DBpedia¹⁷, an RDF version of information from Wikipedia, which contains data derived from Wikipedia's infoboxes, category hierarchy, article abstracts, and various external links, for a total of over 100 million triples.

In order to facilitate access to data, Linked Data applications have been categorized into 3 main sets according to their structure and access constraints: 1. LOD applications on the infrastructure level (such as Virtuoso Sponger¹⁸); 2. low-level access and manipulation of LOD applications, mainly targeting de-

¹⁶<http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

¹⁷<http://dbpedia.org/About>

¹⁸<http://www.w3.org/wiki/VirtuosoSpongerMiddleware>

velopers (such as Tabulator¹⁹); and 3. end-user LOD applications (such as BBC Music Beta). There are also many useful vocabularies in LOD which can be re-used, such FOAF which describes people (identities, affiliations, social networks) and Geo-Names which describes places. Applications that demonstrate how Linked Data is used within wiki-environments include Shortipedia and the Semantic MediaWiki - Linked Data Extension.

2.4 Semantic-based resources in the Healthcare domain

2.4.1 Healthcare Ontologies

During the last few years, thanks to the Semantic Web perspective, the collaboration between the areas of Healthcare Informatics and Knowledge Representation has generated a set of new methodologies and tools for improving healthcare systems, and in particular medical terminologies and the clinical mappings integrated in these systems were translated into more formal representations using ontology methodologies and languages (for instance, the partial formalization of the Foundational Model Anatomy [101] or the National Cancer Institute's Thesaurus [86] into OWL). The use of SW technologies in this field has delivered promising results in particular for the issue of information integration across heterogeneous resources (Ruttenberg *et al.*, 2007 [103]). The usefulness of ontologies in the biomedical and in particular in the healthcare domain can be easily understood if we consider that the treatment of a patient may involve several practitioners from different healthcare institutes, or that care delivery and patient monitoring are becoming more common, and also that there is a need to access patients' healthcare records electronically wherever they are stored: building the infrastructure enabling the sharing of electronic health records of a patient is currently the first priority of the national eHealth roadmaps of many countries. The use of ontologies here facilitates informa-

¹⁹<http://dig.csail.mit.edu/2007/tab/>

tion integration, data exchange, search of healthcare data, and other critical knowledge-intensive tasks.

Here is a practical example of the representation of some diseases and their anatomical membership and taxonomic relationships, as ontological statements (facts):

Pericardium *is-a* Tissue *and containedIn*.Heart
 Pericarditis *is-a* Inflammation *and hasLocation*.Pericardium
 Inflammation *is-a* Disease *and actOn*.Tissue
 Disease *and hasLocation*.containedIn.Heart *is-a* HeartDisease *and Needs* Treatment

Table 2.3: Disease representation as ontological facts in SNOMED CT

Knowledge Representation techniques allow, for example, a system to provide the automatic suggestions on how to manage a patient's condition, tests that have to be carried out, what medication or treatment should be considered, etc. However, information is complex to manage: there are ambiguities and organizational and cultural issues, conflicts of interest and uncertainty. Furthermore, in healthcare most of the work is designed with the following purposes in mind: (i) computer-based reasoning about facts (e.g. determining from a health record that a patient is at risk of a heart problem); and (ii) aggregation, search and retrieval of data from diverse original source systems, which necessitates rationalization / mapping of vocabularies used in the original data. In this scenario ontologies become relevant if integrated into an EHR, which manages an increasing volume of narrative data (clinical notes pertaining to admission, patient progress, shift change, follow-up, consultation, procedures, etc.), to allow: structuring and semantics of the recorded information (e.g. a record of an abdominal examination is likely to include at least some anatomical terms and characterizations, which should not violate what we know about anatomy, and therefore, should be compatible with ontologies of biomedical reality such as an anatomy ontology); and references to concepts from ontologies of the first

kind (terminologies), for instance ICD 10/9 or SNOMED. (Ceusters *et al.*, 2005 [32]). Much effort has been devoted to the creation of Biomedical Ontologies (Rubin *et al.*, 2007 [102]), but here we can mention, apart from the already cited FMA and NCIT, the logical formalization of SNOMED CT (Rector and Brandt, 2008 [98]) or GALEN (Rector and Rogers, 2004 [97]) into OWL-DL, the work of Lee and Geller, 2006 [72] who performed a Semantic Enrichment of medical ontologies using the two level structure of UMLS and WordNet SUMO, and finally the attempt of Heja *et al.*, 2005 [54] of a formal representation of ICD10, providing an OWL encoding of the first two chapters of ICD10, based on GALEN CMR.

2.4.2 Semantic-based Medical Terminology Integration

Two other important issues to take into account in this context are: *Ontology Mapping* (Noy, 2009 [85]), to show how concepts of one ontology are semantically related to concepts of another ontology; and *Ontology Integration*, which allows access to multiple heterogeneous ontologies. Ontology Integration in general can be realized either by merging ontologies into a single one, or by keeping ontologies separate (Euzenat and Shvaiko, 2007 [43]). The goal of knowledge integration systems is to provide uniform access to multiple heterogeneous information sources, so a global schema provides a unified view for querying the set of local schema.

SW technologies have been used not only for creating mashups of biomedical data (Cheung *et al.*, 2008 [33]), but also for terminology integration purposes. For example, Bodenreider, 2008a and 2008b [19, 20] exploits RDF for comparing formal definitions in SNOMED CT and the NCI Thesaurus, and between LOINC and SNOMED CT. Much work has been done in the healthcare domain for the alignment of different Biomedical Ontologies (Smith and Rosse, 2004 [106]), with concept overlap (e.g. FMA concepts with the corresponding anatomical concepts in SNOMED, ICD10, etc.), and for their integration by

means of medical ontology repositories. Concerning this last point a good example is BioPortal, a Web-based system that serves as a repository for biomedical ontologies developed at the Stanford University by Noy *et al.*, 2008a [87]. BioPortal determines relationships both among those ontologies and between the ontologies and on-line data resources such as PubMed, ClinicalTrials.gov, and the Gene Expression Omnibus (GEO), a work which has been extended providing a community-based method to collect the existing ontology mappings between the sources presented in BioPortal. The BioPortal ontology library enables users to provide and discuss a wide array of knowledge components, from submitting the ontologies themselves, to commenting on and discussing classes in the ontologies, to reviewing ontologies in the context of their own ontology-based projects, to creating mappings between overlapping ontologies and discussing and critiquing the mappings. Also Gangemi *et al.*, 1998 [46] worked on medical ontology integration, describing the ONIONS methodology for ontology analysis and integration, which has been applied to some relevant medical terminologies (e.g. in the UMLS project), with the aims of: (i) developing a well-tuned set of generic ontologies to support the integration of relevant domain ontologies in medicine to permit many tasks, including information retrieval, natural language processing, computerized guidelines generation, data base integration, etc.; and (ii) providing an explicit tracing of concept mappings, constraints and choices in ontology building in order to allow extensions and/or updating.

2.4.3 Publishing Data for Healthcare

During the last two years developments in Semantic Web are changing the way domain researchers are building medical resources, such as medical classification systems, terminologies and ontologies. In fact, researchers have started to apply collaborative methodologies for editing and reviewing these medical resources by large groups of scientists. This is motivated by the fact that on-

tologies are becoming too large in their coverage to be developed by one single centralized group of people, and also health organizations are asking broader users to make contributions. Many tools have been created to facilitate and support collaborative development of ontologies and review and validation processes. To give some examples we can mention here WebProtégé, a Web-based version of the Protégé ontology editor (Noy *et al.*, 2001 [88]) developed at Stanford University²⁰, which provides a simple user interface to edit an ontology using a Web browser, enables multiple users to access and edit the same ontology simultaneously, and allows users to configure their own interface; and MoKI - the Enterprise Modelling Wiki - (Ghidini *et al.*, 2010 [47]) a tool developed at Fondazione Bruno Kessler that supports the creation of articulated enterprise models through structured wiki pages, which enables heterogeneous teams of experts, with different knowledge engineering skills, to actively collaborate by inserting knowledge; transforming knowledge; and revising knowledge at different levels of formality.

A good application of such technologies and tools is related to the work done by Tudorache *et al.*, 2010a and 2010b [118, 119] at Stanford University in collaboration with the World Health Organization for the eleventh revision of ICD. In fact, they created a set of tools for collaborative ontology editing and publishing, in an integrated platform (composed of WebProtégé and the BioPortal library) for the creation of the entire ICD11 ontology lifecycle. Here ontology authors use Protégé to develop their ontology collaboratively, with discussions among themselves, and use BioPortal to publish the ontology and to ask for feedback from the broader user community. This integration streamlines the workflow for collaborative development and increases integration between the ontologies themselves through the reuse of terms.

Concerning the use of Linked Open Data (LOD) in the clinical domain, some existing effort is already being devoted to mapping archetypes to OWL, and to

²⁰<http://webprotege.stanford.edu/>

improving terminologies, for example the already mentioned OWL representation of SNOMED CT. In this field, other technologies are added to the LD stack, such as the openEHR archetype query language and the open EHR RM (Electronic Health Record Reference Model) and AOM + SNOMED CT vocabulary.

2.4.4 Applications in Healthcare

The benefits promised by the Semantic Web in the healthcare domain include aggregation of heterogeneous data using explicit semantics, simplified annotation and sharing of findings, the expression of rich and well-defined models for data aggregation and search, easier reuse of data in unanticipated ways, and the application of logic to infer additional insights.

The W3C has established the Semantic Web for Health Care and Life Sciences (HCLSIG) Interest Group²¹ to develop, advocate for, and support the use of Semantic Web technologies for biological science, translational medicine and health care. These areas stand to gain tremendous benefits by the adoption of Semantic Web technologies, as they depend on the interoperability of information from many domains and processes for efficient decision support. The principal members of the task force of this group are: BioRDF, for integrating neuroscience knowledge base; Clinical Decision Support, for the expression and use of guidelines; Clinical Observations Interoperability, for example for in recruiting patients for trials; Linking Open Drug Data, for the aggregation of Web-based drug data; Scientific Discourse, in order to build communities through networking; and finally, Terminology, creating Semantic Web representations of existing resources.

HCLSIG has used RDF tools to integrate several large biological and clinical databases. This has simplified access to relational and hierarchical data and enabled third party additions to the database.

²¹<http://www.w3.org/blog/hcls>

A biomedical application that relies on knowledge from more than 200 publicly available ontologies in order to support its users in exploring biomedical resources is the NCBO Resource Index²². A second example of a Linked Data application from this domain is Disease Map²³. This application combines data from various Life Science data sources in order to generate a “network of disorders and disease genes linked by known disorder gene associations, indicating the common genetic origin of many diseases” (Heath and Bizer, 2011 [53]).

²²<http://bioportal.bioontology.org/resources>

²³<http://diseasome.eu/map.html>

Chapter 3

The Problem Statement and the Approach

The chapter closes the part of the thesis dedicated to preliminaries focussing on the limitations of the “state of the art” and the problem statements and then presenting the solution adopted for face the problems. In particular it provides an overview of the Approach describing the steps which will be explained into details in the next chapters.

The chapter is organized as follows: Section 3.1 reviews the problem domain of the thesis, in particular the already mentioned linguistic gap, and overviews insights and motivations of the framework to be adopted. Section 3.2 introduces the overall method used to bridge this gap and to overcome the problem of integration of lay terminologies with the specialized one.

3.1 The Linguistic Gap

As shown in the state of the art many are the improvements in accessing healthcare data from the web, and in formalizing and integrating medical terminologies and finally in linking data on the Semantic Web. Thanks to the development of Knowledge Representation and Integration technologies and thanks to the Web 2.0 perspective applied to the healthcare domain we also recognised the new active role played by healthcare consumers in accessing, updating, and

managing their personal health care data. However, as previously mentioned in Chapter 2 all the information systems available from the web and used by consumers to access and manage their data but also simply to search information on health topics are based on a medical terminologies oriented to professionals and not to lay persons. This increases the communication gap existing between consumers who use a lay terminology and professionals' language. This gap makes it difficult for laypersons to find, understand and act on health information.

Obviously a clarification process for medical vocabulary goes through the creation of a terminology of well-defined and rigorously applied words, which does not imply the prohibition of specialized terms, as no specific language can do without its lexical background. But it is also evident that medical language is overloaded with many obsolete terms still used in communication with healthcare consumers (e.g. "leontiasis", "sclerophthalmia", "euphoria", etc.), and still full of etymological incongruities (e.g. "mycotic embolus" , meaning an embolus infected by fungus and which etymologically brings the word fungus/mycos, instead of using "infected embolus"), semantic ambiguities (we can think about the conceptual versatility of a too frequently adopted term such as "normal"), and of some archaic and useless terms (e.g. in Italian "cinoresia" defined as "ravenous hunger similar to that of dogs"; "cinospasma", that is "spasm of facial muscles which makes the face similar to that of aggressive dogs"; "cipridopatia", "cipridologia" and "cipridofobia" to indicate respectively "venereal disease", "venerology" and "phobia for venereal diseases" - venerofobia) (Iandolo, 1983 [65]).

In this context, beyond the need for creating order in medical language to allow for clear communication between different medical providers, there exists a more important need, that is, the need to also put laypersons in a position to understand messages and information communicated to them by their physicians. Too often, in fact, the inclination of professionals to use "very technical" words makes patients uneasy while trying to understand. Quoting Baldini, 1996 [11],

we can say that physicians “do not talk *with* the patient but *to* the patient”. In this way communication obviously becomes unidirectional, and if professionals want to make medical language understandable, it has to be flexible and adaptive to the capacity of comprehension of the patient. Also Lucchini, 2008 [74] asserted that in medicine two different requirements live together: one is the communication between healthcare providers, which is necessarily technical, and the other is the communication addressed to patients and healthcare consumers (e.g. Patient Information Leaflets, Medical Reports, etc.) which should be clear and simple to understand.

All these consideration become still more relevant if we consider the “dis-intermediated” (not intermediated by physicians) communication between consumers/patients and the recently developed healthcare applications for consumers. There is a need for terminologies and ontologies to address consumers viewpoint and knowledge. But, in many cases, the lack of lay resources (oriented to consumers), implies that specialized terminologies (clinically-oriented) are adopted, creating in this way an obvious barrier to the use of such applications. In this work we want to focus on the nature of this communication gap in the Italian context, where there is a lack of consumer-oriented medical terminologies (as seen in Chapter 2, all previous works have been done for English), and where low health literacy or regional diversity further intensify the problem. In this context the challenge is to know the language used by laypersons and to know how to map this to medical concepts in order to assist healthcare consumers to formulate queries to understand medical documents on the web, but also to help information systems and professionals to deal with patients inputs (for instance, during patient interviews).

Consumer expressions can be used as category names for browsing hierarchies and as suggested words for health text authoring systems for lay audiences. But they can also be used to create a consumer-oriented entry vocabulary which can be integrated in a professional medical vocabulary for mapping

or expanding query terms, and finally, automatically identifying and linking professional or lay medical terms in medical texts or in authoritative resources for consumers. Figure 3.1 shows how a consumer-oriented medical vocabulary can be useful providing translation functionalities, if integrated in healthcare information systems, in two typical scenarios: 1) in the communication from professional to consumer (e.g. interpretation of medical reports received in his PHR), and 2) in the communication from consumer to professionals (e.g. online medical consultations, PHR data entry).

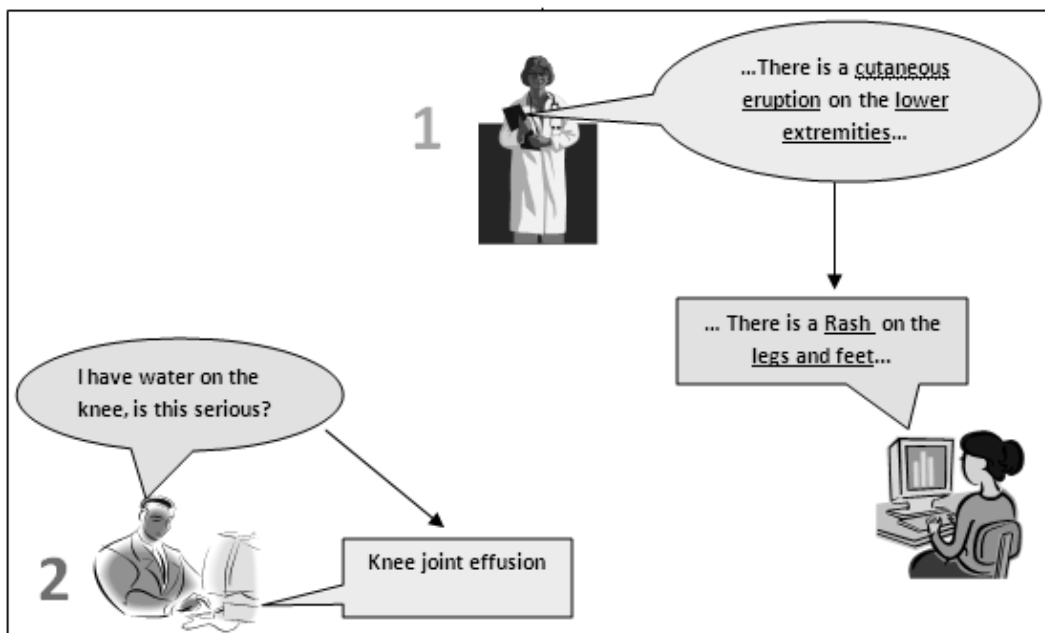


Figure 3.1: Typical scenario for the use of a Consumer-oriented health vocabulary

The integration of consumer-oriented vocabularies together with their mapping to specialized medical terminologies in healthcare information systems provide services which support access to integrated information: (i) helping users in the process of querying and searching healthcare information, (ii) performing inferences on term classification and relations between the different integrated resources, (iii) translating and interpreting professional languages used in clinical notes, procedural results and other documents, and (iv) supporting users in easily describing their problems, complaints and clinical history.

3.2 The Approach: an overview

To contribute to the solution of the aforementioned problems, this thesis presents a hybrid and multi-disciplinary approach, based on the construction of a lay medical vocabulary, focused on Italian, and integrated with standard medical terminologies/ontologies through Semantic Web technologies, in order to have a coherent lexical and semantic medical resource useful both for professionals and for healthcare consumers. The global approach followed in this thesis is divided, in particular, into three macro phases. The first includes the creation of the Consumer-oriented Medical Vocabulary for Italian - ICMV - for collecting common medical expressions and terms used by Italian speaking people to indicate medical concepts in daily life and during their health care and encounters with health care professionals. The second focuses on the formal representation (semantic web-based representation) of medical terminologies and classification systems used in general practice to encode diagnoses, diseases and medical procedures, but also to retrieve information on the web. Finally the third macro phase concerns the integration of these formalized medical classification systems and terminologies with the developed ICMV, using Semantic Web technologies. Figure 3.2 shows the overall architecture of our Integration Framework, starting from the creation of the ICMV and ending with its integration with the medical terminologies presented in the figure. In particular the coloured circles identify the main terminologies we took into account while the dashed boxes put in highlight the 3 phases and the possible application scenarios. In this case the following possibilities:

- PHRs Integration with ICMV. PHRs (in this case “TreC”) are designed for consumers, used and managed by consumers and they need to use consumer-oriented terminologies, in this case “ICMV” to help consumers in filling out their records.
- EHRs involvement. On the other hand PHRs can communicate with physi-

cians' EHRs, for example to receive medical reports, text results, and other documents which on the contrary use specialized medical terminologies (e.g. ICPC2, ICD10, in Italy) so having integrated the ICMV could give the possibility to foster the readability of data deriving from EHRs but also to send to physicians PHR content filled out in lay terms translated in technical language.

- Web searching of health care information and literature. UMLS and in particular MeSH are used for searching purposes and for health-related literature indexing, so having integrated also ICMV could allow lay persons to perform queries on the web for searching medical information.

As shown by the figure, the overall approach is characterized by the following tasks:

Phase 1. Generation of the Italian Consumer Medical Vocabulary. This phase includes two tasks:

1. Knowledge Acquisition/Terminology Extraction. Use of elicitation techniques to acquire all the lay terms, words, and expressions used by laypeople to indicate specific medical concepts, in particular to express symptoms and complaints, medical procedures, diseases, and anatomical concepts. The target chosen for knowledge acquisition analysis is composed of the following samples of people:
 - hospitalized patients and subjects submitted to the triage process in a first aid unit;
 - elderly people in a small community;
 - a community of students and researchers with an adequate level of health literacy;
 - lay people of every age who ask for assistance on healthcare websites.

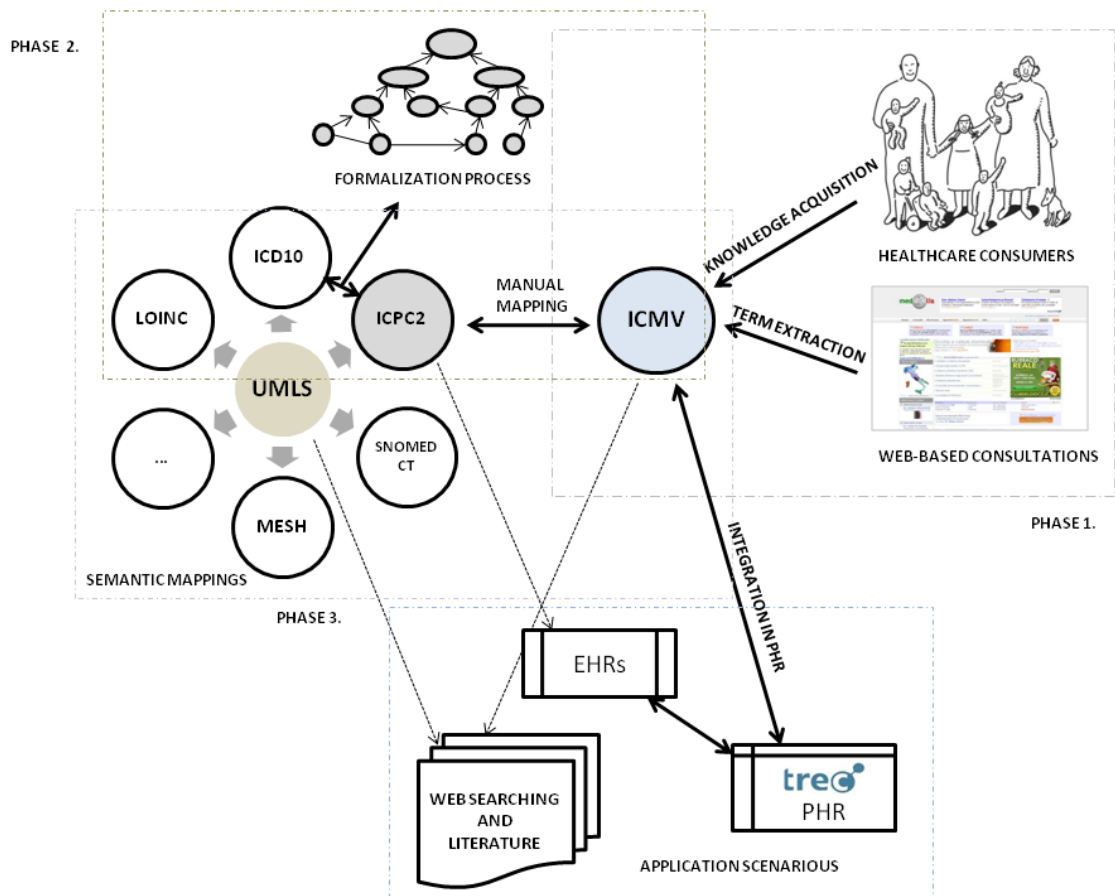


Figure 3.2: Approach Overview

Along with verbal resources (coming from interviews, focus groups, etc.), we also analysed data collected from various sources such as forum postings written by healthcare consumers on Italian websites for asking questions to on-line doctors. We applied NLP techniques for term extraction, parsing, tagging and normalization, to conclude with a statistical analysis (based on term frequency and on the degree of familiarity of the terms expressed by healthcare consumers) and a clinical review performed by a group of physicians, nurses and pharmacists for finding candidate medical lay (considered as synonyms or quasi-synonyms technical terms) which can be included in the ICMV.

2. Manual mapping of ICMV terms to ICPC2. Here physicians are called to find manually a “one-to-one” or “one-to-n” mapping of lay terms with the corresponding medical concepts in ICPC2, to define explicit relationships among them. At the end of the process the collection of candidate terms mapped to medical concepts have been organized in a format that is easy to use and access. Furthermore, in order to make ICMV usable by applications we publish it in the Semantic Web and integrate it with the other medical terminologies as described in the next two phases.

Phase 2. Formalization of specialized and consumer-oriented medical terminologies using Semantic Web languages.

1. Encoding medical classification systems in OWL. In this step two important classification systems used in the General Practice domain, namely ICPC2 and ICD10 (which are clinically integrated) are formalized in OWL ontologies, together with the axiomatization of the existing mappings between them in order on one hand to improve their reuse on a SW perspective and on the other hand to logically analyse their mapping consistency. In this step the tool Protégé (Noy *et al.*,

2001 [88]) and the reasoner Pellet (Sirin *et al.*, 2006 [105]) are used.

2. Encoding of ICMV and of specialized vocabularies in RDF. In this step ICMV and other medical terminologies extracted from UMLS, namely SNOMED CT, LOINC, and MeSH, are encoded as RDF N-Triples¹ (Omelayenko, 2002 [90]) for semantic integration purposes, to guarantee interoperability between the resources taken into account following the efforts and principles of the HCLSIG and BioRDF semantic web groups², and to evaluate Semantic Web technologies in this context.

Phase 3. Creation of the Integration Framework for consumer-oriented health-care applications. This phase includes the following steps:

1. Integration of ICMV and vocabularies in UMLS (MeSH, SNOMED CT, LOINC, ICD10) through ICPC2. We collected all the formalized terminologies/ontologies under consideration in an RDF triple store and we extracted semantic mappings between them using ICPC2 as a pivot to access UMLS vocabularies. We used SPARQL queries to retrieve mappings on the stored graphs. At the end of this process we compared manual mapping between ICMV and ICPC2 performed in Phase 1 and the automatic mapping to UMLS vocabularies via ICPC2, in order to evaluate the best approach to integrate consumer-oriented vocabularies with medical terminologies or classification systems. This Integration Framework will be useful to supply knowledge services to support the development of semantic-based healthcare information systems which need interchanges with patients and health-care consumers in general.
2. Experimental uses of ICMV.

¹<http://www.w3.org/TR/rdf-testcases/#ntriples>

²http://www.w3.org/wiki/HCLSIG_BioRDF_Subgroup

- Publication of ICMV under the form of a consumer-oriented health-care Wiki³, which can be used for browsing and searching purposes. In the future ICMV can also be continuously integrated, in a collaborative way, both by users with new terms and lay synonyms, and by physicians with comments or with new mappings to technical terms in other standardized terminologies.
- Evaluation of the usefulness of this medical resource and integration framework through its application to a new Personal Health Record, developed at Fondazione Bruno Kessler for the Province of Trento, namely TreC (Cartella Clinica Elettronica)⁴. During this step, most of the knowledge extracted during the previous stages was used to conceptualize data included in TreC, and in particular the ICMV together with its mapping to ICPC2 was integrated with TreC to facilitate users in the process of filling out their healthcare data such as symptoms and complaints, their clinical history (pathologies, allergies, intolerances, vaccinations, surgeries, etc.) and their therapies, but also to help them in easily accessing their data, clinical notes and tests results, which have been conveniently categorized.

³<http://ehealthwiki.fbk.eu>

⁴<https://www.trec.trentinosalute.net/pubblico/index.html>

Part II

DEVELOPMENT

Chapter 4

The Italian Consumer Medical Vocabulary*

This chapter opens the part of the thesis dedicated to the acquisition of consumer-oriented terminology/knowledge and to the creation of the Italian Consumer-oriented Medical Vocabulary. We use a hybrid methodology which combines knowledge acquisition techniques with automatic terminology extraction and a final review provided by domain experts to validate acquired knowledge.

The chapter is organized as follows: Section 4.1 introduces some theoretical concepts which give a basis for understanding the practical tasks used to create the ICMV. Section 4.2 outlines the approach to acquire lay medical knowledge/terminology. Section 4.3 describes the automatic term extraction process performed on the document collection. Section 4.4 presents the clinical review approach performed by physicians on the extracted terms in order to find the best candidate to be integrated into the ICMV. The process of associating a familiarity degree to the lay terms in the ICMV is described in Section 4.5, and finally Section 4.6 closes the chapter presenting the manual mapping process performed by physicians to find correspondences between the lay terms in the ICMV and the medical concepts in the International Classification of Primary

***Acknowledgements:** The material of this chapter is based on earlier publications [29, 28, 30]

Care (ICPC2) Results and evaluations are presented at the end of each sections.

4.1 Theoretical background

Before describing the process of generating a consumer-oriented vocabulary in the Italian context, it is useful to introduce some theoretical concepts, starting from a brief explanation of the most important linguistic elements and stating what a vocabulary is (Crystal, 1987 [37]). A vocabulary contains the fundamental building blocks used to convey complex thoughts, including physical objects, abstract ideas, their properties, and their relationships. A basic unit of a vocabulary is the *term*, defined as a lexeme used in a particular domain, that is, the basic linguistic units, composed of form and meaning (or concepts), of technical vocabularies used in Languages for Special Purposes (LSPs). In other words, they represent “atomic nuclei of the knowledge elaborated in a specialized field, and, at the same time, they constitute the rapid and efficient vehicles of communication between domain specialists” (Adamo, 1999 [7]). According to Crystal, 1980 [36] the word “term” can have three common senses:

- Word-Form: An entity or physical object found in written and spoken text
- Lexeme: An abstraction that expresses a set of grammatical variants (e.g., think, thinks, thinking, and thought)
- Word: An abstraction that functions as a fundamental building block of grammar

In this study, “term” refers to the first sense proposed by Crystal in the list above (meaning the pair word-form).

4.1.1 Terminology and Lexicology

Terminology is traditionally associated with the categorization of terms used in discourse within specific domains or fields, also called language for special

purposes (LSP). Protocols and procedures have been developed for the discipline of Terminography², by both academics (e.g., Cabré 1999, and Picht and Draskau, 1985 [24, 93]) and standards bodies, such as ISO/TC37 (1952) [1] and ISO 704 (1987) [2]. These guidelines focus on communication among specialists. However, as Bowker 2001 [21] observed, LSP communication is becoming less and less confined to specialists in that field: it sends its message to the general public through the mass media (Bowker 2001, p. 590 [21]). Lexicology is the study of general language words and their usage. Starting from a formal element, it aims to keep semantic and etymological information, following a path which has been defined as “semasiologic”, that is addressing the meaning of a word. On the other hand, Terminology tries to identify concepts to determine individuality and membership to a systematic structure, following an “onomasiologic” process, intended to name concepts by means of the terms (Cabré 1993, p. 71 [22]). Terminology has the purpose of “individuating and determining objects and concepts which constitute a structured knowledge system proper of a specialized domain, with the aim of naming them univocally through the terms. Cabré 1999 [24] distinguishes between Terminology and Lexicology as follows: Lexicology deals with words in order to account for the lexical competence of speakers; Terminology deals with terms in order to establish a reference to concepts of the real world. The objectives of Terminology are clearly different from those of descriptive Lexicology because terminology does not attempt to provide an explanation of the knowledge that experts have of terms. On the other hand, Terminology is aimed at identifying and naming the concepts belonging to a specific subject (Cabré 1999, p. 36-37

²Terminography uses terms as formal units of specialist language form, which constitute the terminology of a field, studied by the science of Terminology, and is documented by terminographers, whose goal is said to be to explain referents (objects/concepts), and whose approach is (1) onomasiological (meaning that they start from an analysis of concepts, then look for terms to assign/that are assigned to that concept, then enter these terms in the entry for that concept), and (2) systematic (meaning that they represent entries according to theme/some previously developed classification).

[24]). Furthermore, lexicographic work³ has not the aim to serve language as a system, but to serve the discourse, that is the linguistic realization (Adamo, 1999, p. 82 [7]), so Lexicology is at the service of the speaker of a language (Nencioni, 1987, p. 134 [84]). Terminographic production answers, instead, to the needs of an expert public which has recourse to it just for checks and controls, but, above all wants to find equivalent terms in other languages (Adamo, 1999, p. 82 [7]). Another important distinction between Lexicology and Terminology is that the first always keeps track of “polysemy”, considering it as an added value which helps in increasing language, while terminographic works do not use polysemy because a term has to be unambiguous (unique) and rare are the cases where it has synonyms (Adamo, 1999, p. 83 [7]). The choice to produce a lexicographic work is often taken on the basis of the frequency of the use which gives evidence of the concrete distribution and consequently the membership to a language. The frequency of a term, on the contrary, is not relevant in a terminographic production, in fact it is observed that the specialization degree of a term is “inverse proportional” to its use frequency (Adamo, 1999, p. 83 [7]). After these theoretical consideration on Terminology and Lexicology, we can make the following assumption: medical vocabularies used by non-professionals, in concluding consumer-oriented vocabularies such as the one we are going to describe, share characteristics with both a terminographic and a lexicographic product, even if the specialization of the vocabulary makes it nearer to Terminology than Lexicology. Thus, terminographic procedures were used as a starting point for exploring consumer-oriented medical vocabularies. However, the methodology was refined and iterated throughout this study. The final procedure, reported in this chapter, may be useful in informing

³Lexicography uses words as formal units of general language form, that constitute the vocabulary of that language, which is studied by the science of Lexicology, and is documented by lexicographers, whose goal is said to be to explain word meaning, and whose approach is (1) semasiological (meaning that they start from a lexeme, then look for all the senses of that lexeme, then enter these senses in the entry for that lexeme), and (2) alphabetical (meaning that they represent entries according to the sequence of the alphabet).

future research oriented at characterizing non-professional vocabulary usage in specialized domains.

4.1.2 Generating Vocabularies

Generating Vocabulary typically refers to the process of creating a controlled vocabulary, defined as “a way to insert an interpretive layer of semantics between the term entered by the user and the underlying database to better represent the original intention of the terms of the user”⁴. In this thesis, generating a vocabulary refers to collecting and organizing a set of terms representative of a vocabulary assumed to exist. For example, we consider ICMV to include all terms that Italian speaking consumers use to discuss medical topics. Thus, generating the ICMV is shorthand for creating a representative set of words based on specified criteria. Creating a vocabulary involves extracting terms that describe domain-specific concepts or entities from relevant sources of discourse, such as collections of documents or corpora, interviews, and reason for encounters (RFE). The level of specificity depends on the type of vocabulary and its purpose. Natural language vocabularies consist of lexemes from languages used in everyday discourse, whether LGP (language for generic purposes) or LSP. Controlled vocabularies (also known as thesaurus) that limit terms artificially for specificity, however, are typically used with LSP only. Ideally, each form refers to a single concept and each concept is represented by a single preferred form (i.e., an isomorphism). Forms are normalized to control for variations in strings commonly found in natural language. Overall, according to (Soergel, 1974 [108]) the process of vocabulary building typically involves four stages:

1. Specifying domain boundaries and identifying appropriate sources of terms.
In an ideal world, each domain has clearly marked boundaries, the purposes of a vocabulary are well defined and the intended users are clearly

⁴Fred Leise and Karl Fast and Mike Steckel:http://www.boxesandarrows.com/view/creating_a_controlled_vocabulary

specified.

2. Identifying and extracting terms. In this step the sources equally representative of authoritative documents and user needs are selected, and all relevant terms that are clearly specific to the domain and sufficient to express all technical concepts of that domain are extracted. Alternative forms from common usage among the intended users are included as well.
3. Normalizing forms and mapping them to reference concepts. This step aims at processing the terms, including their normalization, to create an unambiguous one-to-one mapping of forms with concepts. Relationships among forms and concepts need to be defined explicitly.
4. Organizing the terms. The final collection of terms is organized in a format that is both easy to use and access.

In reality, each step involves some degree of uncertainty and requires procedural and policy decisions, since language and subject areas are characteristically “messy” and “fuzzy”.

Different subject areas have different numbers of domain-specific terms. In general, the greater the difference between terms in the target domain and in other areas, the sharper the domain boundary. Moreover, developing policies for term inclusion/exclusion depends on the nature of the target domain: for example, whether non-technical terms are required to express domain-specific concepts or whether the domain is interdisciplinary or boundary-spanning. Non-specific domain terms, generally excluded from core technical vocabularies, are included as required to complete a vocabulary. Developing a vocabulary for interdisciplinary domains, such as medicine or healthcare, may require inclusion of terms from related areas (e.g., anatomy, psychology, pharmacology). The level of detail sought for terminology is dependent on factors such as the intended purpose(s) of the vocabulary, the target audience, and the discipline to

be covered. A vocabulary intended for bibliographic citation retrieval may require a greater level of refinement than one for browsing categories on-line, but less detail than one for knowledge engineering. Furthermore, a greater degree of refinement may be required for a technical audience than for non-professional users in order to maintain precision.

As previously defined, terms are words or phrases specific to a domain or discipline. Identification of terms from natural language sources is often complicated; there is no simple algorithm for identifying terms from words in a text. Identifying multi-word terms or determining which words belong to a term is not straightforward. Manual term selection is labour-intensive and subjective. Although researchers have been exploring the use of automated methods to facilitate term identification, techniques developed to date are not precise and require considerable manual manipulation. Since this chapter is not just devoted to acquiring only terms but also to acquiring knowledge expressed by consumers and related to health topic, in the following subsection we will briefly introduce the concept of Knowledge Acquisition and its use in the medical domain. Furthermore we will describe the practical steps in generating our ICMV starting from the knowledge acquisition process, followed by automatic term extraction and analysis of acquired knowledge by means of NLP tools, to conclude with a clinical review performed by physicians to find best candidate to be included in the ICMV and to map these terms and expressions to a standard medical terminology.

4.1.3 Knowledge Acquisition

Knowledge Acquisition aims at identifying and capturing knowledge assets and terminology to populate a knowledge repository for a specific domain. Central areas of this task are: terminology work, which is relevant for a special subject field, including terminography; content analysis of documents; and extraction of knowledge from various sources. A major part of Knowledge Acquisition is

capturing knowledge from experts, a task that can be made cost-effective and efficient by using knowledge models and special elicitation techniques. These techniques should be used in different phases of the process, since each of them supports the capture of a specific typology of knowledge and the achievement of specific aims.

The most common techniques for Knowledge Acquisition are *Interviews*, direct observation of experts performances to extract procedural knowledge, mostly connected to manual skills, such as *Think Aloud Problem Solving*, *Self-report*, and *Shadowing*. Other techniques, such as *Card Sorting*, *Repertory Grid*, and *Twenty Questions*, are useful for understanding how experts conceptualize knowledge related to their own domain of reference (Milton, 2007 [80]). In the task of Knowledge Acquisition, it is important to identify two main components; knowledge types and knowledge modalities: the first refers to knowledge orientation and domain, and the second refers to the representation medium in which knowledge exists. In Knowledge Acquisition for Healthcare domain, according to Abidi, 2007 [6], many different types of knowledge which directly contribute to clinical decision-making and care planning, can be identified: Patient, Practitioner, Medical, Resource, Process, Organizational, Relationship and, finally, Measurement Knowledge. In this thesis we will only deal with Medical Knowledge and Patient Knowledge. These knowledge types are represented by different knowledge modalities, where each knowledge modality may capture one or more knowledge types as a healthcare knowledge artefact, i.e. objects that allow knowledge to be captured, such as documents, healthcare records, knowledge bases, communication between peers, etc. Examples of knowledge modalities can be tacit knowledge from a practitioner, explicit knowledge, clinical experiences, collaborative problem-solving discussions, social knowledge, etc. For the creation of the ICMV, we are interested in tacit and explicit knowledge, clinical experience and in particular social knowledge. In particular this last modality can be viewed in terms of a community of practice

and the communication patterns, interest and expertise of individual community members.

4.2 Knowledge Acquisition Process

To acquire consumer-oriented medical knowledge (lay terms, words, and expressions) used by Italian speakers, we applied a hybrid methodology which is divided into two main steps: the first aims at the identification of medical terms related to “symptoms”, “diseases”, and “anatomical concepts” using three different elicitation techniques and directly involving patients and consumers by recording their oral communication; the second step consists of the acquisition not only of symptoms, diseases and anatomical concepts, but also of terms related to “medical procedures”, “people involved in the healthcare process”, “medical and healthcare sites and institutes”, and “medical instruments and devices” from large written corpora on the web such as forum postings or on-line medical consultations.

Topicality⁵ was an important dimension of document selection. During project planning we discussed two approaches: either limiting the scope of the documents to prevalent symptoms and diseases or selecting them “blindly”, without considering their topicality. At the beginning we applied the first approach because the latter was selected to provide more breadth in topics, but at the expense of depth of coverage.

During the first step, we considered three different target groups for the application of our approach:

- First Aid patients subjected to a Triage Process⁶ assisted by nurses;

⁵In medicine, the property of a medical concept of being related to a particular topic or topics such as symptoms, diseases, body parts, etc.

⁶The Triage activity has the aim to prioritize, by means of a very brief examination, patients based on the severity of their condition.

- A community of researchers and Ph.D students with a middle-to-good level of healthcare literacy, who used an ad hoc collaborative Wiki system to acquire medical terminology;
- A group of elderly people (aged >65 yrs.) with a modest background and low level of healthcare literacy, who participated in traditional elicitation techniques such as Focus Groups, Concepts Sorting and Interviews.

The second type of acquisition involved users who looked for information on medical websites and in particular forums for asking questions to on-line doctors. Here users cover a wide range of ages and their level of healthcare literacy is extremely varied.

Demographic information of the participants involved in the Wiki-based and the Focus group-based acquisition is provided in Table 4.1. For consumers involved in the Nurse-assisted acquisition this information is not available.

Characteristics	Participants
	n = 56
Gender	
Female	36
Male	20
Age	
Elderly people (>65)	23
Adults (35-65)	15
Young people (18-35)	18
Education	
University	27
College	15
Advanced degree	3
Primary school	11

Table 4.1: Demographic characteristics of consumers

Consumers and physicians involved in the acquisition process were instructed to select terms on the basis of their personal experience, knowledge, use in daily

life, and judgment. The following paragraphs describe in detail the different elicitation techniques used to acquire consumer-oriented medical knowledge and the document collection.

4.2.1 Wiki based Acquisition

The first method for acquiring consumer-oriented medical knowledge is based on the use of a Semantic Media Wiki system⁷, an easy-to-use collaborative tool, allowing users to create and link, in a structured and collaborative manner, wiki pages on a certain domain of knowledge. We created for this task an ad hoc wiki system for collecting medical knowledge, namely the “eHealthWiki”⁸, in which each medical term provided by users is a structured Wiki page (as in Wikipedia) categorized in one of our two proposed categories, *Symptoms* or *Diseases*, including for each page a field for term description, a field for corresponding synonyms, another field for the anatomical localization, and, only for the pages included in the *Diseases* category, a field for the association of possible symptoms to that particular disease. The choice of including only categories for symptoms and diseases in the Wiki was motivated by the fact that symptomatology and pathologies are the ones most identified by consumers with lay terms and expressions. In this system users can add and manually edit a Wiki page related to a lay medical term, using a predefined form containing all the fields mentioned above or they can use an import function that allows for uploading external files. Figure 4.1 shows an example of a Wiki page filled in by a user describing the symptom “Abbassamento della voce” (Absence of voice), providing a definition in lay terms, anatomical localization, synonyms.

The eHealthWiki has been tested over a sample of 33 people (17 females and 16 males, aged between 24 and 56 years) including researchers, Ph.D students, Master students and administrative staff of our research institute (Fondazione

⁷http://semantic-mediawiki.org/wiki/Semantic_MediaWiki

⁸This collaborative system can be consulted at:<http://ehealthwiki.fbk.eu>

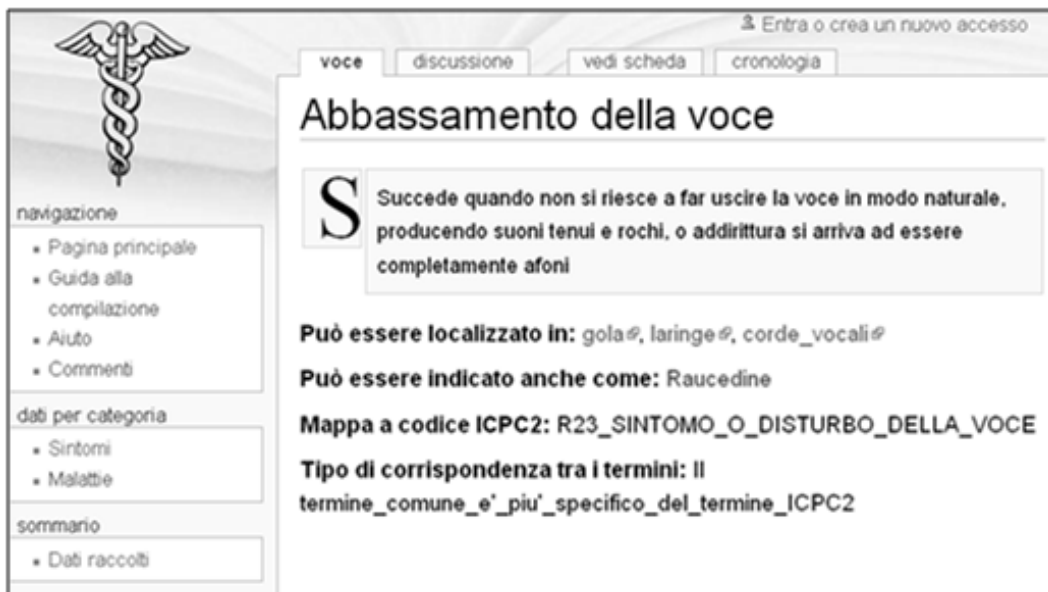


Figure 4.1: Wiki page example created by users to express the symptom Absence of voice

Bruno Kessler) and two other research labs (CELCT⁹ and LabDoc¹⁰) in order to test its efficiency and usability for collecting consumer-oriented medical knowledge. All users had to create an account for accessing the system in order to be able to fill in the Wiki pages and to provide changes. A possible bias of this acquisition task could be testing the tool for medical knowledge acquisition only over a group of people confident with informatics.

Wiki based-acquisition results. In one month (from September to October 2008), we collected 225 wiki pages, 106 for symptoms and 119 for diseases, and a total of 139 synonyms for the inserted terms. During this process, it was very interesting to also test the understanding of the collaborative nature of the Wiki system for this specific task, in fact it gave users not only the possibility to insert medical terms by creating Wiki pages, but also to update and cancel the inserted information by means of corrections and, above all, to modify Wiki pages added by other users, in order to reach a point of convergence on the

⁹<http://www.celct.it/>

¹⁰<http://www.labdoc.it/>

common sense of a medical term. In our case, users were reluctant to modify concepts added by others, even in cases of evident mistakes in definitions or categorization (only 15 out of 33 provided changes to Wiki pages created by other users, for example to add a synonym or additional information in the term description field where needed). Some examples of categorization mistakes that had not been fixed are *Singhiozzo* (Hiccup), and *Mal di testa* (headache), both categorized as “diseases” instead of “symptoms”. In some cases, when users were in doubt about the right categorization of a concept, they inserted it in both categories, as in the case of *Ustione* (Burning). Apart from some technical problems encountered by users in using this collaborative system, the test highlighted the fact that users are not completely able to categorize medical terms and to understand the difference between a disease and a symptom. This is mainly due, on the one hand, to term ambiguity and, on the other hand, to the wrong habit of using medical terms in particular discourse groups during daily life (for instance, the use of *Emicraina* (Migraine) instead of *Mal di testa* (Headache) to express just head pain). Some of the users asked us to have the possibility to start from a predefined list of technical medical terms with their official description in order to later insert the corresponding synonyms in “lay” language, since they found it really difficult to start from scratch and fill in the Wiki, because this meant spending time thinking about possible symptoms and diseases to be inserted which were not already added by other users. Furthermore, for users who did not have experience with healthcare problems (e.g. they only have the typical flu during winter, or simple headache or heartburn) it was really difficult to be creative and add new medical terms, so most of the time they used medical dictionaries or lists of medical terms on-line in order to think about the corresponding lay term to add to the Wiki.

4.2.2 Nurse-assisted Acquisition

The second elicitation technique involved nurses of a First Aid Unit in a Hospital of the Province of Trento¹¹ as a figure of mediation for the acquisition of terminology about patient symptoms and complaints (in particular the “chief complaint”, which is the nurse diagnoses recorded during triage or admission), helping patients to express their problems using the classical subjective examination performed during the Triage Process. In this clinical process collected data can be of two types: objective and subjective. The first type of data, also known as signs or evident information, are observable, perceivable and measurable data (noises, bowel sounds, body temperature, peripheral pulses, rashes, etc., measured by means of instruments and devices or diagnostic exams). The second type of data are subjective data as symptoms or hidden information, patients feelings and assertions about health problems and his clinical history. An example of objective data collected and transcribed by nurses can be the description of a wound, or blood pressure readings (180/110). One example of subjective data collected by nurses is the lay expression “Ho un forte dolore alla spalla” (I have a strong shoulder pain). Subjective data are elicited by means of interviews. In our case, we thought that the best way to collect this type of data would be to transcribe them as a direct citation of the patient, as in the example below:

Patient Number: (*Patient code*) Main problem:

Ho un forte dolore di stomaco dopo mangiato, ogni volta che mi muovo mi viene la nausea.

This acquisition method was tested in a one month period and involved 10 nurses and around 60 patients per day. During this period nurses acquired the

¹¹Specifically, we worked with the Hospital of Cles, “Medicina d’Urgenza e Pronto Soccorso del Presidio Ospedaliero di Cles (Trento)”. For further details see: <http://www.apss.tn.it/Public/ddw.aspx?n=26808>

chief complaints expressed by their patients using “lay” terminology, and transcribed them into a terminological database created ad hoc for the task, which recall the schema of the Triage record. Here nurses transcribed both the term/expression used by patients during the subjective examination (as in the example above) and the corresponding medical concept registered in the Triage record to codify patient data (i.e. the expression “Ho i crampi alla pancia” (I have a stomach ache) inserted together with the corresponding medical concept “Ad-dominalgia” (Abdominal pain)). At the end of the Triage process nurses also transcribed diagnoses assigned by a physician to the corresponding health problem/symptom expressed by patients at the arrival in the First Aid unit. Other methods for collecting chief complaints in Emergency Departments (ED) or evaluate the vocabulary used for expressing Chief Complaints in ED can be found in (Travers *et. al.*, 2001, 2008a, and 2008b [115, 117, 116])

Nurse-based acquisition results. Concerning Nurse-based acquisition, a total of 2,000 Triage records were registered in one month, which were provided to us as a table (as can be seen from Figure 4.2) including for each record the following fields: nosological code assigned in the original Triage record, Triage color assigned to the patient’s problem, the lay expression used by the patient to explain his health problem (most of the time symptoms or injuries); the corresponding medical concept; the diagnosis associated with the expressed problem (in professional language) and finally the date and time of entry. In particular, nurses transcribed 487 lay expressions associated with 1,880 medical concept, which in turn were associated with 1,880 diagnoses.

Transcriptions of patient-nurse encounters during the Triage provided, in particular, authentic utterances and contextual information for analysing patients’ active medical vocabularies, even if this type of approach was deemed too costly administratively (e.g., obtaining institutional approval and recruiting volunteer nurses to take note of their encounters with patients).

Codice Nosologico	Triage	SINTOMO RIFERITO	Sintomo principale	Diagnosi	Dataora ACC
0004965	V	presenza di sangue a livello anale in pregressa rogade	presenza di sangue a livello anale in pregressa rogade	modesta rettorragia da rogade anale	13/03/2009 20:46
0004574	B	REFERITA TOSSE	REFERITA TOSSE	Bronchiolite	16/03/2009 14:19
0005137	V	A RESPIRARE MI FA MALE LA SCATOLA DEL POLMONE	TRAUMA COLONNIA DORSO LOMBARE	TRAUMA CONTUSIVO DORSO-LOMBARE	21/03/2009 10:03
0005293	V	bevo ed urino poco	DEIDRATAZIONE E CHETONURIA	Gastroenterite	03/03/2009 09:31
0005020	V	bruciere agli occhi	BRUCIORE OCCHIO DX E SX	cheratoconjuntivite bilaterale in paziente portatrice di lenti a contatto	01/03/2009 16:52
0005520	V	continuo ad andare in bagno	REFERITA DISSENTERIA	sinrome diarroica persistente. ipertensione in paziente in terapia antipertensiva modesta anemia normocitica.	15/03/2009 23:44
0004680	V	credo che mio fratello abbia bevuto troppo	sospetto abuso etilico	stato di ebbrezza alcolica	18/03/2009 16:06
0004476	V	credo d'avere l'appendice	DOLORE ADDOMINALE DX	idropie della colecisti	04/03/2009 09:41
0005332	V	dai catetere vien fuori sangue	ematuria dopo posizionamento di catetere vescicale	ematuria dopo posizionamento di catetere vescicale	07/03/2009 21:36
0003826	B	DEVO FARE LA TETANICA	RIENTRO CONCORDATO PER RICHIAMO ANTITETANICO	RIENTRO IMPROPRIO PER SOMMINISTRAZIONE DI IG ANTITETANO	14/03/2009 20:50
0005132	V	devo fare le colture del sangue	FEBBRE INVITATA DAL CURANTE PER ESEGUIRE EMOCCULTURA	febbrile in n.d.d	09/03/2009 09:16
0003944	B	DEVO METTERMI IL FESSARIO	dolore in sede di prolasso uterino	prolasso uterino.	06/03/2009 09:48
0004085	V	DIGERISCO MALE	lombalgia persistente e esofagite cronica	LOMBALGIA DI NDD.	12/03/2009 15:43
0004209	V	DIMAGRISCE DI CONTINUO	deperimento organico	GASTRITE CRONICA.	02/03/2009 19:29
0004515	V	dopo che ho preso la botta mi scoppia la testa	cefalea post traumatica	Anorexia, scodimento delle condizioni generali in etilista cronico.	02/03/2009 06:03
0004517	V	e' arrivata l'ora	REFERITE CONTRAZIONI IN GRAVIDA A TERMINE	ematoma extracerebrale	16/03/2009 15:57
0005213	V	e' un periodo che sono stanco	referita astenia vedi richiesta m.c.	CONTRAZIONI IN GRAVIDA PRESSO IL TERMINE	15/03/2009 16:07
0004895	G	FACCIO FATICA A RESPIRARE	difficolta' respiratoria	Astenia marcata, ipertensione e anemia ingravescente in paziente in cht pre neoplasta del colon e secondarismo polmonari	03/03/2009 09:12
0004592	G	FACCIO FATICA A RESPIRARE	dispnea	Crisi asmatica.	03/03/2009 09:57
0004086	V	FACCIO FATICA A RESPIRARE	riferisce dispnea paz. noto	ADDENSAMENTO BRONCOPOLMONARE IN BPCO	15/03/2009 11:40
0005314	G	FACCIO FATICA A RESPIRARE	riferito dolore toracico anteriore e difficoltà respiratoria	INSUF. RESPIRATORIA.	17/03/2009 10:08
0005527	V	faccio pipi' rossa	ipertensione, ematuria dopo autorimozione di catetere vescicale	VERSAMENTO FLEURICO SINISTRO.	12/03/2009 10:55
0005311	V	favevo fatica a respirare per pochi minuti	riferito episodio di difficoltà respiratoria questa mattina della durata di 30" e risoltosi spontaneamente.	EMATURIA DOPO STRAPPAMENTO DI CATETERE VESCICALE.	17/03/2009 10:25
0005460	V	forse e' l'ora	CONTRAZIONI IN GRAVIDA ALLA 39 [^] SETTIMANA	episodio di dispnea in cardiopatico prodromi di travaglio.	02/03/2009 14:20
0003957	V	FORSE QUI PASSA UN NERVO CHE FA MALE	trauma distorsivo polso sinistro	trauma distorsivo del polso sin	24/03/2009 09:58
0004402	B	GO' BRUSSELE DAPPERTUTTO	riferita comparsa di esantema	Dermatite di probabile origine alimentare	15/03/2009 16:58
0004066	V	HA AVUTO CONVULSIONI	REFERITA CRISI EPILETTICA	CRISI EPILETTICA	15/03/2009 17:08
0004647	V	HA AVUTO LE CONVULSIONI	EPISODIO CONVULSIVO FEBBRILE.ORA SI E' RIPRESO.	REFERITO EPISODIO CONVULSIVO	08/03/2009 11:03

Figure 4.2: Nurse-based acquisition table extracted from the Triage records summary

4.2.3 Focus Group Acquisition

The third method used to elicit consumer-oriented medical knowledge consisted of merging three different traditional elicitation techniques: Focus Group, Concept Sorting, and Board Games, in order to allow an environment of interaction and sharing to improve the process of acquisition. The target in this case was a community of 23 elderly people in a Seniors Club (19 female and 4 male, aged from 65 to 83). We used groups activities to acquire lay terms and expressions for symptoms, diseases and anatomical concepts. In particular we distributed our sample in four groups, assigning to each group a specific body part category (i.e. head and neck, abdomen and back, arms and chest, pelvic area and legs) as their main topic in order to acquire knowledge covering all body areas. They were asked to write on little cards all known symptoms and diseases related to the assigned area, comparing their ideas with those of other members of the group to find a common definition for each of the written terms.

Focus Groups acquisition results. About 160 medical terms were collected at a two days meeting, which, at the end of the process, were analyzed together with other groups, creating discussions, exchanging opinions on terms definitions, synonyms, and recording preferences and shared knowledge. In particular, all participants gave preferences for choosing the right body system categorization (digestive, neurological, respiratory, endocrine, etc.) of each of the written concepts. This allowed us not only to extract lay terminology, but also to understand how elderly people define and categorize medical concepts, in order to compare these results with that obtained from the other two techniques mentioned. To give an example of the acquisition process, elderly people in the second group, responsible for the collection of terms related to the body area “abdomen and back”, collected lay terms such as *Fuoco di Sant’Antonio* (Shingles) or corresponding to the medical term “Herpes Zoster”, describing it as a “a painful, blistering skin rash due to the varicella-zoster virus, the virus

that causes chickenpox”, and finally they categorized it as a medical concept belonging to the “Integument System”. Table 4.3 shows the 20 most frequent medical terms extracted from the three corpora.

4.2.4 Web-based medical consultations

In a second step of this acquisition process, to extend the consumer corpus collected through the methods described above, we extracted a large written corpus from a popular Italian medical website, namely “Medicitalia”¹², which collects on-line consultations and patient forum postings. In particular, we collected 80,000 documents (postings where users/patients ask medical questions to on-line doctors) from this medical website, which represents discourse about health topics written by consumers and intended for professionals or other consumers who are interested in the same topic, already classified according to 66 medical topics (specialties from Allergy to Urology) depending on their content. Individual privacy was assured by removing names, email addresses, and other identifying information; only the subject heading, the posting category, and the “cleansed” texts were used. Each document was assigned a unique identification code: “medicitalia-”, followed by a unique number identifying the number of posts in the website section devoted to medical consultations. Figure 4.3 shows an example of a medical consultation on “Medicitalia” website, and the information we took into account. This corpus provided sufficient term coverage in the medical/healthcare domain.

Subsequently, two volunteers (a consumer and a physician) manually annotated a sample of these documents (390 extracted among all the categories, more precisely 5 per category) with semantic tags identifying medical terms, in particular: symptoms, diseases, diagnoses, anatomical concepts, medical procedures, health professionals, and finally structures or institutes providing health services. They annotated the same sub-corpus, and finally, a subsequent

¹²<http://www.medicitalia.it/>

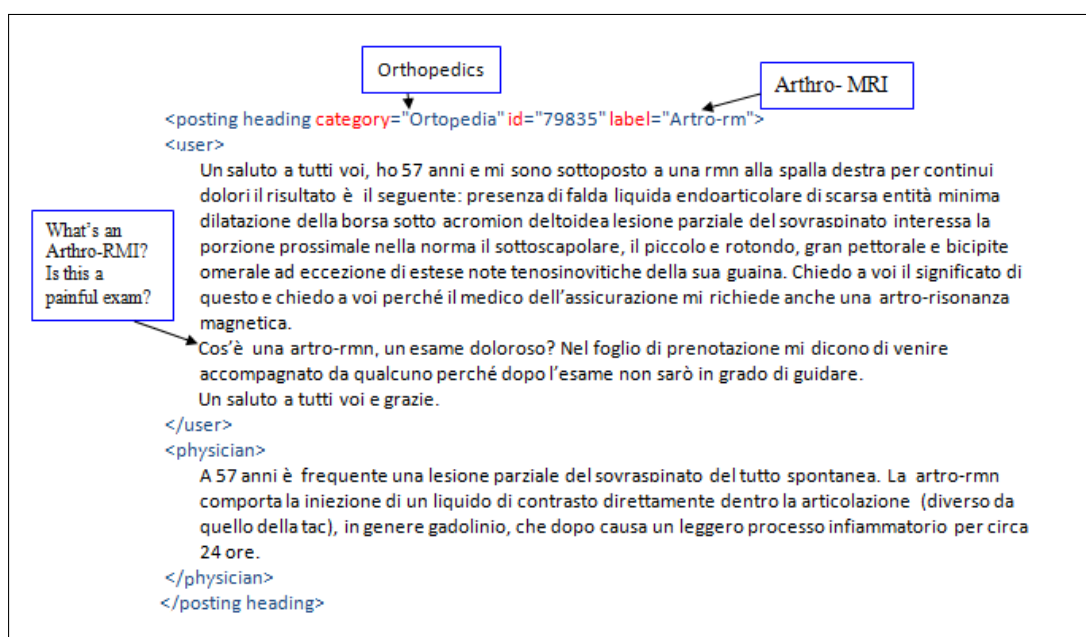


Figure 4.3: XML view of a medical consultation about “Arthro-RMI” in Medicitalia

analysis was performed by another Primary Care physician for reconciliation of the annotations. The tool used for annotating this set of documents was CLaRK (Computational Linguistics and Represented Knowledge)¹³, an XML-based System for Corpora Development, developed at the Linguistic Modelling Laboratory - CLIPPI, Bulgarian Academy of Science. This system facilitates corpus management supporting linguistic work. In fact it includes a tokenizer with a module that supports a hierarchy of token types, morphological and partial parsing analysers, a finite-state engine that supports the writing of cascade finite-state grammars, the XPath query language to support the navigation over the mark-up of a document, and other facilities. It is principally used to markup a corpus. Users work with the XML tools of the system in order to mark-up the texts with respect to an XML DTD (Simov *et al.*, 2001 [104]). Figure 4.4 shows an example of a web-based consultation annotated using CLaRK, where the user who create the posting described his various continuous symptoms such as asthenia, constipation, vomiting, tachycardia, etc. which made him in a desperate

¹³<http://www.bultreebank.org/clark/>

condition.

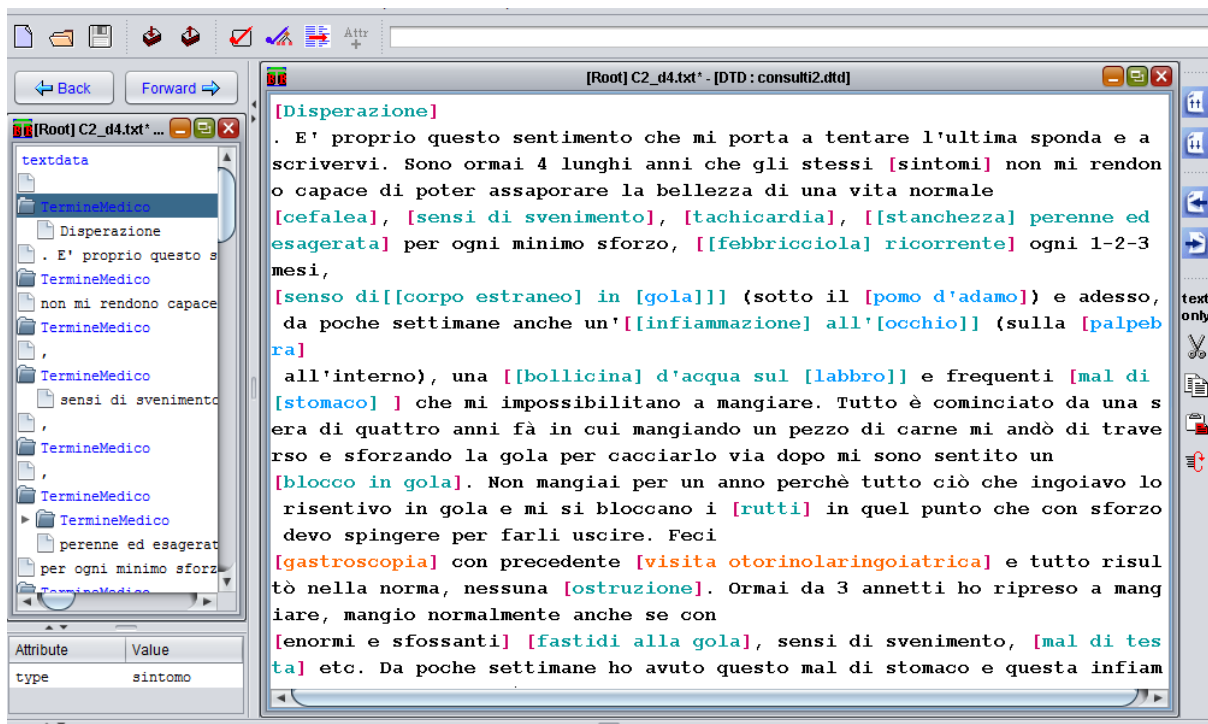


Figure 4.4: Sample marked-up document (Medicitalia-C4d2) from two annotators using CLarK

Guidelines for annotation were very simple. Practically, the two volunteers for each document of the corpus subset, selected the terms belonging to the categories mentioned above, in the figure highlighted in different colors, assigning them also attributes and co-reference relationships (where needed) with other terms in the same texts. After this annotation process, the marked-up set was used to train and test a Term Extraction tool developed at FBK - Irst, namely Keyphrase Extractor - KX - (discussed in details in the next section) in order to automatically extract significant medical terms from our “Medicitalia” corpus (even if the sample of 390 annotated documents could seem extremely small for training a machine learning term extractor this was useful for developing KX and test it for a domain-oriented extraction). We used a new term extractor here because the tool used for automatic term extraction from the other document corpus (the Text-2-Knowledge tool) was not adequate for working on such a big corpus.

After the application of the four acquisition techniques we built a collection of terms to represent numerous lay perspectives deriving from 4 different corpus summarized in Table 4.2.

Corpus	Number of documents
eHealthWiki pages	225
Triage records	2,000
Focus groups/interviews	2
Web-based medical consultations	80,000

Table 4.2: Corpus Summary

The knowledge acquisition process (which took 3 months for collecting the first 3 corpora and 2 months for collecting the web-based medical consultation corpus) was followed by the term extraction phase, which was continued over eight months.

4.3 Term Extraction, Normalization and Candidates Detection

The Term extraction process, as mentioned in the previous section, is divided into two step. The first step concerned the semi-automatic term extraction performed over the first three corpora (Wiki pages, Triage records, and Focus Groups Interviews), where the dedicated software T2K (Text-2-Knowledge) has been used. In the second step, on the other hand term extraction over the “Medicitalia” corpus was performed using the tool KX (Keyphrases Extraction).

Term Extraction using T2K.

T2K is a tool developed at the ILC (Institute of Computational Linguistic) of Pisa ¹⁴ designed for terminology extraction and ontology learning. The candi-

¹⁴<http://www.ilc.cnr.it>

date terms detected by T2K can be either single or multiword terms, and represent the terminology index of the analysed domain. The hybrid architecture of T2K, based on the combination of NLP techniques with statistical techniques, provides structured data supporting the conceptual-terminological indexing of documents (Montemagni, 1996 [82]).

T2K is particularly efficient in working on Italian corpora since the computational analysis system adopted by the tool includes a specific plug-in for the analysis of Italian. The final output provided by T2K is a term-based vocabulary, including term frequency, whose added value is represented by the terms' semantic and conceptual information regarding the vocabulary itself. These terms in fact, either single or multiword terms, are organized in a hierarchical hyponym/hypernym relation depending on the internal linguistic structure of the terms (Bartolini *et al.*, 2005 [13]); that is, by sharing the same lexical head. The steps followed to create the term-based vocabulary are shown in Figure 4.5.

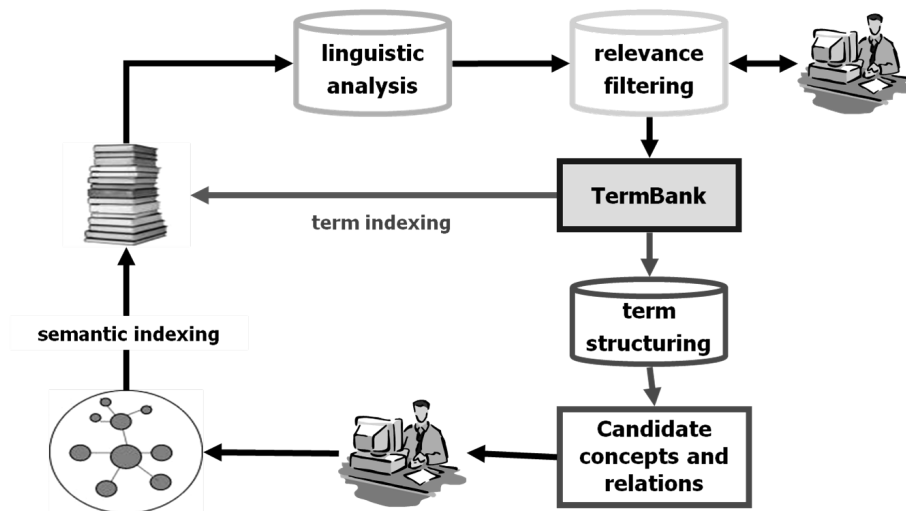


Figure 4.5: T2K at work

The final Term Bank (list of candidate single terms ranked by frequency of occurrence plus list of candidate multiword terms ranked by lexical association strength) is built by setting thresholds which can be interactively selected by users on the basis of the size of the document collection on the one hand, and

of the typology and reliability of expected results on the other hand. Generally for a medium dimension corpus the minimum frequency threshold is set as 7. Since dimensions of our corpus were limited, we set the minimum frequency threshold to 3, in order to improve results but at the expense of low recall.

By means of T2K we detected single and multiword medical terms; and a basic semantic structure defining relations between the extracted terms (BT, NT, RT). Single terms are extracted following a simple process which can be listed in 3 steps: 1. NLP Analysis on the input text and Chunking; 2. Extraction of candidate single terms considering the nominal head of chunks; and 3. Frequency computation. The final output is the list of candidate terms ranked by frequency. Multiword term extraction shares step 1 with single term extraction, and then step 2 is the extraction of candidate bigrams (chunk heads) and the application of Complex NP Grammar, concluding in step 3 with a lexical association strength, computed by applying lexical association measures (e.g. Mutual Information, log-likelihood). Example of the most frequent single and multiword medical terms extracted with T2K are shown in Table 4.3.

This table put in evidence the fact that overlaps concerning the most frequent terms extracted from the three datasets are really rare (at least considering the first 20 extracted terms), in fact we can notice that there is an overlap only between the terms “dolore” in the first column and “male” in the second column both referred to “pain”, and for the term “infiammazione”, present both in the first and the third column. This is due to the different sample used for acquiring consumer-oriented medical terms (frequent terms are different depending on the age of people used for extracting terms and also for type of elicitation techniques).

Concerning the semantic structuring step, extracted single and multiword medical terms have been structured into fragments of taxonomical chains reconstructed from the internal linguistic structure of the terms, head sharing (e.g. *abrasione corneale* (corneal abrasion) IS-A *abrasione* (abrasion)) An example

Wiki pages		Triage records		Focus groups	
Term	TF	Term	TF	Term	TF
PELLE	71	TRAUMA	1052	ERNIA AL DISCO	28
DOLORE	54	GINOCCHIO	382	MAL DI SCHIENA	28
OCCHI	35	DOLORE	365	OSTEOPOROSI	24
STOMACO	31	MANO	304	REUMATISMI	22
MALE	29	DITO	201	ARTROSI	22
FEBBRE	28	FERITA	178	TORCICOLLO	20
PERDITA	27	DISTORSIONE	164	CALCOLI RENALI	17
TESTA	26	SPALLA	163	MALE AI RENI	17
NASO	24	POLSO	147	ASMA	16
POLMONI	19	FRATTURA	137	DIABETE	15
SANGUE	19	EMORRAGIA	127	BRONCHITE	14
GOLA	17	MALE	113	GASTRITE	14
RAFFREDDORE	17	CONTUSIONE	106	COLITE	14
VISTA	17	FEBBRE	100	BRUCIORE	13
BOCCA	16	TRAUMA CONTUSIVO	87	POLMONITE	12
BRUCIORE	16	TRAUMA CRANICO	83	VERTEBRE SACRALI	12
INFIAMMAZIONE	16	CRAMPI	82	ECZEMA	12
PARTE	15	PIEDE	82	GELONI	11
TOSSE	14	DOLORE TORACICO	78	CISTITE	10
CORPO	13	CAVIGLIA	75	ALLERGIA	8
GONFIORE	13	DOLORE ADDOMINALE	69	INFIAMMAZIONE	8
PIEDI	13	GAMBA	68	ULCERA	8
PRURITO	13	VOMITO	56	PRESSIONE ALTA	8

Table 4.3: List of the 20 most frequent terms extracted with T2K from the three datasets

of hierarchical relations among the extracted medical terms is shown in Table 4.4, where we can see BT and NT relations for the terms “Cold”, “Loss” and “Inflammation”.

Broader Term	Narrower Term
RAFFREDDORE	RAFFREDDORE ALLERGICO RAFFREDDORE STAGIONALE
PERDITA	PERDITA MOMENTANEA PERDITA DI SANGUE PERDITA DELLA MEMORIA PERDITA DI CAPELLI
INFIAMMAZIONE	INFIAMMAZIONE DEL PERICARDIO INFIAMMAZIONE DELLA PLEURA INFIAMMAZIONE AI TENDINI INFIAMMAZIONE ALLA ZONA LOMBARE

Table 4.4: Fragments of taxonomical chains

Finally T2K allowed for clustering semantically related terms inferred through dynamic distributionally-based similarity measures using a context-sensitive notion of semantic similarity (computed with respect to the most relevant co-occurring heads). Examples are provided in Table 4.5 where respectively “Contusion” is related to sprain, and injury; “Hand” is related to finger, leg knee and forearm, and finally “Sprain” is related to injury, fracture, and contusion.

To conclude the analysis of terminology extraction using T2K, we summarize statistical results in tables 4.6, showing for each corpus submitted to T2K the number of documents collected, the number of extracted terms and the number of semantic relations between terms.

Term Extraction using KX.

Since the T2K term extractor, in our experience, does not work efficiently with large corpora, we decided to process the “Medicitalia” corpus composed of 80,000 web-based medical consultations with another tool with similar charac-

Term	Related Term
CONTUSIONE	DISTORSIONE TRAUMA
MANO	AVAMBRACCIO DITO GINOCCHIO GAMBA
DISTORSIONE	TRAUMA FRATTURA CONTUSIONE

Table 4.5: Clusters of related terms

Datasets	N. Docs	Terms	Broader Terms	Narrower Terms	Related Terms
Wiki pages	225	962	251	354	0
Triage records	2000	2389	507	1039	102
Focus group	2	321	94	157	24
Total	2227	3672	852	1550	126

Table 4.6: Summary of T2K Term Extraction on the three datasets

teristics, the already mentioned Keyphrase Extractor “KX”, developed at FBK, which exploits basic linguistic annotation combined with simple statistical measures to select a list of weighted keywords from a document/corpus (Pianta and Tonelli, 2010 [92]).

Keyphrases are expressions, either single words or phrases, describing the most important concepts of a document. In our task consumer-oriented medical terms can be considered as keyphrases to be extracted from consumer postings on medical consultations. We choose KX because of his flexibility and adaptability to the domain and to Italian (it has been tested on Italian corpora related to specific domains and has provided very good results).

KX has a very simple architecture based on 4 steps which can be summarized as follows:

1. Extract from corpus C the list “NG-c” of corpus n-grams (where an n-gram is any sequence of tokens in the text), for instance “the body systems”. The maximum length of the selected n-grams can be set by the user. In our task we selected 2-, 3-, and 4-grams.
2. Select from the list NG-c a sub-list of multiword terms “MW-c”, that is, combinations of words expressing a unitary concept, for instance “varicose vein”;
3. For each document in C, recognize and mark the multiword terms. Calculate the inverse document frequency (IDF) for all words and multiword terms in the corpus;
4. Given a document d, count all words and multiword terms and rank them.

Concerning step 1, in our case, even if n-grams occur a few times, they are very likely to be useful for keyphrase recognition from our corpus. During this step a blacklist was used in order to exclude n-grams containing any of the language and domain specific stopwords in the list. In Step 2, those n-grams that matched certain lexical patterns were selected as multiword terms. An example of a lexical pattern is the following:

[N] [O] [NASPGLU] [NAU]

This pattern means that a 4-gram is a candidate multiword term if it is composed of a Noun followed by *di* (with all its variations) - “of” - or *per* - “for” - defined as O, followed by either a Noun, Adjective, Singular noun, Past participle, Gerund, punctuation (L) or Unknown word, followed by either an Adjective, Noun or Unknown word. This is matched for example by the following 4-gram:

ciclo[S] *di*[O] *chemioterapia*[N] *adiuvante*[A]
 “adjuvant[A] chemotherapy[N]”

Furthermore, to recognize multiword terms by local (document) and global (corpus) evidence, only simple frequency has been used as a selection criterion.

In this step two frequency thresholds were set: MinCorpus, which corresponds to the minimum number of occurrences of an n-gram in a reference corpus (in our task it has been set initially to 14 and then to 10 to improve results), and MinDoc, which is the minimum number of occurrences in the current document (in our task set as 2). This means that KX marked an n-gram in a document as a multiword term if it occurred at least MinCorpus times in the corpus or at least MinDoc times in the document (Pianta and Tonelli, 2010 [92]). Examples of marked multiword terms after this step are:

Distorsione della caviglia destra

Distorsione della caviglia

Caviglia destra

To perform step 4, a new document *d*, not included in Corpus *C*, was taken into account to extract keyphrases. First, multiword terms were recognized and marked, through the same algorithm used in Step 3. Then, frequencies of words and multiword terms in *d* were counted to obtain a first list of keyphrases, ranked according to frequency. The re-ranking of the frequency-based list of keyphrases was performed using some parameters such as normalized IDF, keyphrase length, position of first occurrence, and other parameters. The final output delivered by KX is a list of keyphrases (consumer-oriented medical terms) ranked by relevance, as shown in table 4.7.

As mentioned in the previous section, among the 80,000 documents in the web-based medical consultations corpus, we marked-up set of 390 documents (including 6,866 annotated terms for an average of 17,6 terms per document). This sample then was used to train and test KX. More precisely, we split these marked-up documents into a training/development set of 196 documents and a test set of 194 documents. In the evaluation task the system ran on the test set with the best performing parameters combination, and in addition we used a gold standard of keyphrases, created by physicians and consumers, as a baseline.

keyphrase	Relevance score
risonanza magnetica	631.21
medico di base	529.96
analisi del sangue	310.97
dolore	303.24
risonanza magnetica aperta	289.82
fuoriuscita di sangue	275.68
esami del sangue	256.08
attacchi di panico	247.63
punti neri	237.19
episodi di tosse	221.63
mandibola	195.09
rotazione della tibia	194.93
mezzo di contrasto	189.02
mancanza di forza	185.65
dolore al petto	184.07
risonanza magnetica alla spalla	171.26
ernia mediana	161.22
perdita di sangue	161.13
dolore al polso	151.65
intervento chirurgico	151.43
...	...

Table 4.7: Example of KX output list of keyphrases

Concerning the automated extraction of consumer-oriented medical terms from the “Medicitalia” corpus, the evaluation results on the test set showed a low value of Precision and Recall, both at 29%. This was due to the high threshold applied to the system (14-10) but also to the numerous non-medical terms in the corpus. Even if this result could seem negative and if there are some limitations on the evaluation process (very small sample used for testing the system), considering the particular type of term we wanted to extract and considering literature in evaluations of term extraction systems on Italian medical corpora we see that the the most promising results do not exceed 50% of Precision and Recall. Among 2,239 terms extracted by KX, only 651 were rec-

ognized as medical keyphrases by physicians. But after a manual review of KX output, performed by consumers, the number of candidate medical terms increased to 989. These medical terms are categorized as follows: 261 are related to anatomy, 74 are medical procedures, 30 are related to medical devices and instruments for patient care, 17 are Healthcare structures and institutes which provide health services, 17 measurement parameters, 143 are symptoms, 360 are diseases, and 87 are diagnoses.

In spite of the advantages of the automatic extraction process (both using T2K and KX), allowing for extraction of many compound terms, such a procedure has demonstrated that a large number of terms, certainly representative of consumer medical terminology, were not automatically extracted, probably due to both the quantitative limits of the corpus dimensions and to domain specificity. Consequently, we performed an additional manual extraction to take into account these rare terms, usually mentioned by only a single participant. Manual extraction in this study provided useful insights that may be used in the future to improve automated extraction algorithms for non-professional terms.

4.4 Clinical Review

Medical terms extracted using KX (989 terms) have been integrated with the one extracted using T2K (1,938 terms if we consider only terms related to medicine), so 3,000 terms were further reviewed by 5 physicians and 3 pharmacists (chosen for their experience as mediators with respect to physicians in dealing with healthcare consumers and patients) to find mistakes and incongruities in categorization and synonymy. Manual review of all the collected terms by physicians served principally for quality assurance. First, the quality of the extractions could be evaluated: potential forms not selected by the term extractors but deemed relevant could be added; ambiguities in categorization could be solved; long strings (i.e., > 6 words) could be shortened; and some ir-

relevant or non-medical terms deleted (e.g., forms pertaining to family relationships, depending on context). Possible and categorizing types of mismatches from automated mechanisms include:

- Misspellings (e.g., “Erpes” and “Ictus celebrale”);
- Truncation, such as the name in an eponymic term, rather than the full term (e.g., “Down” for “Morbo di Down”);
- Abbreviations (e.g., Dist. al ginocchio);
- Clippings or word fragments;
- Lexical variation, in general (addressed by normalization);
- Non-medical domain forms (e.g., “pausa pranzo” - Lunch break)

In particular, many mistakes were found by physicians in the first set of terms (Wiki-based), where a wrong categorization was assigned to 25 terms, and where wrong synonyms were expressed for 8 terms. They found similar incongruities in the third set (Elderly people), where wrong categorizations were assigned to 40 terms, e.g. “Giramento di Testa” or “Vertigini” (Vertigo or Dizziness), categorized in the Cardiovascular System instead of the right Neurological one. Wrong categorization and misspellings were also recognized analysing the term related to web-based medical consultations extracted with KX. Concerning the nurse-assisted data set, clinical review was directly performed first by a nurse and then by a physician during the process of Triage.

Approximately 5% of all terms from each corpus were modified and about 100 terms were deleted. these terms concerned were related above all to specifications of anatomical parts (4-grams such as “lato destro del piede”), diagnoses (such as “riduzione della frattura” and “quadro clinico compatibile”), units of measurement, combinations such as “50 mg”, and finally all that general symptoms expressed by superlative adjectives to highlight the intensity of pain (such as “fastidio fortissimo”, “fortissimo dolore”, “bruttissima caduta”).

4.5 Consumer Familiarity with Health Terminology

After the review performed by physicians, a familiarity degree was assigned to each consumer-oriented medical term (2,400 in total), which represents the level of understandability and use of a certain medical term for healthcare consumers. Term familiarity is not binary (i.e., known versus unknown) some terms may be understood by 90% of the target lay audience, whereas others may only be understood by 50%. Additionally, it is likely that the level of understanding of the meaning will vary considerably, as a single term may be well understood by some people but only partially understood by others. In order to estimate consumer familiarity with medical terms collected during the knowledge acquisition and the term extraction steps, we used a mechanism to quantify health terms as being more likely or less likely to be understood by typical members of the lay public related to the use of the terms instead of another with the same meaning. We submitted the collected terms to a sample of 80 people distributed around Italy and covering a wide range of ages, who had been contacted by means of social networks and who were asked to assign to each term a familiarity degree on a scale from 1 to 5 (1 means not familiar at all, “a term rarely used”, while 5 means very familiar, “a term used very often in daily life to identify a certain concept with respect to other synonymous terms”). At the end of this analysis, for all the terms in the collection, we computed the average of each of the 80 familiarity values assigned by users in order to have the candidate familiarity degree for each term, to be added as another term attribute in the ICMV (in this case we didn't considered variation or standard deviation). Results showed that 942 terms were considered very familiar and usually used by consumers, having “5” as familiarity degree, 592 terms well known and often used by consumers (with familiarity degree “4”), 350 terms were assigned degree “3”, 358 terms were considered not familiar and only used by people with a good healthcare literacy, and finally the rest of terms (130) were consid-

ered too specific of the domain to be used by consumers. Other approaches to estimate consumer familiarity with medical terms can be found in (Keselman *et al.*, 2007a [67]) where consumer familiarity was predicted using contextual information.

4.6 Finding ICMV correspondences in ICPC2

As final step of the generation of ICMV, a second clinical review was performed by physicians who, this time, were asked to find ICMV correspondences in a specialized terminology. More specifically, we asked them to map a term/medical concept pair by using the above mentioned International Classification for Primary Care 2nd Edition (ICPC2-E, electronic version) [89], which has been widely distributed and received great praise and attention within the European Union and world wide. It addresses fundamental parts of the healthcare process: it is used in particular by general practitioners for encoding symptoms and diagnoses. As we will see in details in the next chapter, it has a biaxial structure that considers medical concepts, related to symptoms, diseases and diagnoses, and medical procedures, according to 17 Problem Areas/Body Systems, for a total of 682 rubrics (concepts) divided as follows: 319 are symptoms and complaints, 40 are medical procedures, and 363 are diagnoses and diseases. While for the terms extracted through the three elicitation techniques (Wiki, Triage records, and Focus groups with elderly people) we considered only symptoms and diseases for the mappings (as shown in the results), for the terms extracted from the “Medicitalia” corpus physicians included also ICPC2 procedures in the process of manual mappings.

By means of this mapping between ICMV “lay” terms and ICPC2 concepts we want to reconstruct the meaning (concept) inherent in the lay usage of a term, and then to show that compatibility between lay and professional terms exists on the basis of this deeper meaning, rather than on the basis of the lexical

form. We identified five different types of relations between consumer terms and ICPC2 medical concepts:

- Exact mapping between the pairs; this occurs when the term used by a lay person can be found in ICPC2 rubrics and both terms correspond to the same concept. For instance, the lay term “Febbre” (Fever) would map to the ICPC2 term “Febbre”, and both will be rooted to the same concept.
- Related mapping; it involves lay synonyms and occurs when the lay term does not exist in the professional vocabulary, but corresponds to a professional term that denotes the same (or closely related) concept. For instance, the lay term “Sangue dal Naso” (Nosebleed) corresponds to “Epistassi” (Epistaxis) in ICPC2.
- Hyponymy relation; this occurs when a lay term can be considered as term of inclusion of an ICPC2 concept. For example, the lay term “Assenza della Voce” (Absence of Voice) is included in the more general ICPC2 concept “Sintomo o disturbo della voce” (Voice Symptom/Complaint).
- Hypernymy relation; in this case the lay term is more general than one or more ICPC2 concepts, so it can be considered as its/their hypernym. For example, the term “Bronchite” (Bronchitis) is broader than “Bronchite Acuta/ Bronchiolite” (Acute Bronchitis/ Bronchiolitis) and “Bronchite Cronica” (Chronic Bronchitis) ICPC2 concepts.
- Not Mapped; this includes lay terms that cannot be mapped to the professional vocabulary. These can be legitimate health terms, the omission of which reflects real gaps in existing professional vocabularies, or they can represent unique concepts reflecting lay models of health and disease. For example, the lay term “Mal di mare” (Seasickness).

Finding ICMV correspondences in ICPC2 results.

As we have previously mentioned, our methodology of acquisition allowed us to acquire varied consumer-terminology and to perform an interesting terminological and conceptual analysis. The Tables below present term extraction and mapping evaluations in terms of a statistical analysis. By means of a term extraction process we were able to extract a total of 962 medical terms from 225 Wiki pages, 375 of which were not considered pertinent to our aim. We thus performed mapping analysis only for 587 terms (61% of the extracted terms) as summarized in Table 4.8. We observe that most of the exact mappings with

	Terms	Exact Matches	Synonyms	Hyponyms	Hypernyms	Not Mapped
Symptoms	306	26	50	68	39	123
Diseases	140	42	19	28	17	34
Anatomy	141	88	11	6	4	32
Other	375	0	0	0	0	0
Total	962	156	80	102	60	189

Table 4.8: Wiki term collection

ICPC2 are related to anatomical concepts (56.41%), and that many synonyms in lay terminology and inclusion terms were found for symptoms (respectively 62.5% and 66.66%). Table 4.9 shows the results related to the Triage acquisition data.

From 2,000 Triage records, we extracted a total of 2,389 terms, but about half of these terms were considered irrelevant for our evaluation, so even in this case we present a mapping analysis only for 1,108 terms (46.37%). Contrary to our previous results, here we can highlight on only the high presence of lay terms used for expressing symptoms with exact mappings to ICPC2 (40.93%), but also many synonyms in lay terminology for ICPC2 symptoms and diseases

	Terms	Exact Matches	Synonyms	Hyponyms	Hypernyms	Not Mapped
Symptoms	508	122	157	12	40	177
Diseases	325	70	75	8	27	145
Anatomy	275	106	65	32	12	60
Other	1,281	0	0	0	0	0
Total	2,389	298	297	52	79	382

Table 4.9: Nurse-assisted term collection

(respectively 52.86% and 25.25%). This is particularly related to the context chosen for acquisition, where patients just ask for help about suspected symptoms and complaints.

Table 4.10 shows the results related to the data acquisition from elderly persons. Among 321 medical terms extracted by the Focus Group/Game activity, 243 were considered for analysis by mapping to ICPC2 (75.70%). The results showed that most of the terms in this case were mapped to ICPC2 and only 12 do not have any correspondence. Here it is interesting to observe that all the symptoms extracted had a corresponding medical concept in ICPC2 (24.61%).

	Terms	Exact Matches	Synonyms	Hyponyms	Hypernyms	Not Mapped
Symptoms	79	30	39	5	5	0
Diseases	87	23	47	6	7	4
Anatomy	77	48	11	8	2	8
Other	78	0	0	0	0	0
Total	321	101	97	19	14	12

Table 4.10: Focus Group/Games with Elderly Person

Concerning the mapping to ICPC2 of the lay terms extracted by the “Medic-

italia” corpus, we can see from Table 4.11 that physicians didn’t find mappings for about half of them. In fact among 989 terms they found only 526 mappings to ICPC2 (53.18%).

	Terms	Exact Matches	Synonyms	Hyponyms	Hypernyms	Not Mapped
Symptoms	143	54	60	12	8	9
Diseases	360	89	97	50	41	81
Medical procedures	74	18	25	7	3	21
Health devices	30	0	0	0	0	30
Health facilities/professionals	17	0	0	0	0	17
Measurement parameters	17	1	1	3	1	11
Diagnoses	87	6	11	21	11	26
Anatomy	261	1	1	3	1	255
Total	989	169	195	96	66	462

Table 4.11: Mappings between “Medicitalia” lay terms and ICPC2 concepts

Considering that besides symptoms and diseases we extracted from the “Medicitalia” corpus there were also concepts related to anatomy (the majority), health facilities and professionals, health devices and medical procedures, if we map these terms to a coding system such as ICPC2 which includes only diseases and diagnoses, symptoms and few medical procedures, it is expected that in the end the correspondences for the type of concepts mentioned above cannot find a good coverage (only 31.95% exact matches for symptoms, 52.66% for diseases and 43.78% for medical procedures, while concerning synonyms we have only 30.76% for symptoms and 49.74% for diseases). Furthermore, we can observe a great number of synonyms with respect to exact matches, which is a situation occurring most of the time for very lay expressions which have in ICPC2 a technical correspondence (such as “Orecchioni” and “Parotite Epidemica”). Then it is possible to observe also a good percentage of hypernyms, terms which are

more specific than the ICPC2 rubrics (e.g. the term “Controllo ecografico” - Ultrasound examination - mapped to the ICPC2 procedure “Esami radiologici o per immagini” - Diagnostic Radiology/Imaging). This highlights the fact that ICPC2 treating of reasons for encounters in Primary Care has not a high granularity; in fact, some rubrics related to symptoms and diseases (such as Injuries, Neoplasms, Disabilities, Fear, etc.) are very generic and they occur repeatedly across all 17 chapters, and this is confirmed also by the fact that there are many “other” or “Not Elsewhere Classified - NEC” rubrics.

Table 4.12 compares the four data sets together and shows that the most profitable methodology for acquiring consumer-oriented medical terminology was the one assisted by Nurses. However, the limit of this method is that it is time-consuming for nurses who have to report all patient “lay” health expressions. Also the Wiki-based method, even if not exploited for the collaborative characteristic, has demonstrated good qualitative and quantitative results. Concerning the third method we can say that, in order to be compared with the other two in terms of quantitative results, it must be applied more than twice (more meetings with elderly persons are needed to collect a comparable number of lay terms). On the contrary we have interesting results related to the mapping process, because almost all the terms extracted are covered by ICPC2 terminology.

Sources	Total Terms	Mapped	Not Mapped
Wiki-based	587	398	186
Nurse-assisted	1108	726	382
Focus-Group	243	231	12
Web-based consultations	989	526	462
Total	2927	1881	1042

Table 4.12: Results overview

The fourth approach related to the mapping of lay terms extracted from the “Medicitalia” corpus to ICPC2 has the advantage of consuming less time and cost since terms are extracted directly from the web, but on the contrary, it

becomes less reliable from a qualitative point of view, first of all because of the percentage of errors occurring in the medical postings, compared to the ambiguities which cannot be solved by the term extractor. Furthermore, this approach required the longest clinical review by physicians.

As a final evaluation concerning the process of lay medical terminology/-knowledge acquisition we can observe that the most frequently appearing forms and concepts in a consumer-oriented medical vocabulary represented symptoms and anatomical parts, while those in specialized terminologies are more related to epidemiology. Furthermore, different types of documents within genres varied in the number of tokens and topics covered. Medical consultations tended to be longer and covered more topics with respect to Wiki pages, Focus Groups or Triage encounter transcriptions. To conclude, we have to highlight that when comparing the first three sets of extracted terms, the overlap is of only 60 relevant consumer medical terms, while comparing these three sets with the fourth one the overlap reached 579 terms, whereas the total number of overlaps regarding the mapping between lay terms extracted from our datasets and ICPC2 is about 360. These overlaps have been deleted in order to have a final number of unique ICMV terms and mappings pairs, having at the end a collection of 2,348 validated consumer-oriented medical terms to be included in our ICMV and a total of 1,521 mappings to ICPC2. On the other hand, among a total of 682 ICPC2 rubrics, 508 mapped at least one time to our ICMV concepts. This means that all the other mapped terms can be considered synonyms or quasi synonyms of ICPC2 concepts. The large number of unmapped terms and the low overlap between the first three datasets demonstrate that we extracted a very wide range of medical terms, many compound terms and expressions, which can be representative of the corresponding technical terms present in standard medical terminologies, and which can be used as a candidate for the construction of our Consumer-oriented Medical Vocabulary for Italian.

Chapter 5

Formalizing Medical Terminologies in Semantic Web Languages*

The chapter introduces the second part of the thesis which is dedicated to the formal representation of standard medical terminologies and our ICMV, using Semantic Web languages, namely RDF and OWL. The Principal aim of this formalization process is to allow the semantic “Integration” of these terminologies and ontologies, as outlined in Section 2.1 and to improve their usability and sharing in a Semantic Web context.

The chapter is organized into two main Sections which describes different approaches for formalizing medical terminologies or classification systems. In fact Section 5.1 presents a method for encoding medical classification systems in OWL light-weight ontologies providing also the logical analysis of their clinical mappings. In this case we applied this process to ICPC2 and ICD10. On the other hand Section 5.2 introduces a second approach which consists in the conversion of a subset of standard medical terminologies and classification systems using RDF. In this case we applied this approach on one hand to a subset of UMLS vocabularies, in particular LOINC, ICD10, ICPC2, SNOMED CT and MeSH (using RDF N-triples format), and to our consumer-oriented vocabulary

***Acknowledgements:** The material of this chapter is based on earlier publications [26, 27, 25]

ICMV (using RDF/S format). Finally Section 5.3 closes the chapter showing some experimental results for the two formalization approaches.

5.1 Encoding ICD10 and ICPC2 in OWL

As mentioned in the State of the Art, Section 2.3.1, during the last few years, the need has arisen for representing in a formal language knowledge contained in medical resources and for establishing unambiguous mappings between different coding systems to guarantee their interoperability. Digitalization of medical coding systems in health records improves the accessibility, exchange, and analysis of medical data. Here, Semantic Web languages, and in particular the Web Ontology language - OWL (Bechhofer *et al.*, 2004, Patel-Schneider *et al.*, 2004 and Smith *et al.*, 2004 [14, 91, 107]), provide an opportunity for the medical community to build formal, sound and consistent medical terminologies, and to provide a standard web accessible medium for interoperability, access and reuse. Before describing the process of OWL conversion of the medical terminologies taken into account we provides in the paragraph below a brief introduction to OWL.

Web Ontology Language (OWL) overview.

As mentioned in Section 2.3.2, OWL is the standard Semantic Web language for representing ontologies. Currently, two releases of the Web Ontology Language are recommended by W3C: OWL 1.0 as of February 2004², used for our approach, and OWL 2 since 2009³. OWL includes three increasingly powerful sublanguages which can be used according to the formalism we want to give to our ontology and to the performances in reasoning and inference we want to obtain: OWL Lite, OWL DL, and OWL Full. The first one is the least powerful OWL sublanguage, in fact it allows for representing taxonomies, while OWL

²<http://www.w3.org/TR/owl-features/>

³<http://www.w3.org/TR/owl2-overview/>

DL (Description Logic) is a sublanguage of OWL Full which imposes restrictions on the use of OWL/RDF constructors. OWL Full is the most expressive one using all the OWL language primitives. All these varieties of OWL 1.0 use RDF for their syntax, so instances are declared using RDF descriptions and typing information, and some of the OWL constructors, such as *owl:Class*, *owl:DatatypeProperty* and *owl:ObjectProperty* are specializations of their RDF/S counterparts.

The OWL 1.0 language is built on formalisms of the Description Logic (DL) family and therefore allow reasoning and inference. Reasoning is the act of making implicit knowledge explicit. For example, an OWL knowledge base containing descriptions of students and their parents could infer that two students exhibited the brother relationship if both were male and shared one or more parent. No explicit markup indicating the brotherhood relationship need ever have been declared. A Reasoning Engine is computational machinery that uses facts found in the knowledge base and rules known a priori to determine Subsumption, Classification, Equivalence, and so on (Staab and Studer, 2004 [111]). F-OWL, FaCT, and Racer are examples of such engines.

5.1.1 ICPC2 and ICD10 Background

The work done in this phase addresses one aspect of the challenge described above, that is how to incorporate structured lexical information such as definitions, synonyms, usage notes, inclusion and exclusion criteria, etc., into the OWL 1.0 ontology model in a standardized and consistent way. For this purpose we selected and analysed two important international coding systems which received growing interest from countries of the European Union during the last years: the International Classification of Primary Care - 2nd edition (ICPC2), which is our reference medical terminology in this study, and the most used coding system all over the world, the International Classification of Disease - 10th edition (ICD10), directly mapped to ICPC2. Both classification systems

are presented in detail in subsection 5.1.1. The choice of these two classification systems is not random but is due to the fact that ICPC2 and ICD10, in particular the first one, are close to the vision of healthcare consumers, in terms of semantics and structure, since they address fundamental parts of the healthcare process, such as symptoms, diseases, diagnoses, and healthcare procedures. In particular, ICPC2 classify medical concepts at a general level and it is used by general practitioners in many European countries to encode primary care reasons for encounters - RFE - (symptoms, diseases and medical procedures), while ICD10 is used mainly by clinicians in hospitals for encoding diseases. ICD10 and ICPC2 are clinically integrated, in fact a technical mapping between them has been made in Wood *et al.*, 1992 and Okkes *et al.*, 2000 [121, 89], allowing primary care physicians to implement ICD10 as a reference nomenclature within the classification structure of ICPC2 and leading to a substantial increase in the diagnostic potential of ICPC2.

The general idea pursued in this step is the following: (1) encode ICPC2 and ICD10 systems into lightweight OWL ontologies, (2) encode the existing ICPC2-ICD10 clinical mappings as inter-ontology axioms, (3) apply logical reasoning to analyse their coherence and consistency by means of reasoning, using Pellet OWL Reasoner⁴. Next paragraphs briefly recall the main characteristics of the two classification systems to be encoded in OWL 1.0.

The International Classification of Diseases 10th Revision - ICD10.

ICD10 is the tenth revision of the International Classification of Diseases published, as already mentioned, by the World Health Organization (WHO)⁵. The goal of the system is to allow the systematic collection and statistical analysis of morbidity and mortality data from different countries around the world. Due to its importance, ICD acts as a *de facto* reference point for many healthcare terminologies.

⁴<http://pellet.owldl.org>

⁵<http://www.who.int/en/>

The ICD10 is structured as a multi-axial classification, which includes 21 chapters, diseases and diagnostic categories, including a total of 14.000 rubrics⁶ distributed among the chapters. At its core, the basic ICD is a single list of rubrics, identified by an alphanumeric code with 3 positions, organized by category, from A00 to Z99 (excluding U codes which are reserved for research, and for the provisional assignment of new diseases of uncertain aetiology). The first character of the ICD code is a letter associated with a particular chapter.

The 3-character code allows to divide each chapter into homogeneous blocks reflecting different axes of classification. Each ICD category can be further divided into up to 10 subcategories, if more details are required, using a fourth numeric character after a decimal point. This is used, for example, to classify histological varieties of neoplasms. A few ICD chapters adopt 4 characters to allow further sub-classification along different axes. For example, as shown in Figure 5.1, in Chapter II – Neoplasms (letters C and D) the first axis is the behaviour of the neoplasm, and the next is its site. A neoplasm of pancreas, belonging to the class *Malignant neoplasms of digestive organs* (code range C15-C26), has code C25, and in turn has other subcategories representing specifications of the localization of the neoplasm. In this case C25 subcategories, identified with the fourth digit code are C25.0 “Head of Pancreas”, C25.1 “Body of Pancreas”, etc. until C25.9 “Pancreas, Unspecified”. Generally the .9 subcategory is used for classifying “other” or “unspecified”, while the .8 subcategory is used for “not elsewhere classified”.

The International Classification of Primary Care 2nd Edition - ICPC2.

In primary care, many symptoms and non-disease conditions are difficult to code in ICD, which has a disease-based structure. For this reason, the International Classification of Primary Care (ICPC) was created to codify the vague and ill-defined conditions for which patients contact their general practitioner. ICPC was published in 1987 by WONCA, the world organization of

⁶A rubric in a classification system roughly corresponds to a class or a category

Chapter I	Infectious and parasitic diseases
Chapter II	Neoplasms
Chapter III	Diseases of the blood and blood forming organs and certain disorders affecting the immune mechanism
Chapter IV	Endocrine, nutritional and metabolic diseases
Chapter V	Mental and behavioral disorders
Chapter VI	Diseases of the nervous system
Chapter VII	Diseases of the eye and adnexa
Chapter VIII	Diseases of the ear and mastoid process
Chapter IX	Diseases of the circulatory system
Chapter X	Diseases of the respiratory system
Chapter XI	Diseases of the digestive system
Chapter XII	Diseases of skin and subcutaneous tissue
Chapter XIII	Diseases of musculoskeletal system and connective tissue
Chapter XIV	Diseases of the genitourinary system
Chapter XV	Pregnancy, childbirth and the puerperium
Chapter XVI	Certain conditions originating in the perinatal period
Chapter XVII	Congenital malformations, deformations and chromosomal abnormalities
Chapter XVIII	Symptoms, signs and abnormal clinical and laboratory findings
Chapter XIX	Injuries, poisoning and certain other consequences of external causes
Chapter XX	External causes of morbidity and mortality
Chapter XXI	Factors affecting health status and contact with health services of a person not currently sick

Table 5.1: The ICD10 chapter headings

Chapter	Blocks	Title	
I	A00-R99	Chapter	Description
II	C00-D48	Chapter II	Neoplasms
III	D50-D89	(C00-C75)	Malignant Neoplasms
IV	E00-E90	(C00-C14)	Lip, oral cavity and pharynx
V	F00-F99	(C15-C26)	Malignant neoplasms of digestive organs
VI	G00-G99	C25	Malignant neoplasm of pancreas
VII	H00-H59	C25.0	Head of pancreas
VIII	H60-H95	C25.1	Body of pancreas
IX	I00-I99	...	
X	J00-J99	C25.9	Pancreas, unspecified
XI	K00-K93		
XII	L00-L99		
XIII	M00-M99		
XIV	N00-N99		
XV	O00-O99		
XVI	P00-P96		
XVII	Q00-Q99		
XVIII	R00-R99		
XIX	S00-T98		
XX	V01-Y98		
XXI	Z00-Z99		
XXII	U00-U99		

External causes of morbidity and mortality
 Factors influencing health status and contact with health services
 Codes for special purposes

Figure 5.1: Example of taxonomy in ICD10 for the concept “Neoplasm of pancreas”.

general practitioners, to allow the classification of three important elements of the healthcare encounter: reasons for encounter (RFE), diagnoses or problems, and process of care. In problem-oriented medical records, these elements allow to classify an episode from the beginning with an RFE to its conclusion with a more defined problem. The current revision is ICPC2, issued in 1998; an electronic version of ICPC2, referred to as ICPC2-E, was released in 2000 (Okkes *et al.*, 2000 [89]), and updated in 2005 to ICPC2-R. This last version is the one we used in this task.

ICPC has a biaxial structure with 17 chapters, identified by a single alpha code, divided into seven components (medical entities), identified by a range of two-digit numeric codes that are not always uniform across chapters. The structure of ICPC is reported in Table 5.2.

Unlike in ICD10, which includes separate chapters for different diseases, in ICPC2 the reason for encounters are distributed among chapters, depending on the body system to which they belong. Component 7 (*Diagnoses and diseases*) is further organized in five subgroups, which are not numerically uniform across chapters: infectious diseases, neoplasms, injuries, congenital anomalies, and

Components	Chapters																
	A	B	D	F	H	K	L	N	P	R	S	T	U	W	X	Y	Z
1. Symptoms																	
2. Diagnostic, screening, prevention																	
3. Treatment, procedures, medication																	
4. Test results																	
5. Administrative																	
6. Other																	
7. Diagnosis, disease																	

- | | | |
|-------------------------|------------------------------------|-------------------------------|
| A. General | L. Musculoskeletal | U. Urinary |
| B. Blood, blood forming | N. Neurological | W. Pregnancy, family planning |
| D. Digestive | P. Psychological | X. Female genital |
| F. Eye | R. Respiratory | Y. Male genital |
| H. Ear | S. Skin | Z. Social |
| K. Circulatory | T. Metabolic, endocrine, nutrition | |

Table 5.2: The structure of ICPC

other diseases. For example, the rubric corresponding to malignant neoplasms of pancreas (code D76) belongs to component 7 (*Diagnoses and diseases*), subgroup *Neoplasms*, and chapter D (*Digestive*).

ICPC2 components 2 to 6 are common throughout all chapters, each rubric being equally applied to any body system.

ICPC-ICD Relationship .

A technical mapping between ICPC and ICD10 has been made by specialists of the domain in collaboration with WICC-WONCA and WHO, allowing primary care physicians to implement ICD10 as a reference nomenclature within the classification structure of ICPC and leading to a substantial increase of the diagnostic potential of ICPC (Wood *et al.*, 1992 [121]). This mapping was revised with the release of ICPC2 and its electronic edition ICPC2-E (Okkes *et al.*, 2000 [89]). Because ICD10 and ICPC were designed for different purposes, in many cases the mapping could not be done on a one-to-one basis, considering also that the diagnostic ICD10 classes at the three digit level are far more

specific than any primary care classification. In making the clinical mapping (which is a manual mapping), three situations arose.

1. A set of three-digit ICD10 rubrics were compatible on a one-to-one basis with a three-digit rubric in the first or seventh component of ICPC.
2. A set of ICD10 three-digit rubrics had to be broken open into four-digit-rubrics for at least one compatible mapping to one or more ICPC rubrics.
3. To allow the compatible mapping to the remaining rubrics of ICPC, the remaining of ICD10 rubrics, either on the three- or four-digit level, had to be grouped into a combination of classes.

In Tables 5.3, 5.4 and 5.5 it is possible to see an example of the three situations and the corresponding mappings. In general, the result is that one ICPC rubric may be mapped to n ICD10 (three or four-digit) rubrics and one ICD10 (three or four-digit) rubric may be mapped to m ICPC rubrics.

ICPC	ICD10
A71 Measles	B05 Measles
D76 Malignant neoplasm pancreas	C25 Malignant neoplasm of pancreas
S81 Angiomatous birthmark, portwine stain	D18 Haemangioma and lymphangioma, any site

Table 5.3: One-to-one mappings between ICPC and ICD10

5.1.2 ICPC2, ICD10 and their clinical mapping in OWL

To support further formal analysis of ICPC-ICD mappings, we encoded ICPC2 and ICD10 classifications using OWL. This additionally allows us to re-use a range of techniques and implemented tools for reasoning with OWL.

In the conversion of our medical coding systems to OWL we had to preserve two important properties of the classifications: the disjointness of nodes and the

ICPC	ICD10
Y14 Family planning male, other	Z30.0 General counselling and advice on contraception Z30.8 Other contraceptive management Z30.9 Contraceptive management, unspecified
W14 Contraception female, other	Z30.0 General counselling and advice on contraception Z30.8 Other contraceptive management Z30.9 Contraceptive management, unspecified
X10 Postponement of menstruation	Z30.9 Contraceptive management, unspecified

Table 5.4: Breaking open of ICD10 rubrics and mapping to one ICPC rubric

ICD10	ICPC
K80 Cholelithiasis K81 Cholecystitis K82 Other diseases of gallbladder K83 Other diseases of biliary tract K87.0 Disorders of gallbladder and biliary tract in diseases classified elsewhere	D98 Cholecystitis/cholelithiasis

Table 5.5: Mapping of a group of ICD10 classes to one ICPC rubric

exhaustiveness. Class disjointness reflects the principle that any given individual should be classified in a single classification code (one class). To guarantee this, we explicitly define OWL sibling classes to be disjoint. On the other hand to guarantee the exhaustiveness of classifications we introduced a residual class “Other” at every stage of subdivision. This later property is modelled in OWL through a closure definition of any subdivision class as to be equivalent to a disjunction of all its child classes including “Other”:

$$C \equiv C1 \sqcup C2 \sqcup \dots \sqcup Cn \sqcup Other$$

And consequently Other is defined as:

$$Other \equiv C \sqcap \neg C1 \sqcap \neg C2 \sqcap \neg \dots \sqcap \neg Cn$$

This is important above all considering the two classifications taken into account. In fact, ICPC2 contains many rag-bag rubrics, which are further mapped to multiple ICD10 rubrics, e.g., Y99 (Genital disease male other) is mapped to 31 ICD10 rubrics.

We can distinguish two general strategies for knowledge representation in the biomedical domain: the linguistic approach (which is focused on describing term meanings) and the ontological approach, that most of the time is realistic (focused on describing things in the reality itself). We have to specify that it is also possible to do a formal ontological analysis inspired by linguistic distinctions, without borrowing a realist view. In general the latter is preferred for the formalization of the domain, mainly because the ontological approach is more intuitive and self-consistent. We have many example of this second type of approach, such as FMA Ontology, Galen, and Gene Ontology. If one chooses to follow the ontological approach, it becomes obvious that biomedical research almost deals with classes of biological entities, not with individuals (this is at least true of the terminologies and ontologies, but not from datasets produced by biomedical research). So following this principle, we have chosen to formalize ICPC and ICD10 terminologies at a class level. The OWL ontology which

we generated is based on data available via the websites of the organizations responsible for maintaining the respective ICPC2 and ICD10 versions. Concerning the English version of ICPC2 we used the official WICC website to obtain the needed source files, in particular the version related to ICPC2e-v.3.0 2005 (the latest available one when we started this research) plus other ICPC2 manuals both for English and Italian. For the English version of the ICD10 we used the official WHO website, and as additional source we used the ICD10 manual [4]. The data which is publicly available on the Internet is well suited to generate a rich formal model of these two classifications, even if in this step we are more interested in a formal representation which could help us during the process of integration.

To develop the ICPC2 and ICD10 ontologies we used the already mentioned ontology editor Protégé 4.0 (Noy *et al.*, 2001 [88]), developed at Stanford University, while for evaluating the coherence of the ontologies we used its plug-in Pellet Reasoner (Cuenca Grau *et al.*, 2004 [49]).

ICPC2 Ontology.

In encoding ICPC2 we reproduced its biaxial structure in OWL. To this end, we created a class for each ICPC2 component and a class for each ICPC2 chapter related to the components. At the highest level we created two disjoint sibling classes. The first superclass was used to represent ICPC2 “chapters”, named as *Problem Area*, a class including all the possible anatomical areas or body systems where these problems such as symptoms, diseases and diagnoses are localized but also other areas (not anatomical) which the problems are related to, such as Pregnancy, Social, etc. The use of a unique axis for chapters in ICPC2 is explained by the fact that developers wanted to avoid the situation present instead in ICD, where chapters are distributed to different axis, from body systems (Chapters III, IV, V, VI, VII, VIII, IX, X, XI, XIII and XIV) to aetiology (Chapters I, II, XVII, XIX, XX) and to others (Chapters XV, XVI, XVIII, XXI). Since this mixture of axes creates confusion, because diagnostic entities can with equal logic be classified in more than one chapter (for example influenza can be classified either in the infections chapter or the respiratory chapter, or both), in ICPC it is asserted that chapters are all based on body systems, following the principle that localization has precedence over aetiology. This explain our choice in maintaining one class including both body systems and problem areas, which in our ontology are renamed keeping the initial letter of the ICPC2 chapter at the end of the class name as follows:

The second superclass represents ICPC2 components, which in ICPC2 encode reasons for encounters, that is symptoms, disorders, requests or concerns expressed by the patient when seeking care, and which are registered by physicians in a patient record after an encounter. We need to make it clear that here “encounter” means a possible encounter (whether occurred it actually or not) with the primary care structure/physician. Based on these definitions we named this class *Reason for Encounter*. In particular, this class includes in turn three subclasses: *Symptom and Complaint*; *Medical Procedure*; and *Disease*. The

ICPC2 OWL problem areas	
Unspecified area_A	Psychological system_P
Blood_B	Urinary system_U
Digestive system_D	Pregnancy and Family planning area_W
Eye_F	Female genital_X
Ear_H	Respiratory system_R
Circulatory system_K	Skin_S
Musculoskeletal system_L	Metabolic Endocrine and Nutrition area_T
Neurological system_N	Social area_Z
Male genital_Y	

Table 5.6: ICPC2 Chapters names in ICPC2 Ontology

first class contains all the classes which represent the symptoms in ICPC2, from code 00 to 29, prefixed by the code of the chapter which the symptom refers to (from A to Z). The *Medical Procedure* class is composed of five subclasses: *Administrative Procedure*, *Diagnostic Screening and Prevention*, *Test Result*, *Treatment and Medication*, and *Other Medical Procedure*. The ICPC2 procedures (from 30 to 69) are the same for all chapters; consequently their codes don't contain any additional information. The third class *Disease* is divided into five subclasses according to the type of disease: *Infection*, *Injury*, *Neoplasm*, *Congenital Anomaly*, and *Other Disease*. These subclasses include all the ICPC2 rubrics from 70 to 99 prefixed by the letter which identify the chapters, as for *Symptom and Complaint*.

The addition of disjointness statements between siblings was a straightforward task except for several particular situations. In the chapter *Skin* for example, some diseases coincide with symptoms, reflecting the fact that some diseases of the skin are immediately evident. This is the case of symptom S12 (Insect bite/sting) or symptom S09 (Infected finger/toe), which are considered both an injury disease and an infection disease, respectively. Consequently, the disjointness of all siblings in OWL leads to an inconsistency. To solve this problem, allowing these subclasses to be included both in the *Symptom* class and in

the Disease class without remaining inconsistent, we defined the disjointness by means of the followings axiom between these two classes:

$$\text{SymptomAndComplaint} \sqcap \neg(\exists \text{isRelatedToTheProblemArea.Skin}_S)$$

and the corresponding axiom for the *Disease* class. To connect Reasons for Encounter (ICPC2 components) with the relative Problem areas (ICPC2 chapters) - cardinality 1, we created axioms at the level of Reasons for Encounter using the property *isRelatedToTheProblemArea*. It is important to stress that in this work we didn't consider the ontological principles used for formalizing anatomical concepts as done for example in the construction of FMA ontologies and many others related to Anatomy, which obviously consider meronymic and hierarchical relationships between anatomical classes (Rosse and Mejino, 2007 [101]). Here we just need to represent our Reasons for Encounter related to the area, anatomical (e.g. a particular body system such as Digestive) and not (e.g. social area), where they typically occur. For instance, to say that class A08 (Swelling) is a symptom localized in an unspecified area (Chapter A in ICPC2) because it involves the entire Body, we say: *swelling and located to some body system area*, adding the following existential and closure restrictions to class A08:

$$\begin{aligned} &\exists \text{isRelatedToTheProblemArea.Unspecifiedarea}_A \sqcap \\ &\forall \text{isRelatedToTheProblemArea.Unspecifiedarea}_A \end{aligned}$$

Furthermore, since ICPC provides a field “Consider”, which relates a rubric to a (set of) rubric(s) that a physician has to consider when classifying a clinical element, we added as many restrictions to each class having this relation in our ontology as the related classes, by means of the property *TakesIntoAccount*. For instance, we added the following restriction to the class A71 (Measles), which

takes into account (considers) concepts such as S07, A76, and A03:

$$\exists \text{TakesIntoAccount.S07} \sqcap \exists \text{TakesIntoAccount.A76} \sqcap \\ \exists \text{TakesIntoAccount.A03}$$

This means that: when coding Measles, the physician may also want to code rash (S07), fever (A03), and exanthem (A76). Figure 5.2 shows the formalization of concept A71 in Protégé.

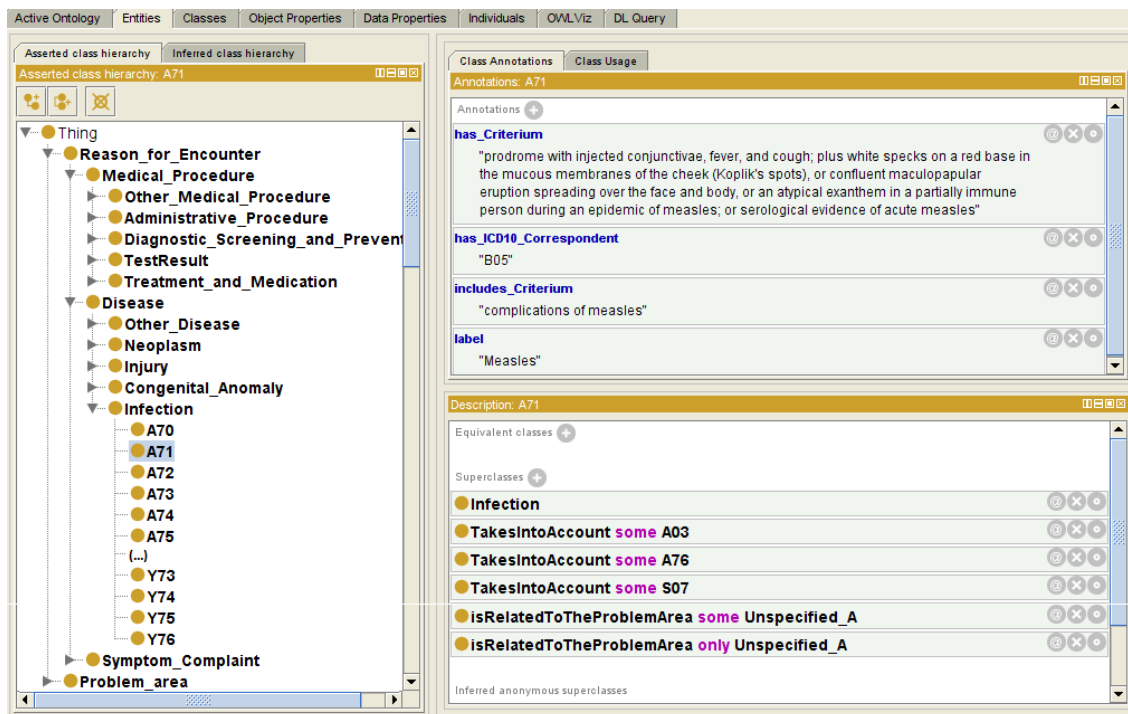


Figure 5.2: OWL encoding of the ICPC2 class A71 “Measles” in Protégé

Concerning other ICPC2 attributes such as concept description, inclusion criteria, exclusion criteria and mapping to ICD10 we translated them into the ICPC2 ontology as Annotation properties:

- Label. We used the `rdfs:label` property to encode the string information, and in this case to associate the ICPC name to each class in the ontology.
- Has Criterion. This property provides a definition of the ICPC concept, but does not appear for all the concepts in the ontology, because in the

perspective of the ICPC terminology, used by physicians, it is necessary to define just those concepts which can generate ambiguities.

- **Excludes Criterion.** This property is associated with those concepts which need a further explanation in order to avoid ambiguities. It consists of a list of similar conditions which should be codified using another ICPC rubric, and thus have to be excluded by a physician during the encoding of the referred concept (e.g. when a physician is codifying the concept “Conjunctivitis infectious” F70 he has to exclude the similar concepts “allergic conjunctivitis with/without rhinorrhea” F71, “ash burn” F79, and “trachoma” F86).
- **Includes Criterion.** This property was added to improve the consistency of the encoding, in order to provide the physician with related terms, hyponyms and hypernyms which can help him during the encoding of a symptom or a disease (e.g. for the previous concept F70, the physician can also include the concepts “bacterial/viral conjunctivitis” and “conjunctivitis NOS”).
- **Has Italian Label.** This property has been added in the ontology even if not present in the ICPC2 DB we took into account for OWL encoding, in order to associate an Italian translation with each ICPC2 class (e.g. Fever HasItalianLabel Febbre).
- **Has Italian Synonym.** This property was added to extend the ICPC2 terminology (at the moment only for the Italian context), in order to provide Italian synonyms for the ICPC2 concepts (e.g. the ICPC2 concept “Infectious mononucleosis” which has an Italian label “Mononucleosi Infettiva”, has Italian synonym “Malattia del bacio”). Data inserted in this field comes from the process of terminology acquisition explained in the previous chapter.
- **Has ICD10 Correspondence.** This property associates with each ICPC2

concept one or more ICD10 concepts, providing mappings between the two classification systems (e.g. ICPC2 concept “Fever” A03 Has ICD10 Correspondence “Fever of other and unknown origin” R50).

In addition, it is possible that a note is provided for an ICPC2 rubric to support human beings in interpreting the ICPC2. Since these contain only continuous text, we represented them using the owl:AnnotationProperty “Has Note”.

ICD10 Ontology.

Starting from the available XML format of ICD10, a file which is structured using the Classification Markup Language (ClAML), designed to represent classification hierarchies (Hoelzer, 2002 [57]), we developed our ICD10 Ontology automatically, creating a class for each ICD10 chapter, a class for each range of ICD10 subchapters, and a class for each of the three or four-digit ICD10 rubrics. We defined only the subsumption relations (represented by the *owl:subClassOf* axioms) in such a way that each range is a subclass of the corresponding chapter, the three digit code class is a subclass of the corresponding range class, and each four-digit code class is a subclass of the corresponding three-digit code class. For example, we created a class for Chapter II (Neoplasms) with a subclass for the chapter block (range) C15-C26 (Malignant neoplasms of digestive organs), which has a subclass for code C25 (Malignant neoplasm of pancreas), which in turn has a subclass for code C25.0 (Head of pancreas). The name of the classes corresponds to the rubric (or rubric range) code. Each class is identified by a URL, which consists of a specific ICD10 name space and the special code as the URL anchor. For text strings associated with the rubric (e.g., description), as for the ICPC2 Ontology, we used the owl:AnnotationProperty “has_Description”. Figure 5.3 shows an (abbreviated) example of the generated class hierarchy.

The work presented by (Moeller *et al.*, 2010 [81]) last year tries to complement our efforts by providing a formal model of additional relations within ICD10, for example the axiomatization of exclusion criteria (we have to high-

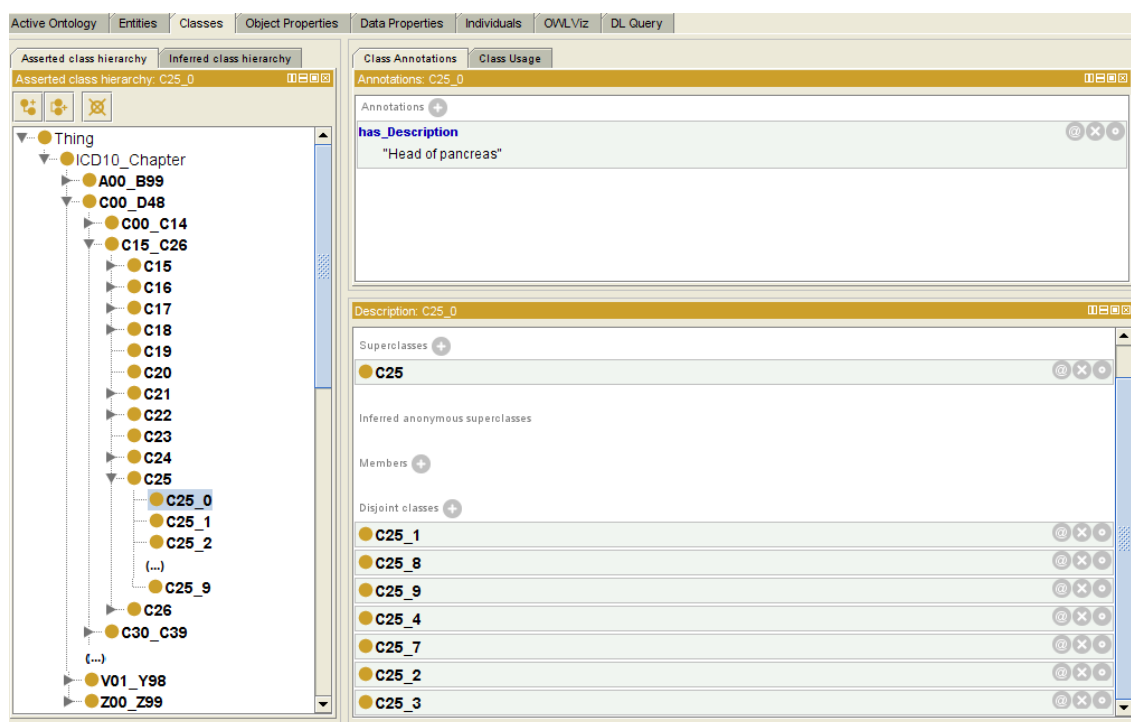


Figure 5.3: OWL encoding of ICD10 class C25.0 “Head of pancreas” in Protégé

light that in ICD “exclusions” are (ontological) subcategories that ICD purposely chooses to eliminate from the subtree and place somewhere else, as for “gestational diabetes mellitus”, although a kind of diabetes mellitus, ICD arbitrarily requires it be coded with the pregnancy disorders, not endocrine disorders).

Mapping ontology

Having constructed the light-weight OWL formalization of ICPC2 and ICD10, we then provided the formal encoding of mappings between them. This is a very important task because formal analysis of mappings and their clear logical encoding is a prerequisite for the correct integration of different healthcare coding systems. A clear logical model of mappings can advance the whole mapping assessment process by allowing the formal evaluation of the quality of mappings, verification of coherence of the newly added mappings or modifications to the mappings, debugging the causes of contradictions, and others. A number

of approaches have been reported in the literature for representing mappings between heterogeneous representations [42]. In this work, we pursued the simplest approach, which consists of expressing mappings as OWL axioms.

Given two heterogeneous representations, a mapping can be viewed as a triple $\langle e, e', r \rangle$, where e, e' are the entities (e.g., formula, terms, classes, etc.) belonging to the different representations, and r is the relation asserted by the mapping. Due to the idea of encoding ICPC-ICD mappings as OWL axioms, the entities in the mapping correspond to ICPC2 and ICD10 classes and expressions, while the relation r is given a set-theoretic meaning by using subsumption and equivalence.

Since historically the goal of establishing an ICPC-ICD mapping was in allowing primary care physicians to implement ICD10 as a reference nomenclature within the classification structure of ICPC2 we decided to start formalization by having exactly this vision in mind. Taking the ICPC-ICD mapping, as a first trial we performed the following encoding: for each ICPC2 class we selected all ICD10 classes to which it can be translated. Since many ICPC2 classes cannot be mapped on the 1:n basis via for example the equivalence relation, we used the subsumption relation for formalizing ICPC2 and ICD10 mappings converting them as m:n mappings by taking the union of the corresponding ICPC2 classes that map to more than one ICD10 class and stated the subsumption relation to mapping. Additionally the other way around subsumption axioms from ICD10 to ICPC2 classes have been introduced. For example, given the mappings in Table 5.4, we constructed the following axioms:

$$\begin{aligned} Y14 \sqcup W14 &\sqsubseteq Z30.0 \sqcup Z30.8 \sqcup Z30.9 \\ X10 &\sqsubseteq Z30.9 \\ Z30.9 &\sqsubseteq Y14 \sqcup W14 \sqcup X10 \\ Z30.0 \sqcup Z30.8 &\sqsubseteq Y14 \sqcup W14 \end{aligned}$$

In total we constructed 3,698 subsumption axioms. Next, we integrated the map-

ping axioms with the OWL formalizations of ICPC2 and ICD10 into a combined knowledge base and analysed it. Figure 5.4 shows an extract of the generated IPCP2-ICD10 mapping ontology in OWL which show the integration of ICPC2 class D76 “Malignant Neoplasm Pancreas” and the corresponding ICD10 class C25 “Malignant Neoplasm of Pancreas” with its subclasses C25.0, C25.1, etc. all inherited by the reasoner as subclasses of ICPC2 D76.

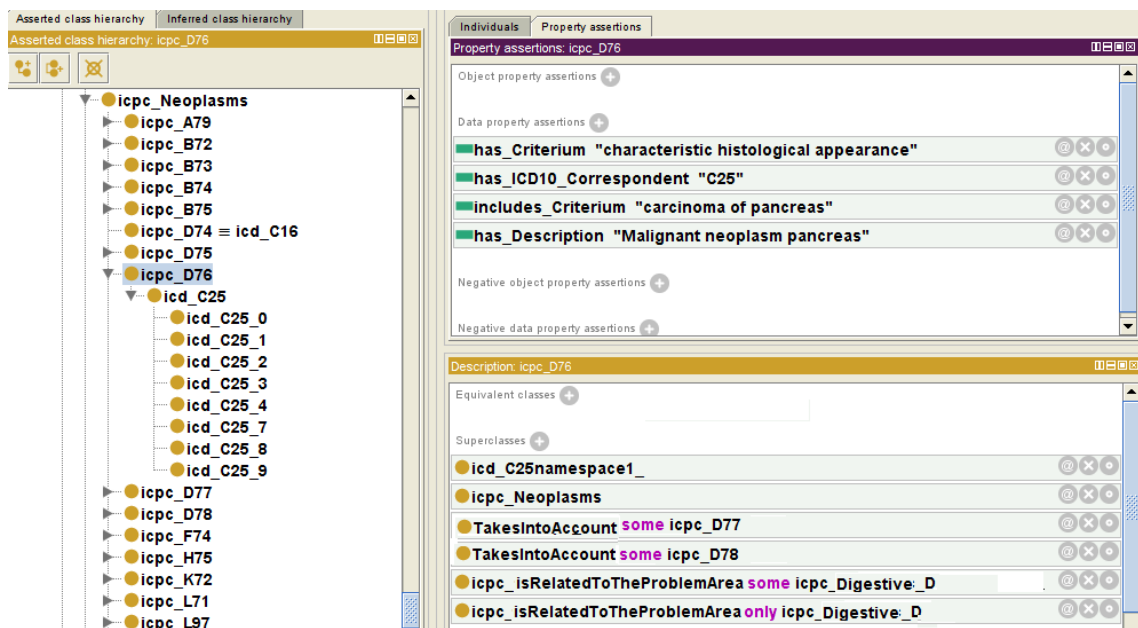


Figure 5.4: Extract of ICPC2-ICD10 mapping ontology in Protégé

The idea behind the analysis was to load the mapping into the OWL reasoner and further see possible logical shortcomings. We used the state of the art Semantic Web reasoner Pellet, which we used to detect whether there were any unsatisfiable classes coming out of the proposed mapping encoding. This analysis allows to also understand what kind of interactions between axioms standing for the mappings and axioms of ICPC and ICD10 can be problematic and, in presence of inconsistencies, how the mapping axioms can be repaired to eliminate the unsatisfiability. After performing this analysis, the reasoner discovered no unsatisfiabilities, meaning that the selected vision was appropriate for the correct representation of ICPC2-ICD10 mappings.

To evaluate the accuracy of this result we applied a different type of axiomatization to the mappings by using, this time, the equivalent relation mentioned above, that is, declaring the mapping axioms by stating the OWL equivalence between the ICPC2 class and the disjunction of the ICD10 classes it is mapped to. For example, for the ICPC2 class X70 (Syphilis Female) which is mapped to ICD10 classes A50 (Congenital syphilis), A51 (Early syphilis), A52 (Late syphilis), A53 (Other and unspecified syphilis), and A75 (Nonvenereal syphilis) we constructed the following axiom:

$$X70 \equiv A50 \sqcup A51 \sqcup A52 \sqcup A53 \sqcup A75$$

This time 686 mapping axioms have been constructed, and following the process of integration with the ICPC2 and ICD10 ontologies and applying again the Pellet reasoner for verification of the mapping coherence, we found inconsistencies (as expected). More precisely, 423 classes derived as unsatisfiable. To diagnose these unsatisfiable classes we applying the OWL debugging technique recently proposed in Kalyanpur *et al.*, 2005 [66], and implemented in Pellet. During the diagnosis, for each of the unsatisfiable classes we used Pellet to detect sets of conflicting axioms that together caused the unsatisfiability and we found that in formalizing mappings with the equivalent relation some of the axioms declared in the two ontologies, in particular in ICPC2, become inconsistent. Some example of problematic interactions between axioms in the ICPC2 classifications are given in Section 5.3.

5.2 Encoding Standard Medical Terminologies in RDF

This section describes the second approach used for formalizing medical terminologies, that is encoding UMLS vocabularies and our ICMV in RDF. In particular we start with a brief introduction to RDF language and then we describe the process of encoding of LOINC, ICPC2 and ICD10 (serialization of their

OWL version presented above) in RDF N-triples and of ICMV in RDF/S. On the other hand RDF N-Triples version for SNOMED CT and MeSH (the other two terminologies we took into account) were already available, created at the National Library of Medicine (NLM) by Bodenreider, 2008ab and Hernandez *et al.*, 2009 [19, 55], and we reused them in the integration process explained in the next chapter. For this reason in this section we also provide explanation for the RDF conversion of these two medical terminologies.

Resource Description Framework (RDF) overview.

As mentioned in Chapter 2, RDF is the Semantic Web language that extends the linking structure of the Web to use URIs to name relationships between things as well as the two ends of the link (usually referred to as a triple as shown in Figure 5.5).

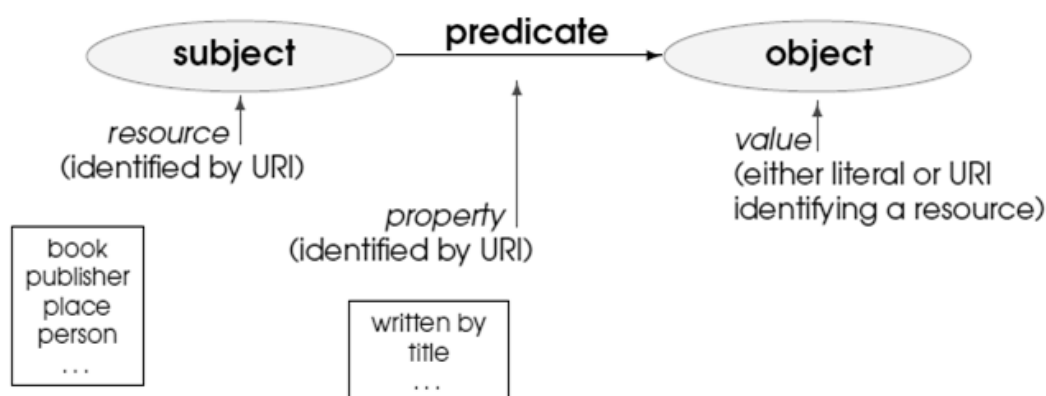


Figure 5.5: RDF Triple format

This linking structure forms a directed, labelled graph, where the edges represent the named link between two resources represented by the graph nodes. This graph view is the easiest possible mental model for RDF and is often used in easy-to-understand visual explanations. RDF graphs can be stored in specialized databases called triple stores (i.e. Sesame, OpenLink Virtuoso, Jena, and many others) and can be queried using the SPARQL query language. RDF has several serialization formats including RDF/XML and N-Triples. This lat-

ter format is the one we used in this chapter for encoding medical terminologies and is a line-based, plain text serialization format for RDF, such as the triples showed in the previous example.

As already mentioned, one RDF extension is RDF Schema (RDF/S) which provides a mechanism for describing specific domains and introduces schema vocabulary (e.g., `rdfs:subClassOf`, `rdfs:Class`, `rdfs:label`, `rdfs:type`, etc.) in RDF. Graphs produced in RDFS are interpreted in structures that are similar to RDF interpretations. In fact, it simply adds extra semantic conditions that give meaning to the schema vocabulary. Usually it is said that RDFS is a primitive ontology language, because it offers certain modelling primitives with fixed meaning, and as with other ontology languages it allows one to define concepts such as class, subclass relations, property, subproperty relations, and domain and range restrictions. However RDFS, unlike OWL, cannot describe resources in sufficient detail, because lacks of important functionalities such as the application of range and domain constraints (e.g. to say that the range of a property `hasChild` is person when applied to persons and for example bear when applied to bears), as well as existence and cardinality constraints, or the declaration of transitive, inverse or symmetrical properties. Because of these limitations, RDF/S is minimally expressive and its support for reasoning is limited.

Even though, as explained at the beginning of the chapter, our task is about encoding medical terminologies and classification systems which do not present high formal expressivity, we want to formalize them in order to retrieve semantic mappings between them. Therefore RDF was deemed suitable for our task because it supports the creation of named graphs which can be aggregated and queried to extract mappings between them.

5.2.1 Materials

In this subsection we briefly review the main characteristics of the terminologies under investigation in this RDF encoding step, except for ICMV described in

the previous chapter, ICD10 and ICPC2, already explained above and adding, on the other hand, a brief introduction to the UMLS, here treated as a repository from which we extracted data for creating RDF N-triples for LOINC.

Unified Medical Language System (UMLS)

UMLS is a compendium of a large number of biomedical national and international vocabularies and classification systems, developed at the National Library of Medicine⁷, to facilitate the development of computer systems which need to understand the meaning of the language of biomedicine and health. In particular, UMLS provides clinical and semantic mapping among all the included terminologies, which are available at the UMLS website as well as the UMLS Knowledge Sources (databases), along with their associated software tools for use by system developers in building or enhancing electronic information systems that create, process, retrieve, integrate, and/or aggregate biomedical and health data and information, as well as in informatics research⁸ (Nelson *et al.*, 2001 [83]). UMLS is composed of three main knowledge sources: the Metathesaurus; the Semantic Network; which provides a broad categorization of all concepts represented in the UMLS Metathesaurus; and the Specialist Lexicon, an English lexicon containing common and biomedical terms with their syntactic, morphological, and orthographic information. In our approach we only focused on the use of the UMLS Metathesaurus (the current release is 2010AA), which is a multi-purpose and multi-lingual vocabulary database containing information about biomedical and health related concepts, their lexical variants, and the relationships among them. It is built from the electronic version of many different resources, such as thesauri, classifications, code systems, and lists of controlled terms used in patient care, health services billing, public health statistics, and indexing of biomedical literature, among many others. The main UMLS Metathesaurus characteristic is that knowledge is organized by

⁷<http://www.nlm.nih.gov/>

⁸http://www.nlm.nih.gov/research/umls/about_umls.html

concept, in particular, synonymous terms deriving from the vocabulary sources are clustered together to form a concept and concepts are linked by means of various types of inter- and intra- concept relationships, which can be hierarchical (e.g. “isa”, “part of”); associative (e.g. “caused by”); or statistical. For further information about UMLS relationships, Semantic Network and the Specialist Lexicon we refer the reader to Bodenreider, 2004 [18]. From a quantitative point of view, UMLS integrates 10 million names for some 2.2 million concepts and 10 million relations, from more than 100 families of biomedical vocabularies in 20 different languages, including Italian. Principal elements of the UMLS Metathesaurus are concepts, terms, strings and atoms. A concept represents a single meaning and contains all atoms from any source that express that meaning in any way, whether formal or casual, verbose or abbreviated. All of the atoms within a concept are synonymous. Each concept is assigned at least one semantic type (broader categories such as “Disease or Syndrome”, “Finding”, etc.). Every concept in the UMLS Metathesaurus is assigned a Concept Unique Identifier - CUI - which uniquely identifies this single concept. In the UMLS Metathesaurus a “term” is the class of all strings that are lexical variants of each other (e.g. “Eye”, “eye”, “eyes” represent only one term). Terms are identified by the Lexical Unique Identifier - LUI - which provides a compact representation of normalized terms (e.g., singular and plural forms of the same term share the same LUI). There are also other elements included in the UMLS Metathesaurus such as strings (a sequence of characters forming a word or phrase in a particular language) which are identified by the String Unique Identifier - SUI - (any difference in upper or lower case, word order, punctuation, or other form would indicate a separate string and receive a different SUI), and finally the atoms, the smallest units of naming in a source, identified by the Atom Unique identifier - AUI. Strings and atoms are not taken into account in this study. Figure 5.6 shows an example of the use of all these elements in representing the UMLS concept “Atrial Fibrillation”: In this work we used

Concept (CUI)	Terms (LUIs)	Strings (SUIs)	Atoms (AUIs) * RRF Only
C0004238 Atrial Fibrillation (preferred) Atrial Fibrillations Auricular Fibrillation Auricular Fibrillations	L0004238 Atrial Fibrillation (preferred) Atrial Fibrillations	S0016668 Atrial Fibrillation (preferred)	A0027665 Atrial Fibrillation (from MSH)
		S0016669 (plural variant) Atrial Fibrillations	A0027667 Atrial Fibrillation (from PSY)
	L0004327 (synonym) Auricular Fibrillation Auricular Fibrillations	S0016899 Auricular Fibrillation (preferred)	A0027930 Auricular Fibrillation (from PSY)
		S0016900 (plural variant) Auricular Fibrillations	A0027932 Auricular Fibrillations (from MSH)

Figure 5.6: Example of the use of UMLS identifiers for the concept Atrial Fibrillation

the UMLS 2009AB release (downloaded by means of a free license from the UMLS Knowledge Source Server, now UMLS Terminology Server)⁹, the last release at the moment we started our study, in particular taking in account the MRCONSO.RRF and MRREL.RRF source files for extracting Semantic Enrichment for our terminologies (UMLS mappings and synonymy, hierarchical relationships), in a RDF N-triple format. All the terminologies under investigation in the thesis approach are integrated in the version 2009AB of UMLS Metathesaurus.

SNOMED Clinical Terms (SNOMED CT)

SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms) is a reference terminology for clinical concepts that was developed by the College of American Pathologists and is now managed by the International Health Ter-

⁹<https://uts.nlm.nih.gov/home.html>

minology Standards Development Organization (IHTSDO)¹⁰. It provides clinical content and expressivity for clinical documentation and reporting. It also includes concepts, terms and relationships with the objective of precisely representing clinical information across the scope of health care. More precisely, it includes 308,000 active concepts with formal logic-based definitions (it is based on the Description Logic system KRSS), distributed among 19 hierarchies (e.g Body structure, Clinical finding, Procedure, Event, etc.), 791,000 active English-language descriptions, and 951,000 logically-defining relationships [5]. In this work, we use the July 2009 version of the international release of SNOMED CT.

Medical Subject Headings (MeSH)

The MeSH Thesaurus is a controlled vocabulary developed by the National Library of Medicine (NLM) for the indexing and retrieval of biomedical literature, including MEDLINE. The main MeSH entities are the *Descriptors*, mostly used to indicate the subject of an indexed item in NLM's MEDLINE and other databases (also known as main headings, 25,186 in the current version) and organized in a hierarchical structure. Descriptors are assigned *Qualifiers* or sub-headings (there are 83 in the current version) indicating particular aspects of a descriptor (e.g. Pathology, Diagnosis, Classification, History, etc.) and used for indexing and cataloguing in conjunction with Descriptors. MeSH also contains Concepts called also *Supplementary Concept Records* (180,672). Finally, each concept consists of a set of *Terms*. Each MeSH entity has a unique Identifier composed of a capital letter indicating the entity (D for Descriptors, Q for Qualifiers, M for Concepts, and T for terms) plus a 6 digit numerical code. Figure 5.7 gives an example of MeSH records for representing the concept. In this work, we use the 2009 version of MeSH.

Logical Observation Identifier Names and Codes (LOINC)

LOINC is a terminology system for laboratory tests and clinical observa-

¹⁰<http://www.ihtsdo.org/snomed-ct/>

MeSH Heading	Heartburn		
Tree Number	C23.888.821.525		
Concept 1 (Preferred)	Heartburn		
	Concept UI	M0009980	
	Scope Note	Substernal pain or burning sensation, usually associated with regurgitation of gastric juice into the esophagus.	
	Semantic Type	T046 (Pathologic Function)	
	Semantic Type	T184 (Sign or Symptom)	
	Term (Preferred)	Heartburn	
		Term UI	T019251
		Date	01-JAN-1999
		Lexical Tag	NON
		Thesaurus	NLM (1964)
	Term	Pyrosis	
		Term UI	T019252
		Date	30-MAR-1974
Lexical Tag		NON	
Thesaurus		UNK (19XX)	
See Also	Gastroesophageal Reflux		
Allowable Qualifiers	BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU VE VI		
History Note	64		
Date of Entry	19990101		
Unique ID	D006356		

Figure 5.7: MeSH record for the concept Heartburn - D006356

tions (current release v. 2.34, December 2010), developed by the Regenstrief Institute and the LOINC Committee¹¹. Its purpose is to facilitate the exchange and pooling of clinical or laboratory results for clinical care, outcomes management, claims attachment, and research by providing a set of universal codes and names. LOINC is composed of two main types of entities: lab test and observation “concepts” on the one hand, and “part” concepts used to support the description of the tests and observations on the other [78]. The fully specified name of a test result or clinical observation has five or six main parts including: the name of the component or analyse measured (e.g., glucose, propranolol), the property observed (e.g., substance concentration, mass, volume), the timing of the measurement (e.g., is it over time or momentary), the type of sample (e.g., urine, serum), the scale of measurement (e.g., qualitative vs. quantitative), and where relevant, the method of the measurement (e.g., radioimmunoassay,

¹¹<http://loinc.org/adopters/regenstrief-institute-inc.html/>

immune blot). These can be described formally with the following syntax (McDonald *et al.*, 2010 [77]).

```
<Analyte/component>:<kind of property of observation or
measurement>:<time aspect>:<system (sample)>:<scale>:<method>
```

The colon character, “:”, is part of the name and is used to separate the main parts of the name. Table 5.7 shows for example the representation in LOINC of the concept B28013 - Pain in back, given by the concatenation of all its LOINC parts: Component, Time, Method, System, Scale and Property.

Component	Prop	Time	System	Scale	Method
Pain	Imp	Pt	Back	Ord	Observed. ICF

Table 5.7: Parts description in LOINC for “Pain in back”

LOINC codes are different according to the type of entity, in fact LOINC Concepts have an associated numeric code including “-” before the last number, while the LOINC Parts have an alphanumeric code starting with “LP” which stand for LOINC Part. LOINC is often used in conjunction with SNOMED. In this case the diagnostic tests are named by LOINC numbers and the results are described by the SNOMED concepts. In this work, we used version 2.27 of LOINC (June 2009) extracted from the UMLS Metathesaurus, which includes 50,809 tests and observations and 44,314 part concepts.

5.2.2 RDF N-Triples Encoding

In this step medical terms and their inter-relations are represented using RDF N-Triples

```
<subject> <predicate> <object> .
```

As already mentioned, two resources, SNOMED CT and MeSH, had already been converted to an RDF representation for other projects at NLM (Bodenreider, 2008 and Hernandez *et al.*, 2009 [19, 55]). In particular, MeSH was converted starting from the XML 2008 version. Here an XSLT was created for

each MeSH record whose output was the creation of RDF N-triples for each MeSH Descriptor. In the resulting triples, the subject is most often a SNOMED CT concept or a MeSH descriptor. The predicates correspond to concept properties, which include type (concept or relationship), preferred name (label), and relations to other concepts (e.g., subclassOf). The object of these triples is either a literal corresponding to a property (e.g., the concept name) or a node representing another concept. Figures 5.8 and 5.9 show examples of RDF triples for SNOMED CT and MeSH, respectively.

```
<http://mor.nlm.nih.gov/SNOMEDCT#25064002> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.w3.org/2000/01/rdf-schema#Class> .
<http://mor.nlm.nih.gov/SNOMEDCT#25064002> <http://www.w3.org/2000/01/rdf-schema#label> "Headache" .
<http://mor.nlm.nih.gov/SNOMEDCT#25064002> <http://mor.nlm.nih.gov/SNOMEDCT#363698007> <http://mor.nlm.nih.gov/SNOMEDCT#69536005> .
<http://mor.nlm.nih.gov/SNOMEDCT#25064002> <http://www.w3.org/2000/01/rdf-schema#subclassOf> <http://mor.nlm.nih.gov/SNOMEDCT#301365009> .
<http://mor.nlm.nih.gov/SNOMEDCT#25064002> <http://www.w3.org/2000/01/rdf-schema#subclassOf> <http://mor.nlm.nih.gov/SNOMEDCT#406122000> .
```

Figure 5.8: SNOMED CT N-triples for the concept 25064002 - Headache

Unlike in SNOMED CT, in MeSH subjects of the triples can be not only the main descriptors, but also MeSH concepts, terms, and qualifiers. Moreover, each concept is associated with a UMLS CUI (concept identifier in UMLS), by means of the predicate MeSH:UMLS_CUI. Other approaches for encoding MeSH into Semantic Web languages, in particular OWL and SKOS, are described respectively in Soualmia *et al.*, 2004, Van Assem *et al.*, 2004 and 2006 [110, 10, 9].

```
<http://nlm.nih.gov#MeSH:D006261> <http://purl.org/dc/terms/identifier> "D006261" .
<http://nlm.nih.gov#MeSH:D006261> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://nlm.nih.gov#MeSH:Descriptor> .
<http://nlm.nih.gov#MeSH:D006261> <http://nlm.nih.gov#MeSH:descriptorClass> "1" .
<http://nlm.nih.gov#MeSH:D006261> <http://www.w3.org/2000/01/rdf-schema#label> "Headache" .
<http://nlm.nih.gov#MeSH:D006261> <http://nlm.nih.gov#MeSH:concept> <http://nlm.nih.gov#MeSH:M0009824> .
<http://nlm.nih.gov#MeSH:M0009824> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://nlm.nih.gov#MeSH:Concept> .
<http://nlm.nih.gov#MeSH:M0009824> <http://nlm.nih.gov#MeSH:isPreferredConcept> "Y" .
<http://nlm.nih.gov#MeSH:M0009824> <http://www.w3.org/2000/01/rdf-schema#label> "Headache" .
<http://nlm.nih.gov#MeSH:M0009824> <http://purl.org/dc/terms/identifier> "M0009824" .
<http://nlm.nih.gov#MeSH:M0009824> <http://nlm.nih.gov#MeSH:UMLS_CUI> <http://nlm.nih.gov#UMLS_MT:C0018681> .
<http://nlm.nih.gov#MeSH:M0009824> <http://www.w3.org/2004/02/skos/core#scopeNote> "The symptom of PAIN in the cranial region."
<http://nlm.nih.gov#MeSH:M0009824> <http://nlm.nih.gov#MeSH:semanticType> <http://nlm.nih.gov#MeSH:T184> .
```

Figure 5.9: MeSH N-triples for Descriptor D006261 - Headache

As presented above, two other resources, ICPC2 and ICD10, had already been converted to an OWL representation. In this process, OWL resources have been serialized in RDF and are therefore directly compatible with other RDF

resources. For this automatic serialization into RDF N-triples we used an on-line tool, Close Word Machine (CWM)¹². For each class in the two ontologies we created N-triples where the subject is the concept itself (a class in ICPC2 and ICD10 ontologies), the concept properties (object and datatype properties) are the predicates (e.g. “Takes into account”, “has label”, etc.) and the objects are either literals corresponding to a property (e.g. name of a concept) or a node representing another class in the ontology. However, unlike for SNOMED CT and MeSH, the subjects of triples from an OWL representation can be blank (anonymous) nodes, used for the representation of restrictions of the classes, in particular related to disjunction and quantifiers (SomeValuesFrom, AllValuesFrom). Furthermore we provided some processing for normalizing ICPC2 and ICD10 RDF, for example adding IDs to the superclasses in the first case (Problem Area, Reason for Encounter, Symptom and Complaint, Disease, etc.). Figure 5.10 and 5.11 show examples of N-triples for ICPC2 and ICD10 concept headache.

```
<http://dkm.fbk.eu#ICPC2E:N01> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.w3.org/2002/07/owl#Class> .
<http://dkm.fbk.eu#ICPC2E:N01> <http://www.w3.org/2000/01/rdf-schema#label> "Headache"^^<xsd:string> .
<http://dkm.fbk.eu#ICPC2E:N01> <http://www.w3.org/2000/01/rdf-schema#subClassOf> <http://dkm.fbk.eu#ICPC2E:1> .
<http://dkm.fbk.eu#ICPC2E:N01> <http://www.w3.org/2000/01/rdf-schema#subClassOf> _:bnode910 .
_:bnode911 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.w3.org/2002/07/owl#Restriction> .
_:bnode911 <http://www.w3.org/2002/07/owl#onProperty> <http://dkm.fbk.eu#ICPC2E:is_a_symptom_belonging_to_the_problem_area> .
_:bnode911 <http://www.w3.org/2002/07/owl#allValuesFrom> <http://dkm.fbk.eu#ICPC2E:CN> .
<http://dkm.fbk.eu#ICPC2E:N01> <http://www.w3.org/2000/01/rdf-schema#subClassOf> _:bnode911 .
<http://dkm.fbk.eu#ICPC2E:N01> <http://dkm.fbk.eu#ICPC2E:has_Italian_label> "CAFALEA"^^<xsd:string> .
<http://dkm.fbk.eu#ICPC2E:N01> <http://dkm.fbk.eu#ICPC2E:has_ICD10_Correspondent> <http://dkm.fbk.eu/ICD10#R51> .
<http://dkm.fbk.eu#ICPC2E:N01> <http://dkm.fbk.eu#ICPC2E:has_Italian_Synonym> "Mal di testa"^^<xsd:string> .
<http://dkm.fbk.eu#ICPC2E:N01> <http://dkm.fbk.eu#ICPC2E:has_Exclusion_Criterion> "cervicogenic headache L83" .
<http://dkm.fbk.eu#ICPC2E:N01> <http://dkm.fbk.eu#ICPC2E:has_Exclusion_Criterion> "migraine N89" .
<http://dkm.fbk.eu#ICPC2E:N01> <http://dkm.fbk.eu#ICPC2E:has_Inclusion_Criterion> "post-traumatic headache"^^<xsd:string> .
```

Figure 5.10: ICD10 N-triples for the concept N01 - Headache

```
<http://dkm.fbk.eu/ICD10#R51> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.w3.org/2002/07/owl#Class> .
<http://dkm.fbk.eu/ICD10#R51> <http://dkm.fbk.eu/ICD10#has_Description> "Headache"^^<file:/tmp/><xsd:string> .
<http://dkm.fbk.eu/ICD10#R51> <http://www.w3.org/2000/01/rdf-schema#subClassOf> <http://dkm.fbk.eu/ICD10#R50-R69> .
<http://dkm.fbk.eu/ICD10#R51> <http://www.w3.org/2002/07/owl#disjointWith> <http://dkm.fbk.eu/ICD10#R52> .
<http://dkm.fbk.eu/ICD10#R51> <http://www.w3.org/2002/07/owl#disjointWith> <http://dkm.fbk.eu/ICD10#R53> .
<http://dkm.fbk.eu/ICD10#R51> <http://www.w3.org/2002/07/owl#disjointWith> <http://dkm.fbk.eu/ICD10#R54> .
<http://dkm.fbk.eu/ICD10#R51> <http://www.w3.org/2002/07/owl#disjointWith> <http://dkm.fbk.eu/ICD10#R55> .
<http://dkm.fbk.eu/ICD10#R51> <http://www.w3.org/2002/07/owl#disjointWith> <http://dkm.fbk.eu/ICD10#R56> .
<http://dkm.fbk.eu/ICD10#R51> <http://www.w3.org/2002/07/owl#disjointWith> <http://dkm.fbk.eu/ICD10#R57> .
```

Figure 5.11: MeSH N-triples for the concept R51 - Headache

¹²<http://infomesh.net/2001/cwm/>

Finally, to create the LOINC RDF version we used a program written in Java to create RDF triples for LOINC from data in UMLS Metathesaurus. More specifically, we extracted from the UMLS table MRCONSO.RRF label, type, and identifier for each LOINC “concept” and “part” entity, where SAB (Source Abbreviation Vocabulary identifier) was equal to “LNC” (LOINC). Similarly we extracted relations among LOINC entities of the type “concept to concept”, “part to part” and “concept to part”, from the UMLS table MRREL.RRF. To extract relationships we maintained, where present, the label given by UMLS (the RELA field), which became the predicate in our triples, while where the type of relations in UMLS were “PAR” (ParentOf) and “CHD”(ChildOf), we put as predicate of the LOINC triple “subClassOf”. Figure 5.12 shows an example of RDF triples for LOINC.

```
<http://nlm.nih.gov#LOINC:LP74908-2> <http://purl.org/dc/terms/identifier> "LP74908-2" .
<http://nlm.nih.gov#LOINC:LP74908-2> <http://www.w3.org/2000/01/rdf-schema#label> "Headache" .
<http://nlm.nih.gov#LOINC:LP74908-2> <http://nlm.nih.gov#LOINC:hasType> "LOINC_Part" .

<http://nlm.nih.gov#LOINC:55466-7> <http://www.w3.org/2000/01/rdf-schema#subClassOf> <http://nlm.nih.gov#LOINC:MTHU000121> .
<http://nlm.nih.gov#LOINC:55466-7> <http://nlm.nih.gov#LOINC:has_component> <http://nlm.nih.gov#LOINC:LP14239-5> .
<http://nlm.nih.gov#LOINC:55466-7> <http://nlm.nih.gov#LOINC:has_method> <http://nlm.nih.gov#LOINC:LP6464-4> .
<http://nlm.nih.gov#LOINC:55466-7> <http://nlm.nih.gov#LOINC:has_system> <http://nlm.nih.gov#LOINC:LP7735-6> .
<http://nlm.nih.gov#LOINC:55466-7> <http://nlm.nih.gov#LOINC:has_time_aspect> <http://nlm.nih.gov#LOINC:LP6960-1> .
```

Figure 5.12: LOINC N-triples for the concept 55466-7, Influenza virus A and for the part “Headache” LP74908-2

5.2.3 Encoding ICMV in RDF Schema

To generate the RDF Schema model of our ICMV we created a Java program taking as input the plain text file created in Excel which resulted from the term extraction process and by manual review provided by physicians and consumers. Figure 5.13 reports a small portion of the input file.

As shown by the Figure the plain text version of the ICMV is composed of several columns that represent on one hand the medical category that each term belongs to (e.g. “disease”, “symptom”, “medical procedure”, etc.), and on the other hand various attributes associated with each term (e.g. “ICMV ID”,

A	B	C	D	E	F	G	H	I
ID	Termine	Categoria	Livello di utilizzo	DESCRIZIONE	SINONIMI	TERMINI CORRELATI	LOCALIZZAZIONE	TERMINE ICPC2 CORRISPONDENTE
ICMV01	Abbassamento della vescica	PATOLOGIA	3	Discesa verso il basso e talvolta fuori dall'introito vaginale della vescica.	Cistocele Prolasso vescicale Prolasso della	Abbassamento dell'utero	vescica uterovagina	X87_Prolasso utero-vaginale U89_Altra r
ICMV02	Abbassamento della voce	SINTOMO	5	Succede quando non si riesce a far uscire la voce in modo naturale, producendo	Calo di voce Voce bassa Raueoedine Afonìa		gola laringe corde voce	R23_SINTOMO O DISTURBO DELLA VC
ICMV03	Abbassamento dell'utero	PATOLOGIA	3	manifesta quando si verifica un indebolimento della muscolatura e dei				X87_Prolasso utero-vaginale
ICMV2516	Abbassamento di pressione	SINTOMO	5		Pressione bassa			K88_IPOTENSIONE ORTOSTATICA
ICMV04	Abbassamento di udito	SINTOMO	4	Non sentire bene da uno o entrambe le orecchie	Sentire di meno			H86_Sordita'
ICMV06	Abbondante versamento articolare	DIAGNOSI	2	Abbondante formazione di liquido nella cavità articolare				L20_Sintomo o disturbo ad una articolazi
ICMV07	Aborto	PATOLOGIA	5	Interruzione prematura di una gravidanza				W82_ABORTO SPONTANEO W83_ABO
ICMV08	Aborto incompleto	PATOLOGIA	3	Espulsione solo di parte del prodotto del concepimento				W82_ABORTO SPONTANEO
ICMV09	Abrasioni	PATOLOGIA	4	Leggera escoriazione della pelle o delle mucose, o lesione della parte più esterna della pelle, provocata da un trauma, o da uno sfregamento con un oggetto che colpisce di striscio la superficie	Escoriazione della pelle			S17_Abrasione graffio vescica
ICMV10	Abrasioni della cornea	PATOLOGIA	3	Lesione di una porzione della superficie corneale con perdita di tessuto superficiale	Abrasioni corneali Escoriazione della cornea			F79_Altra lesione traumatica dell'occhio
ICMV12	Abrasioni al volto	PATOLOGIA	4	Leggera escoriazione della pelle del viso	Lesione cutanea al viso			S17_Abrasioni graffio vescica
ICMV14	Abuso di alcool / alcol	PATOLOGIA	5	Condizione di chi avverte la necessità irrefrenabile e frequente di assumere bibite alcoliche malgrado i danni che tale assunzione comporta	Alcolismo Dipendenza da alcol Abuso etilico			F15_Abuso alcolico cronico
ICMV15	Abuso di droga	PATOLOGIA	5	Condizione di chi avverte la necessità irrefrenabile e frequente di assumere droga (stupefacenti) malgrado i danni che tale assunzione comporta	Tossicodipendenza			F19_Abuso di droga

Figure 5.13: Excel view of ICMV

“Description”, “Synonym”, “ICPC2 Correspondence” etc.). In the process of converting to RDF each term was translated into a class of the RDF model and each category into a superclass, while attributes associated with the terms became properties. In particular, column B *Termine* (Term) and C *Categoria* represents the main classes of the RDF model. Moreover, each class from column B is a subclass of the corresponding class in the column C. For example, the term ABBASSAMENTO DELLA VESCICA (Bladder prolapse) in RDF is a subclass of the category PATOLOGIA (Disease). We translated the other columns containing the values for the RDF properties associated with each ICMV class as follows:

All these properties are datatype properties except for “hasSynonym” and “hasICPC2Correspondence”.

To give a clearer example, we can look at the triples in RDF/S shown in Figure 5.14 for the conversion of the ICMV lay term CUORE IN GOLA (Palpitations):

...

Column name	RDF property name
ICMV ID	hasID
GRADO DI FAMILIARITA	hasFamiliarityDegree
DESCRIZIONE	hasDescription
SINONIMI	hasSynonym
TERMINI CORRELATI	hasRelatedTerm
LOCALIZZAZIONE	isLocalizedIn
TERMINE ICPC2 CORRISPONDENTE	hasICPC2Correspondence
TIPO DI MAPPING A ICPC2	hasMappingType

Table 5.8: ICMV properties in RDF

```

<http://dkm.fbk.eu/ICMV#Cuore_in_gola> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.w3.org/2000/01/rdf-schema#Class> .
<http://dkm.fbk.eu/ICMV#Cuore_in_gola> <http://www.w3.org/2000/01/rdf-schema#subClassOf> <http://dkm.fbk.eu/ICMV#SINTOMO> .
<http://dkm.fbk.eu/ICMV#Cuore_in_gola> <http://dkm.fbk.eu/ICMV#hasDescription> "Avere il cuore che batte forte" .
<http://dkm.fbk.eu/ICMV#Cuore_in_gola> <http://dkm.fbk.eu/ICMV#hasICPC2Correspondence> "K04"^^<http://www.w3.org/2000/01/rdf-schema#ID> .
<http://dkm.fbk.eu/ICMV#Cuore_in_gola> <http://dkm.fbk.eu/ICMV#hasFamiliarityDegree> "5"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://dkm.fbk.eu/ICMV#Cuore_in_gola> <http://dkm.fbk.eu/resources/ICMV.owl#hasICMV_Code> "ICMV559"^^<http://www.w3.org/2000/01/rdf-schema#ID> .
<http://dkm.fbk.eu/ICMV#Cuore_in_gola> <http://dkm.fbk.eu/ICMV#hasMappingType> "SINONIMO TERMINE ICPC2"> .

```

Figure 5.14: Extract of the RDF/S triples for the ICMV lay term “Cuore in gola”

5.3 Results and Evaluation

OWL Encoding.

In this section we presents some quantitative results for the formalizations provided respectively in OWL and RDF. Concerning resources encoded in OWL, Table 5.9 shows some metrics in terms of classes, properties, axioms, and expressivity.

As explained in Section 5.1.2 for the part of logical analysis of ICPC2-ICD10 mappings, we observed that formalizing the mappings as equivalence relations the Reasoner returned us 423 inconsistencies. Problematic interactions between mappings and axiom of ICPC2 have been caused for example by the declaration of disjointness of siblings in the classification, by the classes “Other and Not Specified” and by the declaration of axioms including the property “takes into account”. For illustration, let us look at some typical examples:

- ICPC2 classes Y14 and W14 in Table 5.4 both map to the same ICD10

Metrics	ICPC2	ICD10	Mapping	All
OWL classes	758	14,502	4,041	15,600
Datatype properties	8	2	0	15
Object properties	7	0	0	9
Subclass axioms	2536	15,501	3,698	20,592
Equivalent class axioms	0	0	0	0
Disjointness axioms	170	53,639	54,903	278,542
Hidden CGI	0	0	317	0
Entity annotation axioms	3,662	14,523	0	18,185
DL expressivity	ALC(D)	ALC	ALU	ALC(D)

Table 5.9: Metrics for ICPC2, ICD10 and Mapping ontologies

classes. The reasoner here found that disjoint Y14 and W14 can't be mapped to the same terms in ICD10. While in the reality this is perfectly legal because ICD10 does not distinguish between different aspects of contraception distinguished by general practitioners in ICPC2.

- ICPC2 classes Y06 (Prostate symptom/complaint) and Y99 (Genital disease male other) both map to two ICD10 classes: N42.8 (Other specified disorders of prostate) and N42.9 (Disorder of prostate, unspecified). This is a m:n map, in which the unsatisfiability is due to the disjointness of the symptom Y06 and the rag-bag disease Y99.
- The “Takes into Account” restriction forms an additional channel for propagating an already derived unsatisfiability. For example, ICPC2 classes N06 (Sensation disturbance other) and S01 (Pain/tenderness of skin) both map to the ICD10 class R20.8 (Other and unspecified disturbances of skin sensation). Since S01 is found to be unsatisfiable “Takes Into Account” additionally renders the connected S70 (Herpes zoster) unsatisfiable.

RDF encoding.

Concerning RDF encoding, at the end of the process we created 6 RDF graphs for a total of 2.1M RDF triples. In particular, 97,457 derived from

ICD10; 18,650 from ICPC2; 1.9M from LOINC; and 38,500 from ICMV. Considering also the available RDF N-triples from SNOMED CT (1.8M triples), MeSH (16.6M triples) we reached a total of 20.4M RDF triples to be used for the process of integrating consumer-oriented and medical terminologies.

To conclude this chapter, we can observe the benefits which derive from a logic-based formalization of medical classification systems, above all following the first approach described in this chapter. In fact, formalizing medical terminologies in OWL (in our case ICPC2 and ICD10) allowed to perform reasoning on the expressed semantic and consequently to evaluate the coherence of the mappings between them. ICPC2 and ICD10 ontologies together the mapping ontology could be useful if integrated into a healthcare applications for improving the process of encoding symptoms and complaints during an episode of care, but also if used in on-line application for improving searching of medical information related to the primary care domain. We also took advantage of the RDF language for representing medical terms and their inter-relations. Concerning this point, we could have been used also OWL for the representation of our terminologies, but we would not have been able to take full advantage of its expressivity, due to the underspecification of most of the source vocabularies, which have a simple structure composed mostly of a general hierarchical level (is-a relation and part-of) and few attributes assigned to each medical term. In the next chapter, where we describe the integration process, all these semantic-based resources will be loaded into an RDF repository (Triple Store) which allow us to perform queries among the triples in order to extract semantic mappings between them.

Chapter 6

Integrating ICMV with Standard Medical Terminologies*

This chapter continues the development and elaboration of mappings between medical terminologies/ontologies formalized in the previous chapter.

In this phase of the thesis we try to close the semantic gap that exists between medical terminologies/classification systems which overlap and need to be integrated, by applying the results that have been recently reached in the area of AI and the Semantic Web. In particular, we consider here all the terminologies formalized in the previous phase in order to integrate them with the ICMV (Italian Consumer-oriented Medical Vocabulary) following a similar approach to (Bodenreider, 2008ab and Hernandez *et al.*, 2009 [19, 20, 55]) which consist of retrieving semantic mappings between RDF graphs in a repository using SPARQL queries.

The chapter is organized as follows. Section 6.1 resumes the step of manual mapping of ICMV to ICPC2 (International Classification of Primary Care) performed by physicians. Section 6.2 opens the task of automatic integration of all our resources encoded in RDF Schema or RDF N-Triples by using Se-

***Acknowledgements:** The activity described in this chapter was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), Bethesda MD, and the material of this chapter is based on earlier publications [27]

semantic Web technologies with the aim of integrating our ICMV with the UMLS (Unified Medical Language System) vocabularies by using ICPC2 as a pivot for extracting mappings. This section includes the process of enrichment of RDF graphs with UMLS attributes, the process of Loading and Querying the graphs using SPARQL queries on an RDF triple store and the evaluation of the extracted mappings. The chapter closes with a comparison of the results obtained using the manual approach for mapping ICMV to ICPC2 with the ones obtained using the automatic approach.

6.1 Mapping ICMV to ICPC2: back to the physicians review

As explained in Section 4.6, ICMV has been mapped from a clinical point of view to the ICPC2 coding system thanks to the efforts of 5 physicians who found one by one ICMV lay term/ICPC2 concept pairs with which they also associated the type of correspondence (exact match, synonym, hyponym, and hypernym). At the end of the mapping process, among a total of 2348 lay terms in ICMV, 1521 mappings to ICPC2 were created. Table 6.1 show a short extract of the mappings found manually by physicians:

Among the manual mapping between ICMV and ICPC2, more than 700 are exact matches (e.g. Fever - Fever), 669 are synonyms, 269 hyponyms, and 219 hypernyms. Most of the exact matches are related to anatomical concepts, while synonyms and hypernyms in lay terminology are mostly related to symptoms. Most of the diseases mapped to ICPC2 derive from the Triage records corpus and from the “Medicitalia” corpus, while for example all the symptoms extracted from the Focus Group activity with elderly people have a correspondence in ICPC2. On the contrary, among the 989 terms extracted by the web consultations corpus only half of them were mapped to ICPC2, this is due to the fact that this corpus covers a wide variety of medical topics and terms related for example to health facilities and professionals, health devices and medical

ICMV Term	Category	ICPC2 concept	Type of Mapping
Evidenza di lesioni	Diagnosi	A81_Politraumatismo/lesioni multiple	Hypernym
Frequenza cardiaca	Measurement parameter	31_Esame clinico parziale	Hyponym
Sangue	Parte anatomica	34_Esami ematologici, NAC	Hypernym
Acne	Patologia	S96_Acne	Corrispondenza esatta
Colpo della strega	Patologia	L84_Sindrome dorso lombare senza irradiazione	Synonym
Esame delle urine	Procedura medica	35_Esame urine	Exact match
Addome gonfio	Sintomo	D25_Distensione addominale	Synonym
Ematoma	Sintomo	S16_Contusione/ecchimosi	Synonym

Table 6.1: Example of manual mapping between ICMV and ICPC2

procedures didn't find any correspondence to the ICPC2 concepts, due to the fact that ICPC2 only includes symptoms, diagnoses, and medical procedures at a general level. Remarkably, the number of mappings through synonyms and through exact matches is roughly equivalent, due to the fact that lay expressions often have a technical correspondence in ICPC2 (such as "Orecchioni" and "Parotite Epidemica"). Finally we can remember that after this manual mapping performed by physicians most of the ICPC2 concepts mapped at least one time to our ICMV terms.

With this consideration in mind, we tried to reuse those mappings to test a new approach for extending mapping between our ICMV and other medical terminologies, because ICPC2 can be used, as described below, as a point of access to existing mappings with other terminologies, especially in UMLS, which provides a knowledge base of mappings between terminologies.

6.2 Mapping ICMV to UMLS Vocabularies (Automatic)

In this process, ICPC2 serves as a pivot between our consumer-oriented vocabulary, ICMV, and other professional vocabularies integrated with UMLS. This is possible because ICMV was mapped to ICPC2 and ICPC2 is integrated in the UMLS, along with many other professional vocabularies. For this reason, we use ICPC2 as an entry point into UMLS in order to find mappings to concepts from SNOMED CT, MeSH, LOINC and ICD10, which one also integrated in UMLS. The Integration Framework is shown in Figure 3.2. The major steps of the integration approach can be described as follows:

1. Enrich of each terminology, already formalized into RDF, with UMLS attributes.
2. Load these resources into a Triple Store (Virtuoso), and Explore relations among terms in the medical resources, related to synonymy and hierarchi-

cal relations, through SPARQL queries.

3. Evaluate the quality of the mappings between ICPC2 and UMLS vocabularies and ICMV and ICPC2, in order to test the suitability of ICPC2 for representing ICMV by directly mapping the “lay” ICMV terms to UMLS.

6.2.1 The UMLS Enrichment Process

Since in the previous step we converted our medical terminologies into RDF N-triples, the next step is to integrate these triples with UMLS information which is needed for extracting the mappings. Before that we cleaned up and normalized our N-triples according to the type of integration of each terminology in the UMLS database, looking at the codes used, at the type of attributes, etc. For example, for ICD10 we performed processing on the converted file ICD10.nt to add for each Range Class (E.g. A00-B99, which is a chapter in ICD10) the owl equivalence to map these IDs to the same entities as the one used in UMLS for naming the ICD10 concepts. In UMLS in fact, the ranges IDs for ICD10 have the form “range + .9”, e.g. A00-B99.9). We provided further processing for the file ICPC2.nt in order to make it compatible with the one integrated in UMLS, which also gives codes to the ICPC2 components and chapters, not only to single ICPC2 rubrics (e.g. the component “Symptom and complaint” is codified in UMLS with the code “7”, that is the seventh component. So we renamed our nodes related to ICPC2 superclasses as the one in UMLS.

Then, in order to facilitate term comparisons among vocabularies, we enriched each terminological resource with UMLS attributes. These attributes include the concept unique identifier (CUI), the source abbreviation (SAB), the lexical unique identifier (LUI) and term type (TTY). Of particular importance are the lexical unique identifiers, which provide a compact representation of normalized terms as mentioned in the previous chapter. We extracted this information from the corresponding fields of the table MRCONSO.RRF (this is

available with all the other UMLS knowledge sources on the UMLS website) and created triples as described previously. Figure 6.1 shows an example of these triples for enrichment in ICPC2.

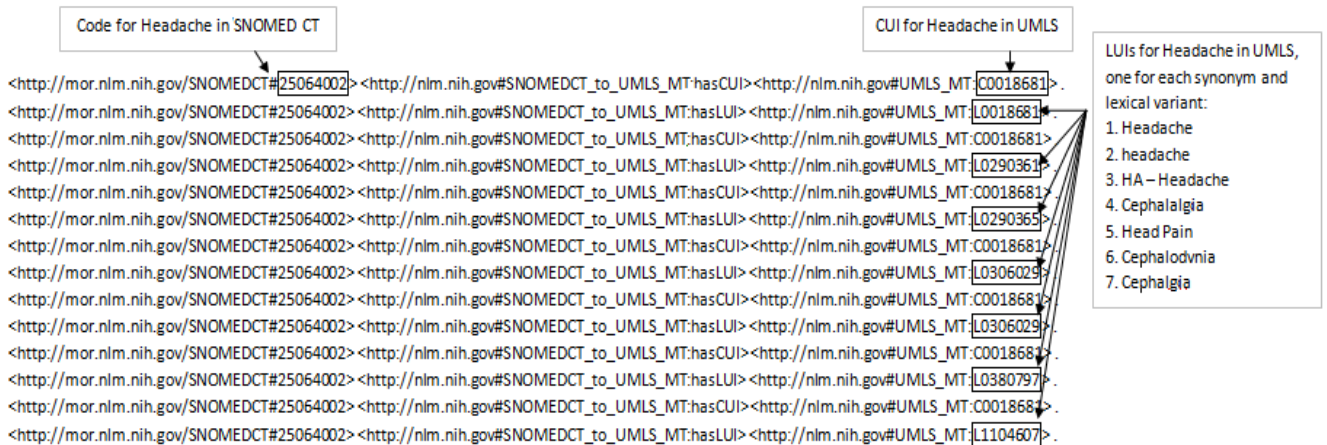


Figure 6.1: SNOMEDCT N-triples enriched with UMLS CUI and LUIs corresponding to “Headache”

We created a Java Program to extract this information from the MRCONSO table and created the enriched RDF N-triples for SNOMED CT, MeSH, LOINC, ICPC2 and ICD10.

6.2.2 Loading and Querying Process

In this phase all the N-triples created in the previous steps (both the graphs related to the original resources and the ones related to the corresponding UMLS enrichment) were loaded into a triple store, which by the end of the process comprised 13 graphs. We used the OpenLink’s Virtuoso RDF store², v. 6.0, an open source triple store. Virtuoso provides an interactive interface (ISQL), which allows users to execute queries and scripts against the triple store. To access Virtuoso and to query the graphs, we used the SPARQL query language, mainly for comparing the representation of concepts in the different terminolo-

²<http://virtuoso.openlinksw.com>

gies. Before explaining the process of Querying RDF graphs we give some details about SPARQL in order to explain its syntax, functionalities and advantages.

SPARQL.

To query an RDF graph or an OWL ontology, the Semantic Web proposes the use of standardized SPARQL, a syntactically-SQL-like query language based on matching graph patterns against RDF graphs. The SPARQL query language consists of the syntax and semantics for asking and answering queries against RDF graphs, including features such as basic conjunctive patterns, value filters, optional patterns, and pattern disjunction. It also supports constraining queries by source RDF graph and extensible value testing. SPARQL 1.0 became a W3C standard in January, 2008, while SPARQL 1.1 is in-progress. SPARQL, which is both a query language and a data access protocol, is considered a query mechanism for the Semantic Web, i.e. Web 3.0 applications. Among various functionalities it allow us to: (i) Pull values from structured and semi-structured data; (ii) Explore data by querying unknown relationships; (iii) Perform complex joins of disparate databases in a single, simple query; and finally (iv) Transform RDF data from one vocabulary to another.

In a SPARQL query performed on RDF Triples (subject, predicate, object), URIs which identify resources can be abbreviated by using prefix names. Objects can be nodes or literals such as strings, integers, booleans, etc. A SPARQL query comprises, in order: 1. a prefix declarations, for abbreviating URIs; 2. a dataset definition, stating what RDF graph(s) are being queried; 3. a result clause, identifying what information to return from the query; 4. the query pattern, specifying what to query for in the underlying dataset; 5. query modifiers for slicing, ordering, and otherwise rearranging query results.

The use of SPARQL as a Semantic Web query language has many benefits. First of all the fact that it has an implicit joint syntax, in fact, because all rela-

tionships in RDF are of a fixed size and data lives in a single graph, SPARQL does not require explicit joins that specify the relationship between differently structured data. Another benefit is the strong support provided by SPARQL for querying semi-structured and ragged data (i.e., data with an unpredictable and unreliable structure). Furthermore, because RDF represents all data as a collection of simple binary relations, most data can be easily mapped to RDF and then queried and joined using SPARQL. Often, these mappings can be performed on the fly, meaning that SPARQL can be used to join heterogeneous data at a higher level than that of the native structure of the data.

On the other hand, SPARQL presents some drawbacks if compared to other query languages such as SQL and XQuery. These limits consist principally in the lack of wide spread deployment, since SPARQL is relatively young and immature, and as such there are not many data stores which can be directly queried with it as compared with SQL or XPath. While SPARQL is designed to query RDF graphs, it has no facilities for easily querying transitive relations or hierarchical structures within a graph.

Querying the graphs.

We created template queries, which we populated with values of interest before sending the queries to the Virtuoso SPARQL engine. In practice, we used a Java program to automate the submission of batch queries to Virtuoso and collect the results. We executed the following types of queries for each concept in ICPC2:

1. Find concepts in SNOMED CT, MeSH, ICD10 and LOINC corresponding to a particular ICPC2 concept, using the UMLS CUI as a bridge.
2. Find preferred terms, synonyms, and lexical variants in SNOMED CT, MeSH, ICD10 and LOINC corresponding to a particular ICPC2 concept, using the UMLS CUI and the UMLS LUI as a bridge.
3. Find shared hierarchical relations between ICPC2 concepts and the other

terminologies, using the UMLS CUI as a bridge.

Below we have an example of a query and its results, which allows extracting concepts in the other terminologies which correspond to a given ICPC2 concept, using the CUI as a term of comparison.

```
PREFIX ICPC2E: <http://dkm.fbk.eu#ICPC2E:>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX UMLS_MT: <http://nlm.nih.gov#UMLS_MT:>
SELECT ?icpc2_id ?label ?cui ?code
from <http://nlm.nih.gov/ICPC2E_to_UMLS_Enrichment>
from <http://dkm.fbk.eu/ICPC2E>
from <http://nlm.nih.gov/SNOMED_CT_to_UMLS_Enrichment>
from <http://nlm.nih.gov/LOINC_to_UMLS_Enrichment>
from <http://nlm.nih.gov/ICD10_to_UMLS_Enrichment>
from <http://nlm.nih.gov/MeSH_Enrichment>
WHERE
{
    ?icpc2_id rdfs:label ?label .
    ?icpc2_id UMLS_MT:hasCUI ?cui .
    ?code UMLS_MT:hasCUI ?cui .
    filter(?icpc2_id = ICPC2E:N01)
};
```

The FROM keyword in SPARQL lets us specify the target graph in the query itself. By using ?... (e.g. ?cui) as an object of one triple and the subject of another, we traverse multiple links in the graph. Concerning FILTER constraints, they simply use boolean conditions to filter out unwanted query results. In this query we are asking:

In the graphs

```
<http://nlm.nih.gov/ICPC2E_to_UMLS_Enrichment>,
<http://dkm.fbk.eu/ICPC2E>,
<http://nlm.nih.gov/SNOMED_CT_to_UMLS_Enrichment>,
...
```

find me all subjects (?ICPC2_ID, so all the ICPC2 concepts) which have as object (?cui, the UMLS CUI) linked with the UMLS_MT:hasCUI predicate. Then return all the values (?code) which share the same object (?cui) and which thus have the same UMLS CUI.

Triple patterns in the query are just like triples, except that any of the parts of a triple can be replaced with a variable. The SELECT result of a clause returns a table of variables and values that satisfy the query. Results of a query are returned as an HTML table.

A query of this type produced as result the following mappings for the ICPC2 concept N01 “Headache”:

ICPC2 Concept	Label	Has UMLS CUI	Mapped Source concept
ICPC2E:N01	Headache	UMLSCUI:C0018681	SNOMED CT:25064002
ICPC2E:N01	Headache	UMLSCUI:C0018681	LOINC:LP74908-2
ICPC2E:N01	Headache	UMLSCUI:C0018681	ICD10:R51
ICPC2E:N01	Headache	UMLSCUI:C0018681	MeSH:D006261

Table 6.2: Query result for the ICPC2 concept Headache

We created a Java program (with Virtuoso APIs to access the triple store) to automate the submission of batch queries to Virtuoso and collect the results. To give an example of these batch queries we show an extract (the query part) of the program which automates the query seen above and consequently finds the mappings to SNOMED CT, MeSH, LOINC and ICD10 via UMLS CUI for all 681 ICPC2 concepts. The first part of the Java code looks for the input file, which contains all the ICPC2 IDs, then we declare the query to be performed as follows:

```
...
private static ArrayList<String[]> getResult(String icpc2_id, \\
int num_selects) throws Exception
{
String query =
    "sparql " +
    "PREFIX ICPC2E: <http://dkm.fbk.eu#ICPC2E:> " +
    "PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> " +
    "PREFIX UMLS.MT: <http://nlm.nih.gov#UMLS.MT:> " +
```

```

    "select ?icpc2_id ?label ?cui ?code " +
    "from <http://nlm.nih.gov/ICPC2E_to_UMLS_Enrichment> " +
    "from <http://dkm.fbk.eu/ICPC2E> " +
    "from <http://nlm.nih.gov/SNOMEDCT_to_UMLS_Enrichment> " +
    "from <http://nlm.nih.gov/LOINC_to_UMLS_Enrichment> " +
    "from <http://nlm.nih.gov/ICD10_to_UMLS_Enrichment> " +
    "from <http://nlm.nih.gov/MeSH_Enrichment> " +
    "where " +
    "{ " +
        "?icpc2_id rdfs:label ?label . " +
        "?icpc2_id UMLS.MT:hasCUI ?cui . " +
        "?code UMLS.MT:hasCUI ?cui . " +
        "filter(?icpc2_id = ICPC2E:" + icpc2_id +") " +
    }";
    ArrayList<String[]> results = getQueryResults(query,
    num_selects);
    return results;}
    ...

```

In the FILTER part the program automatically changes the ICPC2 ID to be used as our subject for finding the UMLS CUIs and then the mappings to the other sources. The next part of the program is about the connection to the Virtuoso database, the execution of the query, and then the return of the results. Finally, in order to perform the queries related to the extraction of shared hierarchical relations between ICPC2 concepts and the ones in SNOMED CT, MeSH, LOINC, and ICD10, we computed Transitive closure (TC) for each graph and we created separate graphs for them, defined as follows:

Having a directed graph $G=(V,E)$, where V is the set of vertices and E is the set of edges (so a binary relation), the transitive closure of G is a graph $G^+ =(V,E^+)$ such that for all v,w in V there is an edge (v,w) in E^+ if and only if there is a non-null path from v to w in G .³

Transitive closure can be thought of as establishing a data structure that makes it

³<http://www.cs.hut.fi/~enu/tc.html>

possible to solve reachability questions efficiently. Figure 6.2 shows a practical example of hierarchical binary relations (solid arrows) among medical concepts in the ICPC2 graph and the transitive relations (dashed arrows) resulted after the computation of the transitive closure.

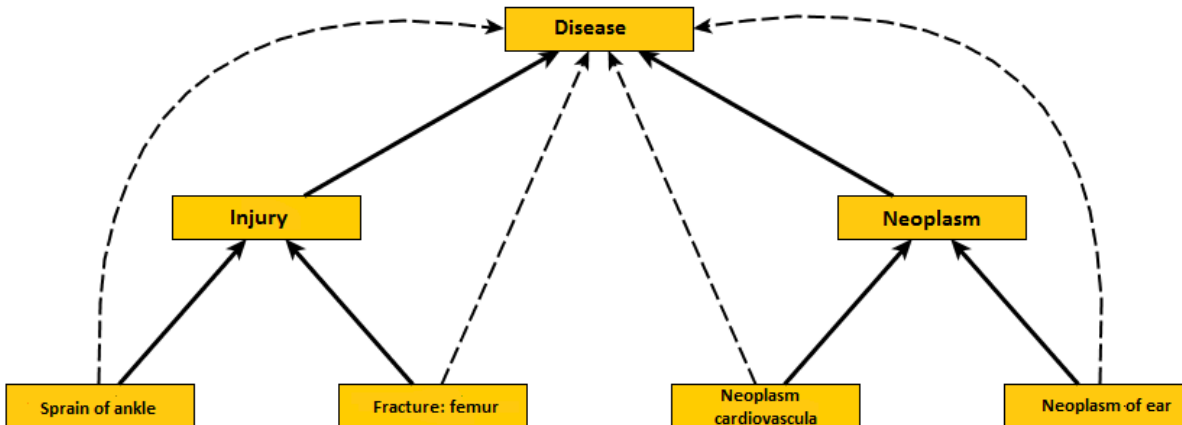


Figure 6.2: Example of Transitive Closure constructs output for some ICPC2 concepts

In order to compute TC for each RDF graph in our RDF triple store we used SPARQL queries as the one below. In particular, since we want to find shared hierarchical relations among graphs, we analysed the predicates which define this type of relation (e.g. in the case of SNOMED CT, ICD10, ICPC2, and LOINC the `rdfs:subClassOf` relation; in MeSH the `MeSH:upperDescriptor` relation).

```

SPARQL
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
select ?des ?anc
FROM <http://dkm.fbk.eu/ICPC2E>
WHERE {
    ?anc ?p ?o .
    {
        SELECT ?des , ?anc
        WHERE {
            ?des rdfs:subClassOf ?anc .
        }
    }
}

```



```

OPTION (TRANSITIVE, T_IN(?des), T_OUT(?anc), T_MIN(1), T_DISTINCT) .
}
}
;

```

This query was used to extract TC relation from the ICPC2 graph, and it asked for all descendant (?des) and ancestors (?anc) where there is a triple composed of “ancestor” (?anc) “predicate” and “object”, and where there is a binary relation of the type “descendant subclassOf ancestor”. Next, we created by the query output new RDF N-triples representing TC relations and we loaded into Virtuoso, in order to allow for SPARQL queries related to shared common ancestors.

6.2.3 Quality Assurance of the Mappings

In order to evaluate the quality of the mapping shown in Section 4.4 between ICMV and ICPC2, as well as the coverage of the Italian version of ICPC2, we mapped all terms from ICMV to the Italian terms in the UMLS Metathesaurus, using an exact match, as supported by the UMLSKS (stands for “UMLS Knowledge Source Server” throughout this section) application programming interface. For each ICMV term, we recorded whether the term was found in the UMLS and, if so, in which source vocabularies:

Mapping found: ICMV term + UMLS CUI + Preferred Term + Source + Code)

A review of other studies that involved mapping to UMLS concepts provided additional insights into potential mapping problems. Travers, 2001 [115] used three rounds, relaxing the criteria with each successive round, to maximize the mapping of forms extracted from Emergency Department chief complaints:

- Punctuation. Including expanding abbreviations (e.g., blood pressure for b/p) and coordinate constructs (e.g., hip pain and thigh pain as separate forms for hip/thigh pain), reformulating non-regular constructs (e.g., groin

rash for rash/groin) and separating series of forms into individual forms (e.g., fever and cancer for fever; cancer).

- Acronyms, Abbreviations, and Truncations. Expansion into the full forms.
- Qualifiers and Modifiers: Deletion of qualifiers, words that specify the meaning of clinical terms (e.g., history of) and modifiers, words that indicate severity, location, or acuity (e.g., acute).

If no closely related UMLS concepts were found, an approximately related (narrower or broader) UMLS concept was sought. Forms for which not even approximate UMLS concepts could be found remained unmapped.

After the direct mapping between ICMV terms and UMLS Italian terms we compared these mappings with the one found using ICPC2 as a pivot to access to UMLS mappings and finally to the one found by means of the manual clinical review performed by physicians.

In particular, after having the relation ICMV term/UMLS CUI, we could be able to create also for the ICMV graph (until now excluded from the process of integration) a UMLS enrichment graph to be loaded in Virtuoso. This new graph includes both, triples representing the UMLS CUI relation and Triples representing the ICPC2 relation to ICMV terms.

Once loaded also this graph in Virtuoso we performed the first type of query to find mappings between ICMV terms and the other resources using the UMLS CUI as a bridge. Obviously this time the input file for the batch query was the one including all the ICMV terms. On the other hand we repeated the inverse process starting from the ICPC2 codes and asking for all the mappings to the ICMV via UMLS CUI.

6.2.4 Results and Evaluation

In addition to the RDF graphs obtained by the process of RDF encoding described in the previous chapter (in total 2.1M triples), including the UMLS

graph which is composed of 50M from UMLS triples, considering the creation of 6 new RDF graphs related to the UMLS enrichment (in total 15,369,152 triples) and 5 new graphs related to the Transitive Closure (5,801,418 triples) we reached at the end of the integration a total of 73,270,570 unique triples loaded into Virtuoso. In terms of performance, loading all these triples into Virtuoso took a few seconds for each graph (the larger UMLS graph was already loaded), and only 5 minutes for the UMLS Enrichment graph. Execution time for batch queries was also short (3-5 minutes for a query on each concept in ICPC2). Poor performance was registered for queries related to Hierarchical mappings (hours).

Results for Query 1. Finding ICPC2 concepts in other terminologies.

Results show that 77% of the ICPC2 concepts (587 of 760 concepts) are integrated in the UMLS. Of these 587 UMLS concepts, 251 are specific to ICPC2 and 336 are common to other terminologies. In particular, as shown by Figure 6.3 we found 257 mappings to ICD10, 663 mappings to SNOMED CT, 201 mappings to MeSH, and 68 mappings to LOINC, that is, only 8.9% of the ICPC2 concepts. In contrast, each ICPC2 concept maps to at least one con-

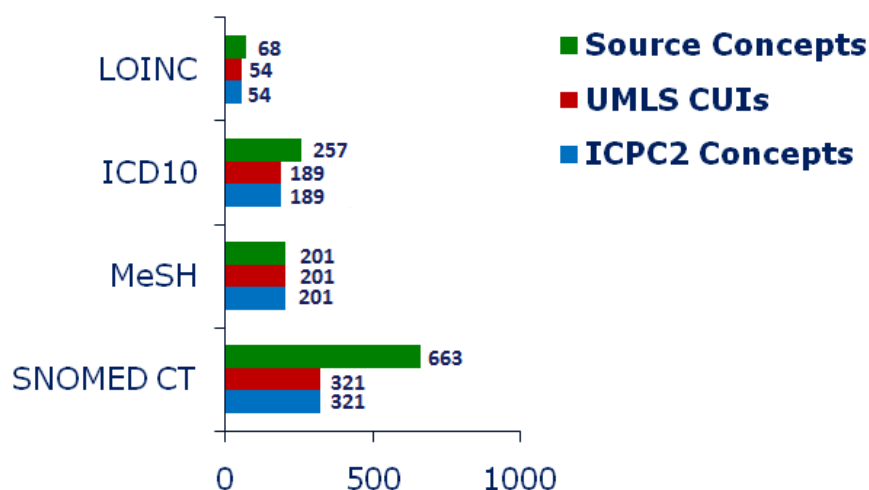


Figure 6.3: ICPC2 mappings to the other resources

cept in ICD10. The overlap with SNOMED CT and MeSH is 87% and 26%,

respectively. More statistical results are shown in Table 6.3.

Concerning overlaps between the resources, results show that among the 336 ICPC2 concepts mapped to the other resources, 130 map to three terminologies, 90 map to two terminologies, 33 map to all the four terminologies (E.g. A03 - Fever, F93 - Glaucoma) and 83 map to one terminology. In particular 74 map only to SNOMED CT (E.g. A18 - Concern about appearance), 4 map only to MeSH (e.g. Speech disorder), and 5 map only to ICD10 (H77 Sprain/strain of ankle), as shown in the Chart below.

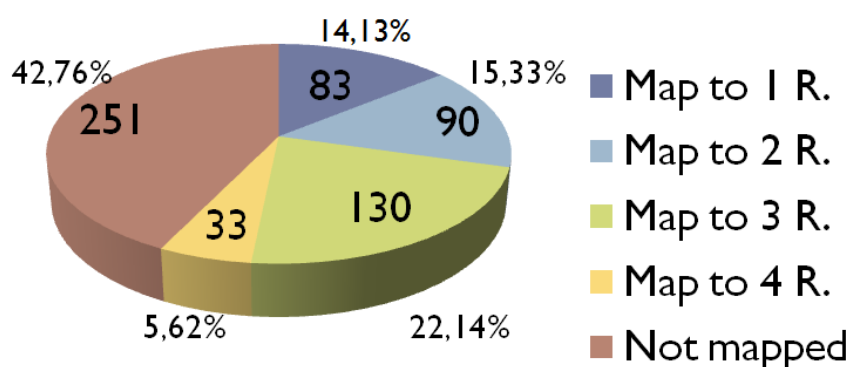


Figure 6.4: Overlap among resources

We observed a large number of multiple mappings between ICPC2 and other terminologies. In particular, 40% of the ICPC2 concepts map to at least three SNOMED CT concepts and two ICD10 concepts. Figure 6.5 shows an example of multiple mappings found for the ICPC2 concept Malaria and Glaucoma. The greatest number of mappings for an ICPC2 concept is for “B72”, Hodgkin’s disease/lymphoma (12 SNOMED CT concepts, 2 ICD10 concepts, 1 MeSH descriptor and 1 LOINC concept).

Results for Query 2. Finding preferred terms, synonyms and lexical variants.

Regarding mappings to synonyms and lexical variants, we found that of the 587 ICPC2 terms mapped through the CUI, 311 are used as the preferred term in the other terminologies, and no synonyms are found for these concepts. These

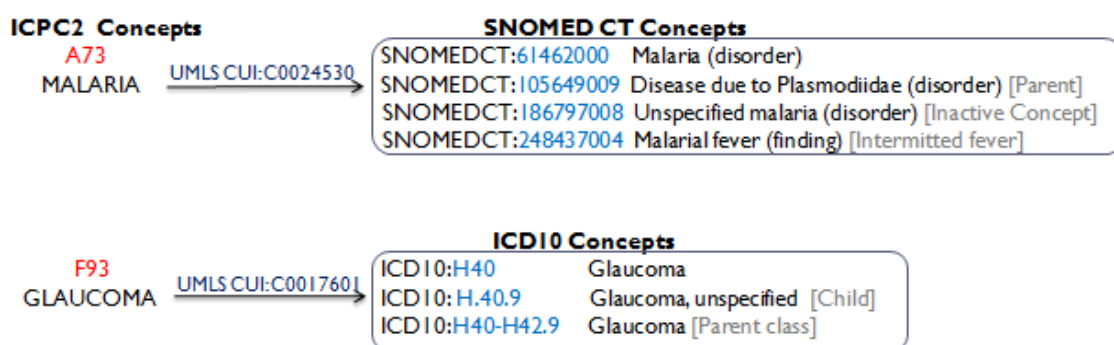


Figure 6.5: Examples of multiple mappings between ICPC2 and other terminologies

are concepts related to the “social problems”, “skin”, “musculoskeletal”, “female genital”, “general” symptoms and diseases. A total of 2818 (1745 unique values) were found for the ICPC2 concepts, among which 1197 are mapped to 663 SNOMED CT concepts, 363 mapped to 196 MeSH descriptors, 156 to 156 ICD10 concepts and 62 to 68 LOINC parts. Using the LUI as bridge to extract mappings, we found a total of 949 synonyms for 422 ICPC2 concepts. The ICPC2 concept with the greatest number of synonyms is Incontinence Urine, code “U04” (14 synonyms), as shown by Figure 6.6. Finally, 739 additional names were found. Most of these lexical variants were contributed by SNOMED CT.

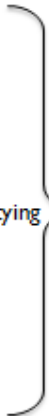
ICPC2 U04 PT: <u>Incontinence Urine</u>	UMLS LUIs	NEW SYNONYMS	
	<i>L0021170</i>		
	L0042024	Urinary Incontinence	
	L0005685	Absence of bladder continence	
	L0308837	Involuntary urination	
	L0527266	Unable to control bladder	
	L0527384	Weak bladder	
	L0583725	Unable to hold urine	
	L0005692	Bladder incontinence	
	L0527301	Unable to prevent bladder emptying	
	L0574730	Unable to hold fluids	
	L0586619	UI - Urinary incontinence	
	L0748747	Bladder: incontinent	
	L0042024	Incontinence, urinary	
	L0527264	Lack of bladder control	
	L0527265	Loss of bladder control	
	L0590897	Leaking of urine	
			SNOMED CT 165232002 PT: <u>Urinary Incontinence</u>

Figure 6.6: Exemple of extended synonyms for ICPC2 U04 *Incontinence Urine* by mapping to SNOMED CT

Results for Query 3. Finding shared ancestors.

Results for the query related to hierarchical relations among the terminologies under investigation showed that only 220 ICPC2 concepts share “parent” relations with the other terminologies, of which 193 are shared with SNOMED CT and 27 with MeSH. No hierarchical relations are shared with LOINC and ICD10. That means that ICPC2 concepts do not share any parent concepts with the corresponding concepts in ICD10 and LOINC. In terms of CUIs we found that 84 unique parent CUIs are shared between ICPC2 and SNOMED CT and MeSH. We need to highlight that hierarchical mappings were found only for ICPC2 concepts related to diseases. Figure 6.7 shows an extract of the mappings concerning shared ancestors between ICPC2 and SNOMED CT (e.g. Leucemia which has the ancestor “Neoplasm” both in ICPC2 and SNOMEDCT).

This process allowed us to discover some inconsistencies in terms of classification of symptoms and diseases among terminologies. For example we found that some ICPC2 symptoms are classified as diseases or diagnoses in the other terminologies. This is the case of the ICPC concept “warts” classified as disease in SNOMED CT, MeSH and ICD10.

Table 6.3 shows a final statistical overview of the various type of mappings found between ICPC2 and the other resources in UMLS by means of SPARQL queries.

Mapped Entities	SNOMED CT	MESH	ICD10	LOINC	UMLS
Unique Resource IDs	663	201	257	68	1773
UMLS CUIs	321	201	189	54	1773
UMLS LUIs	1197	363	156	62	1420
Preferred Terms	703	149	257	68	–
Synonyms	824	125	0	0	–
Shared parent relations	193	27	0	0	–
ICPC2 concepts	321	201	189	54	587

Table 6.3: Statistical summary of the mappings among resources

ICPC2 code	ICPC2 name	SNOMED CT code	SNOMED CT Name	SNOMED CT Ancestor	SNOMED CT second Ancestor	UMLS CUI	UMLS ancestor name
B73	Leukaemia	87163000	Leukaemia	Neoplasms	Neoplasm	C0027651	Neoplasms
B73	Leukaemia	93143009	Leukaemia	Neoplasms	Neoplastic disease	C0027651	Neoplasms
B73	Leukaemia	188767008	Leukemia NOS	Neoplasms	Neoplastic disease	C0027651	Neoplasms
B73	Leukaemia	188762002	Leukaemia of unspecified cell type	Neoplasms	Neoplastic disease	C0027651	Neoplasms
B76	Ruptured spleen traumatic	43756009	Traumatic rupture of spleen	Injuries	Traumatic AND/OR non-traumatic injury	C0175677	Injury
B76	Ruptured spleen traumatic	43756009	Traumatic rupture of spleen	Injuries	Traumatic injury	C0175677	Injury
B80	Iron deficiency anaemia	87522002	Iron deficiency anaemia	Other Diseases	Disease	C0012634	Disease
B80	Iron deficiency anaemia	191137004	Iron deficiency anemia NOS	Other Diseases	Disease	C0012634	Disease
B80	Iron deficiency anaemia	191133000	Unspecified iron deficiency	Other Diseases	Disease	C0012634	Disease
D70	Gastrointestinal infection	128398001	Infectious disease of digestive tract	Infectious	Infectious disease	C0021311	Infection
D71	Mumps	36989005	Mumps	Infectious	Infectious disease	C0021311	Infection
D71	Mumps	240526004	Mumps parotitis	Infectious	Infectious disease	C0021311	Infection
D72	Viral hepatitis	3738000	Viral hepatitis	Infectious	Infectious disease	C0021311	Infection
D72	Viral hepatitis	186642009	Unspecified viral hepatitis	Infectious	Infectious disease	C0021311	Infection
D76	Malignant neoplasm pancreas	363418001	Malignant tumour of pancreas	Neoplasms	Neoplastic disease	C0027651	Neoplasms
D76	Malignant neoplasm pancreas	187800001	Malignant neoplasm of pancreas	Neoplasms	Neoplastic disease	C0027651	Neoplasms
D76	Malignant neoplasm pancreas	269556009	Ca pancreas NOS	Neoplasms	Neoplastic disease	C0027651	Neoplasms

Figure 6.7: Example of shared parent relation between ICPC2 and SNOMED CT via UMLS Parent CUI

Quality assurance of the mappings.

The quality of the mapping between ICMV and ICPC2 was evaluated by mapping all terms from ICMV to the Italian terms in UMLS, using an exact match, as supported by the UMKS application programming interface. Among 1659 ICMV terms a total of 1232 mappings (655 unique ICMV terms) were found to Italian terms in the UMLS, associated with 690 unique UMLS CUIs. Table 6.4 shows an extract of the results of this mapping process.

More than 55% of the mappings derive from the Italian version of MedDRA, 40% from the Italian version of MeSH, and only 4% from the Italian version of ICPC2 in UMLS, as shown in the following chart 6.8:

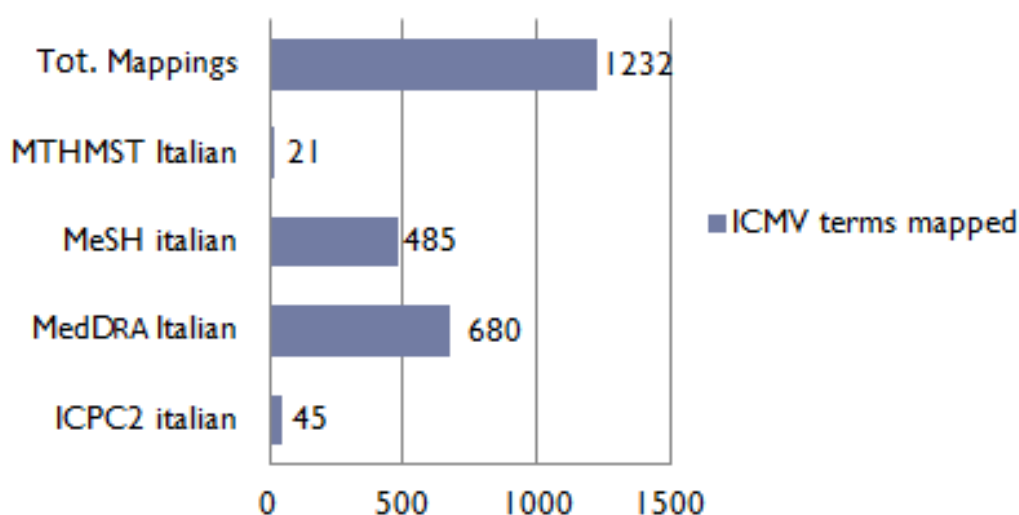


Figure 6.8: Mapping to UMLS Italian Resources

The tool provided by NLM, the Knowledge Source Server (KSS) was an important tool for supporting the mapping of non-regular forms, such as those composed entirely of general language words. However, many cases of ambiguity were found using this approach (e.g., mapping to acronyms). One of the best example for this issue is the case of the Italian term “Anca” (Hip), for which the tool provided both the mapping to the UMLS Italian concept “Anca” (Hip), from MedDRA Italian and the mapping to ANCA - Anticorpi antineutrofili an-

ICMV Term	UMLS CUI	Concept Name	Source Name	Concept ID
COLICHE ADDOMINALI	C0000729	Dolori crampiformi dell'addome	MDRITA	10000058
BALLO DI SAN VITO	C0152113	Corea reumatica	MSHITA	D002819
ACNE	C0001144	Acne volgare	MSHITA	D000152
AIDS	C0001175	Sindrome da immunodeficienza acquisita	MDRITA	10000565
DEMENZA SENILE	C0011268	Demenza senile	MDRITA	10039966
DEMENZA SENILE	C0002395	Malattia di Alzheimer	MSHITA	D000544
MAL DI SCHIENA	C0004604	Dolore dorsale	MSHITA	D001416
BOLLE	C0005758	Vescicola	MSHITA	D001768
CALLO	C0006767	Callo osseo	MSHITA	D002146
MUGHETTO	C0006849	Candidosi orale	MSHITA	D002180
GUANCIA	C0007966	Zigomo	MSHITA	D002610
CRISI EPILETTICA	C0014544	Epilessia	MSHITA	D004827
FUOCO DI SANT' ANTONIO	C0014714	Ergotismo	MSHITA	D004881
ATTACCO CARDIACO	C0027051	Infarto del miocardio	MSHITA	D009203
FEGATO INGROSSATO	C0019209	Epatomegalia	MSHITA	D006529
FECI NERE	C0025222	Melena	MSHITA	D008551

Table 6.4: Example of mappings between ICMV terms and UMLS Italian concepts

ticitoplasma (Antineutrophil Cytoplasmic Antibodies) from MeSH Italian.

Concerning the extension of the original mappings between ICMV and ICPC2, we found that by integrating ICMV with UMLS by using ICPC2 we were able to associate more 1,167 new mappings to our ICMV lay terms, in particular 650 derived from SNOMED CT, 197 derived from MeSH, 250 derived from ICD10 and 70 derived from LOINC.

By means of the direct mapping to UMLS Italian terms, as explained above, only 40% of the terms were mapped to ICPC2 Italian terms, all of which overlaps with the manual mapping created by physicians during the clinical review. So taking into consideration the other mappings to UMLS Italian sources, as shown by Figure 6.8 we found 485 new mappings to MeSH, 680 to MedDra and 21 to UMLS proper codes. Finally, considering the CUIs mapped to ICMV (690 unique CUIs) and performing queries among the other resources (both English and Italian) by using the CUI as a bridge, we extended our mappings to 1990 shared concepts in the other terminologies, distributed as follows: 1059 derived from SNOMED CT, only 90 derived from ICPC2 (not already explored), 454 derived from MeSH, 190 derived from ICD10 and finally 200 derived from LOINC. Comparing this type of extension with the one obtained by using ICPC2 as a bridge to the other resources we can state that the best way to find new mappings to specialized medical terminologies for our ICMV terms is to use UMLS CUI as a bridge. On the other hand concerning the type of approach used to map ICMV to ICPC2 (Italian terms) we found that the manual mapping performed by physicians, even if time consuming, was the most profitable way to associate ICMV with ICPC2 concepts, not only for the fact that it takes into consideration only Italian terms but also because physicians were able to find mappings to ICPC2 for those terms, in ICMV, which are not related to symptoms, diseases and medical procedures, terms that could not be mapped to any concept in UMLS using the automatic approach.

6.3 Manual vs. Automatic Mapping

Table 6.5 compares all the approaches used in this integration process to map ICMV to other standard medical terminologies, and above all shows the differences between the use of manual the mapping to ICPC2 and the automatic mapping via UMLS.

Type of Mapping	UMLS unique CUIs	ICMV terms	ICPC2 concepts	Other sources
ICMV2UMLS Italian	690	655	45	1187
ICMV2ICPC2 Manual	0	1502	572	0
ICMV2UMLS via ICPC2	336	523	587	1773
ICMV2ICPC2 via UMLS	570	559	90	1903

Table 6.5: Comparing the mapping approaches

Considering that the original mapping of ICMV to ICPC2 comprises 1521 correspondences among 2348 ICMV terms and given the results extracted by mapping ICPC2 to the other UMLS vocabularies, and the ones by mapping ICMV directly to UMLS Italian, and finally the results obtained by mapping ICMV to ICPC2 and other resources via UMLS, we suggest that the best way to integrate the ICMV to specialized medical terminologies is by using UMLS for accessing both to ICPC2 and to the other resources. On the other hand, considering only the mappings to ICPC2, we found that the manual mapping performed by physician was more profitable with respect to the use of UMLS.

Part III

APPLICATIONS

Chapter 7

Experimental use of ICMV*

This Chapter constitutes the third part of the thesis and presents two experimental uses of our ICMV and its integration framework. The first use case for the ICMV is its distribution on the web, under the form of an Italian consumer oriented healthcare Wiki, which can be used for browsing and searching purposes. The second experimental use of the ICMV is its integration with a new Personal Health Record, developed at Fondazione Bruno Kessler for the Province of Trento, namely TreC (Cartella Clinica Elettronica) to evaluate its feasibility and to improve its readability and use.

The Chapter is structured as follows: Section 7.1 describes the publication of ICMV in the format of a Semantic Media Wiki (ICMV Wiki), and Section 7.2 presents the case study related to the integration of ICMV in the TreC PHR. Results and evaluations for the two experimental uses are provided at the end of each section.

***Acknowledgements:** In this chapter we will use materials and data which are part of the project TreC, funded by the Province of Trento and which in part supported this thesis work, and in particular the experimental use of ICMV in the TreC PHR.

7.1 Distribution of ICMV on the web

In order to distribute our ICMV and to integrate it in the Web 2.0 vision, we decided to use the collaborative Media Wiki system², a popular free web-based wiki software application (e.g. Wikipedia), developed by the Wikimedia Foundation. It is very simple to customize and to re-publish. Concerning technical aspects, it is written in the PHP programming language and uses a backend database. During the last ten years, numerous wikis have been created around the world to power websites. An extension of this system is the Semantic Media Wiki³, first released in 2005, that helps³ not only to search, organise, browse, and share but also to tag, evaluate, store, query and the wiki's content. In fact, while traditional media wikis contain only text which computers can neither understand nor evaluate, the Semantic Media Wiki adds semantic annotations that allow a wiki to function as a collaborative database. Some applications of this software can be found in Millard *et al.*, 2006 [79]. While an example of the use of a Media Wiki to publish consumer-oriented medical vocabulary is the Consumer Health Vocabulary Wiki⁴ developed by the Open Access, Collaborative Consumer Health Vocabulary Initiative and in collaboration with the University of Utah, Department of Biomedical Informatics, allowing browsing of the terms in the Vocabulary by using as a key both the lay term or the UMLS CUI (since it is mapped to UMLS), and allowing users to leave comments on the Wiki page.

As explained in Section 4.2, we already used a Media Wiki, namely eHealth-Wiki, as a first methodology to acquire consumer-oriented medical knowledge related to symptoms and diseases. In this step we revised this wiki and we adapted it to ICMV's content in order to produce a site designed to provide an open access consumer-oriented vocabulary for Italian speakers and allow public comment on the concepts and terms included in the ICMV (Vocabolario

²<http://www.mediawiki.org/wiki/MediaWiki>

³http://semantic-mediawiki.org/wiki/Semantic_MediaWiki

⁴<http://consumerhealthvocab.chpc.utah.edu/CHVwiki/>

Medico dei Termini Comuni). This way the ICMV Wiki can be both browsed and continuously updated in a collaborative way, both by users with new terms and lay synonyms, and by physicians with comments or with new mappings to technical terms in ICPC2 or in other standardized terminologies. Figure 7.1 shows the main page of the ICMV Wiki which provides an explanation about the vocabulary.

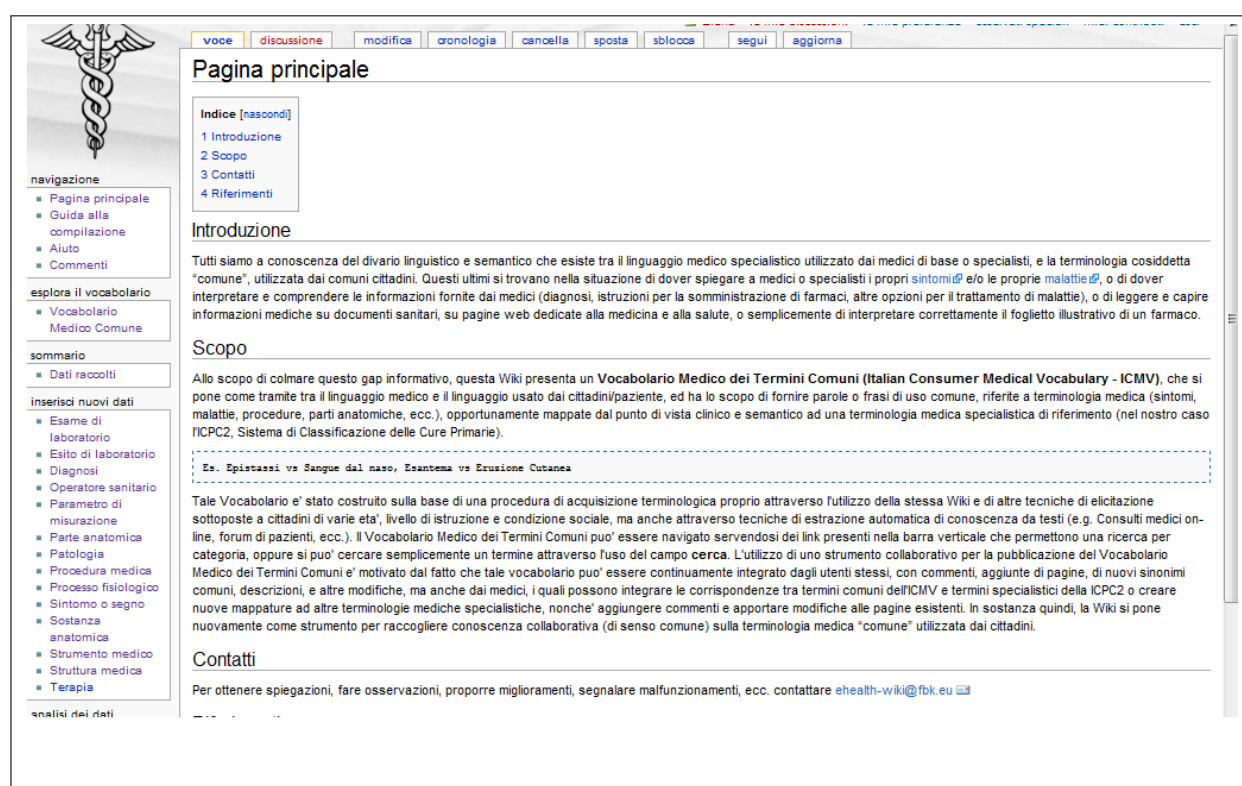


Figure 7.1: The ICMV Wiki main page

To adapt the eHealthWiki to the ICMV, we added many new templates, forms, and properties. First of all it is important to explain that in the Wiki System each ICMV term became a wiki page containing structured content, according to the templates and properties used. In our case, for each page in the wiki we created the following templates, according to the attributes associated

with each term in the ICMV (e.g. ICMV code, Familiarity Degree, Description, Synonym, Related Term, Body Part, Mapping to ICPC2, Type of Mapping to ICPC2, and for the category “Disease” also a field for the associated symptoms). Some of the values of these templates connect to other pages in the wiki (e.g. synonyms, body parts, related terms, and associated symptoms in the case of diseases), while the other ones are simple text, strings or integer values. We also integrated the eHealthwiki with the new categories deriving from ICMV (in the previous version we had only diseases and symptoms): medical procedures, diagnoses, body parts, therapies, medical devices, health and medical facilities, healthcare professionals, bodily functions, and measurement parameters. Figure 7.2 provides an overview of these categories in the Wiki.

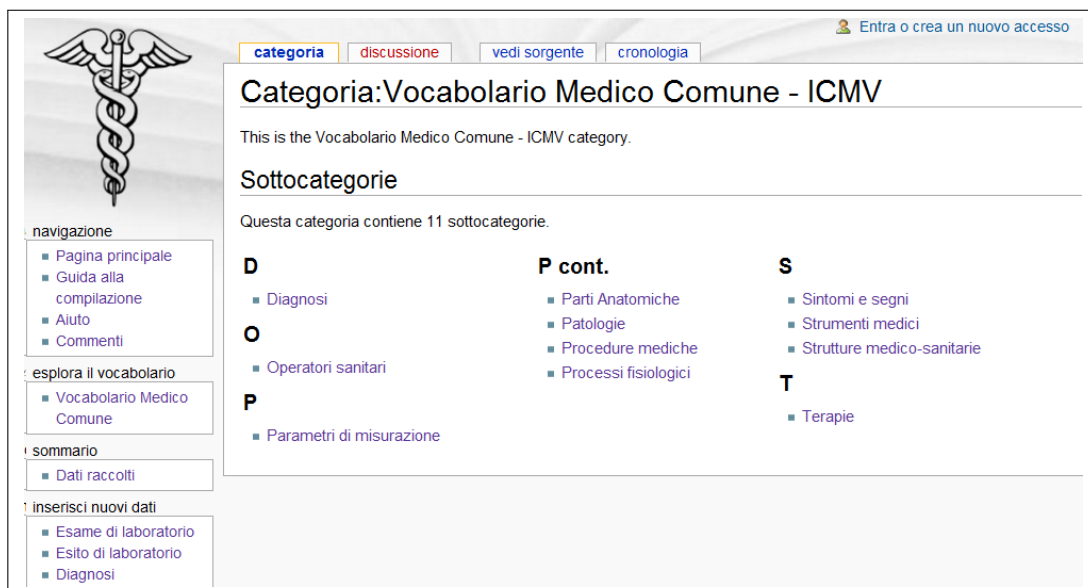


Figure 7.2: ICMV categories shown in alphabetical order in the wiki

We uploaded to the Wiki 11 categories among which the 2348 ICMV terms are distributed. We used a Java program to automatically create the XML MediaWiki file format to import all the ICMV terms. This XML file is composed of a first part containing all the wiki namespaces and then all the pages to be imported which have special tags for specifying the category of the page and the template for visualization, and finally a tag including all the templates and

properties with the corresponding values which compose that particular page. Since 225 wiki pages already existed in the old wiki, we merely updated them by adding the new content deriving from the ICMV for those terms, in particular the ICMV code, new synonyms, and the mapping to ICPC2. To give an example of a lay term represented using the ICMV wiki, Figure 7.3 shows the ICMV wiki page for the symptoms “Ematoma” (Bruise), which is composed of 6 templates including respectively: Description, Synonyms (e.g. Livido, Botta, Bollo blu, etc.), ICPC2 mapping (S16_Contusione/Ecchimosi - Contusion/Ecchymosis) the type of mapping (in this case the term “Ematoma”, described as “raccolta di sangue all’interno di un organo o di un tessuto” is synonym of the ICPC2 concept “Ecchimosi” or “Contusione”), and finally the ICMV code and the degree of familiarity of the term (in this case the highest one, “5” meaning that the term is well understood and used frequently by lay people and used alternatively to “Livido” which also received a “5” as familiarity degree).

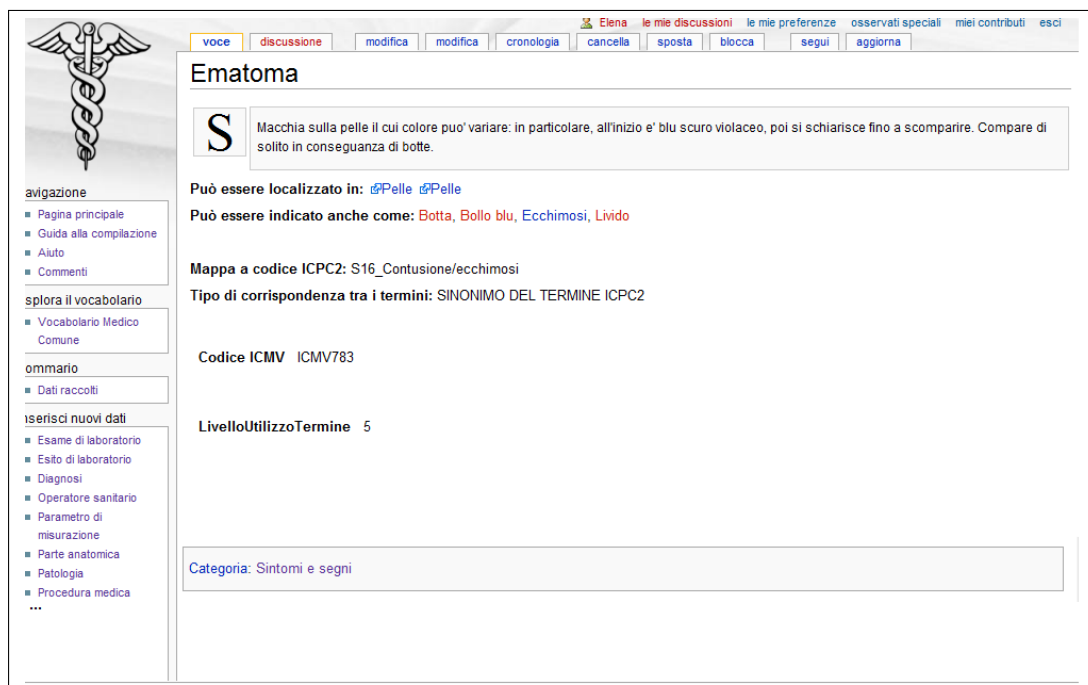


Figure 7.3: Representation of the term “Ematoma” in the ICMV Wiki

We should emphasize that many of the pages in the wiki related to categories

such as body parts, health facilities, health professionals, measurement parameters and health devices do not have any assigned description or synonym or ICPC2 code. This motivated our choice of a wiki system to publish the vocabulary since these “incomplete” pages can be collaboratively updated by ICMV wiki users.

Finally, to conclude this section, we give some statistical results about the content of the ICMV Wiki and about its most visited pages, as shown in Tables 7.1 and 7.2. From the installation of the ICMV Wiki to-date, 35,484 pages have been visited and 7,629 have been created by 50 registered users.

Type of entity	Total
Templates	51
Forms	11
Properties	22
Number of pages	2381

Table 7.1: ICMV Wiki content statistics

Most visited page	Number of visits
Visione dei dati raccolti	5.834
Pagina principale	1.279
Dati raccolti analisi	1.205
Abbassamento della voce	588
Offuscamento della vista	351
Creazione nuovo sintomo	330
Creazione nuova patologia	312
Categoria:Vocabolario Medico Comune - ICMV	274
Guida alla compilazione	252
Pesantezza di stomaco	245

Table 7.2: ICMV Wiki most visited pages

Other statistics, for example those concerning the number of pages which link to other pages in the wiki, or other semantic statistics, can be found directly

in the ICMV Wiki⁵, following the link “Special Pages”.

7.2 ICMV for Personal Health Records

Personal health records (PHR) are easy-to-use applications designed for managing information about a consumer’s health, and add information by the consumer. They are usually as electronic files or records of a consumer’s health information and recent services such as medications, allergies, results of laboratory tests, and doctoral or hospital visits that can be stored in one place, and then are accessible to the consumer and to those authorized by the consumer. The content of each personal health record is captured from various sources such as from a hospital information system, a clinical laboratory, a pharmacy or from the patient him or herself. In this section we will describe TreC, an Italian-language system that implements a PHR for the Province of Trento, funded by the Department of Health and the Department of Research and Innovation of the Autonomous Province of Trento, which involves many partners⁶ including the Fondazione Bruno Kessler, the chief developer of TreC. After an overview of TreC, we will present the integration of the ICMV with this PHR to make it closer to consumers from the terminological point of view, also proving a more understandable user interface, in order to ensure easy and friendly access to and management of their health data. The importance of the readability of PHRs is well described by (Tang *et al.*, 2006 [113]) stating that “In order to be useful to the patient, the PHR must present data and accompanying tools in ways that enable the individual to understand and to act on the information contained in the record. Both terminology and data presentation must be adapted to the

⁵The ICMV can be navigated at:<http://ehealthwiki.fbk.eu>

⁶Fondazione Bruno Kessler, the local health authority (Azienda Provinciale Servizi Sanitari), Department of Sociology and Department of Law (University of Trento), Medical Design Research Unit (IUAV)-Veneto, and finally the companies Argentea - Gruppo GPI, and Attrezzature Medico Sanitarie (Trento). More information about the project are available at:<https://www.trec.trentinosalute.net/web/guest/home>

individual using the PHR, so that they realize optimal benefits”.

7.2.1 TreC - Cartella Clinica del Cittadino

TreC (or 3C), which stands for “Cartella Clinica del Cittadino”, is a Project that started three years ago, and aimed at the development of an Italian PHR for healthcare consumers in the Province of Trento, particularly designed to improve the process of care and patient assistance. TreC allows consumers to directly access their clinical documents produced by healthcare organizations (e.g. laboratory test results, medical reports, clinical notes, and discharge letters) and also to have a health diary to fill in and archive data related to their health conditions such as: Allergies, Diseases, Reasons for Encounter, Vaccinations, Medications and Treatments, Observable parameters, Medical examinations, Hospitalizations, among others. At the moment, TreC PHR is being tested by about 400 citizens in Trento, and for this step only the most important functionalities have been activated (e.g. receiving medical reports from APSS (Azienda Provinciale per i Servizi Sanitari), adding personal and family clinical history, adding allergies, intolerances and vaccinations, consulting their data, and adding medications and treatments. If used for monitoring aims, and if filled in constantly, TreC also provides alerts and reminders to consumers (for instance for drugs consumption in a therapeutic context, for reminder of a medical examination, or for warning patient about drugs or food/drug interactions). The TreC system has been designed not only to allow people an easy and web-based way to manage and store their data, but also to provide interactions between patients and physicians and other health providers in order to facilitate patient monitoring and assistance. This PHR is accessible not only on the web, but also on latest generation Smart Phones and Tablets (Android-based and touchscreen) and can be connected by bluetooth to other devices (scale, BMI marker, blood pressure gauge, and other devices). As shown by Figure 7.4, the different tasks for the development and management of TreC are dis-

tributed among the various stakeholders as follows: the Province of Trento is responsible for user registration on the system; the regional health authority (APSS) is responsible for the data which need to be exchanged (in this case the patient medical reports and patient personal data); and finally Fondazione Bruno Kessler is responsible for the development and maintenance of the TreC Web Portal (based on Liferay). Concerning the technical aspects and the technologies considered in the architecture of TreC, on one hand we have the use of three type of networks to communicate and exchanging data, namely Internet, Telephone and Mobile networks and a security access systems which uses strong authentication systems such as Smart Card CNS and/or OneTimePassword by mobile phone to guarantees high security levels. On the other hand concerning interoperability between systems and data exchanges, we have the Middleware which allow data exchange between APSS Data Bases concerning Medical Report Service and Patient Registry service, and the TreC Web Portal. TreC is based on the HL7 CDA standard for data model structuring, messaging and APSS-physician-patient interaction. TreC is also well integrated with the Web 2.0 vision for its technologies, design and functionalities, and above all for the active role given to the patient/consumer in filling in his PHR, and for the social perspective. In fact, TreC also allows a collaborative management of the PHR by other family members or friends, for instance in the case of children, elderly or disabled persons.

The TreC Data Model is based on the Project Health Design (PHD)⁷ platform components and functionalities, in particular for the following target areas:

1. Identity Management. Resources to help Personal Health Applications (PHAs) manage the list of users and software systems that are authorized to access patient-specific data, authenticate the identity of users and systems

⁷Project Health Design is a research initiative sponsored by the Robert Wood Johnson Foundation and the California HealthCare Foundation to explore the design of advanced Personal Health Applications that address various health conditions and wellness goals:www.ProjectHealthDesign.org

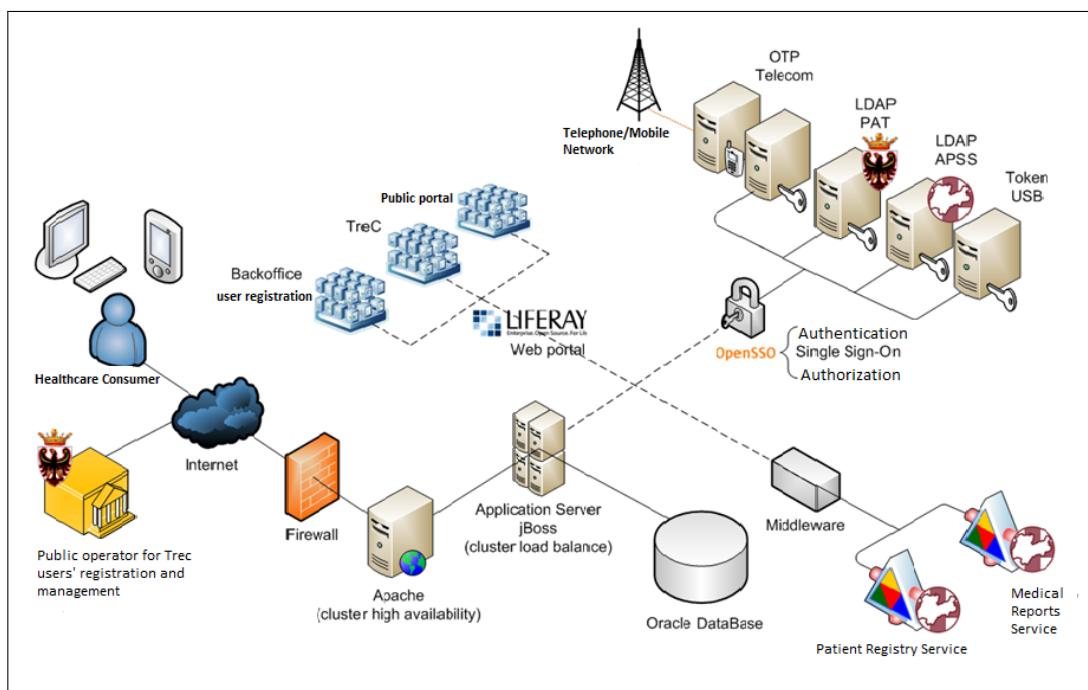


Figure 7.4: TreC Architecture

that are requesting access to this data, and allow patients to monitor and precisely control the access that is provided to their health data.

2. Medication List Management. Resources to help PHAs record, manage, share, and provide advice based on the list of specific medications that a patient takes regularly.
3. Calendaring. Resources to help PHAs record, track, share, and remind patients of specific scheduled events that are relevant to the management of their health or medical conditions.
4. Observations captured in the course of daily living. Resources to help PHAs store, aggregate, analyse, and share data recorded by patients that are relevant to the management of their health or medical conditions and that are captured outside of their encounters with the health care system.

More details about PHD platform components and requirement can be found in [112].

On the user side, the TreC data model is organized in modules involving different aspects of patient health care. These modules are graphically represented in the user interface by Widgets, whose principal task is to “atomize” and make user interaction with the tool itself easier. Practically these widgets are mini-applications which offer dedicated services such as records data entry, their visualization, communication between users or between users and local health organizations. Widgets are grouped in 11 main categories according to the service and function they offer:

- Medication management, which includes the widget devoted to Medications and Therapies management;
- Personal Diaries, which include personal diaries filled in by a user who follows particular programs of care (e.g. “sleep diary”, “pain diary”, “smoke cessation diary”, etc.);
- Observable parameters/Measurements, which include all widgets allowing a user to annotate specific observable clinical parameters (e.g. “blood pressure gauge”, “body temperature”, “BMI”, “LDL cholesterol”, “weight”, and “heart rate”) measured by means of specific devices;
- Medical Reports, which include a specific widget which allows the reception of medical reports such as text results, discharge letters, and other reports directly from the local health organization, and which allows uploading a user’s scanned clinical reports.
- Women’s health, which includes widgets such as “Woman physiological history”, allowing for storage of data such as the day of menarche for young women, pregnancy-related data, menopause data; “Menstrual cycle diary”, for menstruation management, and finally a widget devoted to fertility monitoring.
- Child health, which includes a widget for the constant control of a child’s growth (head circumference, weight and height measurements).

- Familiar Clinical History, which includes a widget allowing the entry of data related to relevant familiar diseases (e.g. Neoplasms, Diabetes, Hypertension, heart diseases, etc.)
- Personal Clinical History, which unlike the previous ones includes widgets devoted to the filling in of personal records related to consumer clinical history, such as current and past “Relevant diseases”, “Allergies and intolerances”, “Vaccinations”, “Medical examinations”, “Surgery”, and “Hospitalizations”.
- Patient Summary, which is a single widget including a summary of the relevant health care data (to be known in case of emergency) deriving from the other records, for instance personal data, medications allergies; relevant diseases; family history, current drugs treatments, and vaccinations.

Figure 7.5 shows an overview of the widgets in the web portal, which can be selected and used by TreC users.

For each widget, the users is asked to fill in a record which in most cases includes the date entry, then depending on the purpose of the widget, the insertion of the disease, of the symptoms, of the allergies, etc. choosing a list of terms presented on a pop-up page which includes a search field, and some filters which allow a categorized representation of the term list. After adding these data, the user is asked to add several other parameters and save the page. In this way all the added data are visible on the summary page “Vedi tutti i dati” and the most important data are saved on the patient summary, namely “Scheda sanitaria”.

In this context, our focus was on the terminological and classification aspects regarding some of the widget listed above, in particular the one concerning personal and family clinical history. More precisely our case study was to use the knowledge acquired during the process of development of ICMV to provide tables including lay terms for specific medical sub-domains, associated with

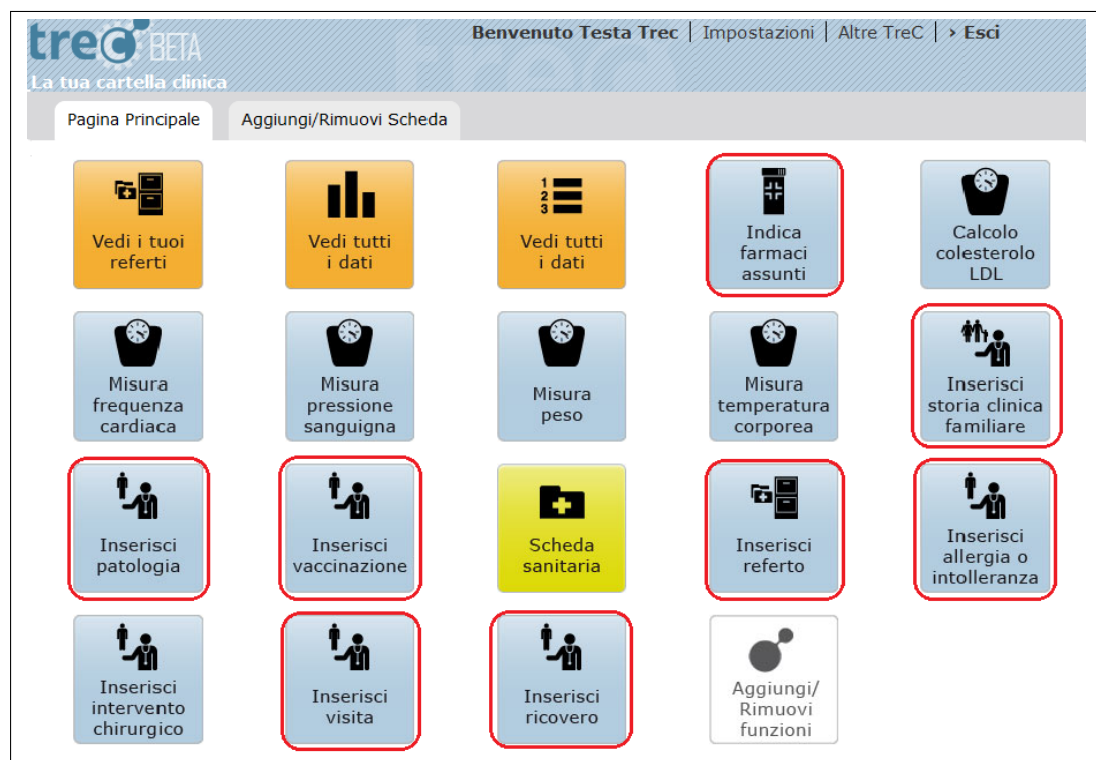


Figure 7.5: TreC widgets in the user interface

synonyms and mapped to the standard ICPC2. The objective of this experimental use of ICMV in TreC is, on one hand, to help TreC users in filling in their personal health records and in accessing their data, and on the other hand to allow data integration between different healthcare information systems once TreC data begins to be exchanged with the general practitioner or other professionals. The methodologies and results of this case study are explained in detail in the next sub-subsections.

7.2.2 Integrating ICMV in TreC

Prior efforts to improve consumer-friendliness of PHR and EHR information have focused on user interface design and/or the links to references or educational materials (i.e., infobuttons) (DeClercq, 2003, and Baorto and Cimino, 2000 [35, 12]). Less attention has been given to the underlying linguistic fea-

tures of PHRs. This case study employs text categorization and simplification as a method for improving the readability of TreC PHR texts and data entry for consumers. Since vocabulary is a key factor in health text readability (Keselman *et al.*, 2007 and Rosemblat *et al.*, 2006 [69, 99]), we focused on the following tasks:

- PHR medical term replacement with those from ICMV,
- Categorization of the list of terms needed in the TreC widgets by using background knowledge for each particular domain treated in these widgets beyond standardized classification systems,
- Integration of these terms with specialized medical coding systems used by physicians during the encoding of primary care reason for encounters and by other professionals in the patient process of care.

For this task the mapping between ICMV and ICPC2 created in this thesis has been integrated into Trec. As mentioned above, TreC includes many widgets whose data model is built around specific lists of terms. In this case study we focused on the improvement of lists of medical terms related to:

- Allergies and Intolerances (in this case we also considered “Adverse Reactions”);
- Reasons for Encounters (Symptoms, Medical Procedures and Diseases);
- Family History (relevant diseases occurring in one or more family member);
- Vaccinations;
- Personal Clinical History;
- Medical Examinations;
- Medical Reports.

In order to create a new list of domain-oriented terms or integrate the existing lists of terms in TreC, we first analysed the domains of the applications mentioned above. This task started with some meetings with domain experts (allergists, laboratory test technicians, primary care physicians, and an immunologist) in order to understand the use of specialized and lay terms for those particular topics and in a care context. During these meetings physicians also provided lists and tables including the most common terms and local classifications used in their departments or laboratories for identifying or coding patients' problems.

After acquiring domain knowledge for the areas of examination we automatically extracted from ICMV all the terms and categories (where available) related to the topics of application, and finally, we integrated them with those included in the lists provided by the above mentioned professionals in order to extend coverage for each topic. Here we have to highlight that many of the terms included in the new lists (especially those technical terms) were new terms, not included in our ICMV. This was particularly observable for example, for those terms related to topics such as Allergies, Intolerances, Vaccinations, and Medical Reports, which refer respectively to allergens, causes of reactions, vaccines, diagnostic, observatory and laboratory tests, all categories not considered during the acquisition of lay terms for the creation of ICMV. We also provided terms and categorizations to be used in widgets such as the "Pain diary" and "Menstrual cycle diary", where users are asked to add the type of pain, the intensity (using a standard pain scale), the location of the pain and the frequency for the first, while for the second users are asked to add symptoms during the cycle, symptom location, and intensity of menstruations. Apart from the type of pain, the pain intensity and the pain frequency for which standard classifications such as Numeric Rating Scale (NRS), Verbal Rating Scale (VRS) and

in particular the Italian Pain Questionnaire⁸ (De Benedittis *et al.*, 1988 [15]) have been applied, for menstrual symptoms and anatomical concepts we used a subset of our ICMV terms. Furthermore, to make it easy to access anatomical concepts, they have been categorized in 5 main filters/categories:

- Head and Neck;
- Chest/Upper Abdomen and Back;
- Upper and Lower Limbs;
- Lower Abdomen/Pelvis and Lower Back;
- General area (Unspecified area).

Once we had the list of medical terms to be used in the TreC widgets for data selection, we integrated them with the ICPC2 codes in order to provide useful mappings for the exchange of data between consumers PHRs and professionals EHRs. This association was automatic for the terms deriving from ICMV, since they are already mapped to ICPC2, while we had to perform the task for the new terms. For some of them (in particular symptoms, diseases and medical procedures) we performed an automatic exact matching process to ICPC2 concepts, while the remaining terms (300) needed to be manually mapped to ICPC2 by domain experts (this new mapping is still in an experimental phase). Here it is important to highlight that most of these remaining terms are related to Allergies, Intolerances, and Medical Examinations, and could not be mapped to ICPC2 since it is designed for classifying primary care encounters and does not provide sufficient support as a referent specialized terminology in this case, because its granularity is not deep enough to cover for example types of allergies or types of blood tests (e.g., etc. In fact, for the categories mentioned above ICPC2 only includes terms such as: “allergies”, “allergic reaction”, “dermatitis contact”; “blood test”, “urine test”, “faeces test”, “histological/exfoliative

⁸This is the Italian translation of the McGill Pain Questionnaire, developed at by Dr. Melzack at McGill University in Montreal Canada to be used to evaluate a person experiencing significant pain

cytology”, and “other laboratory test NEC”.

Due to these problems, other classification systems/terminologies related to the specific domains had to be integrated in TreC beyond our ICMV (e.g. the local classifications for laboratory tests, including 519 entry codes assigned according to the national laboratory tests rate table, for Medical Reports, and for Vaccinations, deriving in this case by the National Institute of Health. As said before, the mapping task for the new terms is still in an experimental phase, so it is not included in the TreC platform at the moment.

7.2.3 Results and Services for TreC

In this subsection we present some statistical results concerning the ICMV terms used in TreC and the categorization provided for accessing data in each widget. The integration of ICMV in TreC produced a set of terminological tables for the different medical topics treated in the TreC widgets, for a total of 837 lay terms, divided as follows: 6 for Allergies and Intolerances, 32 for Adverse Reactions, 258 for Personal Diseases, 33 for Family Diseases, 15 for Vaccinations, 14 for Medical Examinations, 235 for Symptoms, 154 for Anatomical parts, 15 for Menstrual Symptoms, 37 for Medical Procedures, and finally 38 for Medical Reports. Table 7.3 shows more detailed statistics concerning the (consumer- and professional-oriented) terms integrated in TreC, specifying for each widgets how many categories (filters), and subcategories (second level filters) have been created, how many terms have been extracted from ICMV and how many terms were included from other domain-oriented resources.

Some examples of the resulting lists and categories used for helping a user in add data in TreC are shown in Figures 7.6, 7.7, 7.8, and 7.9. In particular Figure 7.6 shows a TreC page which allows the user to look for and add Allergies and Intolerances. On the right of the page user can find filters/categories to access

Widget	Categories	Subcategories	ICMV	other sources	Terms
Allergies and Intolerances	8	15	6	127	133
Adverse Reactions	2	2	32	19	51
Personal Diseases	18	0	258	198	456
Family Diseases	2	0	33	20	53
Vaccinations	4	7	15	6	21
Medical Examinations	2	0	14	14	28
Symptoms	18	0	235	61	296
Anatomical parts	5	0	154	0	154
Menstrual Symptoms	0	0	15	0	15
Medical Procedures	5	0	37	0	37
Medical Reports	12	12	38	519	557
Tot.	76	36	837	964	1801

Table 7.3: Statistics of terminological tables integrated into TreC widgets

the related subset of terms available for this widget. In this example the user selected “inhalation allergies” and in particular “pollen allergies”, to access (on the left of the page) only allergies caused by pollens such as plants (olive tree, beech tree, chestnut, tree, grass family, legumes, etc.).

Figure 7.7 shows the TreC page for searching and adding Vaccinations. Even in this case, we find filters/categories on the right (e.g., Vaccinations for Adult people - Travellers), and the available list of terms for the selected filter on the left (e.g. tetanus, malaria, rabies, streptococcus, yellow fever, etc.).

Figure 7.8 shows the TreC page used by users for searching and adding reasons for encounters in their record, more precisely, symptoms, medical procedures or pathologies (3 main categories which can be chosen on the right of the page). In this case symptoms and pathologies have the same subcategories which refer to their anatomical location (according to the ICPC2 body systems). On the left of the page we can see all the symptoms related to the selected subcategory “General or not specified symptoms” such as *fever*, *general weakness*, *chills*, *swelling*, etc.



Figure 7.6: Searching and adding Allergies or Intolerances in TreC

Finally Figure 7.9 shows the TreC page used by users for searching and adding diseases in their “Personal Clinical History” widget. As for the pathologies included in the “reason for encounter” record, diseases have also been categorized according to ICPC2 body systems. The filter/category selected in the figure is “Eye diseases”.

In addition to the specific filters/categories used for each widget, the user can also use other two filters. One shows the whole list of terms available for a particular widget and another shows only the terms (pathologies, symptoms, allergies, etc.) added by the user in the case where he did not find the term he was looking for.

As mentioned in the previous subsection, TreC PHR is still under going testing, so we are not able to evaluate the usability of ICMV in TreC in terms of use of lay terms instead of the technical ones, or in terms of making easier



Figure 7.7: Searching and adding Vaccinations in TreC

the actions of browsing the system and data entry by using the filters provided for accessing and adding data rather than using simple search or manual editing. The only available test performed during this experimental use of TreC is a usability test taken during the months Novembre-Dicember 2010, when our ICMV was not already integrated with TreC. More information about this test can be found in (Purini and Piras, 2011 [95]). Even if we cannot give results about the usability of ICMV in TreC, we are sure that the use of ICMV for integrating lay terms related to symptoms, diseases and medical procedures, together with the use of other domain terminologies/ontologies where a more specialistic knowledge is needed, can doubtless improve the usability of TreC and help consumers in accessing and managing their healthcare data. This is motivated by the fact that before the integration of ICMV with TreC, the PHR

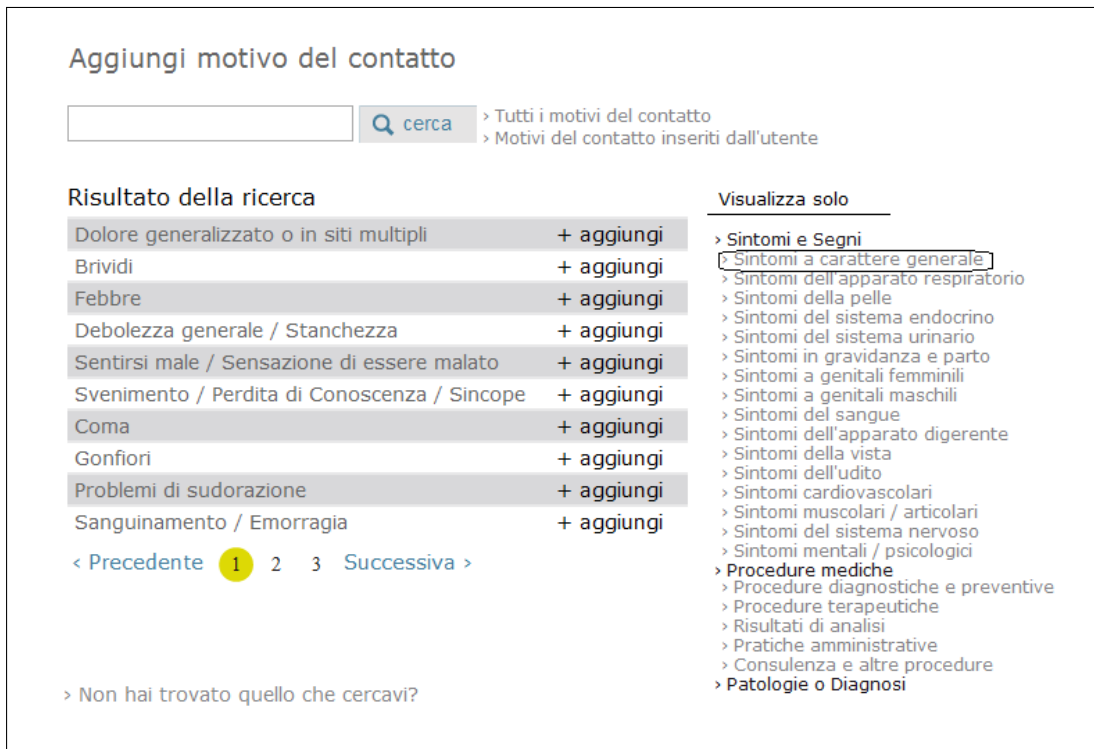


Figure 7.8: Searching and adding Reasons for Encounter in TreC

used only limited flat lists including the most common entries for each medical topic (e.g. the 20 most common pathologies, the 8 most common allergic reactions, and a few others). Looking at the data considered in TreC, we have to admit that during the phase of knowledge acquisition for the creation of ICMV, we could have extended our domain coverage to types of allergies, food intolerances, vaccinations, laboratory tests, and medications and treatments in order to have obtained better results during the phase of ICMV integration in TreC, but the use of other domain sources together with the ICMV, is not completely negative, and indeed can be considered an added value.

Aggiungi patologia

Q
cerca

> Tutte le patologie
 > Patologie inserite dall'utente

Risultato della ricerca

Blefarite	+ aggiungi
Calazio	+ aggiungi
Cataratta	+ aggiungi
Cecità / Perdita della vista	+ aggiungi
Congiuntivite infettiva / batterica	+ aggiungi
Congiuntivite allergica	+ aggiungi
Contusione o emorragia all'occhio	+ aggiungi
Difetto di rifrazione	+ aggiungi
>Miopia	+ aggiungi
>Presbiopia	+ aggiungi

< [Precedente](#) 1 [2](#) [3](#) [Successiva](#) >

Visualizza solo

- > Patologie cardiovascolari
- > Patologie dell'udito
- > Patologie del sangue
- > Patologie del sistema urinario
- > Patologie della vista
- > Patologie a carattere generale
- > Patologie mentali / psicologiche
- > Patologie in gravidanza e parto
- > Patologie della pelle
- > Patologie digestive
- > Patologie del sistema endocrino
- > Patologie genitali femminili
- > Patologie genitali maschili
- > Patologie muscolari / articolari
- > Patologie del sistema nervoso
- > Patologie inserite dall'utente

> Non hai trovato quello che cercavi?

Figure 7.9: Searching and adding Diseases in TreC

Chapter 8

Conclusion

In this chapter we summarize the key contributions of this thesis. We outline the scope of the research performed and contrast it to the achievements conducted in recent related work. We conclude with an outlook for future research.

8.1 Summary and Contributions

In this thesis we addressed the problem of the linguistic gap between “lay” and “specialized” medical terminology in the Italian context through the creation of ICMV and its integration with standard medical terminologies/ontologies. In contrast to traditional approaches which propose the use of specialized medical terminologies to be integrated in consumer-oriented healthcare application, we proposed a new approach based on the creation of a new consumer-oriented vocabulary for the healthcare domain, and we evaluated in the Italian context, opportunely mapped to standardized medical terminologies, in particular taking advantage of the Semantic Web technologies, in order to provide sufficient support during the application of this lexi-ontological resource to consumer-oriented healthcare applications. On one hand, ICMV could help consumers in easily accessing and managing their healthcare data (through translation and interpretation services), while on the other hand, thanks to the integration framework we developed it could help physicians and other healthcare providers in

the process of encoding reasons for encounters (symptoms, diseases, diagnoses and procedures), to automatically interpret their patients clinical history stored in their PHR and to automatically produce clinical notes understandable by healthcare consumers and patients.

The main contributions for each phase are the following:

Phase 1. Generation of the Italian Consumer Medical Vocabulary.

During the first phase of this thesis work, we collected about 2,400 Italian “lay” terms and expressions used by healthcare consumers for describing symptoms, diseases/diagnoses, medical procedures, anatomical concepts, and few other healthcare topics, by using a hybrid methodology of knowledge acquisition and term extraction in the medical domain. This methodology combined the use of traditional elicitation techniques (focus groups, card sorting, and interviews, nurses-patients encounters) to acquire oral and tacit medical knowledge, with automatic term extraction from written documents (such as on-line medical consultations - 80,000 documents, and Triage records collected in an Emergency Room - 2,000 records), and the more exploratory Semantic Web techniques such as the use of ad hoc semantic media Wiki systems to acquire explicit knowledge (we collected about 230 Wiki pages). People involved in this acquisition process were: a sample of 32 people among researchers, PhD students and administrative staff of our research institute, aged between 25 - 56; about 60 patients per day in one month, ranging in age from 20 - 70, for the acquisition of lay terms in the Emergency Room; and finally, a community of 32 elderly people in a Seniors Club, from 65 to 83 years old. Using these techniques together allowed us not only to extract a large number of lay terms but also to extend our medical domain coverage (involving terms belonging to more than 60 medical specializations) and to cover a varied sample of people both in terms of age, background and level of healthcare literacy. We can observe that, using a hybrid approach and merging elicitation techniques to acquire lay medical terms

as we attempted in our approach, and involving a more varied sample of people is useful in improving the results, both from the qualitative and the quantitative point of view. In fact, as shown in Section 4.2, we collected a huge amount of data, consisting not only of the lay terms to be included in the Vocabulary but also in synonyms for them, descriptions, anatomical location for each symptom or disease, etc. In fact during the process of acquisition consumers were asked to provide for each term also other attributes such as a lay description of the term, one or more synonyms both lay and technical, related terms, a familiarity degree, the anatomical location for symptoms and diseases and finally for diseases the possible associated symptoms. As mentioned above, a large number of medical terms were extracted, which cover most of the healthcare topics (e.g. Cardiology, Psychology, Neurology, Gastroenterology, Traumatology, and many others). Furthermore we were able by means of this hybrid approach to compare and validate each type of acquisition looking at the others, in terms of overlaps among the terms, of categorization of the terms and for disambiguation in case of polysemy. On the other hand, we are sure that the use of only one of these acquisition techniques would be not enough to allow a good coverage of the healthcare domain, and would be also limited in terms of accuracy of the extracted terms. The number of extracted terms would be halved (in positive view) considering for example how many terms we extracted only using the Interview techniques with elderly people (only 160 terms in two days meetings), and consequently also the possibility to acquire synonyms or related terms for each extracted term would be decreased. So we think that we followed the right approach to collect consumer-oriented medical terminology.

An important thing to highlight in this context is that all the extracted terms which have been included in the ICMV have been chosen according to various parameters: a statistical analysis performed with the use of NLP tools among our corpora, a familiarity degree (assigned by consumers to each term which represent the level of readability of the term and its effective use in daily life

by consumers); and finally a clinical review performed by general practitioners, nurses and pharmacists, who validated the lay terms as representative of the corresponding medical concepts and helped us to address misconceptions both in the meaning of a term and in its categorization. The clinical review was very useful in our work because physicians analysed more than 3,000 terms finding also mistakes in the use of synonymy and deleting more than 400 terms. Consumer terms, as we know, are highly context sensitive. These general language terms which consumers use to describe medical concepts are more ambiguous than technical terms, so the support of physician and other healthcare professionals to address this issue in the right way is an added value.

First of all we integrated ICMV with ICPC2 by asking a group of physicians to find correspondences between the lay terms extracted through the knowledge acquisition task and the ICPC2 (taken in this work as a benchmark), using a manual approach based on a clinical and collaborative review. Our first results were encouraging because the overlap with ICPC2 was of 508 concepts out of a total of 681 ICPC2 concepts, and a high number of mappings were found (1521), defining also the type of matching, whether they are synonyms, whether there is an exact match between the two, whether there one is hyponym or hypernym of the other. We can highlight that most of the mappings extracted here expressed the synonymy relation.

Furthermore, physicians provided a clinical mapping between the term in the ICMV and the specialized medical terminology taken into account (the Italian version of the International Classification of Primary Care-ICPC2) assigning to each lay term one or more ICPC2 concept and defining also the type of matching (if they are synonyms, if there is exact match between the two or if one is hyponym of the other, etc.). This was a very important contribution for our thesis work, because even if we provided also automatic mapping to ICPC2 and other terminologies using exact matching technique or extraction of semantic mappings from UMLS, the quality of the manual mappings provided by physicians

consider the various contexts where a term can be used, avoid ambiguities and is more precise thanks to the definition of the type of mapping. Reporting some statistic concerning this task, physicians found among more than 1,500 mappings for a total of 2,348 lay terms included in ICMV, in most of the cases exact matching and synonyms. Only for the term extracted from the big written corpus including online medical consultations, physicians found more hyponymy relations between the ICMV terms and the ICPC2 concepts.

Phase 2. Formalization of specialized medical terminologies and ICMV using Semantic Web languages.

Encouraging results were found also when performing the process of encoding terminologies taking into account semantic web language, namely RDF and OWL. In fact we developed consistent light-weight ontologies, in particular for the ICPC2 and the ICD10 coding systems, and RDF graphs (represented as N-Triples) for LOINC and the ICMV. Concerning the first two terminologies, we have constructed and validated OWL representations of ICPC2 and ICD10 systems together with an explicit formal representation of the existing clinical mappings between them, which have a particular importance in healthcare processes in European countries. To formalize ICPC2 and ICD10 we used the Semantic Web standard OWL and applied the open-source OWL Reasoner Pellet to assess the logical coherence of the mappings between them. The results were very positive since the formal interpretation of the mapping we provided was consistent and we showed the benefits of using a logic-based formalization and analysis of mappings between medical classification systems. In fact, by applying and adopting recent developments in the Semantic Web, one can fruitfully reuse available logical methods and system implementations for the analysis of mappings between medical classification systems. In merging the three ontologies together we collected 15,600 classes, 54,903 disjointness axioms, and 20,592 subclass axioms.

Our experience with mapping current ICPC2 and ICD10 coding systems has demonstrated that the process of establishing mappings and their further accuracy analysis is a question of years involving significant human expert effort. We think that by pushing the use of Semantic Web developments, the efforts for upgrading mappings in view of the next publication of the two encoded systems, ICD11 and ICPC3 (WHO and WONCA have started the process of review and restructuring of these coding systems) can be noticeably reduced. To this concern, the WHO already started to provide a more formal representation of ICD for the eleventh revision using Semantic Web languages and tool such as WebProtégé for allowing the collaborative review of the coding systems and the creation of a new more formal and contextualized hierarchy and BioPortal as ontology repository to be used in conjunction with the collaborative tool for providing integration with other terminologies/ontologies (Tudorache *et al.*, 2010 [118]). Additionally, the OWL encoding of mappings allows us to advantageously reuse existing reasoning tools for checking mappings accuracy, as well as for debugging and fixing errors in mappings.

On the other, hand we also need to highlight that the suitability of OWL for representing classification systems has to be motivated by the complexity of the classification system itself, by its level of granularity and by the usefulness that this type of formalization could produce for example of providing reasoning on the classification content or allowing its reuse and interoperability, because some classification systems indeed do not use true subsumption relations and do not need a formalization using a high level of expressivity. This is the reason why much effort has been devoted to the use of less expressive Semantic Web languages to provide a semantic formalization of some medical terminologies and classification systems, such as SNOMED CT, and MeSH, which were encoded in RDF or SKOS (see for example Bodenreider, 2008a and 2008b, and van Assem *et al.*, 2006 [19, 20, 9]). This is also what we had in mind for the conversion of our ICMV in a Semantic Web language, in fact we used the RDF

language, because of the structure of our ICMV and because of the purpose of integrating it with other specialized terminologies expressed into Semantic Web languages. So using RDF syntax was enough to create a graph which can be integrated with other graphs and can be queries to extract semantic information (in our case semantic mappings).

Phase 3. Integration of ICMV with specialized medical terminologies.

We created an integrated framework for the ICMV and other standard medical terminologies. In particular we wanted to compare and improve the mappings found by physicians between ICMV and ICPC2 during the clinical review of the ICMV terms, with an automatic mapping procedure. To this aim, we integrating our ICMV with other medical terminologies, such as SNOMED CT, LOINC, MESH and ICD10 (all of them in English), leveraging existing mappings from the UMLS Metathesaurus and using ICPC2 as a bridge between ICMV and UMLS to extract mappings, since this coding system is already integrated both with ICMV and with the UMLS vocabularies. We took advantage of Semantic Web technologies, using the formalized versions of the resources mentioned in the previous phase, and reusing existing RDF graphs for SNOMED CT, MESH and UMLS Metathesaurus.

Using an RDF triple store for integration purposes (Openlink Virtuoso Triple Store v. 6.0.1), where we collected 18 graphs for more than 73 Million unique triples, we were able to extract mappings between terminologies through simple automatic SPARQL batch queries rather than ad hoc programming, reaching our goal of integrating the ICMV with the other medical terminologies in UMLS via ICPC2. By means of this new approach, we increased the number of mappings between ICMV and specialized terminologies (about 2,000 correspondences were extracted) and new synonyms and lexical variants were found for the lay terms in ICMV. A final evaluation of the quality assurance of these mappings showed that using ICPC2 as a pivot for extracting mappings from

UMLS is a more efficient way with respect to the one which maps all the terms from ICMV to the Italian terms in UMLS Metathesaurus. In fact, applying this second approach by means of an exact match, as supported by the use of the UMKS application programming interface, our results showed that only 655 ICMV terms were mapped to UMLS. Furthermore, more than 55% of these mappings derived from the Italian version of the MedDRA coding system, 40% from the MeSH Italian version and only 5% from ICPC2 Italian version. This results showed that to perform only exact matches to UMLS terminologies it is not so productive, since Italian terminologies and classification systems are not well integrated in UMLS. On the other end by integrating also the ICMV RDF graph, enriched with the UMLS CUI (found during the exact match process) and of the ICPC2 mappings found by physicians we could directly perform SPARQL queries on all the graphs in Virtuoso and using again the UMLS CUI as a bridge for extracting new mappings. Comparing the results of this last process with the other two explained above (using ICPC2 as a bridge to find mappings to other terminologies; and mapping ICMV directly to UMLS Italian terms) we found that the number of mappings to other terminologies which share the UMLS CUI increased to 1990 unique correspondences, and 3,223 if we include also synonyms to our ICMV. On the other hand the mapping to ICPC2 with respect to that found during the manual review by physicians decreased consistently (only 90 considering English and 64 considering only Italian).

Concerning the use of ICPC2 as a bridge to extend mappings between ICMV and other terminologies, one limit could be observed in the fact that ICPC2 is not completely integrated with UMLS. In fact, while we can find in UMLS a complete integration of the ICPC (first release), that means that all the ICPC2 rubrics have assigned a UMLS CUI (so also categories such as “Symptom and complaints”, “Procedures”, and “Diagnoses and diseases” and subcategories such as “Administrative procedures”, “Tests results”, etc.), the ICPC2 is inte-

grated only for the categories “Symptoms and complaints” and “Diseases”. In addition, the ICPC2 category “symptom” is classified under “Intellectual products” in UMLS and is therefore not comparable with symptoms in other terminologies. This explains the fact that we did not find for example common Parent relations between ICPC2 and the other terminologies for the category of symptoms.

A final point to take into consideration concerns the synonymy aspect. Most of the additional synonyms found by using UMLS CUIs and LUIs as terms of comparison are explained by the fact that multiple concepts from a given source (e.g. SNOMED CT, MeSH or ICD10) are collapsed in the same UMLS concept, despite the fact that they are not considered as synonyms in the original source. An example is the mapping we found for the UMLS concept “Malaria” which corresponds to 4 different SNOMED CT concepts: “Malaria”, “Disease due to Plasmodiidae”, “Malaria, unspecified”, and finally, “Malaria fever”. The second concept in SNOMED CT is not a synonym of Malaria but a parent of Malaria in SNOMED CT, while the third and the fourth concepts are classified under different hierarchies.

8.1.1 Comparison with Similar Approaches

Among the existing approaches reported in the literature on the problem of the linguistic gap between consumer-oriented and specialized medical terminologies, and on the problem of formalization and integration of medical classification systems, we would like to give a special emphasis to the framework of consumer-oriented medical vocabularies integrated with specialized medical terminologies and applied to PHRs or EHRs to improve readability of clinical records (Zeng *et al.*, 2006, Soergel *et al.*, 2004, Keselman *et al.*, 2007, and Rosembloom *et al.*, 2006 [123, 109, 69, 100]) as the closest works related to the study presented here.

Concerning the framework of consumer-oriented vocabularies integrated with

specialized medical terminologies, the approaches followed by Zeng *et al.*, 2006, Soergel *et al.*, 2004, Keselman *et al.*, 2007, and Rosembloom *et al.*, 2006 [123, 109, 69, 100], who developed consumer-oriented vocabularies for English, even they have the same goal of bridging the linguistic gap mentioned in the previous sections, differ from our approach because they are focused exclusively on the analysis of large written corpora (forum postings and queries to medical websites), and used only machine learning algorithms and statistical methods (naïve Bayesian classifiers, C-value, etc.) to extract consumer-oriented terminology. On the contrary, we gave more importance to qualitative than quantitative data, focusing on different methods for acquiring medical lay terminology and knowledge directly from consumers in the different scenarios related to General Practice. This allowed us not only to acquire data but also to try to understand how consumers make good or wrong use of medical terminology, how common expressions used every day in health communication really match up with medical concepts used by professionals. In our approach the constant review of physicians on one hand and the confirmation of consumers on the other allowed us to maintain a good balance between the two perspectives. Furthermore, concerning the process of mapping the acquired lay terms to the technical ones, opposed to other approaches which used only a UMLS as Professional Vocabulary, we provided both the creation of clinical mappings performed by physicians and the automatic mappings to ICPC2, and other UMLS vocabularies such as ICD10, SNOMED CT, MeSH, and LOINC, extending the possibility of finding synonyms for the ICMV terms. It is also important to highlight that our study is the first focused on the Italian context and the first showing the use of Semantic Web technologies for integrating consumer-oriented medical vocabularies with those used by professionals. In practical terms, our methodology showed encouraging results because it allowed us to acquire many consumer-oriented terms, a low overlap with ICPC2 medical concepts, and a high number of synonyms and related mappings to the

referent medical terminology and to UMLS.

8.1.2 Current and Potential Impacts

Current Impact

This thesis work can have an immediate territorial impact thanks to the application of the ICMV and its integration with the other terminologies in TreC (Cartella Clinica del Cittadino), an Italian PHR developed by the eHealth research unit of Fondazione Bruno Kessler, whose usability is being tested for more than 300 citizens in the Province of Trento. Our semantic resource was mostly integrated with this PHR, more generally in order to reorganize and categorize the PHR data model from the terminological and semantic point of view, and more specifically to provide each widget of the PHR (e.g. Allergies and Intolerances; Pathologies; Vaccinations; Hospitalizations; Reason for Encounters; Treatments and Medications; Surgery and Diagnostic Procedures; and Laboratory tests) with lists and taxonomies of medical concepts, expressed in lay terms to the users for improving PHR readability, and codified and mapped to Standard medical terminologies.

This work will be a significant added value to integrate the created lexical-ontological resource in TreC, both to improve the management and accessibility of this PHR, and in particular to apply the translation functionality to those sections of TreC which need an interpretative layer between specific medical terminology and that used by the consumer/user: personal-diary, symptoms and pain section, clinical notes, diagnosis and test results section, etc.

Another current impact is given by the possibility of sharing with the community the created resources. On one hand the publication of the ICMV by means of a Media Wiki System (ICMV Wiki) is used for browsing, searching and for the collaborative acquisition of lay medical terminology and collaborative association of new mappings with the terms in Wiki. On the other hand the Semantic Web-based resources we created, in particular the OWL encoding of

ICPC2, ICD10 and their Mapping, has had a positive impact on the community; in fact, our ontologies and mappings have been used by others as baseline for more specific formal representation of these two medical coding systems or for merging and integration with other medical ontologies. In particular, downloads of our resources in three years have been 211 for ICD10, 227 for ICPC2 and 94 for the formal mapping between ICPC2 and ICD10, by various research institute around the world, located principally in US and European Countries. In the case of ICD10, our ontology has been reused to provide a more axiomatized ICD10 ontology treating also the exclusion criteria without starting from scratch (i.e. the work of (Moeller *et al.*, 2010 [81]), which provided the OWL encoding of the ICD10 Hierarchy as well as comprehensive class labels for English and German, including also the modelling of specialties such as “Exclusion” statements, starting from our ICD10 ontology). Concerning ICPC2 ontology and its formal mapping to ICD10 organizations such as WONCA and WICC (developer and promoters of the ICPC2 around the world) are taking into account our formalization in order to improve structuring and semantics of the next release, the ICPC3.

Potential Applications

The ICMV and its integration framework could enable consumer health information systems to link medical information from different sources, such as EHRs, bibliographic databases, and decision support systems, among other applications. This would help consumers and laypersons in different scenarios: 1) searching for healthcare information (e.g., it could facilitate automated mapping of consumer-entered queries to technical terms - searching a bibliographic database indexed with MeSH would produce better search results if the query term used is mapped to MeSH); 2) translating and interpreting clinical notes or test results, which frequently contain jargon (e.g., mappings between a medical vocabulary, such as LOINC or SNOMED CT used in EHRs to a consumer-oriented vocabulary could be useful in providing consumer-understandable names

to help patients interpret these documents); 3) describing their clinical history and their complaints (e.g., in online medical consultations, patients entering consumer expressions, such as “sudden hair loss”, could receive appropriate help from health professionals after the translation of their query into the corresponding technical concept). Additionally, lots of potential applications in the patient empowered health care (e.g. Home care, etc) can be possible.

8.1.3 Future Work

There are several directions for future research stemming from the work presented in the current thesis:

- **Multilingualism.** An added value for the use of ICMV would be its integration with other Consumer Health Vocabularies created for other European Languages, in particular for English, for French, Spanish, and Belgian. It would be very interesting work to map our ICMV to other lay terminologies such as the Consumer Health Vocabulary OAC CHV ¹, and in particular the Multilingual European Medical Glossary for Popular Terms [40]. It is not just a matter of translating from one language to another but, since it involves consumer’s medical expressions and since they depend on many factors, such as age, healthcare literacy, the community they belong to, etc., the consideration of “contexts” in a multilingual perspective is very important. Having a Multilingual ICMV would improve its helpfulness not only if used for browsing and searching, but above all in the actual perspective of integration in TreC PHR and any other PHR or EHR, since it allows, in case of emergency (for example if the user of the PHR is abroad, in one of the European Countries, and needs health care), to translate the content of the PHR, in particular diseases, symptoms, allergies, diagnostic procedures, and all the data which compose the Patient Summary. This is

¹<http://www.consumerhealthvocab.org/>

perfectly in line with recent European Community regulations and efforts to create a unified or at least interoperable access to patient data and to unified process of care.

- Extension of the mappings to other terminologies/ontologies used in Italy in primary care and in other healthcare contexts, such as ICD9-CM, by using automatic mapping extraction (first of all syntactically and then semantically-based).
- Use a more adequate SW format for the ICMV, such as SKOS, to provide the right formal expressivity for a controlled vocabulary which can be used also on the web to access medical information. This is true because of its more appropriate semantics, which includes the definition of hierarchical relations (BT, NT), or associative relations (RT), synonyms, etc., typical for a resource like our ICMV.
- Other possible extensions concern the improvement of ICMV integration in TreC. In particular, it would be interesting to provide automatic annotation of medical reports received in TreC, using as a knowledge base the ICMV and the other integrated ontologies to improve readability. It would be desirable to provide a structured extract for each medical report which explains the content in consumer-friendly terms, using ICMV labels. Furthermore, we can improve the presentation of ICMV terms in the various TreC widgets by allowing users to choose among lay, intermediate and specialized terminologies, according to factors such as age, gender, education, country and occupation, in order to differentiate the user interface depending on the user. In this way we could avoid the use of a lay terminology and categorization of the PHR for those consumers who are confident with the medical/healthcare domain, or on the other hand we could help those lay persons who do not have a good level of healthcare literacy to understand and act as much as possible on their PHR content.

Bibliography

- [1] ISO/TC37. Global group: Historical background. Technical report, International Information Centre for Terminology (Infoterm), Vienna, Austria, 1952.
- [2] ISO 704. Principles and Methods of Terminology. Technical report, International Standards Organization, Geneva, Switzerland, 1987.
- [3] ISO/TC215. Health Informatics - Health Concept Representation. Technical report, International Standards Organization, Geneva, Switzerland, 2000.
- [4] International Statistical Classification of Diseases and related health problems. Technical report, World Health Organization, 2004.
- [5] SNOMED Clinical Terms®(SNOMED CT®) International Release - July 2009. Technical report, The International Health Terminology Standards Development Organisation - IHTSDO, Copenhagen Denmark, 2009.
- [6] S. S. R. Abidi. Healthcare Knowledge Management: The Art of the Possible. In *Proceedings of the Knowledge Management for Health Care Conference K4CARE 2007*, pages 1–20. Springer - Berlin, 2007.
- [7] G. Adamo. Terminología vs. lexicología. *Hieronymus Complutensis*, (8):75–86, 1999.

- [8] G. Antoniou and F. van Harmelen. *Handbook on Ontologies in Information Systems*, chapter Web Ontology Language: OWL, pages 67–92. Springer Verlag, 2004.
- [9] M. Van Assem, V. Malais, A. Miles, and G. Schreiber. A method to convert thesauri to skos. In *In volume 4011 of Lecture Notes in Computer Science*, pages 95–109. Springer, 2006.
- [10] M. Van Assem, M. R. Menken, G. Schreiber, J. Wielemaker, and B. Wielinga. A method for converting thesauri to rdf/owl. In *Proc. of the 3rd Intl Semantic Web Conf. (ISWC04), number 3298 in Lecture Notes in Computer Science*, pages 17–31. Springer-Verlag, 2004.
- [11] M. Baldini. Parlare al paziente, parlare “col” paziente. In *Proceedings of the Conference L’arte medica: Tra comunicazione, relazione, tecnica e organizzazione*, pages 9–25. Scriptorium, Torino, 1996.
- [12] D. M. Baorto and J. J. Cimino. An “infobutton” for enabling patients to interpret on-line pap smear reports. In *Proceedings of AMIA Symposium - AMIA2000*, pages 47–50, 2000.
- [13] R. Bartolini, A. Lenci, S. Marchi, S. Montemagni, and V. Pirrelli. Text-2-knowledge: Acquisizione semi-automatica di ontologie per l’indicizzazione semantica di documenti. Technical report, ILC - CNR, Pisa, 2005. Technical Report for the PEKITA Project.
- [14] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein. Owl web ontology language reference. W3C Recommendation, February 2004. <http://www.w3.org/TR/owl-ref>.
- [15] G. De Benedittis, R. Massei, R. Nobili, and A. Pieri. The Italian Pain Questionnaire. *Pain*, 33(1):53–62, April 1988.

- [16] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web: A New Form of Web Content that is Meaningful to Computers Will Unleash a Revolution of New Possibilities. *The Scientific American*, 284(5):34–43, 2001.
- [17] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.
- [18] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270, 2004.
- [19] O. Bodenreider. Comparing SNOMED CT and the NCI Thesaurus through Semantic Web Technologies. In *In proceedings of the 3rd International Conference on Knowledge Representation in Medicine (KR-MED2008)*. R. Cornet, K.A. Spackman (Eds), 2008.
- [20] O. Bodenreider. Issues in mapping loinc laboratory tests to snomed ct. In *In proceedings of AMIA Annual Symposium, AMIA2008*, pages 51–55, 2008.
- [21] L. Bowker. Terminology and gender sensitivity. *Language in Society*, 30(4):589–610, 2001.
- [22] M. T. Cabré. *La terminología. Teoría, metodología, aplicaciones*. Editorial Antártida/Empúries, Barcelona, 1993.
- [23] M. T. Cabré. Terminology: Theory, methods, and applications. *Terminology*, 1, 1998/1999.
- [24] M. T. Cabré. *Terminology: Theory, methods, and applications*. John Benjamins Publishing Company, 1999.

- [25] E. Cardillo. A Lexical-Ontological Resource for Consumer Healthcare. In *Proceedings of the International Semantic Web Conference'09 - ISWC2009*, pages 949–956. Springer - Berlin, 2009.
- [26] E. Cardillo, C. Eccher, A. Tamin, and L. Serafini. Logical analysis of mappings between medical classification systems. In *Proceedings of AIMSA2008*, pages 311–321. Springer - Berlin, 2008.
- [27] E. Cardillo, G. Hernandez, and O. Bodenreider. Integrating consumer-oriented vocabularies with selected professional ones from the UMLS using Semantic Web Technologies. In *Proceedings of the 3rd International Conference on Electronic Healthcare, eHealth2010*, Casablanca - Morocco, December 2010.
- [28] E. Cardillo, L. Serafini, and A. Tamin. A Hybrid Methodology for Consumer-Oriented Healthcare Knowledge Acquisition. In David Riaño, Annette ten Teije, Silvia Miksch, and Mor Peleg, editors, *Knowledge Representation for Health-Care. Data, Processes and Guidelines*, volume 5943 of *Lecture Notes in Computer Science*, pages 38–49. Springer Berlin - Heidelberg, 2010.
- [29] E. Cardillo, A. Tamin, and L. Serafini. A Hybrid Methodology for Consumer-oriented Healthcare Knowledge Acquisition. In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development KEOD'09*, pages 64–71, 2009.
- [30] E. Cardillo, A. Tamin, and L. Serafini. A methodology for knowledge acquisition in consumer-oriented healthcare. In Ana Fred, Jan L. G. Dietz, Kecheng Liu, and Joaquim Filipe, editors, *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, volume 128 of *Communications in Computer and Information Science*, pages 249–261. Springer Berlin Heidelberg, 2011.

- [31] W. Ceusters, B. Smith, and J. Flanagan. Ontology and Mapping Terminology: Why Description Logics Are Not Enough. In *Proceedings of TEPR2003*, 2003.
- [32] W. Ceusters, B. Smith, and G. De Moor. Ontology-Based Integration of Medical Coding Systems and Electronic Patient Records. In *Proceedings of MIE2005*, 2005.
- [33] K-H. Cheung, V. Kashyap, J. Luciano, H. Chen, Y. Wang, and S. Stephens. Semantic mashup of biomedical data. *Journal of Biomedical Informatics*, 41(5):683–686, 2008.
- [34] C. G. Chute. Clinical classification and terminology: some history and current observations. *J Am Med Inform Assoc*, 7(3):298–303, 2000.
- [35] P. A. De Clercq, A. Hasman, and B. H. Wolffenbuttel. A consumer health record for supporting the patient-centered management of chronic diseases. *Medical Informatics and the Internet in Medicine*, 28(2):117–127, June 2003.
- [36] D. Crystal. *A first dictionary of linguistics and phonetics*. Westview Press, Boulder, CO, 1980.
- [37] D. Crystal. *The Cambridge encyclopedia of language*. Cambridge University Press, Cambridge, UK, 1987.
- [38] N. F. de Keizer and A. Abu-Hanna. Understanding terminological system ii: Experience with conceptual and formal representation of structure. *Methods Inf Med*, 39(1):22–29, Mar 2000.
- [39] R. H. Dolin, S. M. Huff, R. A. Rocha, K. A. Spackman, and K. E. Campbell. Evaluation of a lexically assign, logically refine strategy for semi-automated integration of overlapping terminologies. *J Am Med Inform Assoc*, 5(2):203–213, 1998.

- [40] Heymans Institute of Pharmacology EEC. Multilingual glossary of technical and popular medical terms in nine european languages: Final report (etd/93/a02600/mi/10). Technical report, University of Ghent, and Mercator College, Department of Applied Linguistics, Ghent, Belgium, 1995.
- [41] P. L. Elkin and S. H. Brown. Automated enhancement of description logic-defined terminologies to facilitate mapping to icd9-cm. *Journal of Biomedical Informatics*, 35:5–6, 2002.
- [42] J. Euzenat, F. Scharffe, and L. Serafini. Specification of the delivery alignment format. Deliverable 2.2.6, KnowledgeWeb, 2006.
- [43] J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer, Berlin / Heidelberg, 2007.
- [44] M. M. Foley. Key issues shaping clinical terminology and classification. *Journal of AHIMA*, 77(7), July/August 2006. extended online edition.
- [45] K. W. Fung and O. Bodenreider. Utilizing the umls for semantic mapping between terminologies. In *Proceedings of AMIA Annual Symposium 2005*, pages 266–270, 2005.
- [46] A. Gangemi, D. M. Pisanelli, and G. Steve. *Ontology Integration: Experiences with Medical Terminologies*, pages 163–178. IOS Press, 1998.
- [47] C. Ghidini, M. Rospocher, and L. Serafini. MoKi: a Wiki-Based Conceptual Modeling Tool. In *ISWC 2010 Posters & Demonstrations Track: Collected Abstracts*, volume 658 of *CEUR Workshop Proceedings (CEUR-WS.org)*, pages 77–80, Shanghai, China, 2010.
- [48] A. Gomez-Perez and O. Corcho. Ontology languages for the semantic web. *IEEE Intelligent Systems*, 17(1):54–60, 2002.

- [49] B. Cuenca Grau, B. Parsia, and E. Sirin. Pellet: An owl dl reasoner. In *Proceedings of the 3rd International Semantic Web Conference (ISWC-2004)*, 2004.
- [50] T. R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199–220, 1993.
- [51] N. Guarino. Formal ontology and information systems. In *Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS-1998)*, Trento, Italy, 1998.
- [52] V. Haarslev and R. Moller. Racer system description. In *Proceedings of the International Joint Conference on Automated Reasoning (IJCAR-2001)*, pages 701–706, 2001.
- [53] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space (1st edition)*. *Synthesis Lectures on the Semantic Web: Theory and Technology*. Morgan & Claypool, 2011.
- [54] G. Heja, G. Surján, G. Lukácsy, P. Pallinger, and M. Gergely. Galen based formal representation of icd10. *International Journal of Medical Informatics*, 76(2-3):118–123, 2007.
- [55] G. Jr. Hernandez, R. Ghazzaoui, and O. Bodenreider. Towards a representation of mesh in rdf. 2009.
- [56] P. Hitzler, R. Sebastian, and M. Krötzsch. *Foundations of Semantic Web Technologies*. Chapman & Hall/CRC, London, 2009.
- [57] S. Hoelzer, R. K. Schweiger, R. Liu, D. Rudolf, J. Rieger, and J. Dudeck. XML representation of Hierarchical Classification Systems: from Conceptual Models to Real Applications. In *Proceedings of AMIA Symposium*, pages 330–334, 2002.
- [58] I. Horrocks. *Optimising Tableaux Decision Procedures for Description Logics*. PhD thesis, University of Manchester, 1997.

- [59] I. Horrocks. Description logics in ontology applications. In *Proceedings of the Ninth International Conference on Automated Reasoning with Analytic Tableaux and Related Methods (TABLEAUX-2005)*, pages 2–13, 2005.
- [60] I. Horrocks and P. F. Patel-Schneider. Fact and dlp. In *Proceedings of the Automated Reasoning with Analytic Tableaux and Related Methods (TABLEAUX-1998)*, pages 27–30, 1998.
- [61] I. Horrocks, P. F. Patel-Schneider, and F. van Harmelen. From shiq and rdf to owl: The making of a web ontology language. *Journal of Web Semantics*, 1(1):7–26, 2003.
- [62] I. Horrocks and U. Sattler. A tableaux decision procedure for SHOIQ. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-2005)*, pages 448–453, 2005.
- [63] I. Horrocks, U. Sattler, and S. Tobies. Practical reasoning for very expressive description logics. *Logic Journal of IGPL*, 8(3):239–263, 2000.
- [64] I. Horrocks, U. Sattler, and S. Tobies. Reasoning with individuals for the description logic SHIQ. In *Proceedings of the 17th International Conference on Automated Deduction (CADE-2000)*, pages 482–496, 2000.
- [65] C. Iandolo. *Parlare col Malato. Tecnica, Arte ed errori della comunicazione*. Armando, Roma, 1983.
- [66] A. Kalyanpur, B. Parsia, E. Sirin, and J. Hendler. Debugging unsatisfiable classes in owl ontologies. *Journal of Web Semantics*, 3(4):268–293, 2005.
- [67] A. Keselman, J. Crowell, A. C. Browne, L. Ngo, and Q. Zeng. Assessing Consumer Health Vocabulary Familiarity: An Exploratory Study. *Journal of Medical Internet Research*, 9(1):e5, 2007.
- [68] A. Keselman, R. Logan, C. A. Smith, G. Leroy, and Q. Zeng. Developing Informatics Tools and Strategies for Consumer-centered Health Com-

- munication. *Journal of the American Medical Informatics Association*, 14(4):473–483, 2008.
- [69] A. Keselman, L. Slaughter, C. Arnott-Smith, H. Kim, G. Divita, A. Browne, and Q. Zeng-Treitler. Towards Consumer-Friendly PHRs: Patients? Experience with Reviewing Their Health Records. Unpublished manuscript, 2007.
- [70] H. Kim, Q. Zeng, S. Goryachev, A. Keselman, L. Slaughter, and C. A. Smith. Text Characteristics of Clinical Reports and Their Implications for the Readability of Personal Health Records. In *Proceedings of MED-INFO2007*, pages 1117–1121. IOS Press, 2007.
- [71] B. Kuipers. Multiple ontologies for spatial exploration and mapping. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness, ISA '10*, pages 24–24, New York, USA, 2010. ACM.
- [72] Y. Lee and J. Geller. Semantic enrichment for medical ontologies. *Journal of Biomedical Informatics*, 39:206–226, 2006.
- [73] J. Lomax and A. T. McCray. Mapping the gene ontology into the unified medical language system. *Comparative and functional genomics*, 5(4):354–361, 2004.
- [74] A. Lucchini. *Il linguaggio della Salute: come migliorare la comunicazione con il paziente*. Sperling & Kupfer, Milano, 2008.
- [75] A. Maedche, B. Motik, L. Stojanovic, R. Studer, and R. Volz. Managing multiple ontologies and ontology evolution in ontologging. In *Proceedings of Ontologging. Intelligent Information Processing*, pages 51–63. Kluwer, 2002.

- [76] P. D. Marshall. Bridging the Terminology Gap Between Health Care Professionals and Patients with the Consumer Health Terminology (CHT). In *Proceedings of AMIA2000*, page 1082, 2000.
- [77] C. McDonald, S. Huff, K. Mercer, J.A. Hernandez, and D.J. Vreeman. Logical Observation Identifiers Names and Codes (LOINC), Users' Guide. Technical report, Regenstrief Institute Inc, Indianapolis, Indiana (USA), 2010.
- [78] C. J. McDonald, S. M. Huff, J. G. Suico, J. G. Hill, D. Leavelle, R. Aller, A. Forrey, K. Mercer, G. DeMoor, J. Hook, J. Case, and P. Maloney. Loinc, a universal standard for identifying laboratory observations: A 5-year update. *Clinical Chemistry*, 49:624–633, 2003.
- [79] I. Millard, A. Jaffry, H. Glaser, and B. Rodriguez-Castro. Using a Semantic MediaWiki to Interact with a Knowledge Based Infrastructure. In *Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management*, Podebrady, Czech Republic, October 2006.
- [80] N. R. Milton. *Knowledge Acquisition in Practice: A Step-by-step Guide*. Springer, 2007.
- [81] M. Möller, P. Ernst, M. Sintek, R. Biedert, A. Dengel, and D. Sonntag. Representing the international classification of diseases version 10 in owl. In *Proc. of the International Conference on Knowledge Engineering and Ontology Development (KEOD)*, Valencia, Spain, 25-28 October 2010.
- [82] S. Montemagni. Architecture and Functioning of a System for the Acquisition of Taxonomical Information from Dictionary Definitions. In *Proceedings of the 4th Conference on Computational Lexicography and text - COMPLEX 1996*, Budapest, 1996.

- [83] S. J. Nelson, T. Powell, and B. L. Humphreys. The Unified Medical Language System (UMLS) project. Technical report, NLM, NIH, DHHS, Bethesda, Maryland, 2001.
- [84] G. Nencioni. Verso una nuova lessicografia. In A. Cappelli L. Cignoni C. Peters, editor, *Studies in Honour of Roberto Busa S.J.*, volume IV-V of *Linguistica Computazionale*, pages 133–150. Pisa - Giardini, 1987.
- [85] N. F. Noy. *Ontology Mapping*, pages 573–590. 2009.
- [86] N. F. Noy, S. de Coronado, H. Solbrig, G. Fragoso, F. W. Hartel, and M. A. Musen. *Representing the NCI Thesaurus in OWL DL: Modeling tools help modeling languages*, volume 1, pages 19–23. IOS Press, 2007.
- [87] N. F. Noy, M. Musen, N. Shah, B. Dai, M. Dorf, N. Griffith, C. Jonquet, M. Montegut, D. Rubin, and C. Youn. Bioportal: A web repository for biomedical ontologies and data resources. In *Proceedings of ISWC2008*, Karlsruhe, Germany, 2008.
- [88] N. F. Noy, M. Sintek, S. Decker, M. Crubezy, R. W. Fergerson, and M. A. Musen. Creating semantic web contents with protégé-2000. *IEEE Intelligent Systems*, 16(2):60–71, 2001.
- [89] I. M. Okkes, M. Jamoullea, H. Lamberts, and N. Bentzen. Icp-2-e: the electronic version of icpc-2. differences from the printed version and the consequences. *Family Practice*, 17:101–107, 2000.
- [90] B. Omelayenko. Integrating vocabularies: Discovering and representing vocabulary maps. In *Proceedings of the Workshop on Knowledge Transformation for the Semantic Web at the 15th European Conference on Artificial Intelligence (KTSW-2002)*, 2002.
- [91] P. F. Patel-Schneider, P. Hayes, and I. Horrocks. Web ontology language owl abstract syntax and semantics. W3C Recommendation, February 2004.

- [92] E. Pianta and S. Tonelli. KX: A Flexible System for Keyphrase eXtraction. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 170–173, Uppsala, Sweden, July 2010. Association for Computational Linguistics.
- [93] H. Picht and J. Draskau. *Terminology: An introduction*. University of Surrey Press, Surrey, UK, 1985.
- [94] P. Plessers, O. De Troyer, and S. Casteleyn. Understanding ontology evolution: A change detection approach. *Web Semant.*, 5:39–49, March 2007.
- [95] B. Purini and E. M. Piras. Report cawi. Technical report, Fondazione Bruno Kessler, 2011.
- [96] A. Rector. Clinical terminology: Why is it so hard? *Methods of Information in Medicine*, 38(4):239–252, 1999.
- [97] A. Rector and J. Rogers. Patterns, Properties and Minimizing Commitment: Reconstruction of the GALEN Upper Ontology, in OWL. In *Proceedings of Core Ontologies Workshop (CORONT 2004)*, 2004.
- [98] A. L. Rector and S. Brandt. Why do it the hard way? the case for an expressive description logic for snomed. *Journal of American Medical Informatics Association*, 15(6):744–751, 2008.
- [99] G. Roseblat, R. Logan, T. Tse, and L. Graham. How Do Text Features Affect Readability? Expert Evaluations on Consumer Health Web Site Text. In *Proceedings of MEDNET 2006*, Toronto, CA, 2006.
- [100] T. S. Rosebloom, R. A. Miller, K. B. Johnson, P. L. Elkin, and H. S. Brown. Interface Terminologies: Facilitating Direct Entry of Clinical Data into Electronic Health Record Systems. *Journal of American Medical Informatics Association*, 13(3):277–287, 2006.

- [101] C. Rosse and J. L. V. Mejino. *The Foundational Model of Anatomy Ontology*, volume 6, pages 59–117. Springer - London, 2007.
- [102] D. L. Rubin, N. H. Shah, and N. F. Noy. Biomedical ontologies: A functional perspective. *Briefings in Bioinformatics*, 9(1):75–90, 2007.
- [103] A. Ruttenberg, T. Clark, W. Bug, M. Samwald, O. Bodenreider, H. Chen, D. Doherty, K. Forsberg, Y. Gao, V. Kashyap, and et al. Advancing translational research with the Semantic Web. *BMC Bioinformatics*, 8(Suppl 3):S2, 2007.
- [104] K. Simov, Z. Peev, M. Kouylekov, A. Simov, M. Dimitrov, and A. Kiryakov. CLaRK - an XML-based System for Corpora Development. In *Proceedings of the Corpus Linguistics 2001 Conference*, pages 558–560, 2001.
- [105] E. Sirin, B. Parsia, B. Cuenca Grau, A. Kalyanpur, and Y. Katz. Pellet: A practical owl-dl reasoner. *Journal of Web Semantics*, 2006.
- [106] B. Smith and C. Rosse. The role of foundational relations in the alignment of biomedical ontologies. In *Proceedings of AMIA2004*, 2004.
- [107] M. K. Smith, C. Welty, and D. L. McGuinness. Owl web ontology language guide. W3C Recommendation, February 2004.
- [108] D. Soergel. *Indexing languages and thesauri: Construction and maintenance*. Los Angeles: Melville, 1974.
- [109] D. Soergel, T. Tse, and L. Slaughter. Helping Healthcare Consumers Understand: An “Interpretative Layer” for Finding and Making Sense of Medical Information. In *Proceedings of IMIA2004*, pages 931–935, 2004.
- [110] Lf. Soualmia, C. Golbreich, and Sj. Darmoni. Representing the mesh in owl: Towards a semi-automatic migration. In *Proceedings of the KR*

- 2004 Workshop on Formal Biomedical Knowledge Representation, pages 81–87, 2004.
- [111] S. Staab and R. Studer, editors. *Handbook on Ontologies*. International Handbooks on Information Systems. Springer, 2004.
- [112] V. Sujansky and Associates. *Project Health Design Common Platform Components Functional Requirements*, 2007.
- [113] P. C. Tang, J. S. Ash, D. W. Bates, J. M. Overhage, and D. Z. Sands. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. *J Am Med Inform Assoc*, 13(2):121–126, 2006.
- [114] S. Tessaris. *Questions and Answers: Reasoning and Querying in Description Logic*. PhD thesis, Department of Computer Science, University of Manchester, UK, 2001.
- [115] D. Travers. Identifying umls concepts in emergency department terms using domain knowledge and natural language processing techniques. Technical report, Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, DHHS., Bethesda, MD, 2001.
- [116] D. Travers, J. Dara, J. N. Dowling, G. F. Cooper, and W. W. Chapman. Evaluation of preprocessing techniques for chief complaint classification. *J Biomed Inform*, 41(4):613–623, 2008.
- [117] D. Travers, S. W. Haas, J.E. Tintinalli, D. Pollock, A. Waller, E. Barthell, C. Burt, W. Chapman, W. Coonan, D. Kamens, and J. McClay. Toward vocabulary control for chief complaint. *ACAD EMERG MED*, 15(5):475–482, 2008.
- [118] T. Tudorache, S. M. Falconer, C. Nyulas, N. F. Noy, and M. A. Musen. Will semantic web technologies work for the development of icd-11? In

- Proceedings of the International Semantic Web Conference (ISWC2010)*, pages 257–272, 2010.
- [119] T. Tudorache, S. M. Falconer, C. Nyulas, N. F. Noy, T. Bedirhan Üstün, M-A. D. Storey, and M. A. Musen. Ontology development for the masses: Creating icd-11 in webprotégé. In *Proceedings of EKAW2010*, pages 74–89, 2010.
- [120] Y. Wang, J. Patrick, G. Miller, and J. O’Halloran. Linguistic mapping of terminologies to snomed ct. In *Proceedings of SMCS 2006*, 2006.
- [121] M. Wood, H. Lamberts, J. S. Meijer, and I. M. Hofmans-Okkes. The conversion between icpc and icd-10. requirements for a family of classification systems in the next decade. *Family Practice*, 9:340–348, 1992.
- [122] Q. Zeng, S. Goryachev, A. Keselman, and D. Rosendale. Making Text in Electronic Health Records Comprehensible to Consumers: A Prototype Translator. In *Proceedings of AMIA2007*, pages 846–850, 2007.
- [123] Q. Zeng and T. Tse. Exploring and Developing Consumer Health Vocabularies. *Journal of the American Medical Informatics Association*, 13:24–29, 2006.
- [124] Q. Zeng, T. Tse, G. Divita, A. Keselman, J. Crowell, A.C. Browne, S. Goryachev, and L. Ngo. Term Identification Methods for Consumer Health Vocabulary Development. *Journal of Medical Internet Research*, 9(1):e4, 2007.

