



Clinical Practice Knowledge Acquisition and Interrogation using Natural Language

Aquisição e Interrogação de Conhecimento de Prática Clínica usando Linguagem Natural

David José Murteira Mendes

Tese apresentada à Universidade de Évora
para obtenção do Grau de Doutor em Informática

ORIENTADORA: *Irene Pimenta Rodrigues*

Évora, Março de 2014



INSTITUTO DE INVESTIGAÇÃO E FORMAÇÃO AVANÇADA



Clinical Practice Knowledge Acquisition and Interrogation using Natural Language

Aquisição e Interrogação de Conhecimento de Prática Clínica usando Linguagem Natural

David José Murteira Mendes

Tese apresentada à Universidade de Évora
para obtenção do Grau de Doutor em Informática

ORIENTADORA: *Irene Pimenta Rodrigues*

Évora, Março de 2014



INSTITUTO DE INVESTIGAÇÃO E FORMAÇÃO AVANÇADA

Aquisição e Interrogação de Conhecimento de Prática Clínica usando Linguagem Natural

Sumário

Os conceitos científicos, metodologias e ferramentas no sub-domínio da ***Representação de Conhecimento*** da área da ***Inteligência Artificial Aplicada*** têm sofrido avanços muito significativos nos anos recentes. A utilização de ***Ontologias*** como conceptualizações de domínios é agora suficientemente poderosa para aspirar ao raciocínio computacional sobre realidades complexas.

Uma das tarefas científica e tecnicamente mais desafiante é ***prestação de cuidados*** pelos profissionais de saúde na ***especialidade cardiovascular***.

Um domínio de tal forma complexo pode beneficiar largamente da possibilidade de ***ajudas ao raciocínio clínico*** que estão neste momento à beira de ficarem disponíveis.

Investigamos no sentido de desenvolver uma infraestrutura sólida e completa para a representação de conhecimento na ***prática clínica*** bem como os processos associados para adquirir o conhecimento a partir de textos clínicos e raciocinar automaticamente sobre esse conhecimento.

Clinical Practice Knowledge Acquisition and Interrogation using Natural Language

Abstract

The scientific concepts, methodologies and tools in the **Knowledge Representation (KR)** sub-domain of applied **Artificial Intelligence (AI)** came a long way with enormous strides in recent years. The usage of domain conceptualizations that are *Ontologies* is now powerful enough to aim at computable reasoning over complex realities.

One of the most challenging scientific and technical human endeavors is the daily *Clinical Practice (CP)* of *Cardiovascular (CV)* specialty healthcare providers.

Such a complex domain can benefit largely from the possibility of clinical reasoning aids that are now at the edge of being available.

We research into a complete end-to-end solid ontological infrastructure for **CP** knowledge representation as well as the associated processes to automatically acquire knowledge from clinical texts and reason over it.

Dedico esta obra à primeira e segunda mulheres da minha vida, a minha mãe e a minha irmã

Agradecimentos

I wish to extend my acknowledgments in first place to my tutor, Prof.^a Dr.^a **Irene Pimenta Rodrigues** for all her patience, support and high valued technical and personal tutoring along these last 3 years.

I wish to express my gratitude to my family namely **Alexandre Mendes** and **Rodrigo Mendes** for their coping with my absences and lack of support that they unfortunately surely felt.

My acknowledgments are extended to my wider family for their long standing support namely **Luisa** and **César**.

Last but not the least a special personal note of gratitude goes to **Rosa Silvestre**, the love of my life that was always there, supporting my recurrent ups and downs for all these years and specifically during those when my research in this project went along.

Of course, I acknowledge

- the financial and logistic support of Centro de Investigação em Inteligência Artificial (**CENTRIA**) at the early stages of my research,
- the Departamento de Informática (**DI**) of Universidade de Évora (**UE**) in the person of Prof. Dr. *Salvador Pinto de Abreu* the PhD DI director,
- and Prof.^a Dr.^a **Teresa Gonçalves** the **DI** director for welcoming me in the department,
- Dr. Carlos Baeta for its invaluable support, motivation and devoted hard work,
- Unidade Local de Saúde do Norte Alentejano (**ULSNA**) for the cooperation.
- But most of all, I could only develop my activities thanks to the **Bento de Jesus Caraça** scholarship gently granted by Instituto de Investigação e Formação Avançada (**IIFA**).

To all of them I wish to express my most profound feelings of acknowledgment and gratitude.

Acronyms

ACE	Attempto Controlled English	22
AI	Artificial Intelligence	iii
AOL	Automatic Ontology Learning	28
Apache	Apache Software Foundation	xiv
APE	Attempto Parsing Engine	22
API	Application Programming Interface	7
BO	Biomedical Ontologies	53
BFO	Basic Formal Ontology	54
BioTOP	Top-Domain Ontology for the Life Sciences	39
CAT	Computer Aided Translation	66
CC	Clinical Concepts	63
CCL	Clinical Controlled Language	3
CDSS	Clinical Decision Support Systems	3
CENTRIA	Centro de Investigação em Inteligência Artificial	vii
CERN	Conseil Européen pour la Recherche Nucléaire	7
CH	Clinical History	3
CIDERS	Clinical Integrated Discourse Extended Representation Structure	3
CK	Clinical Knowledge	xiv
CLI	Command Line Interface	73
CNL	Controlled Natural Language	22
CORE	Clinical Observations Recording and Encoding	xiv
CP	Clinical Practice	iii
CP-ESB	Clinical Practice - Enterprise Service Bus	97
CPR	Computer Based Patient Record Ontology	48

CRR	Co-Reference Resolution.....	3
CSI	Computer Semantic Interoperability	xiv
cTAKES	clinical Text Analysis and Knowledge Extraction System.....	xiv
CV	Cardiovascular.....	iii
CVDO	Cardiovascular Disease Ontology	49
CUI	Concept Unique Identifiers	51
CWA	Closed World Assumption.....	80
DBE	Discourse Based Enhancement.....	43
DC	Discourse Controller.....	59
DE	Domain Expert.....	4
DI	Departamento de Informática	vii
DL	Description Logics.....	xvi
DO	Disease Ontology.....	48
DR	Discourse Reasoning.....	44
DRS	Discourse Representation Structure	3
DRT	Discourse Representation Theory	3
EAV	Extraction of Attributes and Values	72
EHR	Electronic Health Record.....	41
EMR	Electronic Medical Record.....	40
ESB	Enterprise Service Bus	xviii
EU	European Union.....	20
FMA	Foundational Model of Anatomy	40
FO	Formal Ontology	17
FOL	First Order Logic	27
GATE	General Architecture for Text Engineering.....	30
GRDDL	Gleaning Resource Descriptions from Dialects of Languages.....	35
GF	Grammatical Framework	95
GO	Gene Ontology	18
HCLS IG	Health Care and Life Sciences Interest Group.....	7
HL7	Health Level 7	44
HTML	HyperText Markup Language	40
IE	Information Extraction.....	19
IIFA	Instituto de Investigação e Formação Avançada	vii
IR	Information Retrieval	19
ISO	International Standards Organization.....	47

Jena	Apache Jena	35
JSON	JavaScript Object Notation	77
KA	Knowledge Acquisition	4
KAB	Knowledge Acquisition Bottleneck.....	3
KB	Knowledge Base	xvii
KR	Knowledge Representation	iii
ML	Machine Learning	4
MS	Manchester Syntax	22
MT	Machine Translation	66
NCBO	National Center for Biomedical Ontologies	xvii
NCOR	National Center for Ontological Research.....	20
NER	Named Entity Recognition	72
NLG	Natural Language Generation	70
NLP	Natural Language Processing	xiv
NLU	Natural Language Understanding.....	70
OASIS	Organization for the Advancement of Structured Information Standards	30
OBO	Open Biological and Biomedical Ontologies	xiv
OD	Ontology Driven	65
ODA	Ontology Driven Expanded Semantic Annotation.....	59
OGCP	Ontology for General Clinical Practice.....	xiv
OGMS	Ontology for General Medical Science	48
OL	Ontology Learning	8
OR	Ontological Realism	xiv
OWL	Web Ontology Language.....	xiii
OWL2	Web Ontology Language v.2.....	7
OWL API	OWL Application Programming Interface	15
OWL DL	OWL Description Language	12
PACE	ACE Parser	80
PDF	Portable Document Format.....	29
PHI	Personal Health Information.....	20
PL	Predicate Logic	13
PLT	Problem List Terms	51
POMR	Problem Oriented Medical Record	40
POS	Part Of Speech.....	72
Protégé	Protégé 4 Ontology Development Tool	xvii

QA	Question Answering	2
RDF	Resource Description Framework	7
RDFS	Resource Description Framework Schema	35
REST	Representational State Transfer	
RIM	Reference Information Model	44
RO	Relations Ontology	39
RTU	Referent Tracking Unit	20
SAM	Sistema de Apoio ao Médico	42
SNOMED-CT	Standard Nomenclature of Medicine - Clinical Terms	xiv
SO	Symptom Ontology	60
SOAP	Subjective, Objective, Assessment, Plan	xvii
SPARQL	SPARQL Protocol and RDF Query Language	7
SW	Semantic Web	9
SWRL	Semantic Web Rules Language	34
TE	Textual Entailment	89
TM	Translation Memory	xvii
TMM	Translation Memory Manager	65
TMX	Translation Memory eXchange	66
UB	New York State University at Buffalo	20
UE	Universidade de Évora	vii
UIMA	Unstructured Information Management Architecture	xiv
ULSNA	Unidade Local de Saúde do Norte Alentejano	vii
UMLS	Unified Medical Language System	xiv
URI	Uniform Resource Identifier	35
URL	Uniform Resource Locator	20
UTS	UMLS Terminology Services	50
VSO	Vital Signs Ontology	60
W3C	World Wide Web Consortium	7
WS	Web Service	7
WSD	Word Sense Disambiguation	72
WWW	World Wide Web	7
XML	eXtensible Markup Language	21
XSLT	XSL Transformations	66

Contents

Sumário	i
Abstract	iii
Table of Contents	xvi
List of Figures	xviii
List of Tables	xix
1 Introduction	1
1.1 Research context and motivation	1
1.2 Research questions	2
1.3 Scientific Innovation	3
2 Tools and Technologies Addressed	7
2.1 Semantic Web	7
2.2 Resource Description Framework	8
2.3 Description Logics and Web Ontology Language v.2	8
2.3.1 Description Logics	9
2.3.2 Web Ontology Language	10
2.4 Reasoning about Knowledge	11
2.5 Reasoning Support for OWL	13
2.6 Web Ontology Language (OWL) Reasoners	14
2.7 Consequence driven reasoning	15
2.7.1 The ELK reasoner	15
2.8 Biomedical Resources	17
2.8.1 Generic and Biomedical Ontologies	17

2.8.2	Biomedical Knowledge Representation	18
2.9	Ontology Learning	18
2.9.1	Knowledge Acquisition through Information Extraction from text Natural Language Processing (NLP)	19
2.9.2	Controlled Natural Language	21
2.9.3	The project: Attempto Controlled English	22
2.10	Tools for specialized NLP	27
2.10.1	Apache Software Foundation (Apache) OpenNLP	28
2.10.2	Apache Tika	29
2.10.3	Apache Unstructured Information Management Architecture (UIMA)	29
2.10.4	Apache clinical Text Analysis and Knowledge Extraction System (cTAKES)	31
2.11	Tools for Ontology manipulation	32
2.11.1	Protégé	32
2.11.2	Protégé Ace View Plugin	34
2.11.3	Protégé OWL API	34
2.11.4	Apache Jena	35
3	Clinical Knowledge (CK)	37
3.1	Knowledge Representation	37
3.1.1	Ontological Realism (OR)	38
3.1.2	Ontological relations for clinical practice	38
3.1.3	Clinical text available sources	39
3.2	Knowledge Acquisition	43
3.2.1	Using the Ontology for General Clinical Practice (OGCP) for Clinical Controlled Language building	44
4	Ontology for General Clinical Practice proposal	47
4.1	OGCP Presuppositions	47
4.1.1	Standard Nomenclature of Medicine - Clinical Terms (SNOMED-CT)	49
4.1.2	Unified Medical Language System (UMLS) Clinical Observations Recording and Encoding (CORE)	51
4.1.3	Suggested representation as Computer Semantic Interoperability (CSI) tool	51
4.2	OGCP ontologies alignment	52
4.2.1	Open Biological and Biomedical Ontologies (OBO) Foundry ontologies	52
4.2.2	OBO Foundry principles	53
4.2.3	Ontological Realism applied to OGCP	53

4.2.4	OBO foundry ontologies integration	54
5	Knowledge Base population	63
5.1	Supervised tutoring	64
5.1.1	Translation Memory as a controlled technical jargon repository	66
5.1.2	Translation Memory Manager tools	66
5.2	Automatic Ontology Learning	68
5.2.1	Clinical Controlled Language translation	72
5.2.2	Knowledge Acquisition through specialized NLP	72
5.2.3	Ontology Driven Expanded Semantic Annotation	76
5.2.4	Smart instance creation	77
5.2.5	Ontological relations formation	78
5.2.6	Pragmatic interpretation in NLP	78
5.2.7	Round Trip Debug and Repair	80
5.2.8	Reasoning with effective logics	82
5.3	Text interpretation	83
5.3.1	Preliminary considerations	84
5.3.2	Ontology structure considerations	84
5.3.3	OGCP enhancements in order to represent healthcare practice episodes	86
5.3.4	DRS rewriting methodology	88
6	Clinical Practice Knowledge Interrogation	91
6.1	Clinical reasoning	91
6.2	Clinical concept guidance	92
6.3	Discourse Based Enhancement	92
6.4	Quality indicators	96
7	Results and Discussion	97
7.1	System Architecture	97
7.1.1	Translation Memories workflow with CP-ESB	98
7.2	OGCP Population examples	100
7.3	Current on-going controlled results	110
7.3.1	Domain experts validation	111
8	Conclusions	113
8.1	Conclusions	113
8.2	Future Work	115

8.3	Comparable high standard formal results evaluation	116
References		119
A	Symbols and terminology	134
B	OGCP Description Logics (DL)	136

List of Figures

2.1	Tasks in TBox, Knowledge Base (KB) and ABox.	9
2.2	Protégé 4 Ontology Development Tool (Protégé).	33
2.3	Protégé ACE View Plug-in.	34
3.1	Subjective, Objective, Assessment, Plan (SOAP) report de-identified sample.	42
3.2	SOAP Points Insertion	43
4.1	Leaf nodes of OGCP.	49
4.2	Ontological structure of Ontology for General Clinical Practice (OGCP).	49
4.3	SNOMED-CT Concept Structure.	50
4.4	Ontology alignment structure in OGCP.	52
5.1	Knowledge Acquisition Phases.	64
5.2	Supervised Translation Memory (TM) training.	65
5.3	Axiom creation from SOAP.	69
5.4	SOAP-5682 Sample complex report.	70
5.5	National Center for Biomedical Ontologies (NCBO) expanded semantic annotation . .	76
5.6	Semantic parsing of ellipsis ill segment	79
5.7	Enrichment of OGCP into a Healthcare Knowledge Base (KB).	82
5.8	OGCP inpatient encounter.	84
5.9	OGCP case history.	85
5.10	Signs and Symptoms in Clinical History	86
5.11	OGCP Analysis	87
5.12	OGCP Medication	87
6.1	ACE View Plugin.	93

6.2	ACE View Plugin snippets list.	94
6.3	ACE View Plugin snippets editor.	94
6.4	Knowledge Base (KB) Enrichment through Interrogation.	96
7.1	CP-Enterprise Service Bus (ESB) Architecture.	97
7.2	TMs workflow with CP-ESB.	98
7.3	OGCP Acquisition Workflow.	99
7.4	SOAP Report example	100
7.5	OGCP Patient	101
7.6	OGCP Physician	102
7.7	OGCP Inpatient encounter	103
7.8	Analysis hierarchy in OGCP	107
7.9	Therapeutic act hierarchy in OGCP	109

List of Tables

4.1	Primitive instance level relations in Relations Ontology	57
4.2	Class-level relations in Relations Ontology	58
4.3	Properties of the relations in the OBO Relations Ontology	59
7.1	XSLT transformations from XML NCBO Annotation into OGCP instance	99
A.1	Symbols and notations	134
A.2	Namespaces for qualified names abbreviation	135

Chapter 1

Introduction

1.1 Research context and motivation

Having followed the Artificial Intelligence (AI) field for the last 14 years it was apparent that the application of *computable reasoning* to the *healthcare* sub-domain of life sciences was at the turn of a corner. In fact, several developments were achieved that raised that possibility.

The sub-domain of Artificial Intelligence (AI) that explores reasoning based in a theoretical representation model of some reality, an **Ontology**, has come a long way in recent years. Several innovative concepts, techniques, methodologies and tools have surfaced recently in literature that induced our research for the last 3 years. Huge multi-billion dollars investments were made, and are in progress, that give consistence to this line of research and have been achieving impressive results in the **Biomedical computing** science domain.

The steady adoption of scientific achievements on *reasoning in the Semantic Web* also bring some new possibilities into our work.

A state-of-the-art of computable knowledge representation in the life sciences domain was developed in the preliminary phases of the work back in 2011 [MR11c]. Being directed at Computer Semantic Interoperability (CSI) application of understanding the meaning of medical texts, it was extended and published in [MR12] and later in [MRB13c]. Still under the theme interoperability studies, a paper regarding issues related to HL7 semantic interoperability is [MR11a]. These works rendered evident that an ontology for the healthcare sub-domain would facilitate the application of Artificial Intelligence to that realm of science.

Actually, in the last decade some *biomedical ontologies* were developed as computable knowledge representation in several sub-domains of life sciences. They were proven to be computationally effective. A multi million dollar investment was made in the *Human Genome Ontology* (GO) beginning in 2001. The first partial model of the Human Genome was done in 2004 and the complete map achieved in 2007. The Human Genome model was built by hypothesis formulation and validation that were ontologies generated by automatic reasoning proving thus **Description Logics** as a fundamental piece

of Artificial Intelligence application in *Life Sciences*.

However, an ontology to model clinical practice does not exist yet due to scarce academic/scientific interest. In fact, scientists don't look after, and investments are few, in the application of Artificial Intelligence in "*mundane activities*" like healthcare.

The second impeaching factor is that the modeling language (*OWL2*) was only standardized in 2009. We shall see how some Description Logics capabilities and correspondent computational reasoning, which are essential to domain modeling in life sciences, did not exist in *OWL* standardized in 2004 but are only defined in *OWL2*.

As the third unfavorable condition, we can observe that the necessary computational reasoning capabilities were only developed and released very recently. The algorithms that allow *clinical reasoning* with acceptable response times appeared in literature only in 2011.

Last, but not the least, there is a distinguishable higher difficulty in development due to the "**Knowledge Acquisition Bottleneck**" that, in our opinion constitutes the major obstacle to the *Semantic Web* development in life sciences. To address this problem we shifted our research from scrapping messages to **text oriented acquisition** after the realization that the largest part of healthcare information resides in clinical notes. The research papers in 2012 [MR12, MRRSB12] illustrate that shifting. These represent the basis of our contribution to solve the Knowledge Acquisition Bottleneck problem.

During 2013 the works focused on ontology development and integration tasks [MR13a, MR13b]. A noticeable contribution was also in the clinical justification of the results of the previous research in [MRB13a] because this justification is fundamental for the acceptance by healthcare professionals.

The summarization of the *OGCP* structure and automatic population appeared in [MRB13b].

This document presents applied Artificial Intelligence developing innovative contributions to Clinical Practice (*CP*) Knowledge Representation (*KR*) and reasoning.

We direct our research into developing a foundation of Horn-*SQIH* ontologies due to their computational advantages both in modeling as in reasoning.

An ontology already pre-populated for very effective reasoning in a *Cardiovascular healthcare* environment is provided. We uncover our options in proposing the ontological framework and the chosen ways to extract significant Clinical Practice (*CP*) knowledge from text.

The acquired knowledge constitutes the support for Question Answering (*QA*) based clinical reasoning aids that will also be presented.

1.2 Research questions

In the present work different research questions had to be addressed but, ultimately, they all collapse into one that embrace all the others.

- How to model such a complex domain like healthcare ?
Of course this is an extremely difficult scientific sub-domain of life sciences to model as it encompasses an enormous number of semantic relations in several sub domains of health like anatomy, physiology, therapeutic procedures, symptomatology among several others.

- How to perform *computational reasoning* over a representation necessarily so complex ?
What reasoning techniques, algorithms and tools are usable that can handle such complex representation is one of the research important questions.
- What are the technologies and tools to be adopted that can lead to generalized acceptance by healthcare professionals ?
It is of significant importance, and particular attention has to be taken to, the acceptance factor that the ultimate final users have to show for the result of this work to reach any impact.
- Can the recent scientific and technological breakthroughs be developed to attain the intended result of *effective clinical computational reasoning*?
This last one can be considered the joint research question because it is the drive shaft of all the PhD endeavors.
We shall research all the most recent scientific developments in the field and their applicability to the former questions to prove our thesis.

1.3 Scientific Innovation

Driven by the research question presented in the previous section some of the results that were obtained can be considered innovative by themselves, we stand out and itemize five of them as the more significant scientific innovations:

1. **Ontology for General Clinical Practice (OGCP)** creation.
An ontology was developed in order to model the healthcare sub-domain of life sciences. Clinical practice healthcare in particular and it was developed in a way that any medical specialty can be incorporated as a movable part.
2. **Knowledge Acquisition Bottleneck (KAB)** problem resolution.
The **KAB** problem is solved by the realization that knowledge figures mainly in controlled medical texts that constitute the source of our automatic knowledge acquisition proposal.
3. Automatic enrichment using **Clinical Controlled Language (CCL)** for any specialty.
The term **CCL** is coined that represents the formalization of "medical language" in accordance to the **OWL** verbalization of the **OGCP** T-Box built for a given specialty.
4. The definition of an extended Discourse Representation Structure (**DRS**), **Clinical Integrated Discourse Extended Representation Structure (CIDERS)**, that increases the scope of Co-Reference Resolution (**CRR**) to the whole of a patients Clinical History (**CH**) using disparate source texts.
Unlike the traditional Discourse Representation Theory (**DRT**) based **DRS** used mainly for anaphoric reference resolution in a single text.
5. **Clinical Decision Support Systems (CDSS)** validated through disambiguation using the acquired clinical practice model.
The **QA** systems that can be developed over the knowledge representation introduced in this work only allow clinically valid questions and answers and constitute then *knowledge oriented CDSS* that are innovative healthcare systems.

In the remainder of this introduction a readers guideline is given, describing the structure of this document.

A careful detail was given to a heavy cross referencing because readers with different expectations at diverse moments will be interested in jumping back and forth to other specific points in text.

The overall sectioning provides a progressive introduction to all the needed concepts intending for a natural sequence reading.

1. A full list of figures, tables and acronyms is presented at the beginning.
All the acronyms in the text refer to the definitions in this preliminary section.
2. In the current chapter 1 the intent of the present work is introduced and the reader is given an explanation about the document structure.
3. In chapter 2 we provide a detailed listing of the needed tools and technologies in all the aspects that are significant to this work. At the beginning a current state-of-the-art is discussed to give a detailed overview of every scientific recent achievements that format our research and support the options taken. Starting in section 2.9.3 we derive from a State-of-the-Art discussion to introduce the relevant aspects of tooling and technology used.
4. We begin the explanation of our work by introducing the preliminary challenges of both Clinical Knowledge Representation (KR) and Knowledge Acquisition (KA) in chapter 3.
5. By introducing and detailing the proposed ontological framework in chapter 4 we justify the adoption of the population and enrichment concepts, methodologies and tools presented in the next chapter.
6. Chapter 5 details a complete overview of all the automatic steps involved in clinical knowledge acquisition:
 - (a) The Machine Learning (ML) process tutoring and bootstrap for controlled translation from natural language SOAP reports to clinical English in the Cardiovascular specialty.
 - (b) The steps for automatic Knowledge Base population from texts are explained.
 - (c) the advanced and practical $\mathcal{EL}++$ inference restrictions, opportunities and techniques that will allow Cardiovascular Domain Experts (DEs) to benefit from.
 finally "Case based Clinical guidance" is also shown in this chapter.
7. Chapter 6 concludes the coverage of our research question itemizing all the steps now possible for Clinical Practice guidance through Knowledge Base natural language interrogation.
8. Some obtained results and discussion around them are to be found in chapter 7.
9. Conclusions and future work suggestions are wrapped up in chapter 8.
10. After the conclusions we find the bibliographic references.
11. For self containment of the work some reading aids are provided in the Symbols and Terminology annex in page 134,

12. An appendix is provided at page 136 with the detailed \mathcal{DL} axiomatic structure of the leaf nodes of Ontology for General Clinical Practice (OGCP).

In this annex the ontology definitions, that are spread along the text, are found for reference using \mathcal{DL} notation.

Chapter 2

Tools and Technologies Addressed

All the tools and technologies that were reviewed leading to the construction of the current proposal are presented.

When adequate State-of-the-Art reviews, regarding specific scientific issues that are to be addressed already exist, they are referred and not transposed here.

When some documents, however, are very important to understand this work, the most significant examples of those documents are reproduced here, with the due credits and references, for self containment.

Every issue that is covered in the current document is presented going from more general themes to more specific ones.

2.1 Semantic Web

The notion of a Semantic Web was first introduced by Tim Berners-Lee the inventor of the World Wide Web (**WWW**) and director of World Wide Web Consortium (**W3C**) at Conseil Européen pour la Recherche Nucléaire (**CERN**) in 2001 [**BLHL01**]. Also known as Web 3.0, the main breakthrough of the Semantic Web is to evolve from the *Web of documents* into the *Web of data*.

This is achieved by turning the concepts managed by computers "*understandable*" by them so that linkage between data in the web can be done automatically [**AAD⁺09**] by systems and software agents. For it to be possible some way of encoding the *concepts meaning* had to be delivered and the Resource Description Framework (**RDF**) protocol was developed.

For manipulation of the data in **RDF** format the language of choice is SPARQL Protocol and RDF Query Language (**SPARQL**) however, in our case, it is seldom used to process ontologies maintained in Web Ontology Language v.2 (**OWL2**) files because we normally process them using either Web Services (**WSs**) or Web Ontology Language (**OWL**) tools and Application Programming Interfaces (**APIs**).

The **W3C** [**W3C11b**] has established the Semantic Web for Health Care and Life Sciences Interest

Group (**HCLS IG**)¹ to help organizations in their adoption of the Semantic Web.

This adoption has been, however, rather slow paced [BZC10] due mainly to the *knowledge acquisition bottleneck* [WLB12] verified in mining knowledge from text in specific scientific domains.

A very recent review addressing the issue of generic knowledge discovery in medicine appeared in [EBMT14].

We are, however, particularly concerned in knowledge harvesting from text for Ontology Learning (**OL**) and reasoning in the **specific sub-domain of healthcare**.

The use of logics in ontologies ranges from sound modeling to practical querying of that knowledge, thus adding a considerable value, so the basic foundations of data representation of ontologies and their Description Logics (**DLs**) are now introduced.

2.2 Resource Description Framework

The Resource Description Framework (**RDF**) is a language for representing information about resources in the World Wide Web [KC06]. The current version **RDF** 1.1 was published in 2014, February 25th [W3C14].

In **RDF**, there is no technical distinction between Tbox and Abox (in some other logic languages there is) and distinguishing between them is just a matter of convention.

TBox (T for Terminology) defines the schema or taxonomy, it is terminological data. In other words, it's the data that defines classes, properties, and relationships in your ontology.

The ABox (A for Assertions) is the data. This is the data where you enumerate the individual instances of your class and describe them. You can't have assertions without using some terminology, and terminology isn't that useful unless you actually use it to make some assertions, so ABox and TBox are just two different but required parts of a knowledge base.

An analogy from the SQL world: "CREATE TABLE ..." creates Tbox data; "INSERT" creates Abox data.

TBox and ABox logic operations differ and their purposes differ. TBox operations are based more on inferencing and tracing or verifying class memberships in the hierarchy (that is, the structural placement or relation of objects in the structure). ABox operations are more rule-based and govern fact checking, instance checking, consistency checking, and the like.

ABox reasoning is generally more complex and at a larger scale than that for the TBox.

2.3 Description Logics and Web Ontology Language v.2

Description logics constitute a family of fragments of first-order logic (nearly all of which are decidable), in which members of this family are primarily differentiated based on the set of allowed logical operators.

For example, some logics exclude negation and universal quantification, which in turn determine the

¹<http://www.w3.org/2001/sw/hcls/>

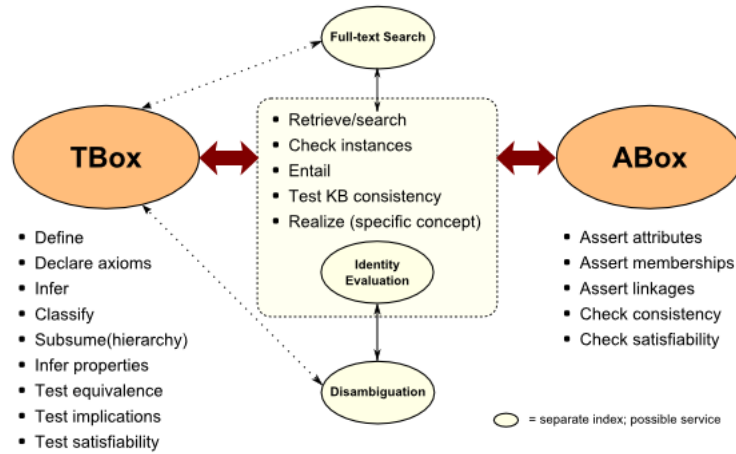


Figure 2.1: Tasks in TBox, KB and ABox.

computational complexity of inference with the language.

Most of the state-of-the-art in DLs was synthesized already in 2003 in *"The description logic handbook: theory, implementation, and applications"* [Baa03], we explore thoroughly the second edition published in 2007 [BCM⁺07]. This comprehensive introduction to DLs in this new edition includes a chapter on ontology languages for the semantic web with full coverage of all aspects of the subject: theory, implementation and applications.

This new chapter is devoted to the new Web Ontology Language (OWL) that was developed and recommended by W3C as the standard web ontology language for the Semantic Web (SW). The Web Ontology Language (OWL) is a standard ontology language that captures the semantics of many description logic languages presented and we introduce ahead in this chapter in 2.3.2 also its state-of-the-art .

2.3.1 Description Logics

As the name Description Logics (DL) indicates, one of the characteristics of these languages is that they are equipped with a formal, logic-based semantics. Another distinguished feature is the emphasis on reasoning as a central service: *reasoning allows one to infer implicitly represented knowledge from the knowledge that is explicitly contained in the knowledge base.* [BCM⁺07]

DL languages are then viewed as the core of knowledge representation systems, considering both the structure of a DL knowledge base and its associated reasoning services. In DL, the important notions of a domain are described by means of concept descriptions that are built from concepts (also referred to as classes), roles (also referred to as properties or relations), denoting relationships between things, and individuals (also referred to as instances).

It is now possible to state facts about the domain in the form of axioms.

- Terminological axioms (*TBox Set*)
make statements about how concepts or roles are related to each other,
- assertional axioms (*ABox Set*) (sometimes also called facts)

make statements about the properties of individuals of the domain [VHH09].

When trying to develop a process of ontology learning by acquiring an ontology from text sources we have to balance all the developments made so far in:

- The expressibility of the DL supporting our ontology in order for the reality to be appropriately represented.
- The reasoning methods and capabilities available for the chosen DL to classify the ontology in practical terms (acceptable response times, tractability).
- The difficulty degree in enriching automatically (ABox) the ontology from text that will, obviously be higher as long as the ontology structure (TBox) is richer in terms of complexity.

As could be expected, all the developments that happened in recent years have been going two-fold simultaneously:

- The complexity of the proposed biomedical ontologies has been increasing by the realization that Ontological Realism (OR) is essential for accurate reasoning and the application of the Open Biological and Biomedical Ontologies (OBO) foundry framework [The14] since it was generally accepted as a cornerstone for coordinated evolution in this scientific domain [SC10].
- Fine tuned subsets of expressive DLs and the reasoning methodologies that have been evolving steadily both for local [BBL05], [LB10] and distributed reasoning [DG08], [UKOVH09], [MMH10], [KKS11].

Web Ontology Language v.2 (OWL2) is the representation language of choice for our proposed ontology so we had to achieve fluency in its details.

2.3.2 Web Ontology Language

The W3C Web Ontology Language (OWL) [W3C11a] is a Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things. The OWL ontology language is based in Description Logics, the family of class-based knowledge representation formalisms presented in the previous section such that knowledge expressed in OWL can be exploited by computer programs, e.g., to verify the consistency of that knowledge or to make implicit knowledge explicit.

OWL documents, known as ontologies, can be published in the World Wide Web and may refer to or be referred from other OWL ontologies.

OWL version 1 was approved by W3C in 2004 for the adoption of the formal evolution of DAML+OIL de-facto standards that had been around in the Knowledge Representation (KR) community for some years then. OWL had serious shortcomings in its ability to represent reality [SCK⁺05].

Evolution led to the approval of the Web Ontology Language v.2 (OWL2) standard and its associated profiles [W3C11a].

With **OWL2** enough expressive possibilities are available to develop the reasoning needed for biomedical and clinical sub-domains and contributions flourished [GHM⁺08]. The most significant, in terms of representativeness and overall quality and coverage, *unified framework* appeared as late as 2010 [CS10]. The application of the proposed framework is fully adopted in our work. Although generating a higher complexity in the enrichment steps, the level of ontological expressiveness will lead our **KBs** to unprecedented reasoning capabilities while maintaining tractability.

2.4 Reasoning about Knowledge

Concerning Knowledge Representation (**KR**) the disadvantages of natural language are its ambiguity, vagueness and potential inconsistency. To represent knowledge in computers people use formal languages. These languages have a well-defined syntax and an unambiguous semantics, and support formal methods, specifically reasoning [SS11, CDGL⁺07].

Uses for Reasoning

Reasoning support is important for

- checking the consistency of the ontology and the knowledge
- checking for unintended relationships between classes
- automatically classifying instances in classes

Checks like these are valuable for

- designing large ontologies, where multiple authors are involved
- integrating and sharing ontologies from various sources

Kinds of Reasoning about Knowledge

Some common ground routines have become usual in the reasoning realm, these can be split into:

- Class membership
 - If x is an instance of a class C , and C is a subclass of D , then we can infer that x is an instance of D
- Equivalence of classes
 - If class A is equivalent to class B , and class B is equivalent to class C , then A is equivalent to C , too

- Consistency
 - X instance of classes A and B, but A and B are disjoint
 - This is an indication of an error in the ontology
- Classification
 - Certain property-value pairs are a sufficient condition for membership in a class A; if an individual x satisfies such conditions, we can conclude that x must be an instance of A

All these abilities are found in the general purpose reasoners that have appeared since the Web Ontology Language v.2 standardization.

OWL Reasoning

The rich semantics of **OWL** provide powerful reasoning capabilities that help build, maintain and query domain models for many purposes. The reasoning capabilities of a system based in any **DL** is strongly dependent in two factors that have been evolving powerfully in recent years:

- The expressiveness of the underlying **DLs**, ranging in the different capabilities introduced above in 2.3.1.
- The processing power of the systems that support the computation, whether is it local or distributed.

The reasoning capabilities depend on the expressiveness of the **OWL** profile which relies on the underlying **DL**. In the current **OWL2** standard there are fragments defined that trade expressive power for favorable properties that are not shared by the full **OWL** language. In particular, several fragments are a subset of OWL Description Language (**OWL DL**), the Description Logics dialect of **OWL**, and aim at high efficiency for reasoning tasks such as subsumption, classification, and satisfiability [W3C11a]. The different profiles are [W3C12]:

- **OWL2 \mathcal{EL}**
- **OWL2 \mathcal{QL}**
- **OWL2 \mathcal{RL}**

The different characteristics of the sub-languages may be verified in the **W3C** page. When trying to select which profile is the most adequate for our representation we studied the trade-off between the availability of the different reasoners with acceptable response times and had to consider the relative sizes of the expected T-Boxes vs. the A-Boxes as explained in [LSS13] and consequently opted to use the \mathcal{EL} expressiveness and profile. Some special attention must be drawn then about the specific computational behavior of the \mathcal{EL} profile: *"OWL2 \mathcal{EL} is particularly useful in applications employing ontologies that contain very large numbers of properties and/or classes. This profile captures the expressive power used by many such ontologies and is a subset of OWL 2 for which the basic reasoning*

problems can be performed in time that is polynomial with respect to the size of the \mathcal{EL} ontology. Dedicated reasoning algorithms for this profile are available and have been demonstrated to be implementable in a highly scalable way. The \mathcal{EL} acronym reflects the profile's basis in the \mathcal{EL} family of description logics $\mathcal{EL}++$, that provide only Existential quantification."

Polynomial time algorithms can be used to implement the ontology consistency and class expression subsumption reasoning problems. As mentioned earlier in section 2.3.1 the recent distributed reasoner implementations leads us to maintain strict $\mathcal{EL}++$ conformance in order for the reasoning steps described in the Advanced Inference section 2.6 to be performed.

In the mere case of diagnosing and its related "cognitive traps", the issue is raised and developed in [LD11], where the estimates place the rate of medical misdiagnoses as high as 15 percent and some 80% of those errors can be accounted for by a cascade of cognitive errors. In the referred paper the hypothesis that by making the underlying reasoning pattern and its inferences explicit, and by helping diagnosticians become conscious of potential pitfalls in reasoning, their skill in making correct diagnoses will improve.

When using a Natural Language Processing (NLP) Question Answering (QA) based system the reasoning becomes explicit and the hypotheses inferred by the system are complete as far as the underlying ontology is, so a significant improvement in the diagnostic accuracy should be expected if the ontology coverage is as wider and accurate as possible in the current state-of-the-art.

2.5 Reasoning Support for OWL

In description logics, decidability of reasoning problems and the provision of sound and complete reasoning algorithms is key [SS11, WZGP04]. Description Logics are formal logics with well defined semantics. The semantics of a description logic is specified via model theoretic semantics, which explicates the relationship between the language syntax and the models of a domain. Semantics are thus a prerequisite for reasoning support. Formal semantics and reasoning support are usually provided by

- mapping an ontology language to a known logical formalism
- using automated reasoners that already exist for those formalisms

OWL is mapped on a description logic, and makes use of reasoners such as *FaCT*, *RACER* and *Pellet*. Description Logics are a subset of Predicate Logic (PL) [Rec03] for which efficient reasoning support is possible and particularly diverse profiles of OWL2 were defined to be computable as viewed in 2.4. In the description logic *SHOIN*, the description logic underlying OWL \mathcal{DL} , we can build complex classes from atomic ones using the following constructors:

- $C \sqcap D$ (intersection), denoting the concept of individuals that belong to both *C* and *D*,
- $C \sqcup D$ (union), denoting the concept of individuals that belong to either *C* or *D*,
- $\neg C$ (complement), denoting the concept of individuals that do not belong to *C*,
- $\forall R.C$ (universal restriction), denoting the concept of individuals that are related via the role *R* only with individuals belonging to the concept *C*,

- $\exists R.C$ (existential restriction), denoting the concept of individuals that are related via the role R with some individual belonging to the concept C ,
- $\geq n R$, $\leq n R$ (qualified number restriction), denoting the concept of individuals that are related with at least (at most) n individuals via the role R .
- $\{c_1, \dots, c_n\}$ (enumeration), denoting the concept of individuals explicitly enumerated.

Based on these class descriptions, axioms of the following types can be formed:

- concept inclusion axioms $C \sqsubseteq D$, stating that the concept C is a subconcept of the concept D ,
- transitivity axioms **Transitive**(R), stating that the role R is transitive,
- role inclusion axioms $R \sqsubseteq S$ stating that the role R is a subrole of the role S ,
- concept assertions $C(a)$ stating that the individual a is in the extension of the concept C ,
- role assertions $R(a, b)$ stating that the individuals a, b are in the extension of the role R ,
- individual (in)equalities $a \approx b$, and $a \not\approx b$, respectively, stating that a and b denote the same (different) individuals.

We can make complex statements, for instance, expressing that two concepts are disjoint with the axiom $A \sqsubseteq \neg B$. This axiom states that A is a subconcept of the complement of B , which intuitively means that there must not be any overlap in the extensions of A and B .

An interpretation consists of a *domain of interpretation* and an *interpretation function* which maps from individuals, concepts and roles to elements, subsets and binary relations on the domain of interpretation, respectively. A knowledge base consists of a set of axioms which act as constraints on the interpretations.

2.6 OWL Reasoners

Well inside the **OWL** period of establishment but way before the standardization of **OWL2** and all its associated profiles, a proposal for a faithful Integration of Description Logics with Logic Programming appeared [MR07] where a methodology for reasoning with any generic \mathcal{DL} is introduced. [Aug05]. Plenty of developments were consecrated ever since that led to a large variety of reasoners that vary in the expressiveness of the underlying \mathcal{DL} that they are able to handle and in the mechanisms for performing the various interesting reasoning tasks as presented in 5.2.8. Reasoning with **OWL** ontologies has high worst complexity but considering the different characteristics of the reasoners and of the ontologies to which they are to be applied a good selection can lead to good results in practice. A very shallow distinction can be made in

1. Classical,
are description logic reasoners based on tableaux algorithms that are able to classify large, expressive ontologies but they usually only provide limited support in dealing with large number of instances (ABox).

2. Database like,
Infer knowledge upfront and are able to handle large amounts of assertional facts, but are limited in terms of the logic they are able to support.
3. Case based reasoners.
Create the set of rules that generate all the inferring possibilities and by extensional application of these rules infer all the **consequences** (*ABox*), normally in polynomial time on the size of the rule set.

advanced inference is normally referred to by that name because we are arguing in favor of the utilization of the latest of the enumerated types, case based reasoners, when performing our classification and knowledge inferring tasks over any **OGCP** Knowledge Base.

We consider that this incarnation of reasoners, being developed most recently by the application of a novel paradigm, fit particularly well in our "advanced" albeit expressively limited ontology. It is a contribution to the naming also the fact that it has been implemented with distributed capabilities rendering the possibility of performing its abilities not just to a limited size **KB** suchlike the present **OGCP** but to something much broader in scope as those that might have all the **SNOMED-CT** inside them.

2.7 Consequence driven reasoning

To handle such big ontologies as those that figure in the Unified Medical Language System (**UMLS**) with millions of instances there is the need for the novel capabilities of distributed reasoning introduced by kazakov et al. as recently as 2011 [KKS11].

Directed to reasoning in an \mathcal{EL} type of \mathcal{DL} environment, to conform to the limitations explained in section 5.2.8, our solution is to restrict **OGCP** to $\mathcal{EL}++$ and have a near polynomial classification time using *ELK*. As explained in the mentioned section, the reason of using an \mathcal{EL} profile ontology is mainly determined by the classification time in an ontology with an expected structure as **OGCP** as shown in [Krö10].

2.7.1 The ELK reasoner

The reasoning rationale for **OWL** \mathcal{EL} is presented here for self containment [VHH09].

ELK is a free and open source reasoner for the lightweight ontology language **OWL2** \mathcal{EL} . It is based on Java and can be controlled using the OWL Application Programming Interface (**OWL API**) through **Protégé** both introduced in section 2.11, the Snow Owl ontology editor, or through a basic command line interface.

ELK is available under the Apache License 2.0.

The goal of *ELK* is to provide a very fast reasoning engine for **OWL** $\mathcal{EL}++$. Currently, the supported **OWL** features and reasoning tasks are still limited (but already sufficient for important ontologies such as **SNOMED-CT**). The aim of the project is to complete the implementation for all **OWL** \mathcal{EL} features and relevant reasoning functions (e.g. for unrestricted use in **Protégé**).

This is being done step-by-step so as to ensure top performance of each new feature.

ELK is very fast, it can classify the **SNOMED-CT** ontology with more than 400,000 classes in a few seconds on a modern laptop. This is achieved by highly optimized consequence-based reasoning algorithms that can also take advantage of multi-core CPUs.

The latest 0.4.0 release features the new incremental reasoning support. If performing reasoning tasks after small ontology modifications, *ELK* 0.4.0 tries to reuse the result of the previous computations as much as possible. This makes reclassification of ontologies in editors like **Protégé** almost instantaneous which is a very important feature in our Clinical Controlled Language (**CCL**) interface for the system to be found acceptable by the end users. Thanks to incremental reasoning, *ELK* 0.4.0 now also supports answering **DL** queries with complex class expressions, as well as finding explanations using the explanation workbench in **Protégé** 4.x and 5.

The main features are:

- Reasoning tasks: classification, consistency checking, class instance retrieval
- OWL API bindings
- **Protégé** plugin
- Command line interface
- Parser for input files in OWL 2 Functional Style Syntax

The following constructs of the OWL ontology language are supported in *ELK* 0.4.0.

- Axiom types:
 - SubClassOf
 - EquivalentClasses
 - DisjointClasses
 - SubObjectPropertyOf
 - EquivalentObjectProperties
 - TransitiveObjectProperty
 - ReflexiveObjectProperty
 - ObjectPropertyDomain
 - ClassAssertion
 - ObjectPropertyAssertion
- Class expressions:
 - owl:Thing
 - owl:Nothing
 - ObjectComplementOf (only positive occurrences, see below)
 - ObjectIntersectionOf
 - ObjectUnionOf (only negative occurrences, see below)

- ObjectSomeValuesFrom
- ObjectHasValue
- DataHasValue (preliminary support, see below)
- Property expressions:
 - ObjectPropertyChain
- Individual expressions:
 - NamedIndividual
- Literal expressions:
 - datatype literals in arbitrary datatypes (preliminary support, see below)

The ObjectComplementOf constructor is supported only in positive positions i.e., in the second concept of SubClassOf axioms, provided it does not occur under another ObjectComplementOf. The ObjectUnionOf constructor is supported only in negative positions, i.e., in the first concept of SubClassOf axioms. In these cases the constructors can be expressed using other constructors.

The lexical-value mapping of data literals is not supported yet, because support for it is still preliminary, so the equality of values in different syntactic forms is not recognized so far. Therefore in our "smart instance creation" introduced in section 5.2.4 we are particularly cautious about always using the same syntactic form when generating the axiom from our annotated text.

2.8 Biomedical Resources

In this section we illustrate the application of the evolutions in ontological engineering and its effect regarding knowledge representation in the biomedical domain of science.

2.8.1 Generic and Biomedical Ontologies

Regarding what is an ontology we must, first of all as suggested by [GOS09], distinguish between the philosophical discipline, namely the branch of philosophy that deals with the nature and structure of reality, and the common meaning in computer science that is what interest us in this work. In the Artificial Intelligence field the term ontology as an information systems artifact was coined back in 1993 by T. R. Gruber [GRU93] and later refined in [Gru95].

Regarding Gruber, *An ontology is an explicit specification of a conceptualization*. We refer to an ontology as a special kind of information object or computational artifact that are a means to formally model the structure of a system, i.e., the relevant entities and relations that emerge from its observation.

Ontologies diverge from terminologies as long as their underlying logical representation allows for reasoning support.

In the Biomedical field a special concern has been evolving around the application of Formal Ontology (FO) [Smi98] to this scientific domain. In the referred paper, that follows Edmund Husserl work [Hus00], the basic concepts of FO are presented:

1. the theory of part and whole,
2. the theory of dependence,
3. and the theory of boundary, continuity and contact.

This concern serves as a principle for Ontological Realism (OR) applied to this scientific domain.

2.8.2 Biomedical Knowledge Representation

In the Biomedical sub-domain of science the most significant amount of work in ontology enrichment and population has been done in the Biomedicine research area as illustrated by [SB⁺08].

Most of the work, as will be evident later, has evolved from the enormous scientific effort poured into the development of the Gene Ontology (GO) in the early years of our century [ABB⁺00]. Several principles adopted in the specific sub-domain of biomedical ontologies developed so far have followed from the foundational work in the GO development.

In concrete terms a solid foundation for harmonized development has been evolving thanks to the efforts of the Open Biological and Biomedical Ontologies (OBO) Foundry [SAR⁺07] that made a systematic adoption of the GO principles as will be detailed in 4.2.2.

With the steady adoption of the Open Biological and Biomedical Ontologies (OBO) ecosystem and the proposals that appeared thereafter like the cited [SB⁺08, CS10] our Knowledge Representation choice was a natural evolution.

2.9 Ontology Learning

Ontology curation has been developing very slowly in the Biomedical field as illustrated in section 2.8.2. The short availability of domain experts and the highly specialized, tedious and error prone characteristics of the mentioned curation tasks have resulted in the slow paced adoption. Ontology editors like Protégé 4 Ontology Development Tool (Protégé) can help the expert formalize his/her knowledge but they are generally very far from an automated procedure. It looks very important then to state explicitly the steps that can be automated in order to alleviate the task of human experts and the burden of knowledge acquisition.

Ontology learning is the application of a set of methods and techniques used for building an ontology from scratch. It uses distributed and heterogeneous knowledge and information sources to induce a reduction in the time and effort needed in the ontology development process. These learning techniques can vary according to the degree of automation (semi-automatic, fully automatic), the ontological knowledge that has to be extracted (concepts, taxonomy, conceptual relationships, attributes, instances, axioms)[ZN10] and finally the purpose (creating ontologies from scratch and/or updating existing ontologies).

The type of knowledge sources seriously affect the Ontology Learning techniques applicable:

- Structured data
Extracting concepts and relations from knowledge contained in structured data, such as databases

- Semi-structured source
Elicit an ontology from sources that have some predefined structure, such as XML Schema
- Unstructured sources
Involves NLP techniques, morphological and syntactic analysis, etc.

We are exploring the realm of **Expressive** Ontology Learning which are that kind constituted by rich ontological relations. The support for reasoning is the major benefit of the use of expressive ontologies grounded in logics.

Reasoning can be used in different phases of the life cycle of an ontology:

1. At development time,
reasoning can be used to validate the ontology and check whether it is non-contradictory.
2. at deployment time,
reasoning allows to derive conclusions from the ontology, like query answering over the ontology and interactively enriching along this process.

It will be shown in section 3.2 that in our case we will use a semi-structured source that albeit being a text source already has an internal clinically oriented structure.

2.9.1 Knowledge Acquisition through Information Extraction from text NLP

A truly important recent review about Information Extraction (IE) is the report of the "Second Strategic Workshop on Information Retrieval" appropriately titled "Frontiers, Challenges, and Opportunities for Information Retrieval" [ACMS12].

Several times in the report the NLP tasks are flagged as both a risk and a research opportunity.

The current "definitive bible" of Ontology Learning from text is [BC08] where an extensive collection of foundational papers in the subject are included. Specially important for the present work is the "Learning Expressive Ontologies" paper that led to the creation of the autonomous book with the same title [VHH09] and the [PP08] about automatic extraction from text of semantically rich ontological relations. The main issues about semantic retrieval from text in the biomedicine field are deeply summarized in [LHC11] and [WLB12] and we illustrate now the specific issues regarding the automated acquisition directed towards the structure of our proposed ontological framework.

Temporal Information Retrieval (IR)

Alonso et al. [ASBYG11] give an overview of the value of temporal information and discuss current research trends in temporal information retrieval.

Identity tracking

In our specific case a form of identity tracking is compulsory because the system will have to provide case based reasoning and we are learning/enriching our knowledge base from several different reports.

The identity tracking is only an issue when several disparate systems have to correlate Personal Health Information (PHI) while preserving the access to confidential information.

Several contributions surfaced to address the problem being the most significant effort the work developed by the National Center for Ontological Research (NCOR) of New York State University at Buffalo (UB) in its Referent Tracking Unit (RTU) [UB 06] where the integration of disparate information resources arising from the different branches of biological research and clinical medicine is studied. The problems that the elicitation of ontological references from free text produce are the subject of this research unit and their contributions to the formation of the formal ontological basement that the OBO rely upon are fundamental.

In their work proposals for accurate ontological relations elicitation are provided [CSKD04, CES06, SKSC06] that ultimately led to the unified proposal [CS10] in 2010.

Ethical and legal issues of clinical information usage

One of the main concerns of everybody working in the Biomedical field is the absolute guarantee of preserving all the ethical and legal restrictions when accessing Personal Health Information (PHI). It must be agreed upon in advance when beginning any work with a healthcare institution for these organisms tend to, and they have to, behave very strictly regarding these issues.

Trying to develop Artificial Intelligence (AI) methods that involve the usage of clinical information affects:

1. The patients sense of privacy regarding their health status and conditions.
2. The clinicians sense of autonomy and professional scrutiny that is normally restricted in a very formal way to their peers.

These aspects must be considered very seriously when starting an endeavor like the one presented in this work. Namely the next items have to be cautiously handled:

Privacy and De-identification

This is an extremely important issue because all clinical data has to be cleansed of the possibility of re-identifying in many of the purposes that may be of interest. For ontology learning is very important that the anonymization is performed automatically because it is impractical to have human based processes due to the big volume of cases to incorporate. A thorough review about automatic de-identification is [MFS⁺10] where a complete study around the U.S. reality is presented. In the U.S. de-identification is due to be in accordance to a specific standard, namely the so-called “Safe Harbor” by the HIPAA² that implies the proper anonymization of 18 patient identifiers including names, all geographical subdivisions smaller than a state, all elements of dates related to the individual, identifying numbers like phone, fax, social security, medical record, health plan, accounts, certificate or license, vehicle identification, device identification or serial numbers, e-mail addresses, Uniform Resource Locators (URLs), IP Addresses, Biometric Identifiers, full face photographs and any other uniquely identifying numbers or codes. In the European Union (EU), in the document -

²Health Insurance Portability and Accountability Act

A comprehensive approach on personal data protection in the European Union [COU11], the Council conclusions present a complete overview about personal de-identification and all the applicable EU regulations. It extends the concerns expressed in the 1995 Data Protection directive to also include Biometric and genetic data. Mainly the document invites the commission to deliver legislation (directives) based in principles in-line with recent technological advances regarding identification through biomedical data and it is then recognized as a work to be developed in cooperation with national data protection supervisors of the individual member states.

Two possibilities may be of concern, whether we are directing pre-processing labors to populate aggregate ontology information and then it seems adequate to have the kind of care suggested by the US Government and similar identifying removal practices must be enforced or the work is directed to other useful endeavors like EHR enrichment through automated reasoning and decision support aids in the clinical ground and then the identity must be removed but the record tagged for follow up purposes. For instance to correlate diagnostic findings to exams and to therapy applied later. In a very recent review study, it was shown that automatic text de-identification minimally reduces the informativeness of clinical notes [MFF⁺14], that only about 1.2–3% of clinical concepts in text are altered by de-identification and also only 0.81% or 1.17% of the annotated clinical concepts fully or partly overlap with Personal Health Information. These results are for generic de-identification but in our case, based in the semi-structured sources we shall use, we will guarantee full de-identification with no possibilities of rebuilding patient related data outside the automated acquisition system or the enriched Knowledge Base. We explore this concept in our benefit and explain how to guarantee this feature ahead in section 3.1.3.

2.9.2 Controlled Natural Language

A complete state-of-the-art in Controlled Natural Language for Knowledge Representation is [Sch10]. It is important to be aware that *"A quick read through of the W3C OWL web pages leaves no room for doubt that the preferred OWL syntax is RDF/XML. Even the OWL guide uses this syntax for the presentation of examples. However, the verbosity of the eXtensible Markup Language (XML), and the fact that it is difficult to write by hand, rule this syntax out for quickly writing and editing class descriptions in a concise manner. An alternative to the RDF/XML syntax is the OWL Abstract Syntax. This syntax is a high level, human readable OWL syntax. However, like the RDF/XML syntax, the Abstract Syntax is also verbose, it has an excessive number of keywords, and typically requires the use of a large number of brackets."* as Horridge et al. referred back in 2006 [HDG⁺06] and this has been representing a major drawback in the Semantic Web paradigm adoption.

Domain Expert are unfamiliar or uneasy with formal languages and formal methods. Furthermore, in order to express domain-specific knowledge in a formal language we need to bridge a conceptual distance. Thus, there exists a conflict between the tendency to use natural languages and the need to use formal languages. **Controlled Natural Languages have been proposed as a way to resolve this conflict.**

To make Web Ontology Language v.2 ontologies more usable for both domain experts, knowledge engineers and users at large, several notations have been developed. The motivation rests in the domain experts finding OWL too difficult to work with. The mainstream notations can be divided in 3 types:

- Human-readable formal syntaxes
Like the **OWL** Manchester Syntax (**MS**) [HDG⁺06].
- Graphical notations
Like UML [BBC⁺10].
- Controlled Natural Languages
like Attempto Controlled English (**ACE**) [KF07] on which our work stands.

A thorough study developed in 2010 based in the fundamental state-of-the-art review of 2008 [SKC⁺08] considered expressiveness, understandability and overall satisfaction by a large group of representative users in order to delineate some guidelines for the future of Controlled Natural Languages (**CNLs**). Some conclusions are eye-openers for anyone working in this field of Knowledge Representation (**KR**):

1. User testing of Manchester Syntax shows <50% comprehension of all structures.
2. **ACE** covers First Order Logic, with a fragment that can be bidirectionally mapped to **OWL** (excluding datatype properties).
3. Often, several possibilities for expressing the same **OWL** axiom exist which is very important regarding that we will have to make syntactic adjustments to maintain controlled expressibility.

We aimed at the active participation of Domain Expert in the ontology creation process. Ontology construction methodologies together with appropriate tools and technologies, such as controlled natural languages, semantic wikis, intelligent user interfaces and social computing, are being proposed to enable the direct input from Domain Expert and to minimize the dependency on knowledge engineers at every step of ontology development. Studying thoroughly [DDH⁺11] and the very recent [SD14] we can assert that the line of research with most promising results, as we will see later, is to base the Domain Expert work in a Clinical Practice Controlled Natural Language based in **ACE** [Att07].

2.9.3 The project: Attempto Controlled English

Attempto Controlled English (**ACE**) is such an important piece of the present work that it was opted to transpose here, for self containment, examples taken with the authors permission from [FKK08b] that illustrate some of the fundamental notions on which the **CNL** processor relies upon.

ACE is designed as a general-purpose controlled English providing a high degree of expressiveness. At the same time, **ACE** is fully interoperable with the Semantic Web standards, since a defined subset of **ACE** can be mapped bidirectionally to **OWL**. **ACE** texts can be mapped to Discourse Representation Structures (**DRSs**) .

The Attempto Parsing Engine (**APE**) translates an **ACE** text into a **DRS**. **APE** is implemented in Prolog as a Definite Clause Grammar (**DCG**) using feature structures. **APE** relies on a lexicon of function words and a full-form lexicon of about 100,000 content words. The resolution of anaphoric references is implemented as a separate module that accepts an unresolved **DRS** with additional conditions for anaphora and potential antecedents and produces a resolved **DRS** as the final output.

APE is publicly available via a remote procedure call, implemented as a REST web service. The service translates an **ACE** text into a **DRS**, and optionally provides information about the tokenisation,

syntax, paraphrase, and classical first-order logic representations of the input text [FKS06]. In the present work the function words lexicon extracted from the OGCP T-Box was uploaded to customize the web service.

In the following we transposed from the 2010 ACE manual [Att10] some advanced cases that can be taken as reference examples to facilitate the process understanding.

The main illustrative features of ACE are:

1. Syntax [Att10].

(a) Vocabulary

Comprises predefined function words (e.g. determiners, conjunctions), predefined fixed phrases (e.g. ‘it is false that’, ‘for all’), and content words (nouns, proper names, verbs, adjectives, adverbs).

(b) Grammar

Is expressed as a set of construction rules and a set of interpretation rules. Defines and constrains the form and the meaning of ACE sentences and texts.

(c) Texts

Are a sequence of declarative sentences that can be anaphorically interrelated. Furthermore, ACE supports questions and commands. Declarative sentences can be simple or composite and must be terminated with a .. Questions have to end with a ? and commands with a !.

(d) Sentence structure

Simple ACE sentences can have the following structure:

subject + verb + complements + adjuncts

Every sentence of this structure has a subject and a verb. [FKK08a] Complements (direct and indirect objects) are necessary for transitive verbs and ditransitive verbs, whereas adjuncts (adverbs, prepositional phrases) that modify the verb are optional. Every sentence of this structure introduces only the object described by the noun phrase. Elements of a simple sentence can be elaborated upon to describe the situation in more detail. To further specify the nouns, we can add adjectives, possessive nouns and prepositional phrases, or variables as appositions.

Composite sentences are recursively built from simpler sentences through coordination, subordination, quantification, and negation.

Coordination by ‘and’ is possible between sentences and between phrases of the same syntactic type.

Coordination by ‘or’ is possible between sentences, verb phrases, and relative clauses.

Coordination by ‘and’ and ‘or’ is governed by the standard binding order of logic, i.e. ‘and’ binds stronger than ‘or’. Commas can be used to override the standard binding order. There are three constructs of subordination: if-then-sentences, modality, and sentence subordination. With the help of if-then-sentences we can specify conditional situations.

Sentence subordination means that a complete sentence is used as an object.

Sentences can be existentially or universally quantified. Existential quantification is typically expressed by indefinite determiners (‘a patient’, ‘3 seizures’, ‘some blood’), universal quantification is typically expressed by the occurrence of ‘every’.

(e) Queries

ACE supports two forms of queries: yes/no-queries and wh-queries. Yes/no- queries ask for the existence or non-existence of a specified situation.

With the help of wh-queries, i.e. queries with query words, we can interrogate a text for details of the specified situation.

From ACE to OWL

The conversion from an **ACE** text to its corresponding **DRS** makes use of a small number of predicates, most importantly *object* derived from nouns and *predicate* derived from verbs [FKK08b].

The predicates share information by means of *discourse referents* (denoted by capital letters) and are further grouped by embedded **DRS** boxes, that represent

- implication (derived from *if...then...* or *every*),
- negation (derived from various forms of English negation),
- disjunction (derived from or),
- conjunction derived from relative clauses, explicit *and*, or the *sentence end* symbol is represented by the co-occurrence in the same **DRS**-box.

DRSs use a syntactic variant of the language of standard first-order logic extended by some non-standard structures for modality, sentence subordination, and negation as failure [FKS06, FKK08b]. The mapping to OWL does not modify the existing **DRS** construction algorithm but only the interpretation of the **DRS**. It considers everything in the toplevel **DRS** to denote individuals or relations between them.

Individuals are introduced by nouns, so that *propernames* map to individuals with type `owl:Thing` and common nouns to an anonymous individual with the type derived from the corresponding noun (e.g. class `Man`).

Properties are derived from transitive verbs. A special meaning is assigned to the copula ‘be’ which introduces an identity between individuals.

An embedded implication-box introduces a *subClassOf* relation between classes: the head of the implication maps to the subclass description, the body to its superclass description.

Transitive verbs introduce a property restriction with *someValuesFrom* a class denoted by the object of the verb, and the copula introduces a class restriction.

Negation and disjunction boxes in the implication-box introduce *complementOf* and *unionOf*, respectively. Any embedding of them is allowed.

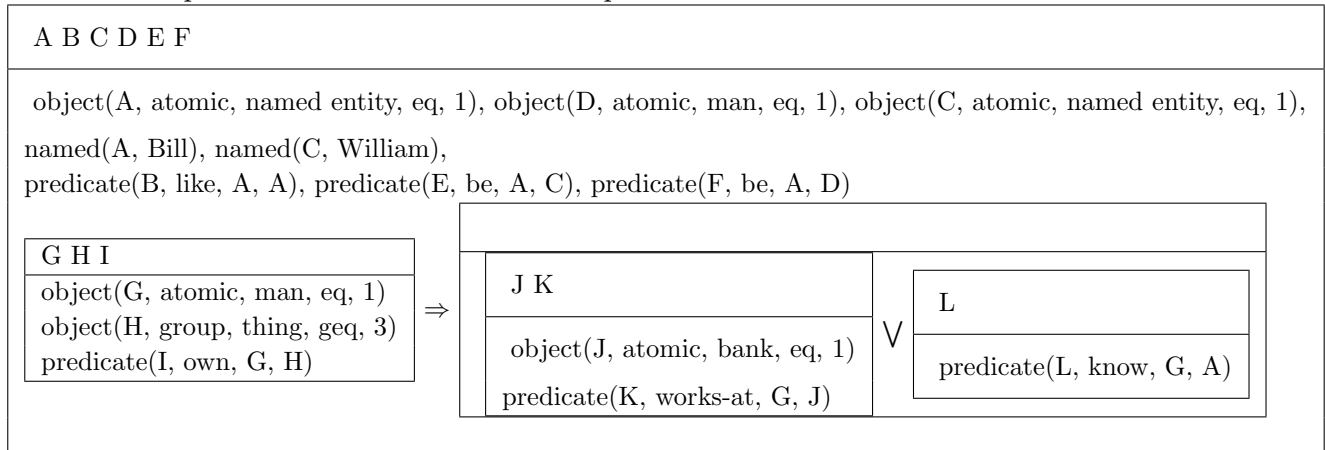
The plural form of a word which can be modified by a number allows to define cardinality restrictions. We illustrate with an example, for clarity and brevity, taken with permission from [KF06b].

The referred example and explanations was only updated to reflect the extraordinary evolution that **ACE** benefited during these 8 years that have gone by.

We use explicitly here the text transposed from the paper [Kal08] that is complex enough to illustrate

the possibilities as well as the limitations of the **DRS** and it's correspondent **DL** representation using **ACE**.

Using the artificial text "*Bill who is a man likes himself. Bill is William. Every man who owns at least 3 things works-at a bank or knows Bill.*" to illustrate concisely several features of **OWL** as expressed in **ACE** the **DRS** can be pictured as:



Thus the DRS of the figure has the following meaning (in Description Logics notation):

$bill \in \top, m1 \in Man, william \in \top,$ $bill = m1, bill = william,$ $likes(bill, bill)$ $Man \sqcap owns \geq 3 \sqsubseteq \exists \text{ works-at Bank } \sqcup \exists knows\{bill\}$

ACE can also describe **OWL** properties (super, inverse and transitivity) [KF06a].

Not all syntactic variants in the **OWL** specification are targeted by the mapping.

disjointWith or **equivalentProperty** cannot be directly expressed in **ACE** but their semantically equivalent constructs can be generated.

Constraining ambiguities

To constrain the ambiguity of full English **ACE** employs three simple means [FKK08a]:

- some ambiguous constructs are not part of the language; unambiguous alternatives are available in their place
- all remaining ambiguous constructs are interpreted deterministically on the basis of a small number of interpretation rules
- users can either accept the assigned interpretation, or they must rephrase the input to obtain another one

In the case of a system that is not offering an interactive interface the solution is to enforce by tutoring the users on the **CNL** usage.

Anaphoric references

In order to explain how anaphora are handled in **ACE**, examples taken from [DCFKK09] are transposed here. Usually an **ACE** text consists of more than one sentence. During the processing of the **ACE** text, all anaphoric references are replaced by the most recent and most specific accessible noun phrase that agrees in gender and number [FKK08b]. What does "most recent and most specific" mean? Given the sentence

A customer enters a red card and a blue card.

then

The card is correct.

refers to the second card, which is the textually closest noun phrase that matches the anaphora 'the card', while

The red card is correct.

refers to the first card that is the textually closest noun phrase that matches the anaphora 'the red card'.

What does "accessible" mean? Like in full English, noun phrases introduced in if-then-sentences, universally quantified sentences, negations, modality, and subordinated sentences cannot be referenced anaphorically in subsequent sentences. Thus for each of the sentences

If a customer owns a card then he enters it. A customer does not enter a card.

we cannot refer to 'a card' with

The card is correct.

Anaphoric references are also possible via personal pronouns

A customer enters his own card and its code. If it is valid then an automated
teller accepts the card.

or via variables

A customer X enters X's card Y and Y's code Z. If Z is valid then an automated
teller accepts Y.

Note that proper names always denote the same object.

ACE to OWL limitations

OWL ACE introduces a number of restrictions: there is no support for intransitive adjectives and some forms of plurals [Sch10]. Furthermore, there are restrictions to the DRS structure which are more difficult to explain to the average user, e.g. disjunction is not allowed to occur at the toplevel DRS ("John sees Mary or John sees Bill."). A further restriction could require the predicates in the implication-box to share one common discourse referent as the subject argument, and not to share the object arguments. This would allow us to exclude sentences like "If a man sees a mouse then a woman does not see the mouse." which does not seem to map nicely to an ontology language but instead to a rule language.

Then again, this restriction is too strong as it would exclude property expressions ("Everybody who loves somebody likes him/her.") and a way to express `allValuesFrom` ("Everything that a herbivore eats is a plant.").

ACE has been subject to very strong evolutions that currently allow the development and installation of an extended structure over the original pre-defined ACE. Being open source, it is possible to develop a tuned lexicon (a `lex` file), morphology, grammar (several grammars have been developed recently in some European research projects), and semantic relations that can form a specific Controlled Natural Language for any domain.

We will use these possibilities to create Clinical Controlled Language (CCL) for Cardio-vascular healthcare.

2.10 Tools for specialized NLP

To cover the explanation of the pipeline proposed for automatic acquisition from texts, in chapter 5, the used tools and methodologies have to be presented. Both the Attempto Controlled English (ACE) set of Natural Language Processing (NLP) tools and the bag of Apache ecosystem of tools for NLP that are used in this work are reviewed.

Attempto Controlled English (ACE) is a precisely defined subset of English that can automatically and unambiguously be translated into First Order Logic (FOL). It is thus both human and machine understandable. We refined the ACE methodology by defining a controlled language to apply to the clinical practice setting and named it Clinical Controlled Language (CCL). Teaching the construction and interpretation rules of Clinical Controlled Language (CCL) to a domain specialist takes about two days [FKK08a].

CCL appears perfectly natural, but it is in fact a formal language. CCL texts are computer-processable and can be unambiguously translated into Discourse Representation Structures (DRSs). DRS derived from CCL texts have been translated into various other languages, for instance into FOL and for stable model semantics. The *Attempto* project [Att07] developed a bidirectional translation of ACE into and from Web Ontology Language v.2 (OWL2). This tool can be used whether locally, via the Protégé plugin described in section 2.11.2, via a Web Service (WS) or through a Web Interface. We have been using it extensively through its various incarnations in our knowledge modeling efforts.

The generic NLP techniques and tools introduced in the state-of-the-art chapter in section 2.9.1 are

refined and developed in our work using the Apache Software Foundation ([Apache](#)) ecosystem of tools for NLP tasks. These tools allow for self-contained, open sourced, far reach systems to be developed in an integrated form. We use the [Apache OpenNLP](#) machine learning toolkit and more specifically the [Apache Tika](#) for content analysis and processing of the source texts presented in section 3.1.3 and [Apache UIMA](#) for all the orchestrations that lead from the basic NLP steps mentioned in section 5.2.2 to the expanded semantic annotation and even the KB assertion creation.

These tools are all Java based and so they fit naturally in the development of our proposed system infra-structure presented in section 7.1. After the NLP phases also a Java based tool, [Apache Jena](#), is used to manipulate the underlying KB through the [OWL API](#). The referred tools were used both in the preliminary development stages in an interactive way as proof of concept but were orchestrated posteriorly to implement the Automatic Ontology Learning ([AOL](#)) described in section 5.2.

2.10.1 Apache OpenNLP

The [Apache OpenNLP](#) library [[Apa14b](#)] is a machine learning based toolkit for the processing of natural language text. It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution. These tasks are usually required to build more advanced text processing services.

Contains a complete set of highly configurable trainable tools that may be invoked through Command Line Interface as its usage invocation renders evident:

```
$opennlp
```

```
OpenNLP 1.5.3. Usage: opennlp TOOL
```

```
where TOOL is one of:
```

Dccat	learnable document categorizer
DccatTrainer	trainer for the learnable document categorizer
DccatConverter	converts leipzig data format to native OpenNLP format
DictionaryBuilder	builds a new dictionary
SimpleTokenizer	character class tokenizer
TokenizerME	learnable tokenizer
TokenizerTrainer	trainer for the learnable tokenizer
TokenizerMEEvaluator	evaluator for the learnable tokenizer
TokenizerCrossValidator	K-fold cross validator for the learnable tokenizer
TokenizerConverter	converts foreign data formats (namefinder,ad,conllx,parse,pos) to native OpenNLP format
DictionaryDetokenizer	
SentenceDetector	learnable sentence detector
SentenceDetectorTrainer	trainer for the learnable sentence detector
SentenceDetectorEvaluator	evaluator for the learnable sentence detector
SentenceDetectorCrossValidator	K-fold cross validator for the learnable sentence detector
SentenceDetectorConverter	converts foreign data formats (namefinder,ad,conllx,parse,pos) to native OpenNLP format
TokenNameFinder	learnable name finder
TokenNameFinderTrainer	trainer for the learnable name finder
TokenNameFinderEvaluator	Measures the performance of the NameFinder model with the reference data
TokenNameFinderCrossValidator	K-fold cross validator for the learnable Name Finder
TokenNameFinderConverter	converts foreign data formats (bionlp2004,conll103,conll102,ad,muc6) to native OpenNLP format
CensusDictionaryCreator	Converts 1990 US Census names into a dictionary
POSTagger	learnable part of speech tagger
POSTaggerTrainer	trains a model for the part-of-speech tagger
POSTaggerEvaluator	Measures the performance of the POS tagger model with the reference data
POSTaggerCrossValidator	K-fold cross validator for the learnable POS tagger
POSTaggerConverter	converts foreign data formats (ad,conllx,parse) to native OpenNLP format
ChunkerME	learnable chunker
ChunkerTrainerME	trainer for the learnable chunker

ChunkerEvaluator	Measures the performance of the Chunker model with the reference data
ChunkerCrossValidator	K-fold cross validator for the chunker
ChunkerConverter	converts ad data format to native OpenNLP format
Parser	performs full syntactic parsing
ParserTrainer	trains the learnable parser
ParserConverter	converts frenchtreebank data format to native OpenNLP format
BuildModelUpdater	trains and updates the build model in a parser model
CheckModelUpdater	trains and updates the check model in a parser model
TaggerModelReplacer	replaces the tagger model in a parser model
Coreferencer	learnable noun phrase coreferencer
CoreferencerTrainer	
CoreferenceConverter	converts muc6full data format to native OpenNLP format

All tools print help when invoked with help parameter
 Example: `opennlp SimpleTokenizer help`

Detailing all the, or any of them for what matters, tools available to us in the **Apache** OpenNLP library is completely out of line in this document. They are very well documented either in the Apache Software Foundation project home page ³ or in the downloaded toolset as well. The wide scope and possibilities are, however, rendered evident by just grasping the comprehensiveness of each tool. Every tool usage may be done by scripting or programmed in a Java environment. They form the basis of most of our Cardiovascular oriented **NLP** tasks.

A complete formal evaluation of individual components usage is planned as outlined in section 8.3. Comparing our fine tuned specialty driven options against what can be considered the higher end set of generic clinical **NLP** tools, the **cTAKES** project, is an ambitious target but is very easily achieved at this stage of work.

2.10.2 Apache Tika

The **Apache** Tika [MZ11] toolkit detects and extracts metadata and structured text content from various documents using existing parser libraries. We use the underlying **Apache PDFBox** parser to manipulate our texts originally in PDF format. In our work we use the metadata juggling facilities of Tika to get rid of the decorating elements of the **SOAP** reports and keep just those that are interesting for our posterior annotation.

We get rid of all the Portable Document Format (**PDF**) structure code. Discard all the elements meant for human consumption like coloring, decorating frames, indentation or the document structure information for readability like page numbering.

To produce the listing in page 71, Tika permits interactive, command line or programmed usage to manipulate our source metadata. We use it progressively in our works to obtain the rendered results.

2.10.3 Apache UIMA

UIMA [SKSBC08, Sou14] is a component architecture and software framework implementation for the analysis of unstructured content.

The Unstructured Information Management Architecture (**UIMA**), is an open-platform middleware for dealing with unstructured information (text, speech, audio, video data), originally launched by

³<http://opennlp.apache.org/documentation.html>

IBM. In the meantime, the Apache Software Foundation has established an incubator project for developing **UIMA**-based software ⁴. In addition, the Organization for the Advancement of Structured Information Standards (**OASIS**) has installed a Technical Committee to standardize the **UIMA** specification. The motivation to develop such a framework was to build a common platform for unstructured analytics, to foster reuse of analysis components and to reduce duplication of analysis development. The pluggable architecture of **UIMA** allows to easily plug-in your own analysis components and combine them together with others. A full analysis task of a solution using unstructured analytics like search or health intelligence applications is often, like in the present case, not a monolithic thing but a multi-stage process where different modules need to build on each other to get a powerful analysis chain. In some cases also annotators from different specialized vendors may need to work together to produce the results needed. The **UIMA** application interested in such results does not need to know the details of how annotators work together to create the results. The **UIMA** framework take care of the integration and orchestration of multiple annotators. So the major goal of **UIMA** is to transform unstructured information to structured information by

- orchestrating analysis engines,
- detect entities or relations,
- build the bridge between the unstructured and the structured world.

The four main **UIMA** services are acquisition, unstructured information analysis, structured information access, and component discovery. Several repositories of **UIMA** components have been sprouting dedicated to different objectives in text structuring like the BioNLP **UIMA** Component Repository [Sou14] that provides **UIMA** wrappers for novel and well-known 3rd-party **NLP** tools used in biomedical text processing, such as tokenizers, parsers, named entity taggers like those devoted to pharmacology software that extracts drug information from Medline abstracts ⁵. Additionally, free analytic tools that can work with **UIMA** include those from the General Architecture for Text Engineering (**GATE**) ⁶ and OpenNLP ⁷. In our project we evaluated the use of **UIMA** as a tool for semantic annotation under the rules defined with the aid of RUTA scripting.

UIMA Ruta Workbench

Apache UIMA Ruta [KTB⁺14, Apa14a] is a rule-based script language supported by Eclipse-based tools. The language is designed to enable rapid development of text processing applications within **UIMA**.

A special focus lies on the intuitive and flexible domain specific language for defining patterns of annotations. Writing rules for information extraction or other text processing applications can be a tedious process. The Eclipse-based tool for **UIMA Ruta**, called the **Apache UIMA Ruta Workbench**,

⁴<http://incubator.apache.org/uima/>

⁵<http://bionlp-uima.sourceforge.net/>

⁶<http://gate.ac.uk/>

⁷<http://opennlp.sourceforge.net/>

was created to support the user and to facilitate every step when writing UIMA Ruta rules. Both the Ruta rule language and the UIMA Ruta Workbench integrate smoothly with Apache UIMA. In our project seminal Ruta rules were originally handcrafted for Cardiovascular domain using the workbench.

2.10.4 Apache cTAKES

Apache clinical Text Analysis and Knowledge Extraction System (cTAKES) [SMO⁺10, SKSBC08, GRDS⁺11] is an open-source natural language processing system for information extraction from electronic medical record clinical free-text. It is a top-level Apache Software Foundation project (as of March 22, 2013).

It processes clinical notes, identifying types of clinical named entities from various dictionaries including the Unified Medical Language System (UMLS) - medications, diseases/disorders, signs/symptoms, anatomical sites and procedures. Each named entity has attributes for the text span, the ontology mapping code, subject (patient, family member, etc.) and context (negated/not negated, conditional, generic, degree of certainty). Some of the attributes are expressed as relations, for example the location of a clinical condition (locationOf relation) or the severity of a clinical condition (degreeOf relation).

Apache cTAKES was built using the Apache UIMA engineering framework and Apache OpenNLP natural language processing introduced in the previous sections. Its components are specifically trained for the clinical domain out of diverse manually annotated datasets, and create rich linguistic and semantic annotations that can be utilized by clinical decision support systems and clinical research.

Apache cTAKES employs a number of rule-based and machine learning methods. Components include:

1. Sentence boundary detection
2. Tokenization (rule-based)
3. Morphologic normalization
4. POS tagging
5. Shallow parsing
6. Named Entity Recognition
 - Dictionary mapping
 - Semantic typing is based on these UMLS semantic types: diseases/disorders, signs/symptoms, anatomical sites, procedures, medications
7. Assertion module
8. Dependency parser
9. Constituency parser
10. Semantic Role Labeler

11. Coreference resolver
12. Relation extractor
13. Drug Profile module
14. Smoking status classifier

It is intended to be modular and expandable at the information model and method level. The **cTAKES** community is committed to best practices and R&D (research and development) by using cutting edge technologies and novel research. The idea is to quickly translate the best performing methods into **cTAKES** code.

We intend to use **Apache cTAKES** only as a comparison platform for the Domain Expert evaluation to be done against what can be considered the standard to be achieved in terms of comparable metrics in Clinical NLP. We are engaging in this evaluation process because we aim to achieve in our specialty oriented (to Cardiovascular) environment comparable results to what was recently open sourced by the **Apache cTAKES** community [ALF⁺13]. All the evaluation plans that are to be engaged very soon are detailed in section 7.3.1 in the results discussion chapter 7.

2.11 Tools for Ontology manipulation

During the various development stages in the project several tools are used.

- For interactive Ontology manipulation
The main tool used extensively in this phase is Protégé 4 Ontology Development Tool introduced in section 2.11.1.
- For programmatic Ontology access
We use mainly Java based tools to access and perform various manipulating tasks in our ontology. We use the **Apache Jena** free and open source framework presented ahead in section 2.11.4 to manipulate the ontologies usually by invoking the **OWL API** library introduced in section 2.11.3.

2.11.1 Protégé

Protégé 4 Ontology Development Tool (**Protégé**)⁸ is an open platform for ontology modeling and knowledge acquisition developed at Stanford Medical Informatics. *Protégé is a free, open-source platform that provides a growing user community with a suite of tools to construct domain models and knowledge-based applications with ontologies.* [Sta14].

Although the development of **Protégé** has historically been mainly driven by biomedical applications, the system is domain-independent and has been successfully used for many other application areas as well.

Developers can integrate the output of **Protégé** with rule systems or other problem solvers to construct

⁸Protégé is supported by grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health.

a wide range of intelligent systems. **Protégé** is based on Java, is extensible, and provides a plug-and-play environment that makes it a flexible base for rapid prototyping and application development. **Protégé** plug-in architecture can be adapted to build both simple and complex ontology-based applications.

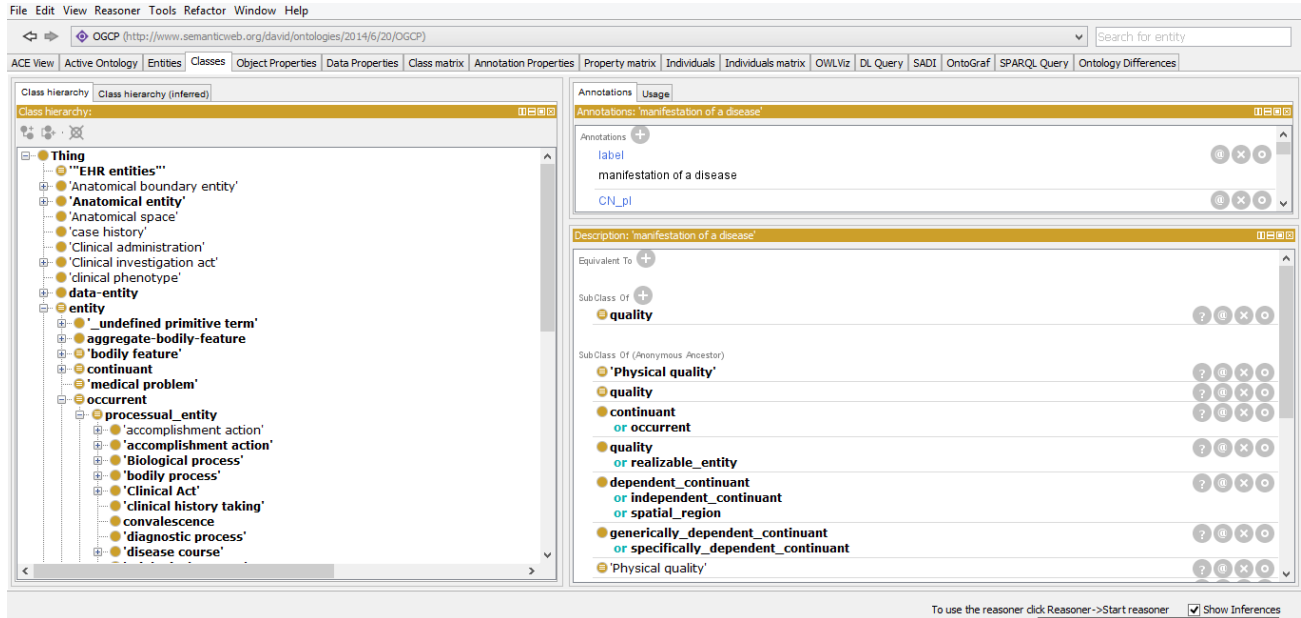


Figure 2.2: **Protégé**.

The most important plug-in that we extensively use in the current project for Controlled Natural Language (CNL) tasks is described next.

2.11.2 Protégé Ace View Plugin

ACE View [Kal08] is an ontology and rule editor that uses Attempto Controlled English (ACE) in order to create, view, edit and query OWL ontologies. Domain Expert can create OWL/Semantic Web Rules Language (SWRL) knowledge bases by working solely in ACE. It can edit OWL Knowledge Bases (i.e. add, remove, modify OWL axioms and Semantic Web Rules Language (SWRL) rules) by switching between the ACE view and the traditional "Protégé views" (forms and description logic formulas). In many cases Domain Expert don't have to know the details of OWL and SWRL — the ACE view hides them. It is possible to open existing OWL ontologies, view and edit them as ACE texts. Using this feature we can verbalize all the OGCP, understand clearly all the clinical information induced by it and even edit it using CCL.

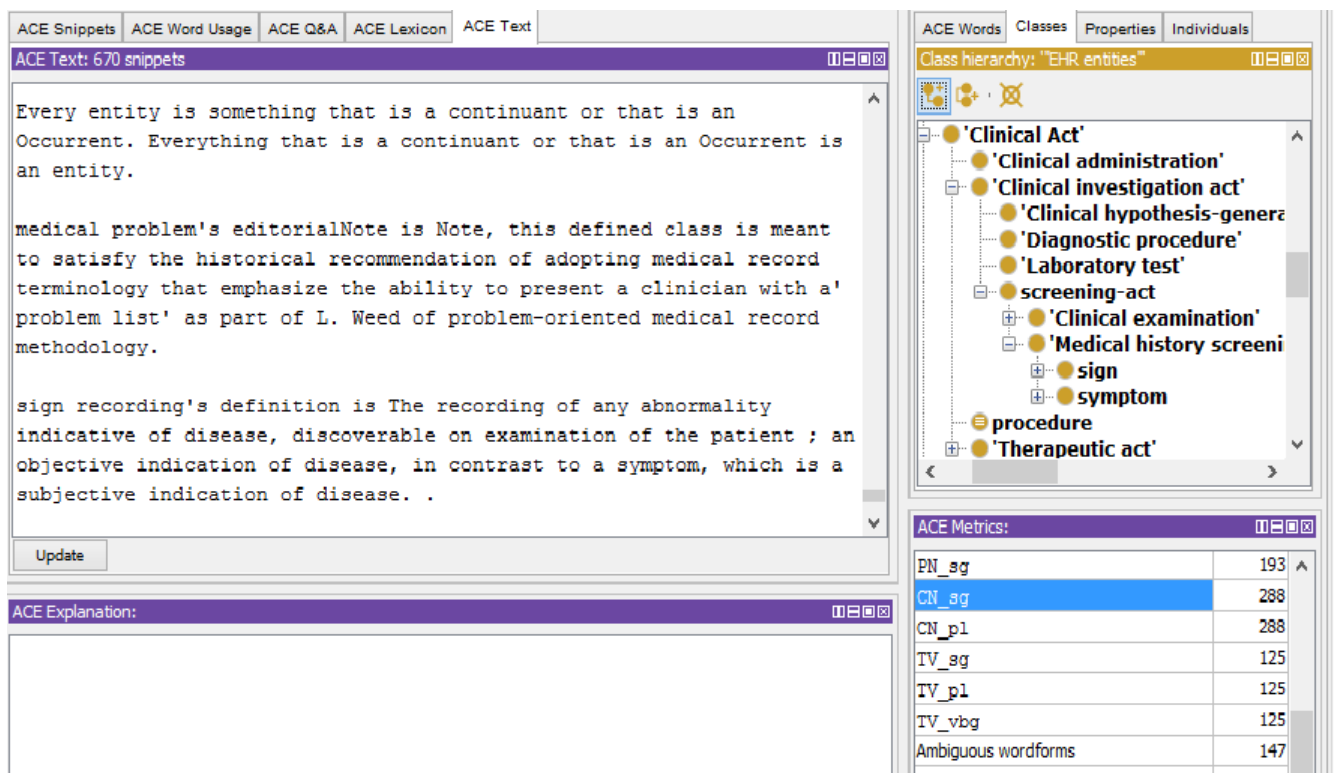


Figure 2.3: Protégé ACE View Plug-in.

The snippet pictured in the previous figure shows clearly (using CCL) the definition of *ogcp:sign* induced and translated from OWL.

2.11.3 Protégé OWL API

The Protégé-OWL API [KFNM04] is an open-source Java library for the Web Ontology Language (OWL) and RDF(S). The Application Programming Interface (API) provides classes and methods to load and save OWL files, to query and manipulate OWL data models, and to perform reasoning based on Description Logic engines. Furthermore, the API is optimized for the implementation of graphical user interfaces. The API is designed to be used in two contexts:

- For the development of components that are executed inside of the **Protégé**-OWL editor's user interface
- For the development of stand-alone applications

Protégé is a flexible, configurable platform for the development of arbitrary model-driven applications and components. **Protégé** has an open architecture that allows programmers to integrate plug-ins, which can appear as separate tabs, specific user interface components (widgets), or perform any other task on the current model. The **Protégé**-OWL editor provides many editing and browsing facilities for OWL models, and therefore can serve as an attractive starting point for rapid application development. The **Protégé**-OWL editor is provided with the standard installation of **Protégé**. Developers can initially wrap their components into a **Protégé** tab widget and later extract them to distribute them as part of a stand-alone application. The **OWL** API enables to perform the following:

1. Working with **OWL** Models
2. Working with Names, Namespace prefixes, and Uniform Resource Identifiers (**URIs**)
3. Understanding the Model Interfaces
4. Creating Named Classes and Individuals
5. Using Datatype Properties and Datatype Values
6. Using Object Properties to Build Relationships between Resources
7. Working with References to External/Untyped Resources
8. Property Domains

That is to manipulate all the possible levels of detail the **OWL** ontologies. When not working inside **Protégé** controlled environment we use **Apache Jena** to programmatically access the ontology representations.

2.11.4 Apache Jena

Apache Jena (**Jena**), a Java **RDF** API and toolkit is a Java framework to construct Semantic Web Applications [Jen07]. It provides a programmatic environment for **RDF**, Resource Description Framework Schema (**RDFS**), **OWL**, **SPARQL**, Gleaning Resource Descriptions from Dialects of Languages (**GRDDL**) and includes a rule-based inference engine. The framework is composed of different Application Programming Interfaces (**APIs**) interacting together to process **RDF** data. **Jena** is a toolkit, using the Java programming language. As explained above we mostly use **Jena** by writing Java programs although there are a few command-line tools to perform some key tasks using **Jena**.

The **Jena** Ontology **API** is language-neutral: the Java class names are not specific to the underlying language. For example, the **OntClass** Java class can represent an **OWL** class. To manage the differences between the various representations, each of the ontology languages has a profile, which lists the permitted constructs and the names of the classes and properties. An ontology model is an extension of the **Jena** **RDF** model, providing extra capabilities for handling ontologies. We use the model to access an abstraction of **OGCP** for very precise manipulations.

Chapter 3

Clinical Knowledge (CK)

We discuss here the issues regarding Knowledge Representation and Knowledge Acquisition in our particular Clinical Practice sub-domain. It's also introduced an overview of the complex pipeline that extracts from semi-structured text into the expressive ontology instance creation.

The knowledge acquisition also encompasses the interrogation process because it is also possible to enrich the ontology with every inferred axiom generated from any interrogation. This step is only detailed in the *Discourse Based Enhancement* section 6.3.

3.1 Knowledge Representation

Back in 2007, Schulz and Stenzhorn [SS07] enunciated ten principles that should guide Clinical Ontologies creation. In this reference work some attention is paid to ontologies as terminological systems or clinical archetypes and we recognize that those principles, mainly I, II, VII and X were thoroughly applied in Standard Nomenclature of Medicine - Clinical Terms (SNOMED-CT) after 2007. Other of the enunciated principles are, however, more important to our work as being a foundation for what a well founded basis for Clinical Practice Knowledge Representation has to be. Namely principle III. "*Ontologies represent universal truths*", where the relation to Formal Ontology (FO) is justified arguing that only *what is assumed to be universally true* should be represented. Formal ontologies make the semantics of terms and relations explicit such that automated reasoning can be used to verify the consistency of knowledge. This has an impressive importance in the reasoning abilities of our ontological framework proposed ahead in chapter 4. These are rendered evident by the examples in the above cited work like "*all instances of the type carotid artery are instances of the type artery*". Or, "*for all instances of the type common carotid there is some instance of the type aorta which it is connected to*". Later in principle V. "*Ontologies organize individual entities – not concepts*", the rationale for a concrete practice ontology is established. This realization influenced or Ontology for General Clinical Practice (OGCP) ontology structure as we found that a realistic approach had to be built and so the principles of Ontological Realism (OR) had to be poured in as we argue in 3.1.1.

Subsequently we found the need to establish some ground methodology to articulate all the ontological pyramid that our representation was to be built upon and it was an addressed issue by Ceusters and Smith back in 2010 [CS10].

Based in the proposed unified methodology the ontology scaffolding is built and the ontological relations extracted from our text sources, and itemized in sub-section 3.1.2, are then surfaced and collected as suggested in [HDO⁺11a] and maintained computationally tractable restricting the \mathcal{DL} representation to $\mathcal{EL}++$ according to [HDO⁺11b, BBL08].

3.1.1 Ontological Realism (OR)

When trying to clear some misconceptions that were wandering in the Knowledge Representation (KR) community numerous efforts had to be developed by ontologists interested in the application to biomedical sciences. In the aftermath of the Gene Ontology (GO) development some guidance was needed to further develop, in a well founded manner, the extensions of the work achieved. Based in foundational philosophical considerations mainly summarized in [SG04] where an account of what formal ontological relations are, an important theoretical corpus was built to define what can be considered the root for current biomedical Knowledge Base (KB) formation. Starting from the realist perspectivalism [Gre03, SB01].

"Ontological Realism is a methodology to avoid mistakes that cannot be detected by logical formalisms alone" [CSKD04].

We still want to highlight the reasoning power that formal ontological relations provide to a carefully crafted ontology given the higher semantic level that these relations comprise [SC10]. The formalization of *Ontological Relations* has been advocated for many years and it succeeded in the development of *"relations that obtain between entities in reality, independently of our ways of gaining knowledge about such entities"* [SCK⁺05].

The Ontological Realism mentors dive into the different equally veridical perspectives to sort out a true orientation that is the one to be rendered by our realistic ontology. A breakthrough was the enlightenment that resulted from the discussion between two major explorers in this area that originated from the developments proposed by Smith in [Smi06] where the duality between *Concepts* and *Universals* was raised trying to enhance the proposal in Cimino's Desiderata [Cim98]. In his refutation [Cim06], Cimino argues that both *Concepts* and *Universals* have to coexist for their different applicable ends. We find that only 4 years later a consensus for both terminologies and ontologies was formally identified and presented in [CS10] where the OR is flagged as the solution to be pursued.

3.1.2 Ontological relations for clinical practice

Relations specifically associated to Biomedicine or Clinical Practice (CP) retain knowledge associated with the clinical domain. Apart from relations such as *is-a* and *part-of*, biomedical ontologies also contain domain specific relations such as *has-location*, *has-manifestation* or *clinically-associated-with*. These relations are, however, nothing but that. That is, relations and this turns them semantically transparent, no specific domain knowledge differentiates these relations from any other given the appropriate definition (cardinality, direction, object, datatype and annotation properties) which for proper computability purposes can be achieved with the adequate OWL Description Language (OWL DL) representation.

Being standardized in 2009 the language of choice, and consequently the associated tools, is Web Ontology Language v.2 (OWL2). OWL2 addresses key expressive and computational limitations of OWL. By adding new constructs to the language, OWL2 more directly supports medical applications. For example, so called “role chains” allow ontologists to express the connection between spatial relations and part-whole relations, e.g., if a fracture is located on a bone which is part of a leg, that fracture is a fracture of that leg.

The expressibility in Biomedical relations has such a standing point in the reasoning possibilities that specific ontologies were developed to represent and enforce their usage: the Relations Ontology (RO) extensively studied in section 4.2.4 and the Top-Domain Ontology for the Life Sciences (BioTOP) also discussed in section 4.2.4. Currently several tools exist for bi-directional converting which can automatically transform Open Biological and Biomedical Ontologies (OBO) ¹ ontologies into the OWL-based format used by the Semantic Web OWL DL [Ber14].

3.1.3 Clinical text available sources

Most of the Information Extraction (IE) research and tools is directed essentially towards free form clinical text like that frequently present in the following use-cases:

- Patient cohort identification
- Clinical decision support
- Health care quality research
- Personalized medicine
- BioSurveillance
- Drug development
- Meaningful Use
- Text Summarization

We are intending however to develop a deep ontological representation of Clinical Practice (CP) reality and studied varied text reports including:

- Demographics
- Clinical notes
- Discharge notes
- Problem lists
- Adverse drug effect lists
- Exams reports

¹<http://www.obofoundry.org>

- Patient histories
- Referral documents
- Prescriptions

However these only represent particular views of a patient or group anamnesis.

Going from clinical episodes free text, that is usually presented in a human friendly format, to one adequate for computer processing involves a fair amount of **IE** to handle problematic situations because:

- Reports aggregate information from different clinical episodes that are not uniquely identified nor even individually dated
- The episode clinician is only identified by his/her name if any identification is made at all
- The information conveyed in free text is intended only meant to be understandable by fellow practitioners or even by the clinician him/herself making use of pragmatic jargon normally plagued with acronyms and nicknames abundant in their specific community
- Text is profoundly intermixed with decorative elements for better legibility, normally in **PDF** or HyperText Markup Language (**HTML**) files
- The time spanning and/or snapping of the processes depicted in natural language is difficult to represent formally
- The clinicians natural language is other than English, without concepts defined in foundational thesaurus like **SNOMED-CT** or ontological references like Foundational Model of Anatomy (**FMA**) for instance, that don't even exist in that particular language

So we searched for some document that could provide a larger picture of the provided healthcare, and available as sources for us, and these are the **SOAP notes**.

The SOAP note was first conceived by Dr. *Lawrence Weed*, MD in the 1970's, under the acronym Problem Oriented Medical Record (**POMR**). At the time, there was not an objective method of documentation, which lead to physicians making unscientific decisions about patient treatment. SOAP notes gave physicians rigor, structure, and a way for practices to communicate with each other. In the early 1970's, the adopters of SOAP notes were able to retrieve all patient records for a given medical problem. Before Electronic Medical Record (**EMR**) software, providers had trouble accessing needed charts. Before standardized **SOAP** notes, providers communicated with each other in unstructured formats, leaving patient care up to great chance. A SOAP note is a documentation method employed by health care providers to create a patient's chart.

There are four parts of a SOAP note: **S**ubjective, **O**bjective, **A**ssessment, and **P**lan.

- **Subjective**
Describes the patient's current condition in narrative form. This section usually includes the patient's chief complaint, or reason why they came to the physician. It's normally the place for the symptoms brought in by the patient and includes:

- Onset (when and mechanism of injury – if applicable)
 - Chronology (better or worse since onset, episodic, variable, constant, etc.)
 - Quality (sharp, dull, etc.)
 - Severity (usually a pain rating)
 - Modifying factors (what aggravates/reduces the complaint – activities, postures, drugs, etc.)
 - Additional symptoms (un/related or significant symptoms to the chief complaint)
 - Treatment (has the patient seen another provider for this symptom?)
- **Objective**
Documents objective, repeatable, and traceable facts about the patient’s status. Associated with signs including:
 - Vital signs
 - Findings from physical examinations, such as posture, bruising, and abnormalities
 - Results from laboratory
 - Measurements, such as age and weight of the patient
 - **Assessment**
The Physician’s medical diagnoses for the medical visit on the given date of a note written.
 - **Plan**
This describes what the health care provider will do to treat the patient – ordering labs, referrals, procedures performed, medications prescribed, etc.

This clinically oriented structuring is well known and widely accepted in the healthcare community and text sources based in it are generally available. Many Electronic Health Record (**EHR**) systems provide a software module to create them.

It’s an identified opportunity for our work if we develop a methodology to enrich our ontology from such an important source.

Cardiology and ICU in Portalegre district

In the Portalegre district in Portugal, the Unidade Local de Saúde do Norte Alentejano (ULSNA)² has as objectives the provision of primary and secondary health care to the population. ULSNA is a healthcare providing regional system that includes 2 hospitals (José Maria Grande in Portalegre and Santa Luzia in Elvas) and the primary care centers in all 15 district counties. A group of clinicians chosen by our trial investigator *Dr. Carlos Baeta* provided us some dozens of clinical reports de-identified according to safe-harbour principles, as reviewed in [MFS⁺10], from the Sistema de Apoio ao Médico (SAM) system in use both in the Primary Healthcare units and in the Hospitals. These clinicians are mainly cardiologists from the hospitals but also general medicine (primary care) physicians that normally use reports like those provided to communicate between them. We used the sample clinical data that is available for us like this de-identified sample:

HEALTH CENTER PONTE DE SOR MAIN OFFICE	Paciente 5689_SOAP	*XXXXXXXX*
Registo Clínico da Consulta	Birth Date XX-XX-XXXX (XX Years)	*XXXXXXXX*
	XXXXXXXXXX	
	XXXX XXXXXXXX	

SPEC.	12/07/2010 18:13	Dr.(a) Carlos Baeta
		Dr.(a) Carlos Baeta

S _{SOAP}	Back-external pain episodes: -Since 2 years. -Without anginal features. - With palpitations and facial flushing Eco - N Has maintained variable hypertension (140/90) Keep episodes of palpitations.
O _{AP}	
S _{AP}	Holter(27/05/10)-RS; 51 a 119: M-75; ESSV infrequent T3, T4, TSH - N; AVM and Catecholamines - N
P _{SOA}	Cordarone - 1 tablet per day Repeat Holter within 6 months for assessment of the need to Arrhythmology query
Comercial Name	Qt.
1 Amiodarona [Cordarone] , 200 mg, Comprimido, Blister - 60 unit(s)	1
Posol.: 1 tablet per day (6 per week)	
CARDIOLOGY	
HOLTER	1

Figure 3.1: SOAP report de-identified sample.

Structure, scope, adequacy

We are extracting from what can be called a semi-structured repository of clinical practice information, the personal SOAP framework reports. When applying the principles of well defined formal ontologies depicted in [SAR⁺07] and trying to avoid the errors mentioned in [CSKD04] we decided to proceed with a pragmatic approach to the representation of disease and diagnostic as illustrated in [SCS09]. There is a clear support for text divided by the 4 pre-defined subsections. For any particular encounter (actually for any Clinical Episode) the text for any of these may be collected in a suitable form for

²<http://www.ulsna.min-saude.pt/>

processing into the **OGCP** and the appropriate suggested slots in the OGCP framing are pinpointed in the following picture taken from the referred paper by Scheuermann et al.:

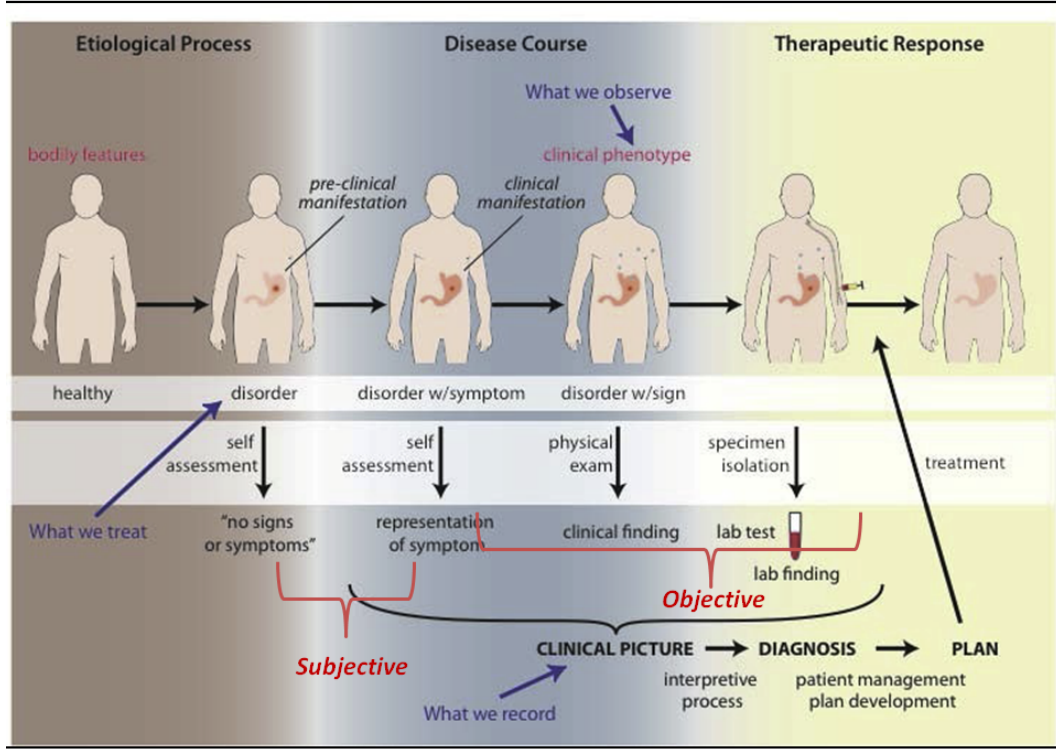


Figure 3.2: SOAP Points Insertion

We apply the developments made based in the Ontology Learning (OL) techniques from semi-structured text introduced in section 2.9 to extract the needed knowledge for our ontology enrichment. We benefit that our text sources are diverse from the system that originated them because we need to guarantee that no Personal Health Information (PHI) is possible to rebuild from our resulting KB. When a SOAP report is processed an internal non-reversible **patientID** is generated based in its ID. No trace back to the PHI can be done but internally to our KB we are able to rebuild in order to apply the needed co-reference between different cases in our case-based reasoning system as referred ahead in both the sections 5.2.2 and 5.2.4.

3.2 Knowledge Acquisition

Having discussed the numerous detailed representation issues in 3.1 we proceed now to present a discussion on a proposed recipe to get our text segments into an adequate knowledge framework, the **OGCP** ontology, that will be revealed in chapter 4. The ontology is to be enriched in a two phased process divided in:

1. Automatic Ontology Learning (AOL), Where a semi-automatic **ABox** learning process ornares **OGCP** from text-sources as illustrated in chapter 5 and,

2. Discourse Based Enhancement (DBE) Where the the ontology is enriched through Discourse Reasoning (DR) mainly using interactive Question Answering (QA) by the domain experts and asserting knowledge in the process illustrated in chapter 6.

In this section, only the general framework and process is presented but, it will be completely detailed in chapter 5 where all the automatic Knowledge Acquisition tasks are scrutinized.

Accepting as evidence the fact that most of the clinical information is maintained in text form, we shifted our focus back in 2012 from the extraction based in semantic models in the Electronic Health Record (EHR)s like the Health Level 7 (HL7) v3 Reference Information Model (RIM) [Hea] [MR13a] to the automated acquisition from clinical reports. This introduced us the problem of acquiring the knowledge necessary for learning ontologies known as the "*Knowledge Acquisition Bottleneck*". This challenging [WLB12] issue remains one of the main barriers for automated acquisition and we tried to circumvent it by using a progressive tutored learning approach.

The amount of Clinical data digitally preserved is colossal, ever increasing and numerous problems when retrieving from text have to be devised and solved as reviewed by Meystre et al. [MSKSH08] and Liu et al. [LHC11].

Most of the clinical data is in text form coming either from typing entry, transcription from dictation or from speech recognition applications. We will figure out a "picture of Healthcare provisioning" through clear identification of the meaning of the available data for reasoning purposes and not only by the capability of cataloging and codifying that huge amount of data.

Facts extracted from documents must refer to a common agreed upon meaning as expressed in some ontology to function as a knowledge enhancement tool. We try to depart from semi-structured text and use the semi-automated translation tasks to generate a controlled *domain specific vocabulary* on which further acquisition tasks build upon [MR12], minimizing ambiguity and redundancy for better reasoning capabilities. When trying to instantiate individuals (populate) in formal heavyweight expressive ontologies like the OGCP we do not normally intend to enrich the ontology but instead turn them from theoretical models of the domain into reasoning able knowledge bases.

3.2.1 Using the OGCP for Clinical Controlled Language building

Our pragmatics processes are based upon ACE tools presented in section 2.9.3. In order to fully understand the usage of the OGCP, the ontology not yet the Knowledge Base, as the foundation for our Cardiovascular model of clinical practice we introduce right now every concept of DRS forming capabilities that we explore to render the model of Clinical Controlled Language (CCL) needed for text interpretation fully explained in section 5.3. For a full description of all the state-of-the-art, reasons, issues and steps involved in a directed *Controlled English* infra-structure in any domain we have recurred to [DCFKK09].

The OWL→ACE mapping allows us to verbalize existing OWL ontologies as ACE texts.

We have recurred to the OWL→ACE mapping to verbalize the OGCP in order to render all the ontological structure, thus the Cardiovascular Clinical Practice Knowledge Representation, able to be immediately queried by Clinical Controlled Language interrogations.

This mapping is not just the reverse of the ACE→OWL as it also covers OWL axiom and expression

types that the **ACE**→**OWL** mapping does not generate. For example

PropertyDomain(write author)

is verbalized as

Everything that writes something is an author.

We first have to explain the ontological structure to be enriched for the assertions acquisition process be more easily understood, so we immediately proceed to it in the next chapter 4.

Later, for the Automatic Ontology Learning, the first tutoring phase will be extensively detailed in section 5.1 and the other recurring phase of Ontology Learning will be accurately presented in 5.2. Finally, for the full enrichment tasks to be covered, in section 6.3 we engage in the Discourse Based Enhancement (**DBE**) explanation .

Chapter 4

Ontology for General Clinical Practice proposal

In this chapter a proposal is made for an Ontology that adequately supports healthcare as a sub-domain of Biomedical science. All the foundational reasons that induced to the creation of Ontology for General Clinical Practice (**OGCP**) are reviewed and all the technological and philosophical justifications are given.

In the early stages of our work, we introduced the proposal of taking advantage of standardization of messaging in **EHR** to develop the tools to finally evolve into “evidence based harmonization” in ontology development meant mainly for clinical practice.

Taking into account the considerations introduced in [MSKSH08, SAMK05], and more recently illuminated by the developments in technology and tools as referred in [MSKSH08, YAM⁺12], the completeness and full coverage of International Standards Organization (ISO)/HL7 27931:2009 Standard [ISO09], that addresses syntactic interoperability in health information systems, grants solutions that do not fall short in particular fields of the different medical specialties. We directed our efforts in trying to find the adequate ontology to cover the broaden field that the referred standard covers. We adopt a Standard Nomenclature of Medicine - Clinical Terms (**SNOMED-CT**) subset as terminological source due to its full coverage, acceptance and applicability to our specific case.

4.1 **OGCP** Presuppositions

To accomplish a successful work the resulting ontologies have to attain the sort of user-friendliness, reliability, cost-effectiveness, and breadth of coverage that is necessary to ensure extensive usage as introduced by Smith and Brochhausen [SB10]. Several factors have to be judiciously handled using all the latest trends in technological and scientific development, among these are the proper selection of what ontologies have to be used for learning/enrichment and all the pragmatic aspects that may render broad usage of the resulting automatically produced knowledge. For all of these we suggested

what were the most promising, or already proved on the field, techniques and ontologies that could lead us to the above presented objectives.

Considering that the amount of Clinical data digitally preserved in **EHRs** is colossal and ever increasing, numerous problems have to be devised and solved as reviewed by Meystre et al. [MSKSH08] and Liu et al. [LHC11]. Most of the clinical data, however, is in text form coming either from typing entry, transcription from dictation or from speech recognition applications. Accurate coding is necessary for comparability, auditability and, last but not least important, accountability. We intend to figure out a “picture of Healthcare provisioning” through clear identification of the meaning of the available data and not only by the capability of cataloging and codifying that huge amount of data.

Ultimately none of the currently existing proposed ontology structures are appropriate for clinical practice knowledge representation and we introduced **OGCP**, a proposal of our own presented in the current chapter.

We explored thoroughly the Computer Based Patient Record Ontology (**CPR**) ontology [Ogb11] and found it to be an adequate framework for full breadth coverage of the clinical practice as suggested in the **ISO/HL7 27931** Standard. It was lacking, however, the ontological relations needed to enforce any model of disease. **CPR** could be seen as an extensible framework to be heavily structured further by any suitable modeling upper ontologies like those that the *OBO Foundry* mandates. With the **CPR** as an healthcare provisioning representation support we still had to enforce any existing model for medical science and we had the Ontology for General Medical Science (**OGMS**) complemented with the Disease Ontology (**DO**) to do so as a general practice model.

Computer Based Patient Record Ontology (**CPR**) derived instances can be viewed in **OGCP** as the *leaf nodes* of our ontology that models a particular Clinical Practice environment with its associated, modeled specialty specific diseases.

OGCP properties model with expressive ontological relations like `has_participant`, `patient_treated`

or `hypothesized_problem` for instance that can be viewed in the following image:

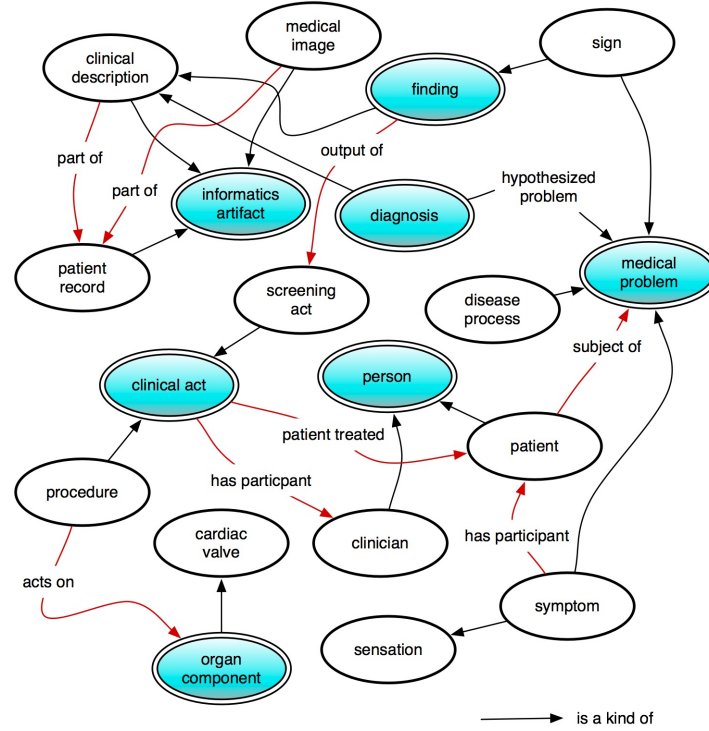


Figure 4.1: Leaf nodes of OGCP.

Finally for a specific specialty use-case, for our ontology to be show-cased, we incorporated the Cardiovascular Disease Ontology (CVDO) that models our sub-domain of interest. Of course our intention is to have OGCP as a generic framework for any clinical specialty that is modeled by some domain specific ontology and so we can consider CVDO as a movable, interchangeable part.

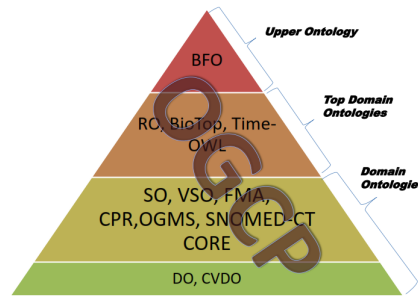


Figure 4.2: Ontological structure of Ontology for General Clinical Practice (OGCP).

4.1.1 Standard Nomenclature of Medicine - Clinical Terms (SNOMED-CT)

The main terminology source for our Ontology Learning system is the Standard Nomenclature of Medicine - Clinical Terms (SNOMED-CT) [TSZI10]. SNOMED-CT has been created by combining SNOMED RT and a computer-based nomenclature and classification known as Read Codes Version 3, which was created on behalf of the U.K. Department of Health and is a Crown copyright. It is

organized in a **Concept** hierarchy that comprises currently 401200 terms (classes), with a maximum depth of 28 levels.

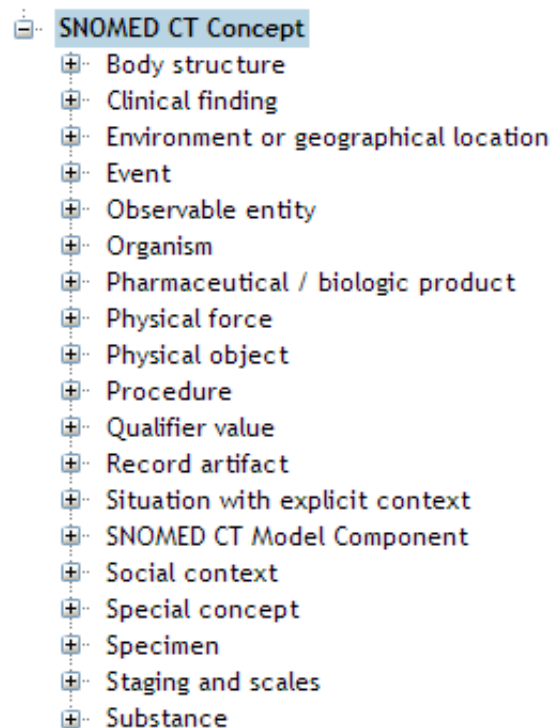


Figure 4.3: SNOMED-CT Concept Structure.

This ontology is made available via the Unified Medical Language System (UMLS). Users of all UMLS ontologies must abide by the terms of the UMLS license. For the current project an academic license was provided by the UMLS Terminology Services (UTS) [UTS14] and an APIkey was issued that we use for Web Services invocation explained in 5.2.3.

SNOMED-CT is the most comprehensive medical terminology. However, its use for intelligent services based on formal reasoning is questionable [HSV08]. Any evaluation would reveal several ontological and knowledge engineering errors which would prevent formal reasoning. The workarounds, however are fairly simple and we just make use of an adequate subset of SNOMED-CT that has a cleansed structure according to the semantically sound Clinical Concepts set induced by Unified Medical Language System (UMLS) CORE subset [DFMSA10a, PSD12, SG10].

Different reasoning techniques and classifiers have demonstrated the capability to handle the roughly 400000 clinical terms in acceptable times like *Snorocket* [LB10] but most of these impose some restrictions on the underlying structure of the terminological basis. We opt for not restricting the data to be classified in terms of the reasoning abilities but by selecting a content view instead [DFMSA10a]. The set of Clinical Concepts that our extended annotation process renders through the BioPortal is already ontology oriented using the UMLS CORE aggregated views applicable to Cardiovascular diseases and the coronary system according to [MBB⁺01].

Furthermore, using the UMLS CORE, only the SNOMED-CT Problem List Subset of frequently used concepts is needed and that restricts the terminology to a controlled problem list identified in the UMLS CORE project [FMS10].

4.1.2 UMLS CORE

Initiated in 2007 the **UMLS CORE** project [FMS10] made an enormous contribution to harmonize and render tractable terminological systems between large healthcare institutions. Using the "Problem List Approach" that covers *'a complete list of all the patient's problems, including both clearly established diagnoses and all other unexplained findings that are not yet clear manifestations of a specific diagnosis, such as abnormal physical findings or symptoms.'* and the authors asked 7 big healthcare institutions to submit their Problem List Terms (**PLT**) together with the actual frequency of usage in their clinical databases. A **UMLS** mapping was carried out in order to achieve two goals:

1. To study and characterize the **PLT** in terms of their size, pattern of usage, mappability to standard terminologies, and extent of overlap.
2. To identify a subset of concepts based on standard terminologies that occur with high frequency in problem list data to facilitate the standardization of **PLT**.

It was verified that the actual usage is concentrated heavily on relatively few terms, which makes the problem more tractable because the need is only to standardize in a relatively small proportion of terms to reap large benefits in data interoperability and reasoning ability. Those heavily used Concept Unique Identifiers (**CUI**)s are also the ones that are commonly shared. A very solid subset of **SNOMED-CT** list of 6776 concepts, (**CUI**)s, the **CORE** Problem List Subset of **SNOMED-CT**, which has been available for download by **UMLS** licensees since June 2009, was further identified on the basis of the **UMLS CORE** Subset. In the current incarnation of **SNOMED-CT CORE** (201311 Version **UMLS** 2013AB) it comprises a total number of 6179 concepts divided by:

Clinical finding:	5,342
Procedure:	566
Situation with explicit context:	210
Event:	61
Total:	6,179

We incorporated the **SNOMED-CT CORE** problem list into our **OGCP**.

4.1.3 Suggested representation as **CSI** tool

Software systems are semantically integrated if their sets of intended models are equivalent. In the area of decision support, the verification of an ontology allows us to make the claim that any inferences drawn by a reasoning engine using the ontology are actually entailed by the ontology's intended models [Gru11]. Leveraging the reasoning capabilities of **OGCP** will lead to use it as a model for Computer Semantic Interoperability given that no unreasonable inferences can be derived from such a solid framework and it provides a common understanding covering the full realistic healthcare sub-domain of science.

4.2 OGCP ontologies alignment

All the underlying ontologies that sustain the **OGCP** and the relations among them may be visualized like this:

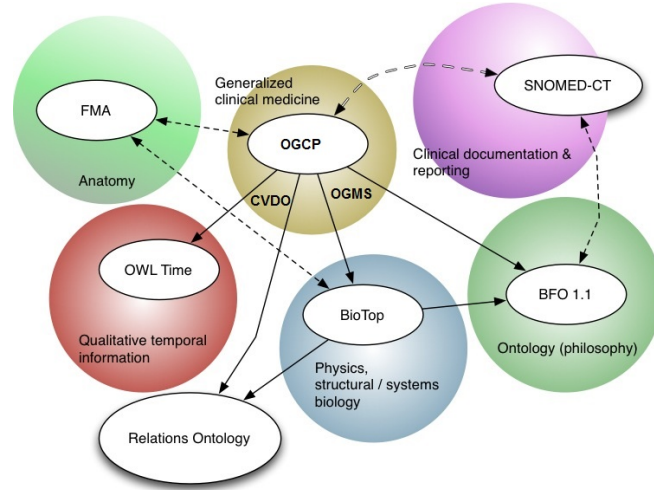


Figure 4.4: Ontology alignment structure in **OGCP**.

All the alignment leverages the Relations Ontology to enforce the Open Biological and Biomedical Ontologies foundry principles application.

4.2.1 OBO Foundry ontologies

Being conservative around ontology engineering when developing our **OGCP**, it emerged as an amalgamated careful reuse of 11 already aligned ontologies from the Open Biological and Biomedical Ontologies (**OBO**) available set.

Ontology reuse involves building a new ontology through maximizing the adoption of pre-used ontologies or ontology components. *"Reuse has several advantages. First, it reduces human labor involved in formalizing ontologies from scratch. It also increases the quality of new ontologies because the reused components have already been tested"* [LED⁺10]. When two ontologies share components through ontology reuse, mapping between them becomes simpler because mappings between their shared components are trivial. In our case we had only to trim and prune according to our intended purposes of representing the Clinical Practice sub-domain of knowledge.

The Open Biological and Biomedical Ontologies (**OBO**) Foundry is an initiative to create a set of well-defined reference ontologies that are designed to work with one another to form a single, non-redundant system. The **OBO** Foundry consortium defines a number of principles for ontology development and ontology developers wanting their ontologies to be members of the OBO Foundry must work to conform to these principles [GNM⁺11].

4.2.2 OBO Foundry principles

The OBO Foundry [The14] *"is a collaborative experiment involving developers of science-based ontologies who are establishing a set of principles for ontology development with the goal of creating a suite of orthogonal interoperable reference ontologies in the biomedical domain"*.

The principles include, for example, that

- The ontology is openly available;
- The ontologies can be expressed in a common shared syntax;
- All the terms in the ontologies have well-formed definitions;
- The ontology has a plurality of users.

Ontology developers who request to have their ontology as an OBO Foundry candidate are expected to work with the OBO Foundry custodians to ensure that their ontology conforms to the OBO Foundry principles. The set of ontologies in the OBO Foundry evolves constantly. New ontologies submitted to "the Foundry" are added to the list of candidate ontologies, and the OBO Foundry community works together to bring these ontologies as close as possible to satisfying the principles. After the custodians decide that an ontology conforms sufficiently to the OBO Foundry principles, it may become a bona fide OBO Foundry member.

4.2.3 Ontological Realism applied to OGCP

In Biomedical Ontologies (BO) the importance of realism is particularly evident. The foundational structures are exemplified here with three examples that are used to illustrate their application:

1. instances vs types,

Instances correspond to individual entities. e.g. The left lung of patient 3643. Types (Classes in OWL) represent the common characteristics of sets of instances (e.g., a kidney is bean-shaped, properties common to all kidneys) [SCK⁺05]. Instances are related to the corresponding types by the relation `instance_of`. For example, my left kidney is an `instance_of` kidney.

2. continuants and occurrents,

`continuants` exist (endure) through time, `occurrents` go through time in phases [Sim00]. Roughly speaking, objects (e.g., a liver, an endoscope) are continuants and processes (e.g., the flow of blood through the mitral valve) are continuants.

3. independent and dependent continuants.

The lung, an object, is an `independent continuant`, whereas its shape is a `dependent continuant` (like all qualities, functions and dispositions, all dependent on their bearers).

Entities organization and consistent ontology development using this foundational concepts is essential for enabling the wide acceptance in the BO science field. The integration of OBO foundry ontologies as upper, top domain or domain ontologies in the structure of OGCP is the guarantee of OR application in our domain.

4.2.4 **OBO** foundry ontologies integration

Two ontologies are aligned when we define the relationships between terms in the different ontologies. Two ontologies are merged when, based on the alignment relationships between the ontologies, a new ontology is created containing the knowledge included in the source ontologies. With this principle in mind we took advantage of the fact that **OBO** foundry ontologies are already aligned to merge different parts of them into the overarching **OGCP**.

In the next sub-sections the 11 ontologies used, along with the rationale for using them is presented. Although they are already consecrated knowledge artifacts in the Biomedical Ontologies (**BO**) domain of computer science, we didn't pay too much attention to the state-of-the-art of the specificities of **OBO** because we will now discuss the articulation among them in every detail that matters to our idea of Ontology for General Clinical Practice (**OGCP**). We cover every one, but with a rather shallow amount of detail for the sake of being space savvy, and not to incur in tiresome reading chores.

Basic Formal Ontology (**BFO**)

Basic Formal Ontology (**BFO**) [**BFO04**] has been the cornerstone of Ontological Realism (**OR**) application to the **OBO** coordinated evolution of ontologies [**SAR⁺07**, **SB10**]. Only in 2010, however, the definitive treaty of the **OR** application surfaced in literature [**SC10**]. Surely a fair amount of acceptance and "new kids on the block", ontologically speaking, had to emerge according to the principles, along with the confirmation of the Semantic Web (**SW**) techniques and tools like the developments in reasoners and standards. **OWL** had to evolve heavily acquiring more powerful expressive features that were basilar to the Biomedical realm as we introduced in sections 2.3.2, 2.4 and 2.4 with the standardization of Web Ontology Language v.2 (**OWL2**) in 2009. Still, the current version of **BFO** is 1.1 despite the v2 draft being evolving already for 2 years now with the 113 page long "BFO 2.0 DRAFT SPECIFICATION AND USER'S GUIDE" reaching a stable condition since late 2012. To assert the evolutionary state of the ontology, the summary of most important changes in BFO 2.0 as compared to BFO 1.1 is very elucidative but we refrain to reproduce its contents here, just the summary itself:

1. Clarification of BFO:object
2. Introduction of reciprocal dependence
3. New simplified treatment of boundaries and regions
 - Revision of treatment of spatial location
 - Treatment of process predications under the heading 'process profiles'
 - Inclusion of relations as part of BFO vs. RO, with changes to relations
 - New relation `exists_at` added.
 - Relation of containment deprecated
 - Relations of parthood disambiguated
 - Revision of Process

The event management in BFO 1.1 is standardized with the usage of the, now deprecated in 2.0 in favor of OWL-Time, *snap* and *span*. For reasoning capabilities improvement in our OGCP we opted to rely already on OWL-Time.

OWL-Time Ontology

OWL-Time [W3C06] is an ontology for describing temporal concepts (content and properties). Based in the former DAML-Time ontology [HFA⁺02], a project led by Jerry Hobbs, who was later tutor for Feng Pan's PhD thesis that led to the development of OWL-Time ontology. The ontology provides a vocabulary for expressing facts about topological relations among instants and intervals, together with information about durations, and about datetime information.

"Extraction and normalization of temporal expressions from documents are important steps towards deep text understanding and a prerequisite for many NLP tasks such as information extraction, question answering, and document summarization.

There are different ways to express (the same) temporal information in documents. However, after identifying temporal expressions, they can be normalized according to some standard format." [SG13].

In the usage of CCL we impose the restrictions as defined in [Aug05].

When using the *smart* assumptions mentioned in section 5.2.4 we leave out explicitly all the situations not properly handled yet by the *OWL-Time* ontology and expressed in the future work section 8.2.

Relations Ontology (RO)

Back in 2005, the evidence that a higher degree of formality led to the proposal and development of an ontology to enhance the treatment of relations in Biomedical Ontologies (BO) [SCK⁺05]. A methodology was advanced for providing unambiguous formal definitions of the relational expressions used. The goal was, and is, to promote interoperability of ontologies and to support the reasoning about spatial and temporal dimensions in BO and particularly of, what matters to us, medical practice phenomena.

Ontologies can be seen as directed acyclic graphs with the *edges* (properties) being the relations between *terms*. Different types of these edges reveal the expressiveness of a given ontology in a certain domain. That is, if an ontology only represents the subsumption, meronomic relations *is_a*, it has to be considered only a *mereology* and lacks the so-called *ontological relations* that expand the reasoning capabilities.

Anatomic ontologies, for example, need a much stronger set of part-hood edges and the formalism introduced in the Relations Ontology contribute steadily to their establishment and acceptance like in the *fly*, *fungus*, *yeast* or *zebrafish* or in our *FMA* that, as all the OBO foundry ontologies do are RO based [RO12]. With Relations Ontology, the relations are no longer incorporated in informal ways that lead to unclear logical interconnections between various ontologies. Based on the requirements presented in [SKK04] the Relations Ontology comes to rectify the defects of inconsistency both within and between ontologies.

A major introductory distinction that must be considered when looking at Relations Ontology is between *Continuants* and *Processes*, these are applied to entities at all levels of granularity. *Continuants* are those entities which endure through time while undergoing different sorts of changes. *Processes* are entities that unfold themselves in successive temporal phases [SCK⁺05]. A continuant

is what changes, a process is the change itself.

To formulate the $\langle \textit{class}, \textit{class} \rangle$ relations definitions a vocabulary is employed that refers to both *classes* and *instances* this vocabulary and terminology is applied throughout the present work and systematized in the A symbols and terminology annex. There is a resemblance, no pun intended, to standard logical notation:

- C, C_1, \dots range over *continuant classes*;
- P, P_1, \dots range over *process classes*;
- c, c_1, \dots range over *continuant instances*;
- p, p_1, \dots range over *process instances*;
- r, r_1, \dots range over *three-dimensional spatial regions*;
- t, t_1, \dots range over *instants of time*;

there will also be the need to incorporate further variables, ranging over temporal intervals, biological functions, attributes and values. One of the important contributions initiated in this ontological proposal is the intended parallelism between the terms applied in the ontology and their $\mathcal{DL}/\text{OWL2}$ counterparts, namely the *logical quantifiers*, leading to a much clearer terminology both for the logicians and computer implementors. Another crucial contribution is the concept of primitive relations which are those self-explanatory hence don't need further regress. The following table itemizes the primitive relations introduced in RO, taken from [SCK⁺05] and used along our work:

$\mathcal{DL}/\text{OWL2}$	Primitive relation
c instance_of C at t	Between a continuant instance and a class which it instantiates at a specific time
p instance_of P	Between a process instance and a class which it instantiates holding independently of time
c part_of c1 at t	Between two continuant instances and a time at which the one is part of the other
p part_of p1, r part_of r1	Part-hood, holding independently of time, either between process instances (one a subprocess of the other), or between spatial regions (one a subregion of the other)
c located_in r at t	Between a continuant instance, a spatial region which it occupies, and a time
r adjacent_to r1	Proximity between two disjoint continuants
t earlier t1	Between two times
c derives_from c1	Involving two distinct material continuants c and c1
p has_participant c at t	Between a process, a continuant, and a time
p has_agent c at t	Between a process, a continuant and a time at which the continuant is causally active in the process

Table 4.1: Primitive instance level relations in Relations Ontology

The Relations Ontology was developed using an incremental methodology where options were picked from *Gene Ontology*, *Foundational Model of Anatomy* and others that were scrutinized by a team of formal ontologists and biologists of the different research groups .

The class-level relations present in Relations Ontology are listed in the referred paper [SCK⁺05], along with a profusion of examples and summarized here:

Relations and Relata	Definitions
C is_a C ₁ ; Cs and C ₁ s are continuants	Every C at any time is at the same time a C ₁
P is_a P ₁ ; Ps and P ₁ s are processes	Every P is a P ₁
C part_of C ₁ ; Cs and C ₁ s are continuants	Every C at any time is part of some C ₁ at the same time
P part_of P ₁ ; Ps and P ₁ s are processes	Every P is part of some P ₁
C located_in C ₁ ; Cs and C ₁ s are continuants	Every C at any given time occupies a spatial region which is part of the region occupied by some C ₁ at the same time
C contained_in C ₁ ; Cs are material continuants, C ₁ s are immaterial continuants (holes, cavities)	Every C at any given time is located in but shares no parts in common with some C ₁ at the same time
C adjacent_to C ₁ ; Cs and C ₁ s are continuants	Every C at any time is proximate to some C ₁ at the same time
C transformation_of C ₁ ; Cs and C ₁ s are material continuants	Every C at any time is identical with some C ₁ at some earlier time
C derives_from C ₁ ; Cs and C ₁ s are material continuants	Every C is such that in the first moment of its existence it occupies a spatial region which overlaps the spatial region occupied by some C ₁ in the last moment of its existence
P preceded_by P ₁ ; Ps and P ₁ s are processes	Every P is such that there is some earlier P ₁
P has_participant C; Ps are processes, Cs are continuants	Every P involves some C as participant
P has_agent C; Ps are processes, Cs are material continuants	Every P involves some C as agent (the C is involved in and is causally responsible for the P)

Table 4.2: Class-level relations in Relations Ontology

The basic semantic interpretations are then defined for each class-level relational expressions and are used as foundation for our biomedical ontological relations referred in 3.1.2 and extracted from text in our Knowledge Base learning process detailed in chapter 5.

Knowing that the set of ontologies that form the basis of our OGCP pictured in section 4.2 are built and structured upon each other we are certain that, for instance, the Foundational Model of Anatomy where the anatomical structures of the patient are defined, and the OGMS where the disease model is enforced have their relations according to the Relations Ontology and this is our guarantee that the properties in the relations presented in the following table, once again taken from [SCK⁺05], apply consistently:

<i>Relation</i>	<i>Transitive</i>	<i>Symmetric</i>	<i>Reflexive</i>	<i>Antisymmetric</i>
is_a	+	-	+	+
part_of	+	-	+	+
located_in	+	-	+	-
contained_in	+	-	-	-
adjacent_to	-	-	-	-
transformation_of	+	-	-	-
derives_from	+	-	-	-
preceded_by	+	-	-	-
has_participant	-	-	-	-
has_agent	-	-	-	-

Table 4.3: Properties of the relations in the OBO Relations Ontology

The logic of *Inverse and reciprocal relations* in [SCK⁺05] is carefully followed in our enrichment process trying to make the Clinical Controlled Language interrogation activity solid and sound in the answers provided by the Discourse Controller (DC), not rendering absurd counter intuitive results following the Knowledge Base enrichment from the clinical texts. Naturally the results obtained so far and presented in chapter 7 are only possible in the controlled restrict Cardiovascular healthcare providing environment of ULSNA within the showcase group.

Top-Domain Ontology for the Life Sciences (BioTOP)

Top-Domain Ontology for the Life Sciences (BioTOP) [SBS07] serves as the glue to connect all the Ontology for General Clinical Practice (OGCP) in an ontologically sound fashion.

With the advances developed in 2009 by Schultz et al. [SBvdH⁺09] aligning the Unified Medical Language System (UMLS) semantic groups with the BioTOP ontology, we are now able to use the semantic infrastructure and groups of UMLS CORE to serve as the mapping for identification of our extracted axioms semantic representation. Concretely, using the BioTOP as enforcer of Ontological Realism articulation, the relations found in accordance to section 3.1.2 and leveraged as detailed in section 5.2.5 through the Ontology Driven Expanded Semantic Annotation (ODA) will provide our Knowledge Base with a very expressive and powerful Discourse Representation Structure.

Foundational Model of Anatomy (FMA)

One of the strongholds of OBO Foundry is the Foundational Model of Anatomy (FMA). This ontology represents the pinnacle of Ontological Realism application for it is capable of representing all the spatial and structural physical components of the human body and their systems (respiratory, circulatory, digestive and others) physical and temporal interrelations [RJ03]. Using some of their more recent OWL incarnations [NR08] the expressiveness of Foundational Model of Anatomy induces complete anatomical inferencing which we take advantage when modeling the healthcare practice. For example, our ontology contains "for free" all the organs hierarchy already as foundation that allows assertions like a "fracture of the femur" being a "broken leg" and the fact that a penis is not part of a woman body or no abortions can be performed to a man.

If Foundational Model of Anatomy were to be aligned with the full SNOMED-CT with its 402000

terms it would render a non computable set with several millions terminological (TBox) assertions. We avoid this *Cartesian Product* explosion problem by carefully using only the **SNOMED-CT CORE** subset of terms as we've seen in section 4.1.1.

Symptom Ontology (**SO**)

The Symptom Ontology (**SO**) [Sch04] was designed around the guiding concept of a symptom being: "A perceived change in function, sensation or appearance reported by a patient indicative of a disease". Understanding the close relationship of Signs and Symptoms, where Signs are the objective observation of an illness, the Symptom Ontology will work to broaden it's scope to capture and document in a more robust manner these two sets of terms. Understanding that at times, the same term may be both a Sign and a Symptom [MPVH06]. The Symptom Ontology was developed as part of the Gemina project starting in 2005 at TIGR and work continues on the project at the Institute for Genome Sciences (IGS) at the University of Maryland.

- The Symptom Ontology is organized primary by body regions with a branch for general symptoms.
- The Symptom Ontology was submitted in July 2008 for inclusion and review to the **OBO** Foundry.
- This project is open to collaborative development, compulsory for **OBO** Foundry consideration.

This ontology interrogation is mainly used for atomic clinical concept identification since it pertains to the most exposed terms in **SOAP** figuring heavily in the **Subjective** (Symptoms) section but also possible in the **Objective** (Signs). As soon as it is recognized (NER) it is queued for classification in the Ontology Driven Expanded Semantic Annotation expanded with semantic adequate symptom types in **UMLS** semantic network.

Vital Signs Ontology (**VSO**)

The Vital Signs Ontology (**VSO**) is an extension of the Ontology for General Medical Science (**OGMS**). **VSO** covers the four consensus human vital signs: blood pressure, body temperature, respiration rate and pulse rate. **VSO** provides also a controlled structured vocabulary for describing vital signs measurement data, the various processes of measuring vital signs, and the various devices and anatomical entities participating in such measurements [GSA⁺11]. In our instantiation of **OGMS**, Vital Signs Ontology enforces terminology and relations knowledge about the objective signs to the more specialized, when applied in Cardiovascular Disease Ontology (**CVDO**) in our case, that is sustained upon it.

Computer Based Patient Record Ontology (**CPR**)

Computer Based Patient Record Ontology (**CPR**) [OBP⁺07] has the representational capabilities needed for healthcare providing processes. In its latest mature definitions [OCF09, Ogb11] it provides full support for a modern healthcare environment. **CPR** does not entail an underlying model of disease

so it is turned usable, with reasoning capabilities, if any model is enforced. Be it either a generalized model like that of Ontology for General Medical Science (OGMS) or a particular instantiation like the Cardiovascular Disease Ontology (CVDO) that we use.

We developed the leaf nodes of OGCP based in the CPR model because that is everything we extract automatically from text using our Automatic Ontology Learning (AOL) set of tasks detailed in section 5.2.

A special detail was given to ensure \mathcal{EL} conformity to the leaf nodes where the clinical cases will be asserted, dropping non valid classes to maintain an efficient \mathcal{DL} as discussed in section 5.2.8 when using the incremental process presented in 5.2.7.

It is possible to get an overview of OGCP external structure looking at its \mathcal{DL} *TBox* that is extensively detailed in OGCP \mathcal{DL} appendix B in page 136.

Ontology for General Medical Science (OGMS)

The Ontology for General Medical Science (OGMS) [OGM10] is an ontology of entities involved in a clinical encounter. OGMS includes very general terms that are used across medical disciplines, including: 'disease', 'disorder', 'disease course', 'diagnosis', 'patient', and 'healthcare provider'. OGMS uses the Basic Formal Ontology (BFO) as an upper-level ontology. The scope of OGMS is restricted to humans, but many terms can be applied to a variety of organisms. OGMS provides a formal theory of disease that can be further elaborated by specific disease ontologies. This theory is implemented using OWL DL, it has OBO Relations Ontology at its foundation and is available in OWL and OBO formats.

OGMS is based on the papers *Toward an Ontological Treatment of Disease and Diagnosis* [SCS09] and *On Carcinomas and Other Pathological Entities* [SKCR05]. The ontology attempts to address some of the issues raised at the Workshop on Ontology of Diseases (Dallas, TX) and the Signs, Symptoms, and Findings Workshop (Milan, Italy). OGMS was formerly called the *clinical phenotype ontology*.

The OGMS project is always interested in application-specific use cases such as those described in the current work.

Existing and planned extensions of OGMS include:

- Sleep Domain Ontology (SDO)
- Infectious Disease Ontology (IDO) and its suite of extensions.
- Ontology of Medically Relevant Social Entities (OMRSE)
- Vital Sign Ontology (VSO)
- Mental Diseases
- Oral Health and Disease ontology
- Cardiovascular Disease Ontology (CVDO)
- Mental Functioning Ontology

- Ontology for Newborn Screening Follow-up and Translational Research
- Drug Ontology
- Model for Clinical Information (MCI)
- Ocular Disease Ontology (ODO)
- Other examples of **OGMS** applied to specific diseases.

Disease Ontology (DO)

The Disease Ontology (**DO**) has been developed as a standardized ontology for human disease with the purpose of providing the biomedical community with consistent, reusable and sustainable descriptions of human disease terms, phenotype characteristics and related medical vocabulary disease concepts. Disease Ontology was developed through collaborative efforts of researchers at Northwestern University, Center for Genetic Medicine and the University of Maryland School of Medicine, Institute for Genome Sciences. The Disease Ontology integrates semantically disease and medical vocabularies through extensive cross mapping of Disease Ontology terms to MeSH, ICD, NCI's thesaurus, SNOMED and OMIM.

We use a pruned version developed in-house with only the **SNOMED-CT** alignments included. No noticeable loss of expressiveness is felt because we don't care about other aspects of medicine besides healthcare practice.

Cardiovascular Disease Ontology (CVDO)

The Cardiovascular Disease Ontology (**CVDO**) is designed to describe entities related to cardiovascular diseases (including the diseases themselves, the underlying disorders, and the related pathological processes). **CVDO** incorporates terms from the **OBO** Foundry. In particular, it is based on **OGMSs** model of disease, and uses the **BFO** as an upper-level ontology. **CVDO** is so far available in **OWL** format. It is being developed at the INSERM research institute (Institut National de la Santé et de la Recherche Médicale).

Chapter 5

Knowledge Base population

When using ontologies as support for Knowledge Representation (KR), Knowledge Base (KB) Population is the process of enriching an ontology with facts (assertions).

After the establishment of the OGCP ontology in chapter 4, we introduce now the elaborate process and tools for clinical information elicitation from text.

We build upon the fundamental 2008 treaty [BC08] by Buitelaar and Cimiano and incorporate many recent contributions like those summarized by Wong in 2012 [WLB12].

The overall process is divided into three very distinct phases:

1. A semi-supervised tutoring phase is done with a small amount (a dozen or less) of texts with representative sample segments for concept extraction process learning. In this step the Domain Translation Memories (TMs) that induce the automated atomic Clinical Concepts (CC) recognition in the subsequent phases are developed.

This is the supervised tutoring phase presented next in section 5.1 and visualized in figure 5.2.

2. The Ontology Learning (Population) tasks, tuned with the developed Translation Memories as Machine Learning artifacts, proceed unsupervised and are the core population process for healthcare Knowledge Base creation.

Now that the ontological framework (Chapter 4) and the controlled translation utilities (Section 5.1) are set we can get to the core of the acquisition processes that are explained in section 5.2 and pictured in figure 5.3.

3. The Discourse Based Enhancement interactive enrichment that relies on clinical interrogation for possible further population is the last phase and won't be detailed until the next chapter 6 where Clinical Practice knowledge interrogation is brought up.

While the first phase has to run only once, it can be successively refined into better, more accurate Clinical Controlled Language (CCL) repositories in order for the second and third phases to get

incrementally better in their task of Knowledge Base population.

The three different steps can be viewed like this with the respective before and after conditions in each phase:

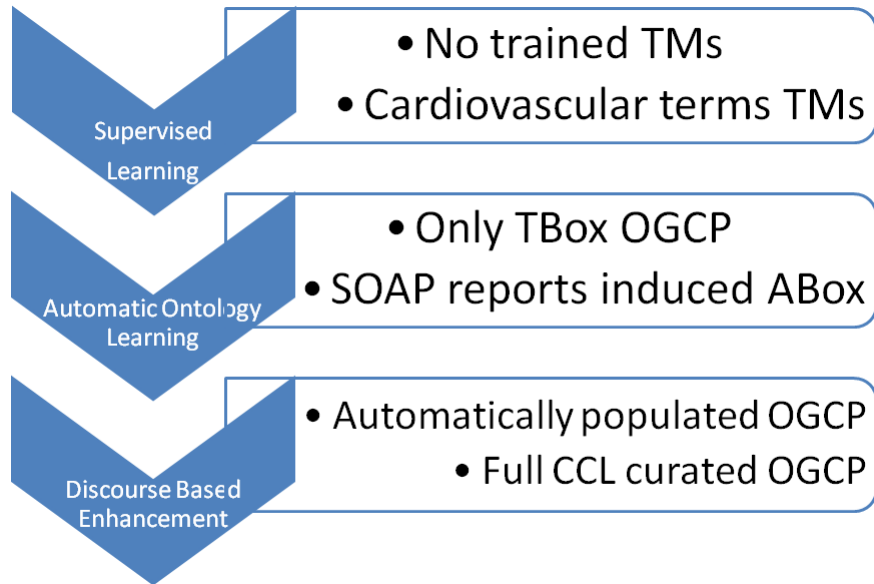


Figure 5.1: Knowledge Acquisition Phases.

5.1 Supervised tutoring

The ontology presented in chapter 4 is an empty structure to be populated into a Knowledge Base with information extracted from the clinical history *SOAP* reports in an automated fashion.

The first step in developing the automatic acquisition from those reports is a preparation process where the system is taught using some similar *SOAP* texts on how to properly identify and classify all the different segments. This preliminary phase needs to be done only once for every clinical reality that is to be subject afterwards of Knowledge Base creation, for a specific doctor, a given service, a period or whatever intended.

The telegraphic form that is common among clinicians also poses some constraints to the usual Natural Language Processing techniques used in other fields. Contextual complex features like negation, temporality, location, granularity, personal form nicknames and event subject identification are crucial for accurate interpretation of the extracted information but renders high ambiguity in free text, most work however has been developed so far, as presented by Demner-Fushman et al. in [DFMSA10b]. For this preliminary step to be accomplished we took inspiration for the supervised training in the development of a *Gold Standard* clinical annotated corpus presented in [RGH⁺09]. In fact, the set of the resulting Translation Memories obtained can be seen as a *Gold Standard* for the accurate clinical term identification. The process can be pictured like this:

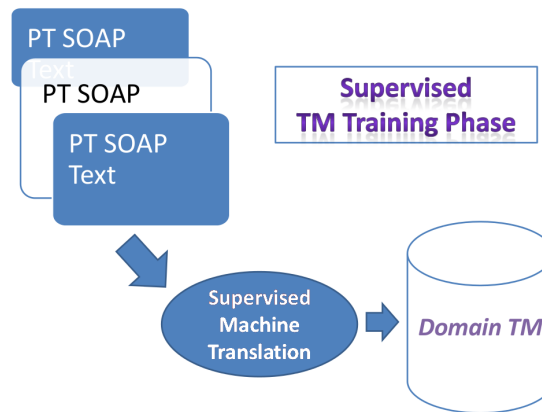


Figure 5.2: Supervised Translation Memory (TM) training.

The learning process is based on the refinement of the Translation Memory applied to that specific acquisition. It can be seen as a "manual translation tutoring" although the focal point is not really the translation from different "natural languages" to English but, in fact, from the original controlled technical jargon to "specialized English".

After the tutoring incremental process, the found acronyms, named entities and clinical terms are ready to be adequately tagged in the process presented in section 5.2 for the correct Ontology Driven (OD) annotation.

As reviewed in chapter 2, the state-of-the-Art for acquisition from clinical text has enjoyed strong developments in recent years. We are exploring here the generic possibility of extracting from free text present in most human interfaces used by clinicians. Going from clinical episodes text, that is usually presented in a human friendly format, to one adequate for computer processing involves a fair amount of text pre-processing to handle different problematic situations because:

1. Reports aggregate information from disparate clinical episodes that are not uniquely identified or not even individually dated.
2. The clinician is only identified by his/her name if any identification is made at all.
3. The information conveyed in free text is intended only to be understandable by fellow practitioners or even by the clinician him/herself making use of pragmatic jargon normally plagued with acronyms and nicknames abundant in their specific community.
4. Text is profoundly intermixed with decorative elements for better legibility, normally in PDF or HTML files.
5. The clinicians natural language may be other than English without concepts defined in a foundational thesaurus like **SNOMED-CT** or **UMLS** for instance that don't even exist in their particular language.
6. The time spanning and snapping of the processes depicted in natural language is often difficult to extract and represent formally.

In the supervised tutoring process we benefit from the fact that we have to translate from jargon to English to customize our centralized Translation Memory Manager (**TMM**) like the *Google translator*

toolkit or *mymemory translation services* enhanced with our own *Translation Memories* and *Glossaries*.

5.1.1 Translation Memory as a controlled technical jargon repository

In the world of professional Computer Aided Translation (CAT), the use of personal terminologies has been subject to standardization using the Translation Memory (TM) [Wik11] model. A TM is a database that stores so-called "segments", which can be acronyms, nicknames, words, phrases and paragraphs that have already been translated, in order to aid human translators through the CAT software. The source text and its corresponding translation in language pairs called "translation units" are stored in the TMs in a standardized exchangeable form. We adopted the use of the second generation TMX standard. Much more powerful than first-generation TMs, they include a linguistic analysis engine, use chunk technology to break down segments into intelligent terminological groups, and automatically generate specific glossaries.

5.1.2 Translation Memory Manager tools

From the huge listing of resources available currently in the Internet we checked carefully those that can handle Translation Memories (TMs) based in the Translation Memory eXchange (TMX) Standard [Wik14]. TMX was originally developed and maintained by OSCAR (Open Standards for Container/Content Allowing Re-use) a special interest group of LISA (Localization Industry Standards Association). The format allows easier exchange of translation memory between tools and/or translators with little or no loss of critical data. Translation Memory Managers (TMMs), that handle creation and maintenance of the TMX files can be desktop centric or centralized TM systems that store on a central server. They work together with desktop TMM and can increase TM match rates by 30-60% more than the TM leverage attained by desktop TMM alone. They can very fruitfully develop Machine Translation (MT) based TMs that can then be exported and further refined by the desktop systems. We explored 2 systems that proved to be interoperable given the standardized formats that are handled: *Google translator toolkit*¹ and *mymemory translation services*² that will be detailed further with an appropriate example.

The first step is the creation of a seminal TMs for some sample documents by using Machine Translation. Clearly all the acronyms and personal defined nicks are not matched but we can then download a base Translation Memory (TM) to be further refined. TM files have an incremental XML structure that can be very easily manipulated through adequate XSL Transformations (XSLT), that can be manually applied or enqueued in an automated workflow as we show in Software Architecture section 7.1. Some tools rendered our work even simpler like the open source Apache Tika included in the OpenNLM framework.

We use the file positioning capabilities of our machine introduced in the mentioned section 7.1 to automate all TM management. The knowledge accumulation for our increasingly accurate translation tasks are just the positioning of the subsequently developed TM files in a specific folder. We have then a TM enrichment workflow that can incrementally match more and more concepts.

¹<https://translate.google.com/toolkit>

²<http://mymemory.translated.net>

TMs can even be orchestrated by service or specialty for example, rendering a fewer amount of work left to be done individually.

5.2 Automatic Ontology Learning

Recall that one of the main contributions of our work is the healthcare Knowledge Base creation that is made through highly specialized Natural Language Processing.

We shall use a **QA** example suggested by our trial investigator to illustrate the problems faced and the solutions adopted in the subsequent **NLP** phases presented ahead.

The intended goal of our system may be show-cased with the following 3 (three) hypothetical user Question Answering with gradually higher complexity that can show the following results after all the enrichment process:

- **Q1:** What is the patients personal history?
- **A1:** Hypertension for 15 years; Diabetes Mellitus type 2 for 10 years; Cholecystectomy 2 years ago; Diabetic father; Obese BMI 26,5; Abdominal perimeter 106 cm.
- **Q2:** What is the suggested diagnosis?
- **A2:** Laboratory routines: lipid profile; HgA1c; Rx thorax; ECG in rest; Echocardiogram; Effort test (Effort proof or Chardiac scintigraphy);
- **Q3:** What is the immediate recommended therapy assuming that AHT and Diabetes are not controlled ?
- **A3:** Rich fiber and vegetable diet; polifraccionate and hiposaline; IECA or ARA II; Calcium Antagonist; Metformine; Estatine;

It can be figured out from the previous set of examples that our work intends to complement both the text understanding as well as text generation in the clinical environment.

We explain in this section the developments incorporated in the various **NLP** steps of the learning pipeline that represent specific advances for the healthcare domain. To achieve the illustrated capabilities in the examples the process still has to rely in the Discourse Based Enhancement (**DBE**) subsequent interactive phase explained in section 6.3

Collecting our information from **SOAP** reports introduced above in section 3.1.3 we take advantage of the fact that the report depicts a clinical encounter in a semi-structured way to direct into a more tractable source. The Subjective, Objective, Assessment, Plan framework, used to structure progress notes to facilitate problem specific, clinical decision making by physicians, is a well known, canonical structure in the medical domain. The underlying construction of the **SOAP** report induces some very important assumptions. We find sections that can be associated with

- Subjective, the symptoms section **S**;
- Objective, the objective section **O** that are sign records that we take as generator for clinical observations (findings) in the Assessment section;
- Assessment, the analysis section **A** which are the clinical investigation acts;
- Plan, the plan section **P** where discharge prescriptions are registered.

The main parser of our text sources has to split the segments found in the structured Subjective, Objective, Assessment, Plan (**SOAP**) reports and direct their parsing into the suitable handler. As it will be demonstrated later, in section 5.2.4, different types of segments have to be handled differently. Some segments don't have a specific location in our sources introduced in section 3.1.3 but, instead, we take advantage of their particularities to develop a methodology that has learning capabilities and thus shows increasing precision and recall.

The more assertions the Knowledge Base contains the less work has to be devoted by the domain expert to assure it's effectiveness. This is commonly referred to as *incremental knowledge acquisition* [RMVGFB⁺11]

We use an Automatic Ontology Learning task to extract an **ABox** set of axioms from text that can be pictured like this:

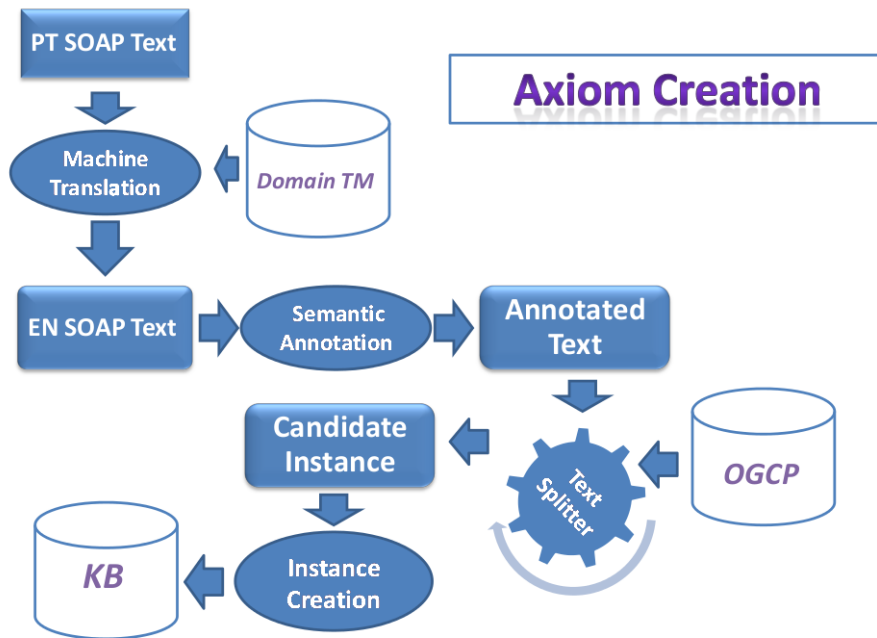


Figure 5.3: Axiom creation from SOAP.

In the following subsections of this important part of the document, several examples will be used to illustrate the different problems faced in the successive steps of the Knowledge Acquisition pipeline. All the examples will use parts of this report:

CENTRO DE SAÚDE PONTE DE SOR SEDE		XXXXXXXXXXXXXXXXXXXX Data Nasc: XX-XX-XXXX (XX anos) XXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXX		*49/448616* *497448616*
Registo Clínico da Consulta				
ESPEC.		12/07/2010 15:55		Dr(a) Carlos Baeta
Dr(a) Carlos Baeta				
S SOAP	Enviada à consulta pela Dra. Isabel Taveira Pinto Episódios de dor retroesternal: -Intensa, não opressiva. -Surge em repouso -Duração > 1h. -Sem sintomatologia acompanhante -Sem irradiação. =>Sem características anginosas. AP: -HTA -Eritema nodoso -AIT há 7 anos Nega DM			
O sO AP	AP - N AC - N			
A SOA AP	(18/05/10)-Hgb-13,3; Htc-39,1; VGM-86,3; HGM-29,4; VS-21; Glicemia-101; Creatinina-0,9; AU-4,1; Colesterol(T-227; HDL-63; LDL-143); TG-94; GOT-15; GPT-12; Na-136; K-4,1; ECO (05/06/10): AE-34; AO-33x16; VE-37/17/54%; SIV-12; PPVE-11 Holter (29/03/10)-RS (61:134; M-80); ESSV esporádica; 1 TSV.			
P SOA P	Cozzar 50 plus; Nebilet 5; Isoptin 120; Ogasto; Ticlidix; Vastarel LM; Xanax 0,25; R/ PE (excluir CAD)			
CARDIOLOGIA ECG COM PROVA DE ESFORÇO				

Figure 5.4: SOAP-5682 Sample complex report.

The first preliminary step is Controlled Natural Language translation that produces a document carefully translated to a correspondent cardiovascular technical English dialect, the already seen **CCL**. Several particularities were defined in order to turn both the Natural Language Understanding (**NLU**) in the Automatic Ontology Learning (**AOL**) phase and the Natural Language Generation (**NLG**) in the interactive Question Answering easier to accomplish.

Applying this step using **Apache Tika** yields the following text (with decorative elements and metadata stripped here for enhanced legibility):

Soap

submitted to the consultation by Dr. Isabel Taveira Pinto
retrosteral pain episodes.

-Intense, not oppressive.

-occurs in home

-duration > 1h.

-Without accompanying symptoms

No irradiation.

= > Without anginal features.

Personal History:

-High Blood Pressure

-Erythema Nodosum

-Temporary Ischemic Accident for 7 years

deny DM

sOap

Personal History - No

Clinical History - No

soAp

(18/05/10)-Hgb-13,3; Htc-39,1; VGM-86,3; HGM-29,4; VS-21; Blood Sugar-101; Creatinine-0,9; AU-Cholesterol(T-227; HDL-63; LDL-143); TG-94; GOT-15; GPT-12; Na-136; K-4,1; ECO (05/06/10): AE-34; AO-33x16; VE-37/17/54%; SIV-12; PPVE-11
Holter (29/03/10)-RS (61:134; M-80); spurious ESSV; 1 TSV.

soaP

Cozzar 50 plus; Nebilet 5; Isoptin 120; Ogasto; Ticlidix; Vastarel LM; Xanax 0,25;
R/ Proof of effort (exclude CAD)

Cardiology

ECG with evidence of effort

Some noticeable points are:

1. the verbs are specifically created in lowercase as an hint for the semantic parser and easier for the clause creation in Predicate Logic in accordance to the **CCL** tokenization ruling present in section 5.2.2.
2. Some stop words are trained to be recognizable like the *Dr.* that indicates a fellow clinical practitioner.
3. Sentence boundaries are not very well defined ranging from a normal sentence delimitation as in natural English, to text segmented in lines, delimitation by periods and/or colons or even several key-value pairs in the same line.
4. Plagued with acronyms that can be either translated or not according to the Gold Standard trained Translation Memory like *Personal History* translated from *AP* (Antecedentes Pessoais) or in the other case *DM* (Diabetes Melitus) which is not translated.
5. Natural occurrence of negative clauses like *Deny DM* or *not oppressive*.
6. Free text section not strictly in accordance to the **SOAP** framework so that it does not induce any specific ontology class generation and needs to be tailored for each specialty. Currently, only cardiology is supported.
7. Numerous acronyms as key for their value pairs which are to be incorporated as individuals in the Knowledge Base

Of course that the **CCL** translation is far from perfect, and we could consider an interactive process of refinement that could act as a sidekick in the Automatic Ontology Learning but it has not been developed because we are in a preliminary proof-of-concept phase and the subsequent steps in the pipeline further refine (automatically) the semantic representation and the representational performance of the Knowledge Base.

One line regarding this possibility will be present in the future work section 8.2.

5.2.1 Clinical Controlled Language translation

The process used in this step is the same engaged for Translation Memory (TM) tutoring presented in section 5.1 with the exception that now the identified text segments aren't any longer rectified but accepted as they are generated.

As it was previously scrutinized in section 3.2.1, the current CCL version offers language constructs like singular and plural countable nouns; mass nouns; existential and universal quantification; generalized quantifiers; indefinite pronouns; relative phrases; active and passive verbs; negation, conjunction and disjunction of noun phrases, verb phrases, relative clauses and sentences; and various forms of anaphoric references to noun phrases. [KF07]. The controlled dialect of cardiovascular specialty is defined by a small number of construction rules that define its syntax and a small number of interpretation rules that disambiguate constructs that in full English might be ambiguous. This grammar is machine learned by induction of the texts in the tutoring phase. Following translation, a content analysis tool (Apache Tika detailed in 2.10.2) extracts the decorative PDF elements and pagination but maintains the structuring stop words: *Soap*, *sOap*, *soAp* and *soaP* rendering the text source as presented in the beginning of section 5.2 (page 71).

5.2.2 Knowledge Acquisition through specialized NLP

Ontology population/enrichment is performed through Information Extraction (IE) from the clinical texts. IE is a specialized sub-domain of NLP that returns pieces of information from text analysis, unlike Information Retrieval (IR) that returns documents. As illustrated in the review by Meystre et al. [MSKSH08] complemented by the review in [LHC11] many IE methodologies are already thoroughly presented and discussed and are, of course, refined and developed in our work.

Aligning the extracted information in form of Clinical Concepts and its relationships in healthcare directed ontologies involves classification into our specific OGCP using several NLP techniques. These tasks form a pipeline that include

1. Metadata extraction and pruning from source files,
2. Tokenization,
3. Part Of Speech (POS) tagging,
4. Named Entity Recognition (NER),
5. Word Sense Disambiguation (WSD),
6. Co-Reference Resolution (CRR),
7. Extraction of Attributes and Values (EAV) and
8. finally expanded Ontology Driven Expanded Semantic Annotation (ODA) for Clinical Concepts (CC) matching being these the CC introduced in [CdKAH06].

The OGCP will then be further refined and improved by reasoning as presented in section 6.3 based in its foundational ontologies in the Biomedical domain. For this purpose different valuable approaches

reviewed in [LHC11], specifically for our particular domain of healthcare, are applied according to the restrictions and opportunities identified in section 3.1.2.

Having evaluated different possibilities in several toolkits based in diverse operating systems, programming languages and paradigms we opted for a complete tool and methodology that handles well the following fundamental characteristics:

- Has to address coherently the whole proposed **NLP** pipeline.
It will permit to maintain a consistent Application Programming Interface (**API**), programming model and language along all the work to be developed with the inherent advantage of avoiding to face different learning curves.
- Has to exhibit both a Command Line Interface (**CLI**) and a Application Programming Interface (**API**) along all the utilities.
They are both needed at different phases. First the **CLI** is heavily used to develop interactively the proof-of-concept techniques that later have to be automated in a programmed form for user deployment.
- Has to be based in a solid community to guarantee future developments.
- Preferably based in Java for it is the programming framework (language, virtual machine, associated dynamic environments) currently most vibrant and full featured in the world.
- Has to be open source, for all the good reasons.

Tokenization

All the tables that systematize the **OWL** generation and verbalization are in [KF07] illustrated with a profusion of examples and we don't find the need to reproduce them here although some rules that are important for the tokenization step are explained in the example based in the mentioned paper. For the correct **OWL2** parsing, generation and verbalization all names used in the ontology are Cardiovascular (**CV**) domain words. Furthermore, individuals are denoted by singular proper names (preferably capitalized), named classes by singular countable nouns, and (object) properties by transitive verbs in their lemma form (i.e. infinitive form).

These restrictions are needed because the names will be used in certain syntactic constructions or will undergo some morphological changes. Proper names are used in the subject and object positions without a determiner, e.g. "Every woman has a heart.", "**Alice** is a woman.". Common nouns are used in the subject and object positions with determiners 'every', 'a', 'at least 2', etc., and can have a plural ending, e.g. "Every heart has at most 4 cavities.". Transitive verbs are often used in singular, but under negation and in plural will stay in infinitive, e.g. "Every person **knows** a child that does not **own** a bike and that has at least 3 friends that **own** a bike.". In some cases, most often when verbalizing the ObjectPropertyRange-axiom, the verb will be turned into a past participle in order to construct a passive sentence, e.g. "Everything that is owned by something is a possession.".

Part Of Speech tagging

Several part-of-speech taggers (POS-taggers) were developed specifically for the biomedical domain. There is evidence that POS-taggers trained and tested on formal text that does not include clinical documents do not achieve state-of-the-art performance. For example, training a POS-tagger on a relatively small set of clinical notes improves the performance of the POSTagger [PCC06] trained on Penn Treebank from 90% to 95% in one study and from 79% to 94% in another study . We use the features in the Apache UIMA POS-Tagger that was explicitly trained with the referred corpus and extended to our Cardiovascular domain during the Supervised Tutoring task introduced in section 5.1.

Named Entity Recognition

Named Entity Recognition involves identifying the boundaries of the name in the text and understanding (and disambiguating) its meaning, often through mapping the entity to a unique concept identifier in an appropriate ontology. [AFT04]. We usually name it here the *Atomic Clinical Concept Identification* and it is done using the UMLS Terminology Services Web Service as explained in section 5.2.3, restricting the invocation to the UMLS CORE in order for the Knowledge Base to be restricted to SNOMED-CT CORE. This modality is a fundamental condition for the tractability of the resulting Knowledge Base due to the small volume of the different asserted controlled terms in the resulting OGCP ABox.

Word Sense Disambiguation

The overall process is based in [NV05] and thus founded in Ontology Driven Expanded Semantic Annotation

Co-Reference Resolution

Co-Reference Resolution (CRR)'s goal is to identify all mentions of a particular concept or entity in a piece of text. When processing a SOAP text we maintain it under our overarching Clinical Integrated Discourse Extended Representation Structure (CIDERS) to be able to scope the reference in our broader Discourse Based Enhancement (DBE) process.

We have an extended scope to deal with in our DBE mentioned previously in section 3.1.3 in identifying internally the co-occurrences of a specific patient to apply case based reasoning to the full extension of his/hers clinical episodes in the Knowledge Base. Either when applying the CRR step described here in the Automatic Ontology Learning or in the a-posteriori interactive DBE phase the references have to be aware that the patient is the same and the co-referencing has to be considered.

Fortunately for us, both the Co-Reference Resolution as well as anaphoric references are rendered transparent due to the reversibility of OWL to CCL and back since the extended DRS CIDERS is maintained by the OWL and ACE representation of our enriched OGCP.

Extraction of Attributes and Values

The correct tunneling process of identifying the accurate attributes and values begin early in the tutoring process explained in section 5.1 where the Domain Expert direct the Translation Memories towards the subset of controlled terms inside the **UMLS CORE** subset. This is the main reason why we tend to name our controlled language as Clinical Controlled Language because we maintain the technicalities restricted from the start.

5.2.3 Ontology Driven Expanded Semantic Annotation

Just before the instance creation we have finally to make use of the highly specialized Web Service provisioned by the National Center for Biomedical Ontologies (NCBO) that provides Ontology Driven Expanded Semantic Annotation. This expanded annotation takes into consideration several semantic driven proximity factors to provide accurate ontological relations concept tagging. It uses, as foundational ontologies, both UMLS and their local Biomedical Ontologies.

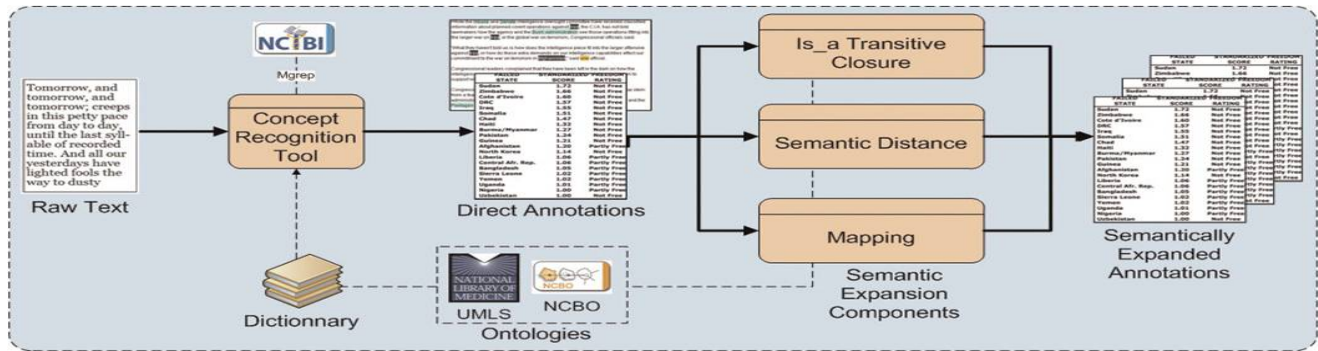


Figure 5.5: NCBO expanded semantic annotation

Having restricted the atomic concept identification to UMLS CORE in the previous process and selecting a content view according to [DFMSA10a] the ontological relations domain is much more confined avoiding thus an exponential growth in the Knowledge Base creation. To get this in an automated manner we have to invoke the Web Service provided by NCBO Biportal imposing in the GET method the restrictions to some predefined terminologies which in our case is SNOMED-CT in category *health*, in the *upper-level* category we use Basic Formal Ontology and Top-Domain Ontology for the Life Sciences, Foundational Model of Anatomy in the *anatomy* category and finally the UMLS from the OBO Foundry group. We select the adequate UMLS semantic types according to the kind of annotation intended to enforce the correct ontological relations identification, e.g.

- In **Subjective** we pick the *T184 - Sign or Symptom*, *T047 - Disease or Syndrome*, *T022 - Body System*, *T042 - Organ or Tissue Function*, *T030 - Body Space or Junction* and *T055 - Individual Behavior* as semantic types,
- in **Objective** we select the *T184 - Sign or Symptom*, *T060 - Diagnostic Procedure*, *T033 - Finding*, *T022 - Body System*, *T030 - Body Space or Junction* and *T034 - Laboratory or Test Result*,
- In **Assessment** we choose *T059 - Laboratory Procedure* and *T034 - Laboratory or Test Result*,
- Finally in **Plan** the semantic types *T059 - Laboratory Procedure* and *T061 - Therapeutic or Preventive Procedure* as suggested by our trial investigator.

As an example we take *retrosternal* in the Subjective section and invoke the Web Service using the ApiKey that was provided to us from the UMLS Terminology Services (UTS) [UTS14]:

```
http://rest.bioontology.org/biportal/search/?query=retrosternal&ontologies=(SNOMED-CT, FMA)
?apikey=YourApiKey
```

The results may be returned whether in **XML** or in JavaScript Object Notation (**JSON**) formats. We retrieve in **XML** format because it is directly consumed by *Apache Tika*. We apply a transformation to pick the returned ID that has a `matchType` of **PREF** which denotes the preferable annotation in that particular choice of semantic types and ontologies.

5.2.4 Smart instance creation

After having annotated the texts based in all the atomic concept identification, going all the way through the above pipeline, there are still some issues remaining that have to be judiciously handled. These quirks and the *smart* ways they are handled are:

- tagging correctly the different types of clinical text segment

We take advantage of the fact that the report depicts a clinical encounter in a semi-structured way to direct into a more tractable source. The Subjective, Objective, Assessment, Plan (**SOAP**) framework, used to structure progress notes to facilitate problem specific, clinical decision making by physicians, is a well known, canonical structure in the medical domain. The underlying structure of the SOAP report induces some very important assumptions to be true. We find sections that can be associated with **Subjective**, the symptoms section **S** where we extract directly into a `ogcp:symptom` record, medications found here are those administered only during the patient visit. **Objective**, the objective section **O** that are sign records `ogcp:sign-finding` that we take as generator for `ogcp:clinical_finding` in the **Assessment** section. **Assessment**, the analysis section **A** which are the clinical investigation acts whose outputs can be clinical artifacts to investigate things that can be consequence of any kind of physiological or pathological processes. Finally **Plan**, the plan section **P** where the `ogcp:therapeutic-act` can be extracted with all the timing, posology and prescriptions registered in that particular clinical encounter, medications here are prescribed for discharge [MWV⁺11]. Aggregating the instances collected so far we finally engage in the more complex `ogcp:medical_problem` instantiation that illustrates the representational complexity of the Cardiovascular Knowledge Base.

- time framing the clinical episodes (*eventuality*)

Trying to identify the correct temporal occurrence of the depicted events we take advantage that all the facts expressed in the Automatic Ontology Learning task refer to Clinical History (**CH**) and we only have to ticket the historical event with a Natural Language Understanding time event instance extracted, if possible, from the text. The smart step applied for solving this is to use the trained Translation Memory in order to generate direct OWL-Time verbalizations of one of the different five temporal categories available in the ontology structure. In our representation of the world we intend always to link the time instance to other things elsewhere in our specific-domain ontology. The description of the event is adequately tagged, according to our previously explained *smart step*, as one of the observed Clinical History episodes, an *eventuality*, and use one of the four possible predicates *atTime*, *during*, *holds* and *timeSpan* to link to the *eventuality* [HP04]. The Translation Memories are personally tailored to impose the mentioned four verbs to correctly instantiate the OWL-Time ontology as described in section 4.2.4. We tailored the Apache UIMA based temporal information extraction system to extract and normalize TIMEX3-based temporal expressions from clinical text referred above. It has been adapted from the

open-source temporal tagger, HeidelTime³ but re-engineered toward the clinical domain and our smart OWL-Time instance creator.

- The current version of **ACE** offers language constructs like countable and mass nouns, collective and distributive plurals, generalised quantifiers, indefinite pronouns, phrasal and prepositional verbs, noun phrase/verb phrase/sentence negation, and anaphoric references to noun phrases through proper names, definite noun phrases, pronouns, and variables [FKS06]. In the **CCL** definition phase of Translation Memory tailoring we refine the verbalization by restricting the capabilities of Attempto Controlled English to what matters to our domain of interest. Our previously taught parser only represents in its **DRS** the entities and constructs that are able to be represented. The verbalization now has to proceed dependent of what **SOAP** section we are in and has to apply the scope of the particular case to the correct patient. That is, it has to pick all the cases already in the Knowledge Base related with that PatientID to enhance the Clinical Integrated Discourse Extended Representation Structure (**CIDERS**), as happened before in the **CRR** step, for the appropriate instantiation.

This is why we consider that **CIDERS** as a broader knowledge scope than the traditional **DRS** in **NLP**.

5.2.5 Ontological relations formation

Clinical concept acquisition encompasses in our view the application of the notion of *non-taxonomic roles extraction*, that is *ontological relations*, from the sources. This theme has been extensively discussed in recent literature [PHT⁺13, PK13, Res99]. *Without non-taxonomic roles, ontology generation boils down to generating taxonomies which lack a lot of crucial semantic information compared to ontologies* [PK13] so we retrieve the ontological relations using this novel mechanism. The referenced work introduced a methodology that fits accurately into our acquisition pipeline. In fact it allows us to elicit the complex non-taxonomic clinical relations through:

1. Pre-annotation of strings with formal concepts: using the previous ontology oriented atomic clinical concept tagging revealed in section 5.2.3
2. Spot the pre-annotated substrings that contain the roles
3. Find the "non-taxonomic" relations between the concepts

5.2.6 Pragmatic interpretation in **NLP**

Semantic parsing is the process of mapping a natural-language sentence into a formal representation of its meaning. A shallow form of semantic representation is semantic role labeling, which identifies roles such as agent, patient, source, and destination with few ontological value. A deeper semantic

³<http://code.google.com/p/heideltime/>

analysis provides a representation of the sentence in Predicate Logic or other formal language which supports automated reasoning.

We trained our semantic parser revealed in section 2.10.3 in the tutoring process by providing a deeper role and entity labeling gold standard when developing our Translation Memories as explained above in section 5.1.

Although the SOAP structure already allows a simplified interpretation of the discourse based in the CCL orientation performed by the users when developing their SOAP note, some examples occur that need the adequate handling of a segment using the expanded DRS introduced to make the right interpretation of the utterance. Most of the issues arrive due to the omitted references because the resolution is trivial for a human reader. This phenomenon of ellipsis can be solved by inference but it incurs in paying a high computational cost [AL94] that we can, and have to, circumvent.

As an example consider the first sentence in case 5682_SOAP after translation:

Submitted to the consultation by Dr. Isabel Taveira Pinto.

The tagging and syntactic parsing processes detailed above render:

Submitted/VBN to/TO the/DT consultation/NN by/IN Dr./NNP Isabel/NNP Taveira/NNP
Pinto/NNP

```
(ROOT
  (VP (VBN Submitted)
    (PP (TO to)
      (NP (DT the) (NN consultation))))
    (PP (IN by)
      (NP (NNP Dr.) (NNP Isabel) (NNP Taveira) (NNP Pinto)))))
```

The shallow semantic annotation using off the shelf standard components from the Lund University found in [Lun14] that use the same (OpenNLP) open source tools as we do, introduced in section 2.10.1 render the following explanatory structure:

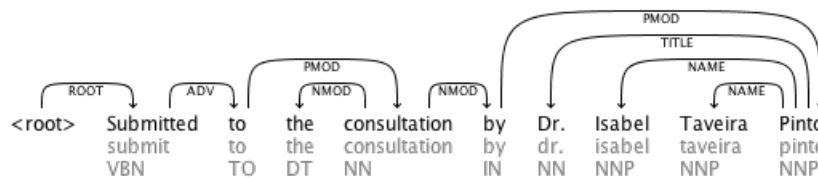


Figure 5.6: Semantic parsing of ellipsis ill segment

it is evident that the omitted patient name, usual in this notes, is referred here to the subject of the SOAP report and we can solve the anaphora as suggested in [AL94] by unifying with the patient ID in the report already mentioned in the Clinical Integrated Discourse Extended Representation Structure (CIDERS) as patientID.

The **CIDERS** in our First Order Logic variant is then (variable names suggested here with adequate acronyms for legibility):

```
patient(PatientID), submitted(PatientID,C,MD), consultation(C), dr(MD), by(C,MD)
```

and our ACE Parser (**PACE**) based Discourse Controller (**DC**) through its reification capabilities in the **CIDERS** unifies easily with the right patient.

Discourse Controller used for enrichment control

The possibility of reusing the **OWL** to **CCL** and **CCL** to **OWL** translators provides the ability of using the parser as a Discourse Controller (**DC**). With the **Protégé** ACE View plugin the assertion of **OWL** constructs is maintained under control of the parser that only allows for precise **CCL** constructs formation under the correct **DRS**.

In a similar way, the *DL Query view* of **Protégé** also permits the creation of Description Logics (**DL**) queries, **OWL** class expressions, strictly consistent with **OGCP** and its incorporation to the ontology requiring however a deep knowledge of the underlying **DL**. The usage of Manchester Syntax is however not that obvious for Domain Experts which does not contribute to our, and our team of trial investigators, goal.

Reasoning with the Discourse Structure

Discourse Reasoning (**DR**) is done by the ACE Parser (**PACE**) according to all the learned cases, the *ABox*, that extends the **OGCP** into the Knowledge Base available to build the **CIDERS** "on-the-fly". Based on this controlled representation of the Cardiovascular healthcare provided to the patient, all the questions posed to the reasoner are also limited to the available knowledge and guaranteed to be well formulated and scientific and technically valid by the underlying **OGCP** ontological structure. All the Clinical Knowledge interrogations are made using the same phrasing that the clinicians are already used to as will be detailed in the clinical interrogation chapter 6.

5.2.7 Round Trip Debug and Repair

The name of this section is closely related to the nature of the algorithm that is used to guarantee the consistency of our **KB** because it is implemented as a **cycle** that tries to express the stated or inferred axioms within \mathcal{EL} using syntactic variations exposed ahead or liminally disregards it if it can not achieve to express it. We implement some "smart tricks" to guarantee that our ontology remains $\mathcal{EL}++$ conformant by disregarding all those candidate axioms that cannot be restricted to the \mathcal{EL} characteristics presented in section 2.4. By classifying automatically the *ABox* in the process of ontology learning, that is the **KB** population, we try to apply an artificial Closed World Assumption (**CWA**) using modalities of negation where appropriate. In our example the symptom **deny DM** is turned from an explicit negation of $\neg \text{DM}$ into an acceptable positive fact of **states**($\neg \text{DM}$).

This is only possible not rendering our **CIDERS** explicitly false yet consistent and maintaining "**clinical CWA**" because we are in the **Subjective** section of our text source. It could not be possible if

we were in the **O**bjective part because signs are supposedly and veritably accurate and no syntactic maneuvers can reverse this clinical truth.

If an inconsistent axiom is ultimately found, and we cannot turn it into a representable one, it is liminally disregarded.

The future work section 8.2 contains an explicit mention to the interest in developing a feature that alerts the practitioner to the impossibility of turning a given assertion representable in the **KB** due to the reasoning abilities restrictions.

Of course, this process of "smart instance creation" based in the round trip cycle is founded in the restrictions of axiom creation that abide to the **OWL** constructs supported by the **ELK** reasoner presented in section 2.7.1.

Knowledge Base instance creation

Every assertion that reaches this stage of the pipeline is guaranteed not to render the **KB** clinically inconsistent so the instance creation is just appending the **OWL** axiom to the **OGCP** file which is done invoking the **OWL API** method presented in section 2.11.3. The overall process:

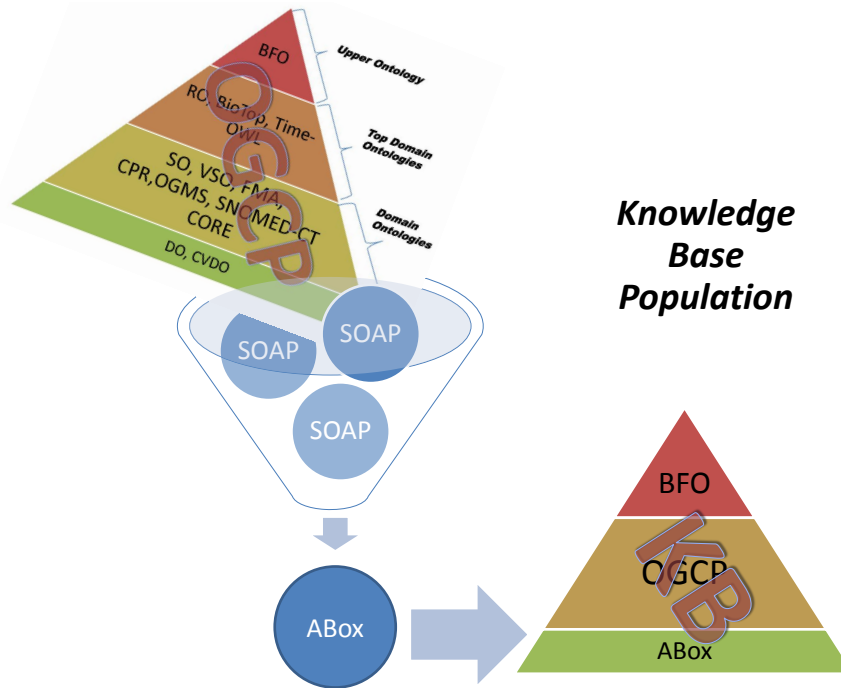


Figure 5.7: Enrichment of **OGCP** into a Healthcare Knowledge Base (**KB**).

As already stated, this invocation can be done

- interactively in **Protégé**, directly in the **ACE** view plug-in while working with Clinical Controlled Language either populating or interrogating,
- inside **Protégé** through the **OWL API** plug-in or yet
- using a Java program that leverages the open source **OWL API** for Java.

5.2.8 Reasoning with effective logics

Web Ontology Language v.2 \mathcal{QL} , \mathcal{RL} and \mathcal{EL} are syntactic subsets of **OWL** Full. The advantage of the \mathcal{EL} syntactic subset is that it can be used via a different formal semantic system than **OWL2** Full (with identical results) and this different semantics ('direct' semantics) has numerous efficient implementations due to the developments in that particular area of logics.

In the **OWL2** \mathcal{EL} syntactic profile specification we find that

- entities are defined in **OWL2 \mathcal{EL}** in the same way as in the structural specification of **OWL2** and **OWL2 \mathcal{EL}** supports all predefined classes and properties.
- The set of supported datatypes has been designed such that the intersection of the value spaces of any set of these datatypes is either empty or infinite, which is necessary to obtain the desired computational properties.
Consequently, the following datatypes must not be used in **OWL2 \mathcal{EL}** : `xsd:double`, `xsd:float`, `xsd:nonPositiveInteger`, `xsd:positiveInteger`, `xsd:negativeInteger`, `xsd:long`, `xsd:int`, `xsd:short`, `xsd:byte`, `xsd:unsignedLong`, `xsd:unsignedInt`, `xsd:unsignedShort`, `xsd:unsignedByte`, `xsd:language`, and `xsd:boolean`.
- **OWL2 \mathcal{EL}** does not support anonymous individuals.
- Inverse properties are not supported in **OWL2 \mathcal{EL}** , so object property expressions are restricted to named properties.
- In order to allow for efficient reasoning, **OWL2 \mathcal{EL}** restricts the set of supported class expressions to `ObjectIntersectionOf`, `ObjectSomeValuesFrom`, `ObjectHasSelf`, `ObjectHasValue`, `DataSomeValuesFrom`, `DataHasValue`, and `ObjectOneOf` containing a single individual.
- A data range expression is restricted in **OWL2 \mathcal{EL}** to the predefined datatypes admitted in **OWL2 \mathcal{EL}** , intersections of data ranges, and to enumerations of literals consisting of a single literal.
- The class axioms of **OWL2 \mathcal{EL}** are the same as in the structural specification, with the exception that `DisjointUnion` is disallowed.

All the full featured **OWL** Full reasoners obviously can handle a subset like **OWL \mathcal{EL}** but only show $\mathcal{EL}++$ performance when restrained to the itemized syntactic restrictions. Some specially drafted reasoners like *Snorocket* [LB10], taking advantage of these restrictions, demonstrate ground breaking capabilities like classifying the full Standard Nomenclature of Medicine - Clinical Terms (**SNOMED-CT**) in few seconds using off the shelf hardware. This reasoner, however, has to be fed with a specifically defined structure of **SNOMED-CT** to perform its duties and our intention is to be able to reason over a "generic" **KB** that by definition is guaranteed to be \mathcal{EL} compliant like those resulting of **OGCP** enrichment. We follow the path of using the **ELK** distributed reasoner [KKS11] to achieve similar results in practice and detail the reasoning in section 6.1

5.3 Text interpretation

In the sequence of fine tuning the **PACE** into a **CCL** controller as introduced in section 3.2.1 the process of text interpretation is limited regarding all the knowledge present in the Clinical Practice

ontology and therefore restrained to Clinical Controlled Language (CCL).

5.3.1 Preliminary considerations

Our SOAP reports depict a clinical encounter of the cardiovascular specialty whether in a primary or secondary care facility. Given this, an instance of an OGCP:inpatient encounter

(Thing/entity/occurrent/processual_entity/health care process/health care encounter/inpatient encounter)

is built to be further populated with all the episodes in the various sub-sections (Subjective, Objective, Assessment and Plan) of the report.

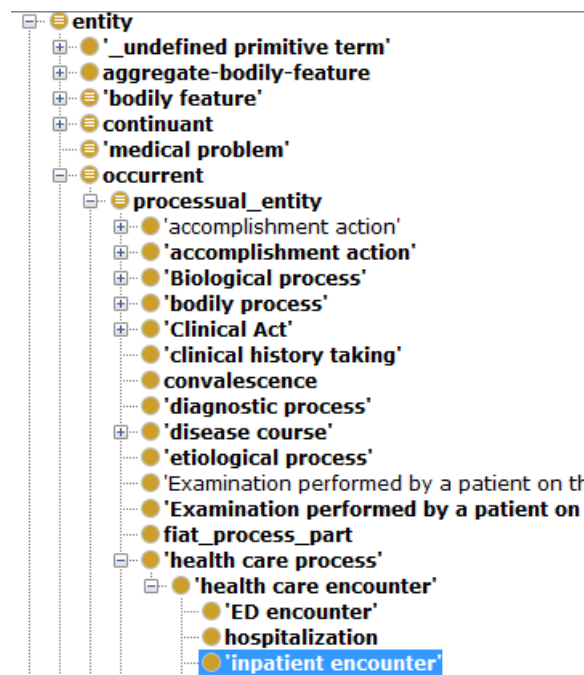


Figure 5.8: OGCP inpatient encounter.

5.3.2 Ontology structure considerations

It is very important to understand the relation of a **SOAP** report and its underlying structure with the proposed ontology hierarchy in order to realize the reasoning capabilities that an adequate population will provide.

The **SOAP** has two very distinctive parts, in one hand the **Subjective** part introduces pointers to the clinical history that are complemented with verifiable **Objective** findings with **signs** registered in the corresponding section. This first part has to be considered as the OGCP:case history which is asserted as equivalent to 'clinical finding' and ('output of' some 'Medical history screening act') in a class hierarchy of

Thing/entity/continuant/dependent_continuant/generically_dependent_continuant/Representational artifact/clinical artifact/
clinical finding/case history

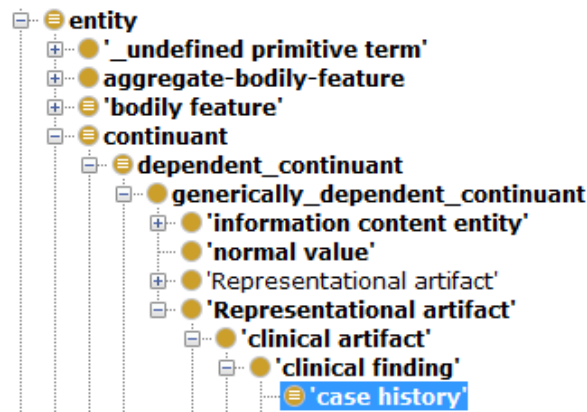


Figure 5.9: OGCP case history.

and all the domain modeling is then done taking this as historical factoids. Apart from `OGCP:case history` the `OGCP:clinical findings` still have 3 disjoint historical classes: `'clinical diagnosis'`, `'laboratory test finding'` and `'record of a clinical sign'`.

The second part however, formed by the `Assessment` and `Plan`, are no longer to be considered as `'Representational artifacts'` because they are part of the activities developed in the current encounter so they are part of

`Thing/entity/occurrent/processual_entity/Clinical Act` and thus they are no longer modeled as `continuants` but `occurents` that occur in the process of the instanced encounter. The main difference from now on is that the `Clinical Act` can be further divided in the `Clinical investigation act` where the instances collected from the `Assessment` subsection are gathered. It is noticeable, however, that these instances have the same kind of representational artifact as output (and thus a `continuant`) as those that were recorded as part of the `case history` which appears to be consistent, the outputs of an occurrence in the current encounter (diagnostic procedure, laboratory test or any kind of `Clinical investigation act`) are naturally to be incorporated in the patients clinical history.

5.3.3 OGCP enhancements in order to represent healthcare practice episodes

OGCP is an ontology that is meant for *Clinical Practice* in general as the name suggests. To be more fine linked with the clinical aspects of healthcare it has to be further developed into an appropriate representation of the *patient single process* usually found in most *Electronic Health Record* systems.

1. We enhanced the **OGMS** Inpatient Encounter with a data property 'has id' to function as a key to identify the particular encounter found in a **SOAP** report.
2. In the same line we introduced the 'has patientID' key to unequivocally identify that patient.
3. Regarding the physician it's sufficient to use his/her name as identification it so we defined a string data property 'has name' to be used as its key.
4. One significant difference from the originally proposed **OGMS** to our **OGCP** is the sign and symptoms positioning in the ontology hierarchy. Originally both of these were subclasses of **Thing/entity**. The appropriate positioning is to make these both sub-classes of **OGCP:Medical history screening act** and take advantage of the reasoning abilities present in the **Thing/entity/ocurrent/Clinical act/Clinical Investigation act/screening-act/Medical history screening act** hierarchy. These are of course occurrences of sign and symptom records that are to be part of the registered patient clinical history unrelated to the original sign and symptom classes originally positioned by the **OGMS** that still remain present.

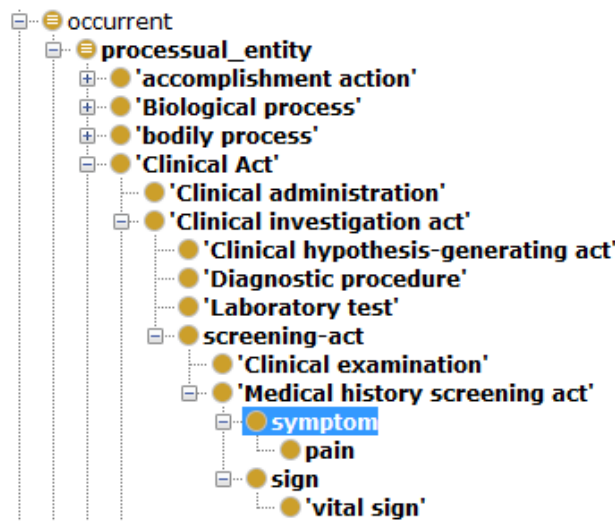


Figure 5.10: Signs and Symptoms in Clinical History

5. For the **Assessment** part of the report we had to develop an `OGCP:analysis` class has a subclass of `OGCP:'Clinical examination'` that will represent every instance found in the **A** subsection of our reports. This positioning enables a large number of inferences to be immediately found by the wealthy ontological structure of the **'Clinical examination'** class. As an example just consider that our *analysis* may have 'has output' some 'clinical artifact' or it investigates some ('etiologic agent' or ('indicated by' some 'Therapeutic act')) already enforced by **OGMS**.

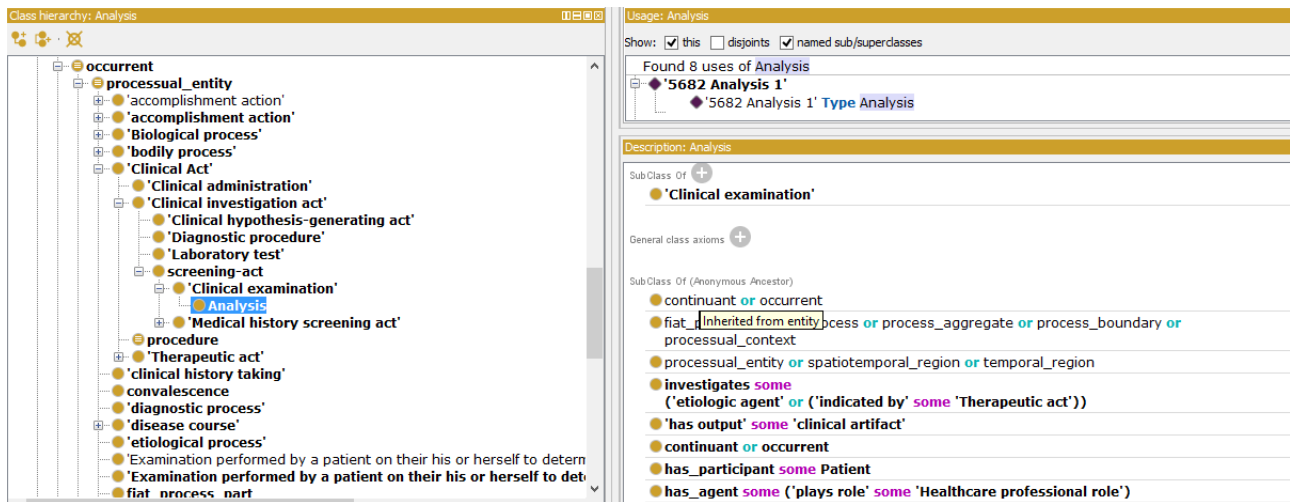


Figure 5.11: OGCP Analysis

6. Regarding the last section, the **Plan**, the support of `OGCP:Therapeutic act` its used, it is sub-classed by `Medical therapy`, `Physical therapy` and `Psychological therapy`. In particular its important to notice that the subclass `OGCP:Substance administration` of `Medical therapy` renders its direct superclass equivalent to a named class "has_participant SOME Medication" which enables the reasoning to be done through the `OGCP:Medication` used.

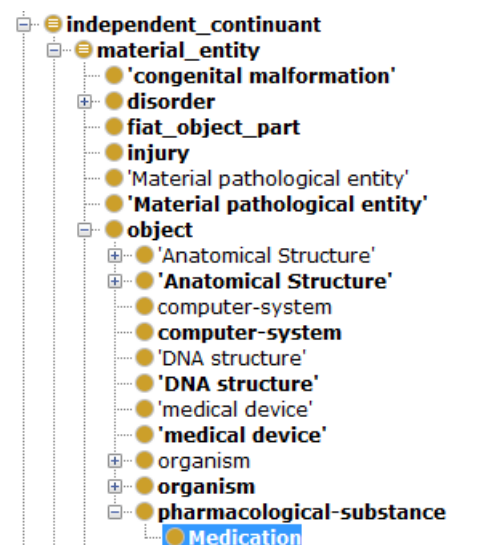


Figure 5.12: OGCP Medication

7. A very important issue to be adequately dealt with is the provenance of each acquired assertion. To maintain a reference back to the *SOAP* report where the instance was collected from we developed properties in *OGCP* to link each of the 4 different possible axioms (**S**ubjective, **O**bjective, **A**ssessment and **P**lan) to the *Inpatient Encounter* where they were extracted from. These have their originating domain in respectively *symptom*, *sign*, *analysis* and *plan* and they all point to (range) the *Inpatient Encounter* individual. Although not immediately evident, these properties have to be dissimilar because there is the need to prevent reasoning to mix signs with symptoms, therapies, exams, and other clinical acts that would occur through ABox classification if the properties were the same.
8. Another important enhancement is the ability to clinically codify the episodes using our chosen standard, *SNOMED-CT CORE*, that enables us the reasoning over its underlying semantic structure. We defined the properties '*XXX with snomed ct code*' where *XXX* stands for, as usual, *symptom*, *sign*, *analysis* or *Therapeutic act* that have their obvious domains and as range the *SNOMED-CT CORE* represented by that particular code. As in the previous case different properties are defined in order to support clear reasoning avoiding the multiple inheritance problem possible in *OWL2*.

5.3.4 DRS rewriting methodology

In the *Clinical Controlled Language* reports that are the subject of our study, the text segments that form the concepts to be acquired and the respective knowledge represented are not necessarily fully well formed natural language sentences.

They are a formally defined mix of acronyms, units and other context driven text segments intended to convey information restricted to domain specialists. This is simultaneously a problem and a benefit to our work. A problem because text segments may not be fully understandable phrases and thus regular English parsers, which are already thoroughly studied and available, might not be suitable for direct usage. An advantage for us because if the used *CCL* is a formally defined jargon regarding its segment structuring we can use commonly available parsers to tag very accurately the token lexical structures which is a fundamental step in our *DRS* rewriting cycle.

To achieve a correct interpretation of each of the candidate lines in the *SOAP* text we proceed by defining a *DRS* representing the tokens (words) in each segment and refine it by rewriting successively into more unambiguous *DRS*. We can think of the steps of the rewriting process roughly ordered as:

1. Controlled Natural Language Text *DRS* (CNL-*DRS*)

In this initial *DRS* the terms that figure in the text are transposed directly to the *DRS*. These possess a substantial degree of ambiguity that will be successively dealt with.

The text segments are isolated in order to be parsed and the lexical structures identified. This identification is a crucial step because the disambiguation algorithm relies heavily in the specific lexical category of the token. The algorithm checks for existing classes when it finds a token tagged with *NN* or *NNS*. For common names the ontology is interrogated for classes named accordingly. In *OGCP* classes are annotated using the *ACElexicon*, which is a superset of our *CNL*, with the singular *CN_sg* for common name singular and *CN_pl* for plural forms of prospective classes. The first step applied here relies on the local *OGCP* ontology interrogation or on Web Service usage for ontology driven generation of candidate classification of terms in

order to render the next level **DRS** already with "coded terms" according to our chosen standards. For instance it will use the local embedded **FMA** to check for anatomical classification but for some correct **SNOMED-CT** code elicitation it has to interrogate **UMLS** through a Web Service in order to disambiguate regarding its semantic relation in the source.

It will make different consultations whether we are checking for subjective, objective, assessment or plan possible candidates. Also the attitude to programmatically select one of the possibilities depends on the source of the candidate line. Ultimately we have to rely on an interactive discourse controller to manually disambiguate.

Regarding the properties they are assigned a sequential **rel.X** prospective name which is then used to interrogate the ontology for its *domain* and *range* to check if they exist and, if found, they are instantiated to the found property and asserted as such.

2. Term interpretation **DRS** (TI-**DRS**)

With every term unequivocally identified with their chosen codes most probably there will be both individuals and properties to be asserted. The former are just direct individual instantiation in the correct class and the later creates the relations between the newly identified individuals. In this step the disambiguation has to be done regarding the identified properties. Once again the local structure of **OGCP** is used. Consulting the domain and range definitions of the properties enables the generation of the different candidate possibilities of the property definition. As an example we can consider the following text: **Cordarone - 1 tablet per day** that rewrites **CNL-DRS** as TI-**DRS**: $\exists P, E, C$ Therapeutic act reported in encounter(P,E), Therapeutic act with snomed ct code(P, SNOMED CT:69236009)

Ultimately our proposal aims at Textual Entailment (**TE**) of Clinical Knowledge from clinical reports. In our case however we aim at getting ahead the severe limitations that traditional Textual Entailment suffers by incorporating reasoning capabilities possible by our clinical and disease model ontology **OGCP**.

So far we have acquired automatically, as much as possible, clinical information. All the ontology population done so far is exemplified in the results on chapter 7, section 7.2.

An important part of the Clinical Knowledge acquisition is, however, yet to be presented because it is the interactive step that relies on the interrogation of our clinical Knowledge Base that is only introduced in the next chapter.

Chapter 6

Clinical Practice Knowledge Interrogation

In our approach we intend to allow the querying of the Knowledge Base (**KB**) to generate inferred valid axioms. This has to be in line with the semantic expressiveness that is leveraged by usage of a Discourse Controller (**DC**) [MRN12] that analyses the question, represents its semantic regarding the expected generation possibilities, and provides the Clinical Controlled Language (**CCL**) answers.

6.1 Clinical reasoning

Our attempts intend to get in the development path of Clinical Decision Support Systems (**CDSS**) based Knowledge Representation (**KR**) built upon Description Logics (**DL**) ontologies. However, several recently identified concerns and limitations about such systems impose that we don't refer to our efforts as directed to **CDSS**, but merely as working aids for healthcare professionals.

At the same time, there is a latent concern about the acceptance by the mentioned professionals of the interference of Artificial Intelligence (**AI**) tools in their modus operandi and their profound sense of scientific, technological and professional independence and ethics. Our proposals will serve, as most of the technological advances do, as a helping tool for the practitioner to have at hand to better perform his/her activities.

The reasoning capabilities demonstrated here will allow only for better, faster, more accurate and safer cross-checking of the innumerable details a healthcare professional has to pay constant attention to. With the enforcement of the modeling capabilities of Ontology for General Clinical Practice (**OGCP**) presented in chapter 4 and the good computational characteristics obtained by the techniques presented in section 2.6 we can have a solid aid for automated support directed at clinical reasoning.

We can distinguish the possibilities in diagnosis help that can be defined as the *estimated identification of the disease by analyzing the signs and symptoms of a patient* [RGLGCP⁺12], and the prescription help provided by the building of an ontology driven "*Clinical Picture*" present in our

system. These guiding steps are illustrated ahead by developing a new concept, Clinical Integrated Discourse Extended Representation Structure (**CIDERS**), along with the reasons why it is included and the description of all the details that have to be considered.

6.2 Clinical concept guidance

Atomic clinical concept recognition, as previously seen, does not need to rely in our semantic representation because it is the result of the Ontology Driven Expanded Semantic Annotation (**ODA**) presented in section 5.2.3. It is, however, a job for our semantic structure to be able to represent the discourse that embraces all the acquired and the inferred knowledge that exists at any given moment. That is, our Discourse Representation Structure (**DRS**) has to be seen as the application of Discourse Representation Theory (**DRT**) to the whole of the Knowledge Base (**KB**) so that the reasoning abilities use it for new knowledge inferring in the ontology enhancement process. This was originally mentioned when introducing the acquisition process in section 5.2, and allows for the implementation of the extended Pragmatic interpretation mentioned in section 5.2.6 that we like to call Clinical Integrated Discourse Extended Representation Structure (**CIDERS**).

CIDERS can be considered as an extension of the semantic representation to the whole set of the, so far acquired, texts while being maintained tractable in **NLP** terms due to its controlled condition [TBC14].

6.3 Discourse Based Enhancement

When we are in the process of Automatic Ontology Learning, explained in section 5.2, all the Knowledge Base is populated from texts.

After the automatic acquisition, the possibility to use the Clinical Controlled Language Question Answering interactive process to further enrich the ontology is opened. The ontology itself is a core part of the system. We have passed the first automated acquisition phase referred in section 3.2, and engage now in the recurring process of interrogation/enrichment.

For the interactive manipulation of the clinical interrogation in Clinical Controlled Language, we make use of **ACE** view plug-in in **Protégé**.

We can, using the available Lexicon and grammar rules, manipulate **ACE** snippets that are **CCL** readable **OWL**.

The snippets are the **CCL** version of our **OWL** representation.

- When only the **OGCP TBox** is represented, we can view it as the generic **CNL** model of Cardiovascular practice.
- When the supervised tutoring phase is completed the resulting **ABox** snippets inform a concrete model of Cardiovascular practice.
- When the Automatic Ontology Learning is taking place, we have a concrete instantiation of that particular practice, we can pose **CCL** interrogations, and eventually enrich the **KB** with further knowledge in the form of clinical snippets.

The overall view of the **ACE** View plug-in is:

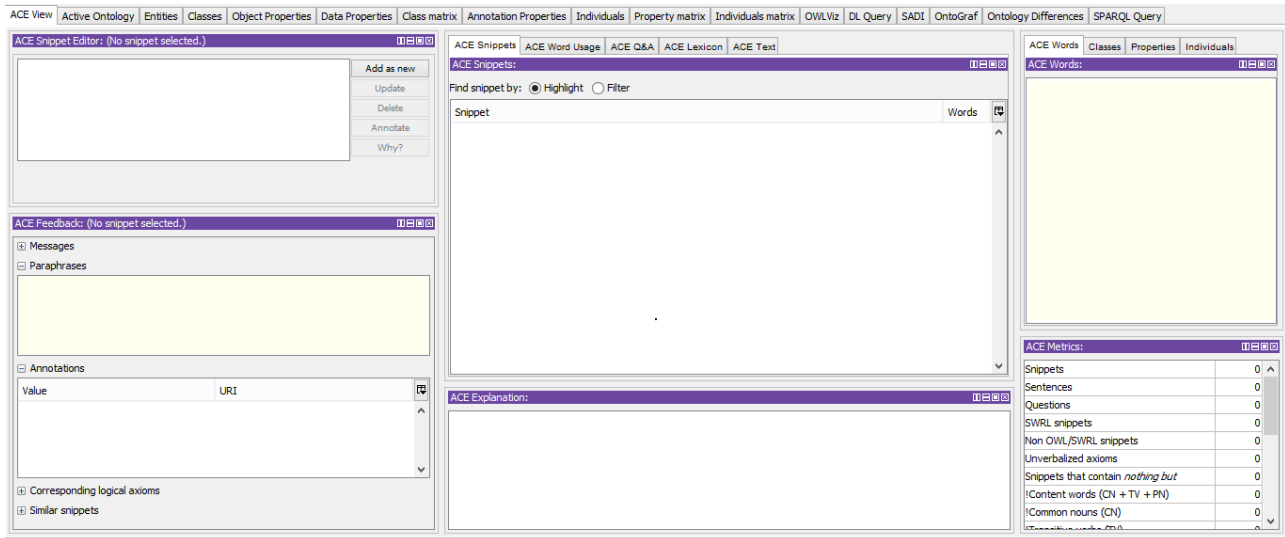


Figure 6.1: ACE View Plugin.

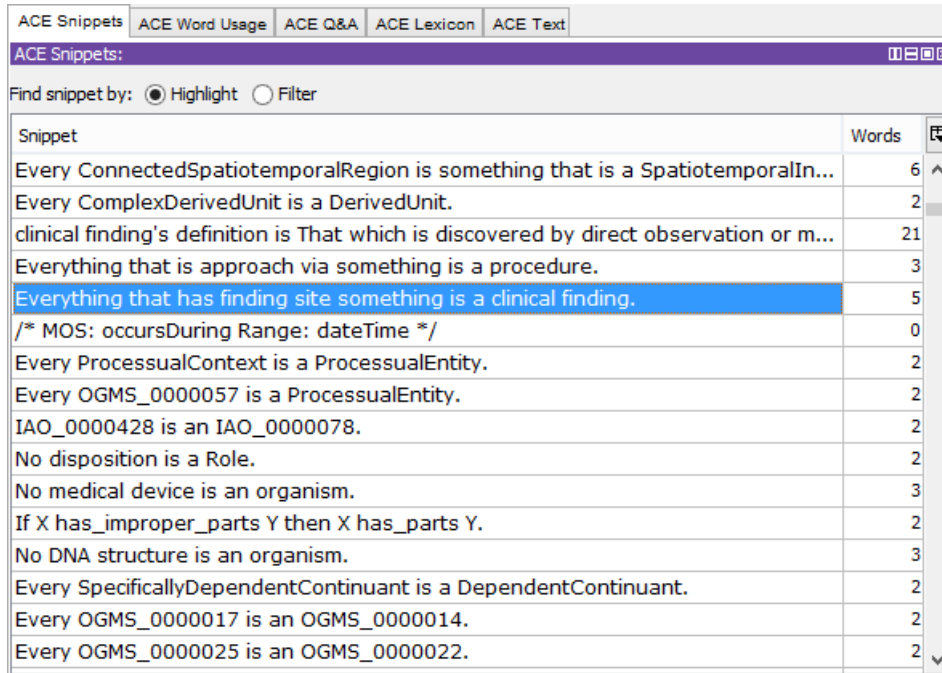
As can be perceived, when the ontology is loaded all the screen parts show very important tools and information.

CNL snippets can be manipulated in the various parts of the plug-in.

We have information about:

- **ACE** Word usage,
- **ACE** Q&A,
Where it's even possible to search for answers that are snippets which contain a word selected from the list of words that appear in the right side of the screen
- **ACE** Lexicon,
Where all the lexical entries are listed with their rendering, lexical type (CN - Common Name, PN - Proper Name or TV - Transitive Verb, for instance), singular and plural forms, verbs past participle and the frequency of each entry in the ontology.
- The **ACE** text
list saturated with all the axioms in the ontology that we can consider to be a human understandable view of our **CIDERS**.

A snippet may be selected in the list:



Snippet	Words
Every ConnectedSpatiotemporalRegion is something that is a SpatiotemporalIn...	6
Every ComplexDerivedUnit is a DerivedUnit.	2
clinical finding's definition is That which is discovered by direct observation or m...	21
Everything that is approach via something is a procedure.	3
Everything that has finding site something is a clinical finding.	5
/* MOS: occursDuring Range: dateTime */	0
Every ProcessualContext is a ProcessualEntity.	2
Every OGMS_0000057 is a ProcessualEntity.	2
IAO_0000428 is an IAO_0000078.	2
No disposition is a Role.	2
No medical device is an organism.	3
If X has_improper_parts Y then X has_parts Y.	2
No DNA structure is an organism.	3
Every SpecificallyDependentContinuant is a DependentContinuant.	2
Every OGMS_0000017 is an OGMS_0000014.	2
Every OGMS_0000025 is an OGMS_0000022.	2

Figure 6.2: ACE View Plugin snippets list.

And then manipulated in the Snippet editor:

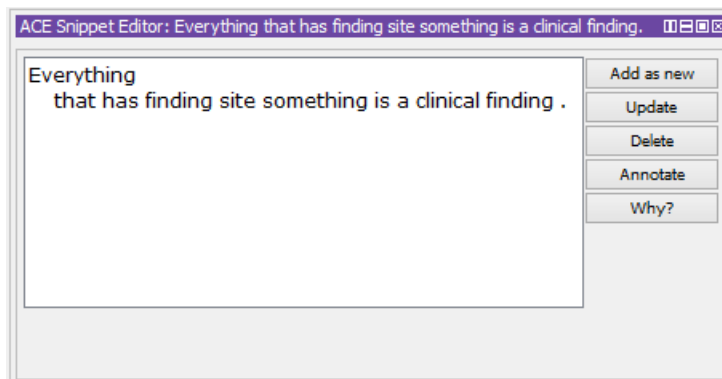


Figure 6.3: ACE View Plugin snippets editor.

Where, as we can see, we can:

- Formulate the **CCL** interrogation and add it as a new axiom enriching interactively the **OGCP** Knowledge Base regardless of it being a terminological *TBox* or a fact *Abox* axiom,
- Update the axiom in a consistent form because only valid **CCL** snippets are allowed,
- Delete it from the Knowledge Base,
- Providing **OWL** annotations to the entry and,

- Get an explanation, that pops in the **ACE** explanation window, about the origin of the snippet.

Both text understanding and generation is strictly related to the acquisition and interrogation enhanced capabilities. For this, we use specially tuned **DRSs** whose development was fully described in section 5.3.

Several considerations about reasoning in our particular domain and chosen \mathcal{DL} have to be previously introduced and we proceed to it in next section.

Controlled Natural Language generation in concrete domains

As we've seen before, the controlled way the **SOAP** reports were built induced the capability of our Clinical Controlled Language semantic processor to provide more easily the answers to the users questions.

As a reminder, terms had from the beginning some formation rules according to [KF07].

CNL has, yet, to be as natural and grammatical as possible for the **CNL** to **OWL** and back translator that we use, inspired in the Grammatical Framework (**GF**) collected from [AR10] that is based in the work in [FKK08a].

Namely the evolutionary usage of the **Protégé** plugin, the Web Service available and finally the *Prolog* based converter that were the tools involved in our work development.

OWL verbalization

The verbalization makes Web Ontology Language (**OWL**) ontologies accessible to people with no training in formal methods. In particular our Domain Experts expressed to feel comfortable when orienting their daily **SOAP** activity reports using our ruling already presented. In section 7.3.1 some preliminary results based in current state-of-the-art work are presented.

Knowledge Base Controlled Natural Language interrogation

We've mentioned already in 6.3 the Discourse Based Enhancement (**DBE**) phase where the Knowledge Base (**KB**) is enriched through interrogation. The same process is used to just query the ontology for clinical guidance. Actually, this one is a part of the other because it just lacks the insertion back to the **KB** of the \mathcal{EL} queries generated in the interrogation process.

Remembering, in the clinical guidance through interrogation the **DC** maintains the formalization of the \mathcal{EL} based in the ontology itself by the realization of the Clinical Integrated Discourse Extended Representation Structure (**CIDERS**). This subprocess is the same as the other one, just without the enrichment step.

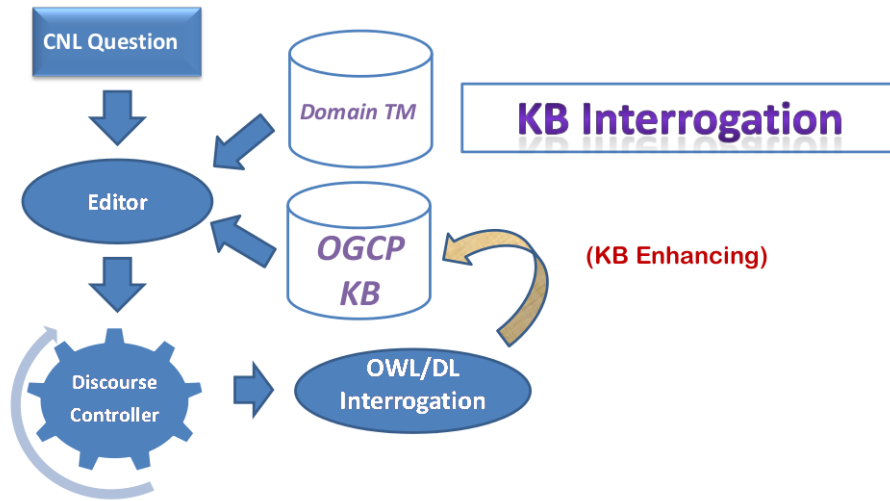


Figure 6.4: Knowledge Base (KB) Enrichment through Interrogation.

6.4 Quality indicators

When trying to achieve comparability between the quality of any work in a given field some measurement scales have to be agreed upon. We focused our work in Knowledge Acquisition from text mining and we shall use the usual methods for comparison here: *Precision*, *Recall* and *F-Measure*. This indexing applied commonly in Artificial Intelligence in the Information Retrieval field will be used along several sub-processes in our work namely for the quality assessment of some of the text processing sub-tasks mentioned in the previous sub-sections but also for measurement in the ontology driven annotation (in section 5.2.3), ontology instance creation (in section 5.2.4) and knowledge inferring (in section 2.4). We use the benchmarking capabilities available in Apache clinical Text Analysis and Knowledge Extraction System (cTAKES) and define a complete plan for this evaluation that provides the quality indicators mentioned in section 8.3.

Chapter 7

Results and Discussion

7.1 System Architecture

The system is a mix of both an ontology **OGCP** and an associated bag of tools for Knowledge Base enrichment from texts.

The ontology is already pre-populated with all the *SHOIN* consistent consequence based axioms derived from the **CVDO** restrictions imposed over the **OGMS** model.

With this expressive knowledge infrastructure in place, the tools for putting up some Clinical Practice reality are provided to Domain Expert to create their "Gold Standard" based in their own **SOAP** reports and posteriorly interact with the **OGCP KB** as a knowledge box.

As presented in [MR11b] we build our proposal based in a lightweight messaging bus that we call Clinical Practice - Enterprise Service Bus (**CP-ESB**).

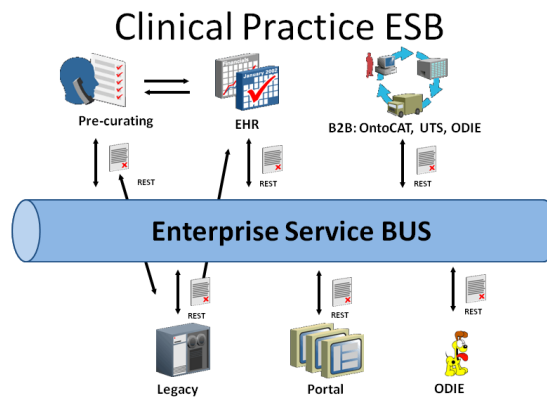


Figure 7.1: **CP-Enterprise Service Bus (ESB)** Architecture.

This RESTful based hub is responsible for orchestrating all the communications between Web Service invocations allowing a high degree of customization and plug-and-play ability that renders our proposal very flexible and future proof.

7.1.1 Translation Memories workflow with CP-ESB

For both the tutoring phase presented in section 5.1 and for the subsequent translation steps in the acquisition from clinical reports we have built a **REST** workflow using the Translation Memory Managers distributed tools mentioned in the 5.1.2 section.

Pictographically this software component can be viewed like this:

TM Workflow with CP-ESB

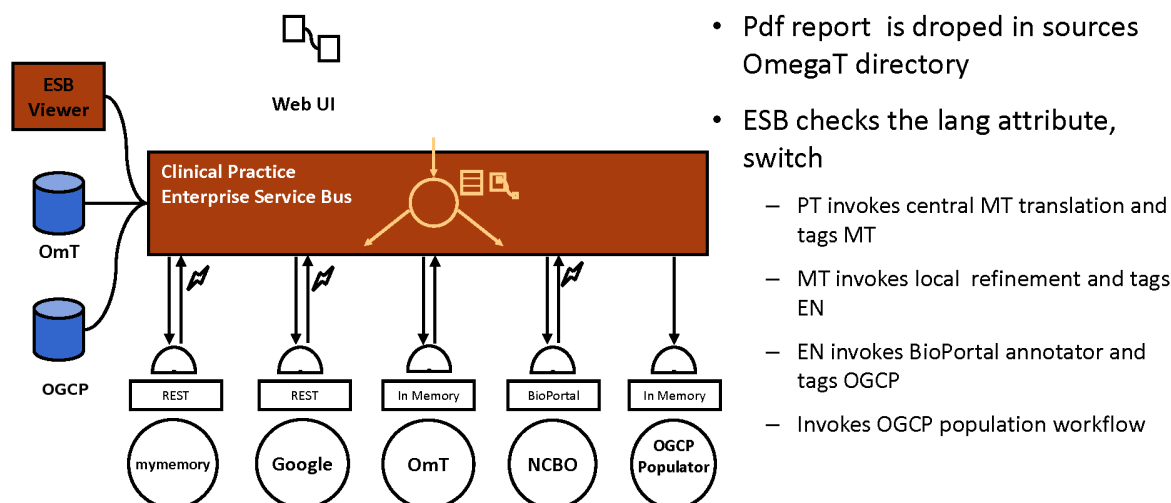


Figure 7.2: **TM**s workflow with **CP-ESB**.

The end-points mentioned are divided in remote (Google translator, MyMemory services and NCBO) and local (OmegaT and OGCP Populator) Web Services. As seen above this is transparent to the Enterprise Service Bus and can be interchanged without incurring in any reconfiguration issues. This provides excellent scalability and implementation independence.

The current **OGCP** populator performs an **XSLT** transformation from the file that results of the extended annotator into the appropriate **OGCP** instance through a Gleaning Resource Descriptions from Dialects of Languages (**GRDDL**) transformation orchestrated by Unstructured Information Management Architecture (**UIMA**). We can view the full acquisition workflow in the following picture where the populator corresponds to the last blue square:

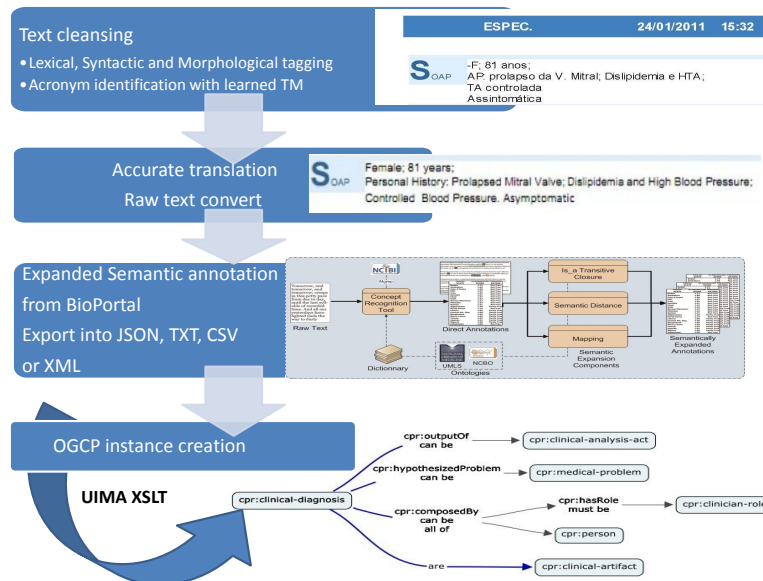


Figure 7.3: OGCP Acquisition Workflow.

Reviewing the referred points in figure 3.2 in section 3.1 from SOAP we can now summarize in the following table the complete acquisition points that fit in the timeline for clinical process acquisition proposed by Scheuermann et al in 2009 [SCS09].

Each identified point corresponds a gleaning entry in `soap_to_ogcp.xslt` transform file summarized in the next table:

Instance	SOAP Section
ogcp:person	H: Header (if not de-identified)
ogcp:patient	H
ogcp:symptom	S
ogcp:sign-finding	O
ogcp:clinical-finding	A
ogcp:clinical_diagnosis	A
ogcp:therapeutic-act	P
ogcp:therapeutic-act	All

Table 7.1: XSLT transformations from XML NCBO Annotation into OGCP instance

The nature of the REST architecture enables us to populate appropriately according to the label inserted in the last step of the TM workflow with the adequate instance tag (`ogcp:person`, `ogcp:patient`, ..., `ogcp:therapeutic-act`).

7.2 OGCP Population examples

To illustrate several significant cases we will use the following report along it's various lines:

HEALTH CENTER
PONTE DE SOR
MAIN OFFICE

Registo Clínico da Consulta

Paciente 5682_SOAP *XXXXXXXX*

Birth Date XX-XX-XXXX (81 Years) *497448616*

XXXXXXXXXX

XXXX XXXXXXXXXX

SPEC.	12/07/2010 18:13	Dr.(a) Carlos Baeta
Dr.(a) Carlos Baeta		
S SOA	Retrosternal pain episodes -Since 2 years. Personal History: - TIA 7 years ago; - Denies Diabetes Melitus.	
O s AP	Blood Pressure- 130/75 mmHg Arrhythmic pulse	
A SOA	Holter(27/05/10)-RS; 51 to 119: M-75; ESSV infrequent T3, T4, TSH - N; AVM and Catecholamines - N	
P SOA	Cordarone - 1 tablet per day Repeat Holter within 6 months for assessment of the need to Arrithmology query	
Comercial Name Qt.		
1 Amiodarona [Cordarone] , 200 mg, Comprimido, Blister - 60 unit(s)		1
Posol.: 1 tablet per day (6 per week)		

Figure 7.4: SOAP Report example

The normal population process proceeds by interrogation of the **Knowledge Base** about any generated axiom, and if no instance (class or individual) is found, it is asserted into the ontology ABox. We describe the DL interrogation/class expressions using the Manchester Syntax for clarity.

1. OGCP:patientID

In the report header we find a number just below the code bar that is opaque to our system, that we don't relate to any personally identifiable information for the purpose of patient anonymization, and we use it to identify the patient. It will be the **patientID** and it is used as a key (functional data property) to all the acquired events related to this patient namely further reports processed in order to enhance the case clinical history for enhanced reasoning capabilities. e.g. 497448616.

patient: 497448616

	A B
DRS:	object(A, atomic, named entity, eq, 1), object(B, atomic, Patient, eq, 1), named(A, 49744861), predicate(B, 'has patientID', A, "497448616")

DL: patient AND 'has patientID' VALUE 497448616^^string

- #A = 0 - patient does not exist
inserts the patient and the associated property: insert patient(A), insert patientID(A, "497448616")
- #A = 1 - patient exists with that patientID
A is instantiated

A is instantiated and its patientID is 497448616.

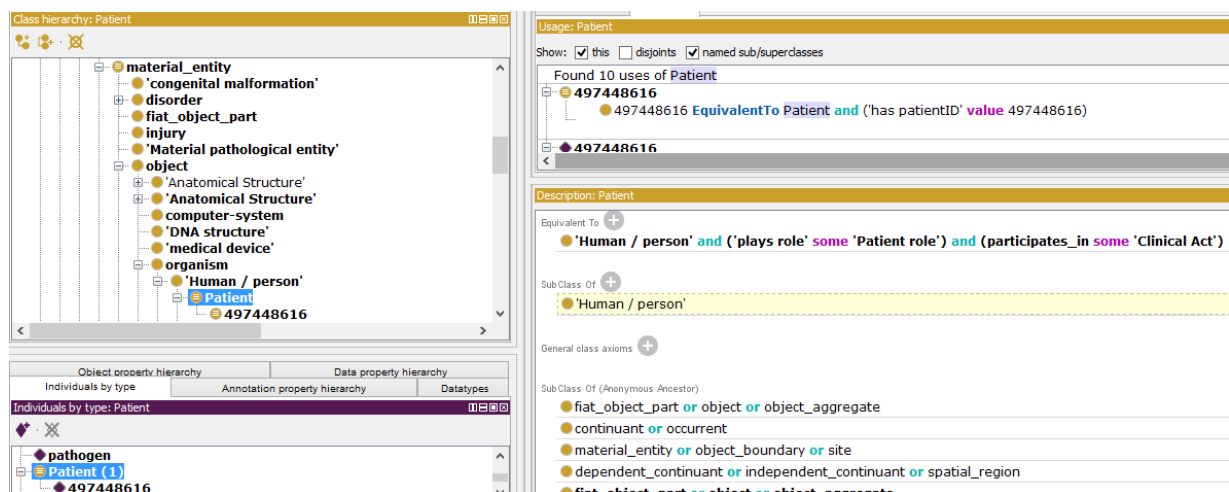


Figure 7.5: OGCP Patient

2. OGCP:Physician

We use the clinicians first and last name to interrogate/populate the clinician in the KB. In all the presented examples, for instance, our trial MD is Dr. Carlos Baeta that is interrogated in our KB and is inserted as an instance of `OGCP:Physician` in the first appearance of his SOAP reports.

physician: Carlos Baeta

	M N
DRS:	object(M, atomic, named entity, eq, 1), object(N, atomic, physician, eq, 1), named(M, "Carlos Baeta"), predicate(N, 'has name', M, "Carlos Baeta")

DL: 'has name' VALUE "Carlos Baeta"^^string AND physician

- $\#M = 0$ - physician does not exist
inserts the physician and the associated property: insert physician(M), insert 'has name'(M, 'Carlos Baeta')
- $\#M = 1$ - physician exists
M is instantiated

M is instantiated with the physicians name

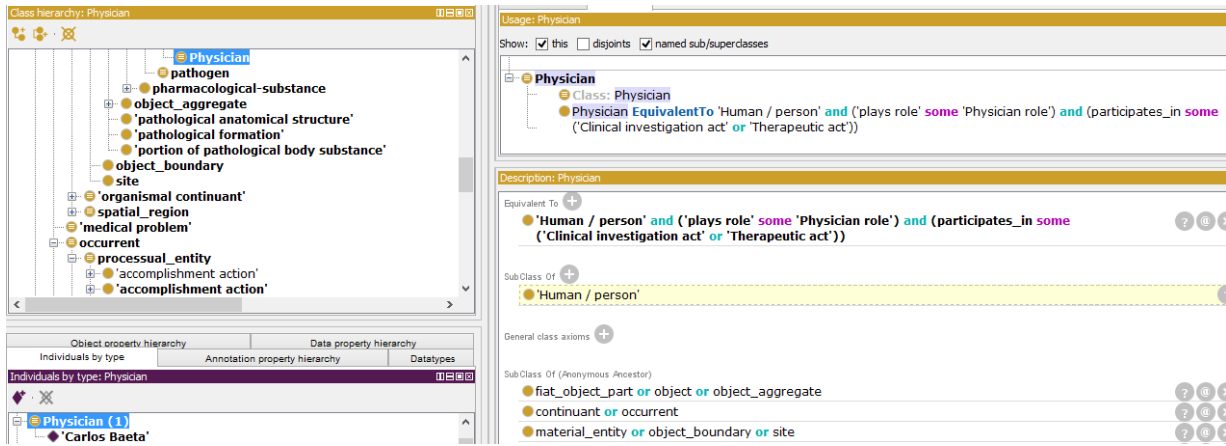


Figure 7.6: OGCP Physician

3. OGCP:inpatient encounter

OGCP:inpatient encounter is a subclass of OGCP:health care encounter that "belongs" to the OGCP:health care process with the full hierarchy as seen ahead in figure 7.7. Each encounter is identified by its EncounterID that's part of the SOAP filename before the string "_SOAP". e.g. 5682.SOAP.pdf is the report for the 5682 inpatient encounter. We fill the encounter data properties by extraction of the DRS variables from the text like the encounter's physician, patient, date and id. We use the encounter properties that exist in our model stating that an encounter has a physician with that specific role, a patient that also exists has a participant and a date which is a data property of the adequate type.

inpatient encounter: 5682

DRS: $\exists E, P, M, D, 'has id'(E, "Inpatient encounter 5682"), inpatient encounter(E, P, M, D)$

E P M D

object(E, atomic, named entity, eq, 1), object(P, atomic, named entity, eq, 1),
object(M, atomic, named entity, eq, 1), named(E, "Inpatient encounter 5682"), named(P, "497448616"),
named(M, "Carlos Baeta"), predicate(E, 'has encounter ID', "5682"), predicate(E, 'has participant', P),
predicate(E, 'has role', M, "Physician"), predicate(E, 'has date', D)

DL: 'inpatient encounter' AND 'has encounter ID' VALUE E^^string AND has_agent SOME M
AND M has_role Physician AND has_participant SOME P AND P has_role patient AND E 'has

date' VALUE D^^string

- #E = 0 - Encounter does not exist
inserts the encounter: insert 'inpatient encounter'(E), insert 'has encounter ID'(E), insert 'has date'(E, D)
- #E = 1 - Inpatient encounter exists
E is instantiated

E is instantiated with the EncounterID

To decorate the individual inpatient encounter the additional assertions are:

has_agent some ('plays role' some 'Healthcare professional role') e.g.

has_agent value ('Carlos Baeta') and has_participant value 497448616.

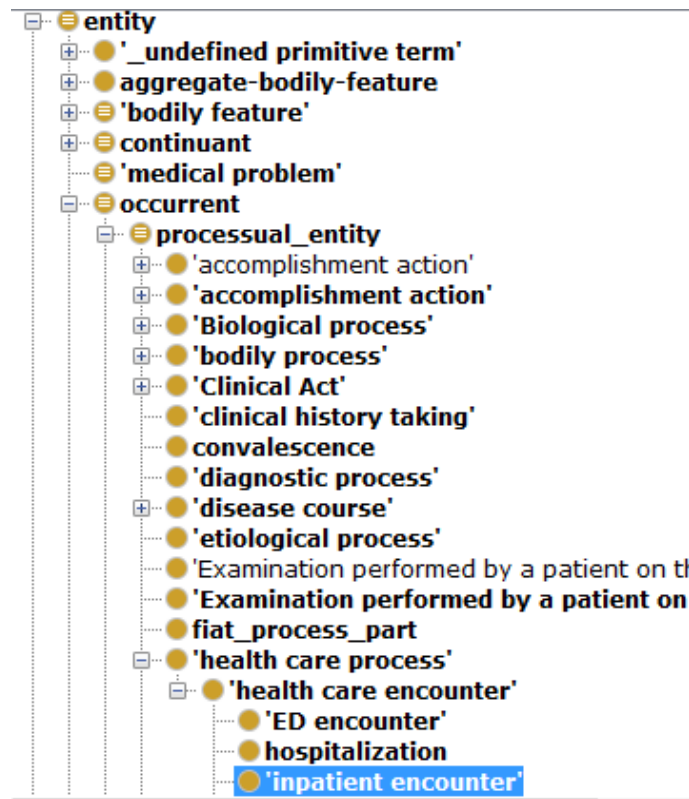


Figure 7.7: OGCP Inpatient encounter

4. Clinical history

One of the main goals that the reasoning procedures, possible with these techniques, provide us is the elicitation of the patients clinical history from the different registered episodes. In a single SOAP report is customary to be verbose about the historical records of signs and symptoms. All these are medical antecedents that can be either OGCP:symptom or OGCP:sign and are to be properly positioned (smart instanced) in their adequate slots. As seen above in the section about the mapping from the SOAP structure to the ontology, sub-section 5.3.2, this first part of the report is used for instantiation of OGCP:Medical history screening act with both it's symptoms and signs. The structure pictured in figure 5.10 reveals the adequate positioning in OGCP to make historical clinical reasoning possible.

(a) OGCP:symptom

A symptom is instantiated serialized by it's found line number in the proper part (Subjective) and the 'found in subjective line' property is filled accordingly.

Then for each different subjective line several possibilities have to be considered but most of them are already ontology driven by the **Controlled Natural Language** approach refined by the tutored translation that occurred when generating the English SOAP report. As examples we illustrate:

- Retrosternal pain episodes

Symptom: Retrosternal pain episodes

CNL-DRS:

S L C T
object(S, atomic, named entity, eq, 1), object(L, atomic, named entity, eq, 1), object(C, atomic, named entity, eq, 1), object(T, atomic, named entity, eq, 1), named(S, "symptom"), named(L, "restroternal"), named(C, "pain"), predicate(T, 'episodes', "≥2"), predicate(S, 'has finding site', L), predicate(S, 'has symptom code', C)

A rewriting step is done for each condition in this **DRS**. Given that 'Retrosternal' is tagged as a proper name, *PN*, the **DRS** indicates that it will have to query the ontology for *OWL individuals* and, in this case, the rewriting of `name(L, 'Retrosternal')` will evaluate `W=SNOMED CT:22253000`. For common names the method evaluates for class expressions taking advantage of the annotations present in the ontology for *CN_sg* and *CN_pl*. In this case *episodes* is found to be a class name in plural form so it will try to evaluate the subsequent relations (properties) for plural " > 1 " cardinality in its range. In our present case, `rel_1(S, L)` is evaluated as a property that relates a *symptom* as domain and a singular *location* as range and it finds the property '**OGCP:has finding site**' as a possibility without no further disambiguation needed. In the case of `rel_2(C, T)` the cardinality, determined previously to be plural by the class name of the occurrence, is found to be only possible with the modality ≥ 2 .

TI-DRS:

S F C
object(S, atomic, named entity, eq, 1), object(F, atomic, named entity, eq, 1), object(C, atomic, named entity, eq, 1), named(S, "symptom"), named(F, "FMA:34688"), named(C, "SNOMED CT:22253000"), predicate(S, 'has finding site', F), predicate(S, 'has symptom code', C)

DL: symptom AND 'symptom reported in encounter' SOME E AND 'has finding site'
 VALUE FMA:34688^^string AND 'has symptom code' VALUE SNOMED CT:22253000^^string

– #S = 0 - Symptom does not exist

inserts the symptom: insert symptom(S), insert 'has finding site'(S,L), insert 'has symptom code'(S, SNOMED CT:22253000), insert 'occurs'(E,"≥2"),
 insert 'symptom reported in encounter'(S, E), insert 'symptom reported in line'(S,
<Line in SOAP report>)

- #S = 1 - Symptom exists
S is instantiated

S is instantiated with the Symptom

In this case the FMA integration in OGCP is interrogated to provide the anatomical code for retrosternal and it is found to be FMA:34688:

```
(Thing/Anatomical Entity/Physical anatomical entity/Material anatomical entity/Anatomical
Structure/Cardinal Organ Part/Organ region/Organ zone/Zone of muscle organ/Zone of pectoralis
major/Zone of sternocostal part of pectoralis major/Sternal part of pectoralis major)
```

Its just asserted the OGCP:has finding site (OGCP:located_in/has finding site) as an object restriction: 'has finding site' some 'Zone of pectoralis major'. 'has finding site' is an object property that relates two individuals, in this case the symptom and the anatomical site individuals, unlike all the previously mentioned that were data properties intended just to decorate the ontology individuals (line numbers, names for physicians or patients, inpatient encounter metadata and so on).

In this case there is no need to use the time framing capabilities of the Episode class through the 'has date' property because it does not have it explicitly defined since it is not clinically relevant. A different situation happens regarding another symptom of the personal clinical history as described ahead in subsection 4a.

Has book keeping efforts we instantiate the 'symptom reported in encounter' and the 'symptom reported in line' data properties.

- Personal history:

As defined in the *CNL* approach, a line ending with a colon like **Personal history:** induces all the subsequent lines until the period is found to be treated as individual symptoms and some noticeable examples are for instance: - **TIA seven years ago;** that indicates a significant historical event that has to be well identified by recurring to its *SNOMED CT CORE* code and time framed by usage of the time framing properties already present in **OGCP**. It is, however, irrelevant where it was first mentioned and consequently instantiated as long has it is found present in the patients history. We use the **Episode** class that has a date framing to instantiate the relationship from an episode to it's timing for it to be clinically relevant.

symptom: TIA seven years ago

DRS:

S C D
object(S, atomic, named entity, eq, 1), object(D, atomic, named entity, eq, 1), named(S, "symptom"), predicate(S, 'symptom with snomed ct code ', C), predicate(S, 'symptom reported in encounter ', E), predicate(S, 'occurs', D)

DL: symptom AND 'symptom with snomed ct code ' SOME C^^string AND 'has date' VALUE D^^string

- #S = 0 - Symptom does not exist
inserts the symptom: insert 'symptom'(S), , insert 'symptom reported in encounter'(S,E),
insert 'occurs'(S, D), insert 'has date'(S, D)

- #S = 1 - Symptom exists
S is instantiated
S is instantiated with the found symptom

(b) **OGCP:sign**

Signs are verified observables that are depicted in the **Objective** section of the **SOAP** report. Correctly positioned now in the clinical history section of our ontology as seen in figure 5.10 the reasoning capabilities are enhanced by the relationships found when several episodes are acquired relating the same patient. As before for the symptoms, for each line found the objective section of our SOAP we create an axiom numbered with the line for adequate filling of the 'sign reported in line' property.

sign: Blood Pressure - 130/75 mmHg

The **SNOMED-CT** integration in **OGCP** is interrogated to provide the code for Blood Pressure and it is found to be **SNOMED CT:75367002** and we insert the value in its preferred unit as a **Dimensional size** data property.

CNL-DRS:

S E C D L
object(S, atomic, named entity, eq, 1), named(S, "sign"), predicate(S, 'sign reported in encounter ', E), predicate(S, 'sign with snomed ct code ', C), predicate(S, 'sign reported in line', L), predicate(S, 'Dimensional size', D),

TI-DRS:

S E C D L
object(S, atomic, named entity, eq, 1), named(S, "sign"), predicate(S, 'sign reported in encounter ', "Inpatient encounter 5682"), predicate(S, 'sign with snomed ct code ', "75367002"), predicate(S, 'sign reported in line', L), predicate(S, 'Dimensional size', "130/75"),

DL: sign AND 'sign reported in encounter ' SOME E AND 'sign reported in line ' SOME L AND 'sign with snomed ct code' SOME "75367002" AND 'Dimensional size' VALUE "130/75"^^string

- #S = 0 - Sign does not exist
inserts the sign: insert 'sign'(S), insert 'sign reported in encounter '(S, E), insert 'sign with snomed ct code'(S, "75367002") , insert 'Dimensional size'(S, "130/75")
- #S = 1 - Sign exists
S is instantiated

S is instantiated with the found sign

sign: Arrhythmic pulse

The **SNOMED CT** integration in **OGCP** is interrogated to provide the code for 'Arrhythmic pulse' and but it is not present as such. A enhanced version of 'find snomed ct code' tries to

find a synonym concept and by interrogation of the integrated VSO ,the *Vital Sign Ontology*, it finds the equivalent concept **Irregular pulse** that is found to be SNOMED-CT:61086009. The value is asserted using 'sign with snomed ct code' along with the original text as an annotation to the found sign as usual.

TI-DRS:

S E C D L
object(S, atomic, named entity, eq, 1), named(S, "sign"), predicate(S, 'sign reported in encounter', "Inpatient encounter 5682"), predicate(S, 'sign with snomed ct code', "SNOMED-CT:61086009"), predicate(S, 'sign reported in line', L), predicate(S, 'Dimensional size', D),

DL: sign AND 'sign reported in encounter' SOME "Inpatient encounter 5682" AND 'sign reported in line' SOME L AND 'sign with snomed ct code' SOME "SNOMED-CT:61086009"

- #S = 0 - Sign does not exist
inserts the sign: insert 'sign'(S), insert 'sign reported in encounter '(S, E), insert 'sign with snomed ct code'(S, "61086009")
- #S = 1 - Sign exists
S is instantiated

S is instantiated with the found sign

(c) OGCP:Analysis

The **Assessment** section of the report is no longer part of the *Clinical History*. Although they are still 'Screening acts', exams and analysis performed in the encounter are not considered part of the 'Medical history'. The correct positioning in the *OGCP* hierarchy for enhanced reasoning possibilities is then:

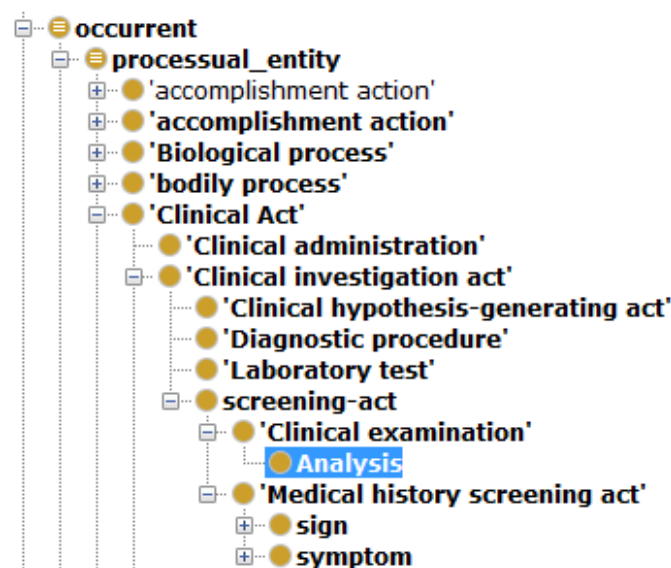


Figure 7.8: Analysis hierarchy in OGCP

Analysis: Holter (27/05/10)

There is not much difference regarding the ontological structure and acquisition point of views from the **signs** and **symptoms** to the **Analysis**. The **SNOMED CT** integration in **OGCP** is interrogated to provide the code. Given that we are in the *A* section we chose the correct result from those found in *SNOMED CT* to be the procedure named *Holter extended electrocardiographic recording (procedure)* as **SNOMED-CT:427047002**. The assertions made in the *ABox* are the usual. In the following **DRS** the **T** is the original text **Holter** with the date omitted because the date can not be entered as such due to the **OWL2** limitation when representing dates that are not in conformance to the usual structure in Portugal: DD/MM/YY.

Everywhere when a date is found, as explained in 4a, it is passed to the optional '**has date**' data property assertion as a string.

CNL-DRS:

A E T C D L
object(A, atomic, named entity, eq, 1), named(A, "Analysis"), predicate(A, 'Analysis reported in encounter', E), predicate(A, 'Analysis reported in line', (Line in SOAP report)), predicate(T, 'find snomed-ct code', C), predicate(S, 'Analysis with snomed ct code', C)

TI-DRS:

A E T C D L
object(A, atomic, named entity, eq, 1), named(A, "Analysis"), predicate(A, 'Analysis reported in encounter', "Inpatient encounter 5682"), predicate(A, 'Analysis reported in line', (Line in SOAP report)), predicate(S, 'Analysis with snomed ct code', "SNOMED-CT:427047002")

DL: Analysis AND 'Analysis reported in encounter' SOME E AND 'Analysis reported in line' SOME L AND 'Analysis with snomed ct code' SOME "SNOMED-CT:427047002" AND 'has date' VALUE D[^]string

- #A = 0 - Analysis does not exist
inserts the Analysis: insert 'Analysis'(A), insert 'Analysis reported in encounter'(A, "Inpatient encounter 5682"), insert 'Analysis reported in line'(A, (Line in SOAP report)), insert 'Analysis with snomed ct code'(A, "SNOMED-CT:427047002"), insert 'has date'(A, D)
- #A = 1 - Analysis exists
A is instantiated

A is instantiated with the found Analysis

Finally,

(d) **OGCP:Therapeutic act**

In the last section of a SOAP report, the **Plan** section, all the planned therapeutic actions are enumerated. In *OGCP* we leveraged the *Therapeutic act* class to include all the actions to be taken. As can be seen in

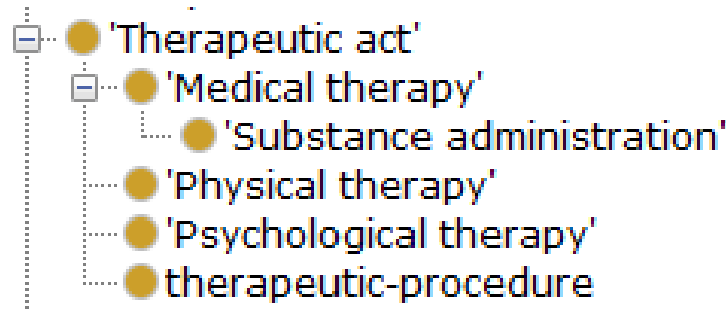


Figure 7.9: Therapeutic act hierarchy in OGCP

Therapeutic act encompasses the various types of therapy administered: Medical (Substance administration), Physical, Psychological and the more generic Therapeutic procedure. In the subclass *OGCP:Substance administration* it is important to notice that it is a subclass of both *OGCP:Medical therapy* and of the anonymous class: *'has-participant SOME Medication'*. This fundamental particularity allows for extended reasoning possibilities when filling the *Medication* range with the correct pharmaceutical code. It is gathered by interrogation of *SNOMED CT* using its active compound name and benefits greatly from the *OGCP:is IndicationFor*, *OGCP:isContraindicationFor* as well as *OGCP:has contraindication* for automatic prevention of adverse reactions and the correct usage of the medication.

Therapeutic act: Cordarone 1 tablet per day

The point interesting to be illustrated here is the periodicity that is included in the ABox by usage of a base framing property already present in *OGCP* due to its integration of the *OWL-Time* ontology. We follow the generic techniques introduced in [XW] and in this example we instantiate the *ogcp:intervalAfter* eventuality [HP04] to link the medication timing (event) to its periodicity. The *SNOMED-CT* integration in *OGCP* is interrogated to provide the code of the named medication. Given that we are in the *P* section we pick the correct result from those found in *SNOMED CT* to be the *Pharmacological Substance* *SNOMED-CT:69236009*. That assertion *per se* induces automatically the drug to be a *Class III antiarrhythmic drug* as defined in the *SNOMED CT* ontological hierarchy granting us, for instance, all the underlying adverse reactions and the correct posology possibilities. The assertions made in the *ABox* are similar to those found in the previous sections. In the following *DRS* *P* is the planned therapeutic act, *E* the inpatient encounter, *T* is the original text "Cordarone" and the periodicity is *F* "one tablet per day".

The substance administration is linked to the inpatient encounter event using *OGCP:afterInterval* because it is assumed that it is a substance to be administered starting immediately after the reported encounter.

CNL-DRS:

P E T C F
<p>object(P, atomic, named entity, eq, 1), named(P, "Therapeutic act"), predicate(P, 'Therapeutic act reported in encounter', E), predicate(P, 'find snomed ct code', T), predicate(P, 'Therapeutic act reported in line', L), predicate(P, 'Therapeutic act with snomed ct code', T), predicate(P, 'intervalAfter', F),</p>

TI-DRS:

P E T C F
<p>object(P, atomic, named entity, eq, 1), named(P, "Therapeutic act"), predicate(P, 'Therapeutic act reported in encounter', "Inpatient encounter 5682"), predicate(P, 'Therapeutic act reported in line', $\langle \text{Line in SOAP report} \rangle$), predicate(P, 'Therapeutic act with snomed ct code', "SNOMED-CT:69236009"), predicate(P, 'intervalAfter', 1),</p>

DL: Therapeutic act AND 'Therapeutic act reported in encounter' SOME E AND 'Therapeutic act reported in line' VALUE L^{integer} AND 'Therapeutic act with snomed ct code' VALUE SNOMED CT:69236009 AND 'Therapeutic act'.intervalAfter VALUE 1^{Integer}

- $\#P = 0$ - Therapeutic act does not exist
 inserts the Therapeutic act: insert 'Therapeutic act'(A), insert 'Therapeutic act reported in encounter'(P,E), insert 'Therapeutic act reported in line'(P, $\langle \text{Line in SOAP report} \rangle$), insert 'Therapeutic act with snomed ct code'(P, "SNOMED CT:69236009"), insert 'intervalAfter'(P, 1)
- $\#P = 1$ - Therapeutic act exists
 P is instantiated

P is instantiated with the found Therapeutic act (Plan)

7.3 Current on-going controlled results

Remembering our illustrative examples extensively refined along the Discourse Based Enhancement sections 6.3 we wrap-up here the achieved controlled results.

- Q: What is the patient's personal history?
 A: Hypertension for 15 years; Diabetes Mellitus type 2 for 10 years; Cholecystectomy 2 years ago; Diabetic father; Obese BMI 26,5; Abdominal perimeter 106 cm.

This is a direct expression of the entailment of having several Clinical History entries in **CIDERS**. The **OWL** verbalizer just tries to present the best possible, with higher usage evaluation in the trained Grammatical Framework, expression in the time framing of the represented events in the *OWL-Time* properties related to that patient.

- Q: What is the suggested diagnosis?
 A: Laboratory routines: lipid profile; HgA1c; Rx thorax; ECG in rest; Echocardiogram; Effort test (Effort proof or Cardiac scintigraphy);

Suggested diagnosis triggers the mark words **Suggested** and **diagnosis**. With **Suggested**, ELK rules are automatically generated that try to satisfy the set of inferences that are consistent in the **KB** for every patient that has the same, or non-contradictory set of conditions modeled by **CVDO** in top of **OGMS**. Diagnosis suggested then are all those already part of the historical practice (**ABox CIDERS**) or enforced by the underlying ontological framework. The verbalizer then proceeds as in the previous example.

- Q: What is the immediate recommended therapy assuming that AHT and Diabetes are not controlled ?

A: Rich fiber and vegetable diet; polifracctionate and hiposaline; IECA or ARA II; Calcium Antagonist; Metformine; Estatine;

Here we have the most challenging inferencing situation. The sole generation of hypothesis based in the set of assertional $\mathcal{EL}++$ facts aren't enough to sustain the question any longer. This situation explore the additional VNN: $\neg \text{controlled}(\text{Diabetes})$ and $\neg \text{controlled}(\text{AHT})$ to trigger further rules in order to generate all the consequences consistent with the new negated hypotheses (that have to be syntactically pre-processed as seen in 5.2.7).

These are, of course, highly controlled results that are only possible so far under the interactive set of **Protégé** and the set of plugins detailed in section 2.11.1.

7.3.1 Domain experts validation

At the time of writing this thesis the team of Domain Experts in Cardiovascular diseases from **ULSNA** are proceeding with an assessment of the merits and usefulness of our proposed architectures and solutions. Recently a contribution regarding the acceptance of **CNL** in specific domains has been provided by Tobias Khun in [Kuh13] where the results show that **CNL** is easier to understand, needs less learning time, and is more accepted by its users. Facing this, which was apparently evident to our collaborators we engaged in an informal evaluation using our controlled results but the plan to formally evaluate our set of tools performance are detailed in section 8.3.

Chapter 8

Conclusions

8.1 Conclusions

We presented our proposal for a knowledge representation infrastructure for Clinical Practice enabling the usage of highly optimized distributed consequence based reasoners that are referred in literature only in 2011.

With these very recent developments it's finally possible to validate a controlled size Clinical Practice Knowledge Base.

This **KB** is created by automatically populating a proposed Ontology for General Clinical Practice (**OGCP**) that relies on extensive, very solid, foundations like Standard Nomenclature of Medicine - Clinical Terms and Foundational Model of Anatomy among others.

We extensively demonstrate a solid approach to overcome the "Knowledge acquisition Bottleneck" by using an automated acquisition process into a highly tractable Knowledge Base directly from clinical text reports based in the well known Subjective, Objective, Assessment, Plan methodology.

Dividing the Knowledge Base between an expressive foundation (*SHOIN*) that relies in the coordination of **OBO** foundry set of ontologies and a less expressive, but highly effective in computational characteristics (\mathcal{EL}) *ABox*, we render a knowledge infrastructure with very interesting properties for Clinical Practice representation and reasoning in the Cardiovascular domain.

We introduce clinical reasoning aids that are based on such breakthrough techniques.

We also show how to maintain the size of the **OGCP** ontology very limited in order to be able to apply these innovative Artificial Intelligence advances and techniques in commodity hardware.

Logical inferencing and clinical facts entailment that is possible through this capability is an interesting contribution to the application of Artificial Intelligence to healthcare.

A number of restrictions surfaced during the development of this work that prevented us to achieve better results

- State of the Art restrictions

1. Foundational ontology weaknesses

Research and development in the different *OBO Foundry* ontologies is undergoing steadily to overcome many limitations that are surfacing as more research and implementations are available. Being the **OGCP** such a complex artifact relying on so many foundational nuts and bolts, it is highly exposed to that evolution, and still has many quirks, with the best solution yet to be devised. Here are some examples that are currently undergoing strong development efforts:

- **OWL-Time** In parallel with the enforced limitations that our Knowledge Acquisition (**KA**) procedures incur, some restrictions arise from the OWL-Time known non completeness. These are well defined in [HP04] also in the future directions section as being:
 - * **Temporal Arithmetic**
e.g. January 31, 2003, plus 3 months,
 - * **Deictic Time**
e.g. now, today, tomorrow night, last year,
 - * **Vague Temporal Concepts**
e.g. "soon", "recently", "late" or a "little late" which require an underlying theory of "vagueness".
 - * **Aggregates or Temporal Entities**
e.g. every Wednesday, although mentioned in [HP04] is already supported in OWL-Time ontology since the works developed in 2005 presented in [Pan05, PH05] are already included in the W3C version of the ontology [W3C06].
- **Symptom Ontology**
The Symptom Ontology is yet in a very infant state regarding the Cardiovascular specialty. There are no strong ontological relations that go further the simple *general medicine symptomatology*. It needs to be heavily curated in order to the specific cardiac system related symptoms associated with heart and coronary diseases enforce specifically that model. So far, it's essentially the **OGMS** model of generic disease complemented with the Disease Ontology (**DO**) and Cardiovascular Disease Ontology (**CVDO**) that provide clinical structure to **OGCP** as intended. It is not, however, a contra-sense to have **SO** as a basis for symptomatology because philosophically, **OGCP** is geared towards general Clinical Practice although for practical reasons we have restricted it to the Cardiovascular specialty in this work.

- **Time or other type of constraints**

During the time this PhD was being developed some time constraints became evident mainly regarding the relations with third parties that had severe scheduling limitations like the possibility of hand validation of the generated Translation Memory by the Cardiology domain specialists. Also, the quality indicators assessment presented in 6.4 is currently undergoing.

- **Resource access limitations**

Initially planned for show-case study, the usage of the data in the *Mozambican Health Integrated Information System* was not available at the moment of this thesis writing and so it could not be used. Possibly it's yet a valid proof-of-concept to be developed.

- Extension of the concept to wider geographical coverage

So far only the **ULSNA** proof-of-concept was developed. In this environment, in the application to the Cardiovascular specialty it is easy to develop a consensus about the usefulness of the ideas and tools showcased because it may be considered fairly limited and controlled.

The author's original idea of widespread acceptance and usage of such a system in **CDSSs** in low income countries of Africa and Asia, for instance, has a huge and difficult obstacle to overcome. This problematic issue is not related with technical or budget reasons albeit with the ethical and professional possible doubts about letting Artificial Intelligence techniques interfere in the medical decision process.

It will, eventually, become a non-issue if the justification framework renders the reasoning process really self-evident. That is, if the knowledge inferred and proposed to the practitioner is presented with all the clinical reasoning clearly justified, the doctors will balance the system advantages versus the possible ethical dilemma of having computer systems interfering in their scientific and professional ability.

- Extension of the concept to different clinical specialties

The disease model in our proposal is generically enforced and maintained by the **OGMS** usage but any application to a given clinical specialty has to be instantiated by a specific domain ontology. In the present case it is the **CVDO** that embodies this instantiation but, as mentioned in chapter 4, we can consider the specific domain ontology to be a movable, interchangeable part for any other specific application within medical science.

Of course that a serious ontology engineering work has to be done to guarantee the assurance of the clinical validity of that application and this involves the deep collaboration between ontological engineers and medical doctors in the given sub-domain.

8.2 Future Work

Mainly all the work that is proposed here represents the authors intention to evolve the proof of concepts developed so far into wider realms of application to show the applicability and importance of this line of research.

- Develop a software module to complement the Automatic Ontology Learning (**AOL**) tasks to allow further interactive refinement of the **CCL** Gold Standard Translation Memories for progressive translation accuracy enhancement. The current Clinical Controlled Language definition is static relative to it's formation from the **OGCP** basis. This can be seriously improved if, in the process of Discourse Based Enhancement, we could use the new acquired clinical expressions and equivalent **OWL** representations to enhance the accuracy of the Controlled Natural Language translations.
- Develop a software feature to alert the practitioner that some assertion is not representable with \mathcal{EL}^{++} .

This should be done as early as possible in the **KB** definition process in order for the situation to be prevented because the later it is solved the more difficult it turns to apply the syntactic variations needed to maintain the consistency. As explained in section 5.2.7 we have to apply modalities to enforce Closed World Assumption turning the cycle increasingly onerous when

it can be easily prevented if the **CCL** expression applied is already conform to the **CIDERS** generation that is $\mathcal{EL}++$.

This is however tricky to be balanced, we simultaneously try to hide the \mathcal{DL} complexities by applying the **CCL** abstraction over it but intend the user to be aware of their intricacies for it to be rendered more effective. This is one of the main reasons why interactive **KB** creation processes like graphical, **QA** cooperative refining or oriented **CNL** building are so far widespread adopted in industry.

8.3 Comparable high standard formal results evaluation

For the usual Natural Language Processing (**NLP**) performance formal results **F-measure**, **Precision** and **Recall** to be usable when evaluating, the comparing platforms need to have the least number of differences in terms of objective, scope, domain, underlying tools, implementation infrastructure and more. Given the set of base tools on which our system is implemented, the Apache Software Foundation (**Apache**) **NLP** suite, comparable systems are immediately those that pertain to a similar genealogy like the most widely implemented in the world so far, directed to Clinical text, the **Apache cTAKES** platform.

Aiming at establishing a high comparable standard in specific clinically oriented **NLP** the **cTAKES** team performed a very serious development effort in 2010 [SMO⁺10]. *This project takes a foundational step towards bringing the field of clinical NLP up to par with NLP in the general domain. The corpus creation and NLP components provide a resource for research and application development that would have been previously impossible.*

All the evaluation parameters are detailed in the cited article thus we refrain to replicate them here but since our intention is to mimic the **Mayo clinic** experiment we will develop the statistically equivalent samples and will evaluate all the respective tools that we have aimed at the Cardiovascular specialty and **SOAP** texts as source.

We shall use a subset of the recently open sourced manually annotated corpus [ALF⁺13] that consists of 13091 sentences containing 1772 distinct predicate lemmas, 28539 named entity annotations spread over 15 **UMLS** semantic groups. The most frequent annotations belong to the **UMLS** semantic groups of Sign or Symptom (12.46%), Disorders (14.74%), Anatomy (12.80%), Procedures (15.71%), Concepts and Ideas (15.10%), and the **UMLS** semantic type of Chemicals and Drugs (7.49%). Obviously we compare only to the first 4 semantic groups when applied to our different **SOAP** origins as explained in section 5.2.3.

Trying to further maintain comparability as accurate as possible we will chose the equivalent individual components:

- sentence boundary detector - accuracy=0.949;
- tokenizer - accuracy=0.949;
- part-of-speech tagger - accuracy=0.936;
- shallow parser - F-score=0.924;
- named entity recognizer -
system-level F-score=0.715 for exact and 0.824 for overlapping spans, and accuracy for concept

mapping, negation, and status attributes for exact and overlapping spans of 0.957, 0.943, 0.859, and 0.580, 0.939, and 0.839, respectively.

Since the project open sourced both the corpus, tools for evaluation and statistical tools and methodology in late 2013 we are now able to perform a comparable evaluation of our tools.

We selected a sample of source SOAP reports and applied the same distribution as the original study, we divided the corpus into training, development, and evaluation sets (85%, 5%, and 10% respectively).

As soon as our clinical partners are available to perform the current plan the results will be available and published presumably in the next pair of months.

References

- [AAD⁺09] Dimitra Alexopoulou, Bill Andreopoulos, Heiko Dietze, Andreas Doms, Fabien Gandon, Jörg Hakenberg, Khaled Khelif, Michael Schroeder, and Thomas Wächter. Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy. *BMC bioinformatics*, 10:28, January 2009.
- [ABB⁺00] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May 2000.
- [ACMS12] James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from swirl 2012 the second strategic workshop on information retrieval in lorne. *SIGIR Forum*, 46(1):2–32, May 2012.
- [AFT04] Sophia Ananiadou, Carol Friedman, and Jun’ichi Tsujii. Introduction: named entity recognition in biomedicine. *Journal of Biomedical Informatics*, 37(6):393–395, 2004.
- [AL94] José Abraços and José Gabriel Lopes. Extending drt with a focusing mechanism for pronominal anaphora and ellipsis resolution. In *Proceedings of the 15th conference on Computational linguistics- Volume 2*, pages 1128–1132. Association for Computational Linguistics, 1994.
- [ALF⁺13] Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F Styler, Colin Warner, Jena D Hwang, Jinho D Choi, Dmitriy Dligach, Rodney D Nielsen, James Martin, et al. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, 20(5):922–930, 2013.
- [Apa14a] Apache Software Foundation. Apache uima ruta, 2014. <http://uima.apache.org/ruta.html/>.
- [Apa14b] Apache Software Foundation. openNLP Natural Language Processing Library, 2014. <http://opennlp.apache.org/>.

- [AR10] Krasimir Angelov and Aarne Ranta. Implementing controlled languages in gf. In *Controlled Natural Language*, pages 82–101. Springer, 2010.
- [ASBYG11] Omar Alonso, Jannik Strötgen, Ricardo A Baeza-Yates, and Michael Gertz. Temporal information retrieval: Challenges and opportunities. *TWAW*, 11:1–8, 2011.
- [Att07] Attempto Project. Attempto Controlled English (ACE), 2007. <http://attempto.ifi.uzh.ch/site/description/>.
- [Att10] Attempto Project. Ace 6.6 syntax report. 2010., 2010. http://attempto.ifi.uzh.ch/site/docs/ace/6.6/syntax_report.html.
- [Aug05] Juan Carlos Augusto. Temporal reasoning for decision support in medicine. *Artificial Intelligence in Medicine*, 33(1):1 – 24, 2005.
- [Baa03] Franz Baader. *The description logic handbook: theory, implementation, and applications*. Cambridge university press, 2003.
- [BBC⁺10] Janis Barzdins, Guntis Barzdins, Karlis Cerans, Renars Liepins, and Arturs Sprogis. Owlgrid: a uml style graphical notation and editor for owl 2. In *OWLED*, 2010.
- [BBL05] Franz Baader, Sebastian Brand, and Carsten Lutz. Pushing the EL envelope. In *In Proc. of IJCAI 2005*, pages 364–369. Morgan-Kaufmann Publishers, 2005.
- [BBL08] Franz Baader, Sebastian Brandt, and Carsten Lutz. Pushing the el envelope further. 2008.
- [BC08] Paul Buitelaar and Philipp Cimiano. *Ontology learning and population: bridging the gap between text and knowledge*, volume 167. Citeseer, 2008.
- [BCM⁺07] Franz Baader, Diego Calvanese, Deborah L McGuinness, Daniele Nardi, and Peter F Patel-Schneider. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2007.
- [Ber14] Berkeley Bioinformatics Open-source Projects. Obolib obo2owl converter, 2014. <http://www.berkeleybop.org/software/obolib-obo2owl-converter>.
- [BFO04] IFOMIS BFO. Basic formal ontology, 2004.
- [BLHL01] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001.
- [BZC10] Aurélien Béné, Chao Zhou, and Jean-Pierre Cahier. Beyond web 2.0... and beyond the semantic web. In *From CSCW to Web 2.0: European Developments in Collaborative Design*, pages 155–171. Springer, 2010.
- [CDGL⁺07] Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. Tractable reasoning and efficient query answering in description logics: The dl-lite family. *Journal of Automated reasoning*, 39(3):385–429, 2007.
- [CdKAH06] R Cornet, N F de Keizer, and a Abu-Hanna. A framework for characterizing terminological systems. *Methods of information in medicine*, 45(3):253–66, January 2006.

- [CES06] Werner Ceusters, Peter Elkin, and Barry Smith. Referent tracking: The problem of negative findings. *Studies in health technology and informatics*, 124:741, 2006.
- [Cim98] James J Cimino. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of information in medicine*, 37(4-5):394, 1998.
- [Cim06] James J. Cimino. In defense of the desiderata. *Journal of Biomedical Informatics*, 39(3):299 – 306, 2006. Biomedical Ontologies.
- [COU11] COUNCIL OF THE EUROPEAN UNION. A comprehensive approach on personal data protection in the european union. Memorandum, European Union, February 2011.
- [CS10] Werner Ceusters and Barry Smith. A unified framework for biomedical terminologies and ontologies. *Studies in health technology and informatics*, 160(Pt 2):1050, 2010.
- [CSKD04] W. Ceusters, B. Smith, A. Kumar, and C. Dhaen. Mistakes in medical ontologies: where do they come from and how can they be detected? *Stud Health Technol Inform.*, 2004.
- [DCFKK09] Juri Luca De Coi, Norbert E Fuchs, Kaarel Kaljurand, and Tobias Kuhn. Controlled english for reasoning on the semantic web. In *Semantic techniques for the web*, pages 276–308. Springer, 2009.
- [DDH⁺11] Ronald Denaux, Catherine Dolbear, Glen Hart, Vania Dimitrova, and Anthony G Cohn. Supporting domain experts to construct conceptual ontologies: A holistic approach. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2):113–127, 2011.
- [DFMSA10a] Dina Demner-Fushman, James G. Mork, Sonya E. Shooshan, and Alan R. Aronson. UMLs content views appropriate for nlp processing of the biomedical literature vs. clinical text. *Journal of Biomedical Informatics*, 43(4):587–594, Aug 2010.
- [DFMSA10b] Dina Demner-Fushman, James G Mork, Sonya E Shooshan, and Alan R Aronson. UMLS content views appropriate for NLP processing of the biomedical literature vs. clinical text. *Journal of Biomedical Informatics*, 43(4):587–594, 2010.
- [DG08] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.
- [EBMT14] Nura Esfandiary, Mohammad Reza Babavalian, Amir-Masoud Eftekhari Moghadam, and Vahid Kashani Tabar. Knowledge discovery in medicine: Current issue and future trend. *Expert Systems with Applications*, 0(0):–, 2014.
- [FKK08a] Norbert E Fuchs, Kaarel Kaljurand, and Tobias Kuhn. Attempto controlled english for knowledge representation. In *Reasoning Web*, pages 104–124. Springer, 2008.
- [FKK08b] Norbert E Fuchs, Kaarel Kaljurand, and Tobias Kuhn. *Discourse representation structures for ACE 6.0*. Department of Informatics IFI, 2008.

- [FKS06] Norbert E Fuchs, Kaarel Kaljurand, and Gerold Schneider. Attempto controlled english meets the challenges of knowledge representation, reasoning, interoperability and user interfaces. In *FLAIRS Conference*, volume 12, pages 664–669, 2006.
- [FMS10] Kin Wah Fung, Clement McDonald, and Suresh Srinivasan. The {UMLS-CORE} project: a study of the problem list terminologies used in large healthcare institutions. *Journal of the American Medical Informatics Association*, 17(6):675–680, 2010.
- [GHM⁺08] Bernardo Cuenca Grau, Ian Horrocks, Boris Motik, Bijan Parsia, Peter Patel-Schneider, and Ulrike Sattler. {OWL} 2: The next step for {OWL}. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(4):309 – 322, 2008. [jce:titleSemantic Web Challenge 2006/2007;ce:title](#).
- [GNM⁺11] A. Ghazvinian, N.F. Noy, M.A. Musen, et al. How orthogonal are the obo foundry ontologies. *J Biomed Semantics*, 2(Suppl 2):S2, 2011.
- [GOS09] Nicola Guarino, Daniel Oberle, and Steffen Staab. What is an ontology? In *Handbook on ontologies*, pages 1–17. Springer, 2009.
- [GRDS⁺11] Vijay Garla, Vincent Lo Re, Zachariah Dorey-Stein, Farah Kidwai, Matthew Scotch, Julie Womack, Amy Justice, and Cynthia Brandt. The yale ctakes extensions for document classification: architecture and application. *Journal of the American Medical Informatics Association*, 18(5):614–620, 2011.
- [Gre03] Pierre Grenon. Knowledge management from the ontological standpoint. In *Wissensmanagement*, pages 415–416, 2003.
- [GRU93] Thomas R GRUBER. Toward principles for design of ontologies used for knowledge sharing. *Technical Report KSL*, pages 93–04, 1993.
- [Gru95] Thomas R Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5):907–928, 1995.
- [Gru11] Michael Gruninger. Verification of the owl-time ontology. In *Proceedings of the 10th international conference on The semantic web - Volume Part I*, ISWC’11, pages 225–240, Berlin, Heidelberg, 2011. Springer-Verlag.
- [GSA⁺11] Albert Goldfain, Barry Smith, Sivaram Arabandi, Mathias Brochhausen, and William R Hogan. Vital sign ontology. *Bio-Ontologies 2011*, 2011.
- [HDG⁺06] Matthew Horridge, Nick Drummond, John Goodwin, Alan L Rector, Robert Stevens, and Hai Wang. The manchester owl syntax. In *OWLed*, volume 216, 2006.
- [HDO⁺11a] R. Hoehndorf, M. Dumontier, A. Oellrich, D. Rebholz-Schuhmann, P.N. Schofield, and G.V. Gkoutos. Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning. *PloS one*, 6(7):e22006, 2011.
- [HDO⁺11b] Robert Hoehndorf, Michel Dumontier, Anika Oellrich, Sarala Wimalaratne, Dietrich Rebholz-Schuhmann, Paul Schofield, and Georgios V. Gkoutos. A common layer of

- interoperability for biomedical ontologies based on owl el. *Bioinformatics*, 27(7):1001–1008, 2011.
- [Hea] Health Level Seven International. HL7 version 3: Reference information model (rim). http://www.hl7.org/implement/standards/product_brief.cfm?product_id=77.
- [HFA⁺02] Jerry R Hobbs, George Ferguson, James Allen, P Hayes, I Niles, and A Pease. A {DAML} ontology of time. *online: http://www.cs.rochester.edu/~ferguson/daml/daml-time-20020830.txt*, 2002.
- [HP04] Jerry R Hobbs and Feng Pan. An ontology of time for the semantic web. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(1):66–85, 2004.
- [HSV08] Gergely Héja, György Surján, and Péter Varga. Ontological analysis of {SNOMED CT}. *BMC medical informatics and decision making*, 8(Suppl 1):S8, 2008.
- [Hus00] Edmund Husserl. *Logische Untersuchungen*. Halle: Niemeyer, 1900.
- [ISO09] ISO. ISO/HL7 27931:2009 data exchange standards – health level seven version 2.5 – an application protocol for electronic data exchange in healthcare environments, 2009. http://www.iso.org/iso/catalogue_detail.htm?csnumber=44428.
- [Jen07] Apache Jena. semantic web framework for java, 2007. <http://jena.apache.org/>.
- [Kal08] Kaarel Kaljurand. Ace view—an ontology and rule editor based on attempto controlled english. In *OWLED*, 2008.
- [KC06] Graham Klyne and Jeremy J Carroll. Resource description framework (rdf): Concepts and abstract syntax. 2006.
- [KF06a] Kaarel Kaljurand and Norbert E Fuchs. Bidirectional mapping between owl dl and attempto controlled english. In *Principles and Practice of Semantic Web Reasoning*, pages 179–189. Springer, 2006.
- [KF06b] Kaarel Kaljurand and Norbert E Fuchs. Mapping attempto controlled english to owl dl. In *3rd European Semantic Web Conference. Demo and Poster Session, Budva, Montenegro*, 2006.
- [KF07] Kaarel Kaljurand and Norbert E Fuchs. Verbalizing owl in attempto controlled english. In *OWLED*, volume 258, 2007.
- [KFNM04] Holger Knublauch, Ray W Ferguson, Natalya F Noy, and Mark A Musen. The protégé owl plugin: An open development environment for semantic web applications. In *The Semantic Web–ISWC 2004*, pages 229–243. Springer, 2004.
- [KKS11] Yevgeny Kazakov, Markus Krötzsch, and František Simancík. Concurrent classification of {EL} ontologies. In *Proceedings of the 10th international conference on The semantic web - Volume Part I, ISWC’11*, pages 305–320, Berlin, Heidelberg, 2011. Springer-Verlag.

- [Krö10] Markus Krötzsch. Efficient inferencing for owl el. In *Logics in Artificial Intelligence*, pages 234–246. Springer, 2010.
- [KTB⁺14] Peter Kluegl, Martin Toepfer, Philip-Daniel Beck, Georg Fette, and Frank Puppe. Uima ruta: Rapid development of rule-based information extraction applications. *Natural Language Engineering*, pages 1–40, 2014.
- [Kuh13] Tobias Kuhn. The understandability of OWL statements in controlled English. *Semantic Web*, 4(1):101–115, January 2013. <http://dx.doi.org/10.3233/sw-2012-0063>.
- [LB10] Michael J Lawley and Cyril Bousquet. Fast classification in protégé: Snorocket as an owl 2 el reasoner. In *Proc. 6th Australasian Ontology Workshop (IAOA’10). Conferences in Research and Practice in Information Technology*, volume 122, pages 45–49, 2010.
- [LD11] Anton E. Lawson and Erno S. Daniel. Inferences of clinical diagnostic reasoning and diagnostic error. *Journal of Biomedical Informatics*, 44(3):402 – 412, 2011. Biomedical Complexity and Error.
- [LED⁺10] Deryle Lonsdale, David W. Embley, Yihong Ding, Li Xu, and Martin Hepp. Reusing ontologies and language components for ontology generation. *Data & Knowledge Engineering*, 69(4):318 – 330, 2010. Including Special Section: 12th International Conference on Applications of Natural Language to Information Systems (NLDB07) - Three selected and extended papers.
- [LHC11] Kaihong Liu, William R Hogan, and Rebecca S Crowley. Natural language processing methods and systems for biomedical ontology learning. *Journal of Biomedical Informatics*, 44(1):163–179, July 2011.
- [LSS13] Domenico Lembo, Valerio Santarelli, and Domenico Fabio Savo. Graph-based ontology classification in owl 2 ql. In *The Semantic Web: Semantics and Big Data*, pages 320–334. Springer, 2013.
- [Lun14] Lund University Language Technology Group. Semantics technologies at lth, 2014. <http://nlp.cs.lth.se/semantics/>.
- [MBB⁺01] Alexa T McCray, Anita Burgun, Olivier Bodenreider, et al. Aggregating UMLS semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, (1):216–220, 2001.
- [MFF⁺14] Stéphane M. Meystre, Óscar Fernández, F. Jeffrey Friedlin, Brett R. South, Shuying Shen, and Matthew H. Samore. Text de-identification for privacy protection: A study of its impact on clinical text information content. *Journal of Biomedical Informatics*, 0(0):–, 2014.
- [MFS⁺10] Stephane Meystre, F Friedlin, Brett South, Shuying Shen, and Matthew Samore. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10(1):70, 2010.

- [MMH10] Raghava Mutharaju, Frederick Maier, and Pascal Hitzler. A MapReduce Algorithm for EL+. In *23rd International Workshop on Description Logics DL2010*, page 456, 2010.
- [MPVH06] Robert Minchin, Fabio Porto, Christelle Vangenot, and Sven Hartmann. Symptoms ontology for mapping diagnostic knowledge systems. In *Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on*, pages 593–598. IEEE, 2006.
- [MR07] Boris Motik and Riccardo Rosati. A faithful integration of description logics with logic programming. In *IJCAI*, volume 7, pages 477–482, 2007.
- [MR11a] David Mendes and Irene Pimenta Rodrigues. The bridg model as the most authoritative resource in shared semantics for ontologies development in healthcare practice. In *Proceedings of the International Conference on Knowledge Management and Information Sharing*, pages 1–9, Paris, 2011.
- [MR11b] David Mendes and Irene Pimenta Rodrigues. Clinical practice ontology development best practices. 2011.
- [MR11c] David Mendes and Irene Pimenta Rodrigues. A semantic web pragmatic approach to develop clinical ontologies, and thus semantic interoperability, based in hl7 v2.xml messaging. In *HCist 2011 - Proceedings of the International Workshop on Health and Social Care Information Systems and Technologies*. Springer-Verlag - book of the CCIS series (Communications in Computer and Information Science), 2011.
- [MR12] David Mendes and Irene Pimenta Rodrigues. Advances to Semantic Interoperability through CPR Ontology extracting from SOAP framework reports. *electronic Journal of Health Informatics*, 2012.
- [MR13a] David Mendes and Irene Pimenta Rodrigues. A Semantic Web pragmatic approach to develop Clinical ontologies, and thus Semantic Interoperability, based in HL7 v2.xml messaging. In Ricardo Martinho, Rui Riço, Maria Manuela Cruz-Cunha, and João Eduardo Varajão, editors, *Information Systems and Technologies for Enhancing Health and Social Care*. IGI Global, 2013.
- [MR13b] David Mendes and Irene Pimenta Rodrigues. OGCP - A new ontology for clinical practice knowledge representation and a proposal for automated population. In *Proceedings of JIUE 2013 - Jornadas do Departamento de Informática da Universidade de Évora - Évora, Portugal*, 2013.
- [MRB13a] David Mendes, Irene Pimenta Rodrigues, and Carlos Baeta. Ontology based clinical practice justification in natural language. *Procedia Technology*, 9(0):1288 – 1293, 2013. CENTERIS 2013 - Conference on {ENTERprise} Information Systems / ProjMAN 2013 - International Conference on Project MANagement/ {HCIST} 2013 - International Conference on Health and Social Care Information Systems and Technologies.

- [MRB13b] David Mendes, Irene Pimenta Rodrigues, and Carlos Fernandes Baeta. Development and population of an elaborate formal ontology for clinical practice knowledge representation. In *Proceedings of KEOD 2013 - International Conference on Knowledge Engineering and Ontology Development - Vilamoura, Portugal*, 2013.
- [MRB13c] David Mendes, Irene Pimenta Rodrigues, and Carlos Fernandes Baeta. Enrichment/Population of customized CPR (Computer-based Patient Record) ontology from free-text reports for CSI (Computer Semantic Interoperability). *IJEHMC - International Journal of E-Health and Medical Communications*, 2013.
- [MRN12] Dora Melo, Irene Pimenta Rodrigues, and Vitor Beires Nogueira. Um Sistema de Pergunta-Resposta para Ontologias OWL. In *INForum 2012*, 2012.
- [MRRSB12] David Mendes, Irene Pimenta Rodrigues, Carlos Rodriguez-Solano, and Carlos Baeta. Enrichment/Population of Customized CPR (Computer-based Patient Record) Ontology from Free-text Reports for CSI (Computer Semantic Interoperability). In Elsevier, editor, *Procedia Technology*, volume 5, pages 753–762, Vilamoura, Portugal, October 2012. Elsevier B.V.
- [MSKSH08] S M Meystre, G K Savova, K C Kipper-Schuler, and J F Hurdle. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, pages 128–44, January 2008.
- [MWV⁺11] Danielle Mowery, Janyce Wiebe, Shyam Visweswaran, Henk Harkema, and Wendy W. Chapman. Building an automated soap classifier for emergency department reports. *Journal of Biomedical Informatics*, (0):–, 2011.
- [MZ11] Chris Mattmann and Jukka Zitting. *Tika in Action*. Manning Publications Co., 2011.
- [NR08] Natalya F. Noy and Daniel L. Rubin. Translating the foundational model of anatomy into {OWL}. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(2):133 – 136, 2008.
- [NV05] R. Navigli and P. Velardi. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(7):1075–1086, 2005.
- [OBP⁺07] Chimezie Ogbuji, Eugene Blackstone, Chris Pierce, World Wide Web Consortium, et al. Case study: A semantic web content repository for clinical research, 2007.
- [OCF09] Chimezie Ogbuji, Cleveland Clinic, and Foundation. Overview of cpr ontology. 2009.
- [Ogb11] Chimezie Ogbuji. A Framework Ontology for Computer-Based Patient Record Systems. In *Proceedings of the ICBO: International Conference on Biomedical Ontology*, pages 217–223, Buffalo, NY, USA, 2011.
- [OGM10] OGMS. Ontology for general medical science, 2010.
- [Pan05] Feng Pan. A temporal aggregates ontology in owl for the semantic web. In *Proceedings of the AAAI fall symposium on agents and the semantic web*, pages 30–37, 2005.

- [PCC06] Serguei V Pakhomov, Anni Coden, and Christopher G Chute. Developing a corpus of clinical notes manually annotated for part-of-speech. *International journal of medical informatics*, 75(6):418–429, 2006.
- [PH05] Feng Pan and Jerry R Hobbs. Temporal aggregates in owl-time. In *FLAIRS Conference*, volume 5, pages 560–565, 2005.
- [PHT⁺13] Sujan Perera, Cory Henson, Krishnaprasad Thirunarayan, Amit Sheth, and Suhas Nair. Semantics driven approach for knowledge acquisition from emrs. *IEEE journal of biomedical and health informatics*, 2013.
- [PK13] Alina Petrova and Maria Kissa. Non-taxonomic role extraction for biomedical concepts. *ESSLLI Student Session 2013 Preproceedings*, page 140, 2013.
- [PP08] Patrick PANTEL and Marco PENNACCHIOTTI. Automatically harvesting and ontologizing semantic relations. *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, page 171, 2008.
- [PSD12] Laura Plaza, Mark Stevenson, and Alberto Díaz. Resolving ambiguity in biomedical text to improve summarization. *Information Processing & Management*, 48(4):755 – 766, 2012.
- [Rec03] Alan L. Rector. Modularisation of domain ontologies implemented in description logics and related formalisms including owl. In *Proceedings of the international conference on Knowledge capture - K-CAP '03*, pages 121–130, New York, New York, USA, 2003. ACM Press.
- [Res99] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.
- [RGH⁺09] Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, and Andrea Setzer. Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics*, 42(5):950 – 966, 2009.
- [RGLGCP⁺12] Alejandro Rodríguez-González, Jose Emilio Labra-Gayo, Ricardo Colomo-Palacios, Miguel A Mayer, Juan Miguel Gómez-Berbís, and Angel García-Crespo. Sedelo: using semantics and description logics to support aided clinical diagnosis. *Journal of medical systems*, 36(4):2471–2481, 2012.
- [RJ03] Cornelius Rosse and José L.V. Mejino Jr. A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of Biomedical Informatics*, 36(6):478 – 500, 2003. <http://www.sciencedirect.com/science/article/pii/S1532046403001278>.
- [RMVGFB⁺11] Juana María Ruiz-Martínez, Rafael Valencia-García, Jesualdo Tomás Fernández-Breis, Francisco García-Sánchez, and Rodrigo Martínez-Béjar. Ontology learning from biomedical natural language documents using {UMLS}. *Expert Systems with Applications*, 38(10):12365 – 12378, 2011. <http://www.sciencedirect.com/science/article/pii/S095741741100532X>.

- [RO12] OBO Foundry RO. Relations ontology, 2012. <http://obofoundry.org/ro/>.
- [SAMK05] I. Spasic, S. Ananiadou, J. McNaught, and A. Kumar. Text mining and ontologies in biomedicine: making sense of raw text. *Briefings in bioinformatics*, 6(3):239–251, 2005.
- [SAR⁺07] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–5, November 2007.
- [SB01] Barry Smith and Berit Brogaard. A unified theory of truth and reference. *Logique et Analyse*, 43(169-170):49–93, 2001.
- [SB⁺08] B. Smith, M. Brochhausen, et al. Establishing and harmonizing ontologies in an interdisciplinary health care and clinical research environment. *Studies in health technology and informatics*, 134:219, 2008.
- [SB10] B Smith and M Brochhausen. Putting biomedical ontologies to work. *Methods Inf Med*, 49, February 2010.
- [SBS07] Holger Stenzhorn, Elena Beißwanger, and Stefan Schulz. Towards a top-domain ontology for linking biomedical ontologies. *Medinfo*, 129:1225–1229, 2007.
- [SBvdH⁺09] Stefan Schulz, Elena Beisswanger, László van den Hoek, Olivier Bodenreider, and Erik M. van Mulligen. Alignment of the umls semantic network with biotop: methodology and assessment. *Bioinformatics*, 25(12):i69–i76, 2009.
- [SC10] Barry Smith and Werner Ceusters. Ontological realism: A methodology for coordinated evolution of scientific ontologies. *Applied ontology*, 5(3-4):139–188, November 2010.
- [Sch04] Lynn Schriml. Symptom ontology, 2004.
- [Sch10] Rolf Schwitter. Controlled natural languages for knowledge representation. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 1113–1121, Stroudsburg, PA, USA, 2010.
- [SCK⁺05] Barry Smith, Werner Ceusters, Bert Klagges, Jacob Köhler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan L Rector, and Cornelius Rosse. Relations in biomedical ontologies. *Genome biology*, 6(5):R46, 2005.
- [SCS09] Richard H Scheuermann, Werner Ceusters, and Barry Smith. Toward an ontological treatment of disease and diagnosis. In *2009 AMIA Summit on Translational Bioinformatics*, pages 116–120, San Francisco, CA, 2009.
- [SD14] Hazem Safwat and Brian Davis. A brief state of the art of cnls for ontology authoring. In *Controlled Natural Language*, pages 190–200. Springer, 2014.

- [SG04] Barry Smith and Pierre Grenon. The cornucopia of formal-ontological relations. *Dialectica*, 58(3):279–296, 2004.
- [SG10] Mark Stevenson and Yikun Guo. Disambiguation of ambiguous biomedical terms using examples generated from the UMLs metathesaurus. *Journal of Biomedical Informatics*, 43(5):762–773, Oct 2010.
- [SG13] Jannik Strötgen and Michael Gertz. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298, 2013.
- [Sim00] Peter Simons. Continuants and occurrents: Peter simons. In *Aristotelian Society Supplementary Volume*, volume 74, pages 59–75. Wiley Online Library, 2000.
- [SKC⁺08] Rolf Schwitter, Kaarel Kaljurand, Anne Cregan, Catherine Dolbear, and Glen Hart. A comparison of three controlled natural languages for owl 1.1. In *4th OWL Experiences and Directions Workshop (OWLED 2008 DC)*, Washington, pages 1–2, 2008.
- [SKCR05] Barry Smith, Anand Kumar, Werner Ceusters, and Cornelius Rosse. On carcinomas and other pathological entities. *Comparative and functional genomics*, 6(7-8):379–87, January 2005.
- [SKK04] Barry Smith, Jacob Köhler, and Anand Kumar. On the application of formal principles to life science data: a case study in the gene ontology. In *Data Integration in the Life Sciences*, pages 79–94. Springer, 2004.
- [SKSBC08] G Savova, Karin Kipper-Schuler, J Buntrock, and C Chute. Uima-based clinical information extraction system. *Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP*, 39, 2008.
- [SKSC06] Barry Smith, Wacław Kusnierczyk, Daniel Schober, and Werner Ceusters. Towards a reference terminology for ontology research and development in the biomedical domain. In *KR-MED*, volume 222, 2006.
- [Smi98] Barry Smith. The basic tools of formal ontology. In *Formal Ontology in Information Systems*, pages 19–28. IOS Press, Washington. Frontiers in Artificial Intelligence and Applications, 1998.
- [Smi06] Barry Smith. From concepts to clinical reality: An essay on the benchmarking of biomedical terminologies. *Journal of Biomedical Informatics*, 39(3):288 – 298, 2006. <http://www.sciencedirect.com/science/article/pii/S1532046405001036>.
- [SMO⁺10] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.
- [Sou14] Sourceforge. BioNLP UIMA Component Repository, 2014. <http://sourceforge.net/projects/bionlp-uima/>.

- [SS07] Stefan Schulz and Holger Stenzhorn. Ten theses on clinical ontologies. *Studies In Health Technology And Informatics*, 127:268 – 275, 2007.
- [SS11] Michael Schneider and Geoff Sutcliffe. Reasoning in the owl 2 full ontology language using first-order automated theorem proving. In Nikolaj Björner and Viorica Sofronie-Stokkermans, editors, *Automated Deduction CADE-23*, volume 6803 of *Lecture Notes in Computer Science*, pages 461–475. Springer Berlin Heidelberg, 2011.
- [Sta14] Stanford University - Stanford, CA. Protégé, 2014. <http://protege.stanford.edu/>.
- [TBC14] Camilo Thorne, Raffaella Bernardi, and Diego Calvanese. Designing efficient controlled languages for ontologies. In *Computing Meaning*, pages 149–173. Springer, 2014.
- [The14] The Open Biological and Biomedical Ontologies. The OBO Foundry, 2014. <http://www.obofoundry.org/>.
- [TSZI10] Donna Truran, Patricia Saad, Ming Zhang, and Kerry Innes. SNOMED CT and its place in health information management practice. *Health Information Management Journal*, 39(2):37, 2010.
- [UB 06] UB State of New York University at Buffalo. Referent tracking unit, 2006. <http://www.referent-tracking.com/RTU/>.
- [UKOVH09] Jacopo Urbani, Spyros Kotoulas, Eyal Oren, and Frank Van Harmelen. Scalable distributed reasoning using mapreduce. In *The Semantic Web-ISWC 2009*, pages 634–649. Springer, 2009.
- [UTS14] UTS. Umls terminology services, 2014. <https://uts.nlm.nih.gov/home.html>.
- [VHH09] Johanna Völker, Peter Haase, and Pascal Hitzler. *Learning expressive ontologies*. IOS, 2009.
- [W3C06] W3C. Time ontology in owl, 2006. <http://www.w3.org/TR/owl-time/>.
- [W3C11a] W3C. OWL 2 Web Ontology Language, 2011. <http://www.w3.org/TR/owl2-overview/>.
- [W3C11b] W3C. The world wide web consortium (w3c), 2011.
- [W3C12] W3C. Owl 2 web ontology language profiles (second edition), 11 2012. <http://www.w3.org/TR/owl2-profiles/>.
- [W3C14] W3C. RDF 1.1 Concepts and Abstract Syntax, 2014. <http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/>.
- [Wik11] Wikipedia. Translation memory - wikipedia, the free encyclopedia, 2011. http://en.wikipedia.org/wiki/Translation_memory/.
- [Wik14] Wikipedia. Translation memory exchange - wikipedia, the free encyclopedia, 2014. http://en.wikipedia.org/wiki/Translation_Memory_eXchange.

- [WLB12] Wilson Wong, Wei Liu, and Mohammed Bennamoun. Ontology learning from text: A look back and into the future. *ACM Computing Surveys*, 44(4):1–36, Aug 2012.
- [WZGP04] Xiao Hang Wang, Da Qing Zhang, Tao Gu, and Hung Keng Pung. Ontology based context modeling and reasoning using owl. In *Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second IEEE Annual Conference on*, pages 18–22. Ieee, 2004.
- [XW] Min Xia and Ju An Wang. Temporal text extraction and automated time-owl population.
- [YAM⁺12] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.F. Chang, and L. Hua. Data mining in healthcare and biomedicine: A survey of the literature. *Journal of medical systems*, pages 1–18, 2012.
- [ZN10] Amal Zouaq and Roger Nkambou. A survey of domain ontology engineering: methods and tools. In *Advances in intelligent tutoring systems*, pages 103–119. Springer, 2010.

Annexes

Annexes A

Symbols and terminology

This is a complete reference of the lettering, symbols and and terminology found in all the present work

Term	Notation	Term	Notation
axiom	α	ontology	\mathcal{O}
I	\mathcal{I}	concept	concept
Concepts	C	Roles	R
set	$\{individuo\}$	tuple x,y	$\langle x, y \rangle$
$\langle x, y \rangle$	tuple x,y	complement - neg	\neg
role	role	individual	individual
the set of all roles	R	the set of all concepts	C
transitive	Transitive (<i>transitive</i>)	Concept inclusion	\sqsubseteq
universal restriction - forall	\forall	existential restriction - exists	\exists
At most n	$\leq n$	At least n	$\geq n$
union	\sqcup	intersection	\sqcap

Table A.1: Symbols and notations

Examples:

Disjoint(parentOf, childOf)

A nominal is a concept that has exactly one instance. For example, $\{\text{john}\}$ is the concept whose only instance is (the individual denoted by) *john*.

Namespace name	Namespace
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
xsd	http://www.w3.org/2001/XMLSchema#
owl	http://www.w3.org/2002/07/owl#

Table A.2: Namespaces for qualified names abbreviation

Annexes B

OGCP Description Logics (DL)

Only the leaf nodes of **OGCP** based in **CPR** that constitute the **OWL** \mathcal{EL} scaffolding of the Knowledge Base (**KB**) are presented

Classes

Continuant
DependentContinuant
Disposition
Entity
Function
GenericallyDependentContinuant
IndependentContinuant
MaterialEntity
Object
ObjectAggregate
ObsoleteClass
ObsoleteClass \sqsubseteq metadata-entity
PhysicalQuality
PhysicalQuality \equiv Quality
Prefix
Prefix \sqsubseteq metadata-entity
Process
ProcessAggregate
ProcessualEntity
Quality
Quality \equiv PhysicalQuality
QualityValue
QualityValue \sqsubseteq data-entity
Role
SpecificallyDependentContinuant
Subset
Subset \sqsubseteq metadata-entity
SurgicalMethod
SurgicalMethod \sqsubseteq ProcessualEntity
Synonym
Synonym \sqsubseteq linguistic-construct
SynonymType
SynonymType \sqsubseteq linguistic-construct

TemporalRegion
 Thing
 UnitOfMeasurement
 UnitOfMeasurement \sqsubseteq data-entity
 adverse-event
 adverse-event \sqsubseteq pathological-process \sqcap sequela
 aggregate-bodily-feature
 aggregate-bodily-feature \sqsubseteq Entity
 aggregate-bodily-feature $\sqsubseteq \exists$ has_part bodily-feature
 anamnesis
 anamnesis \equiv clinical-finding $\sqcap \exists$ outputOf medical-history-screening-act
 anamnesis $\sqsubseteq \neg$ sign-recording
 anamnesis $\sqsubseteq \neg$ clinical-diagnosis
 anamnesis $\sqsubseteq \neg$ laboratory-test-finding
 anatomical-boundary-entity
 anatomical-boundary-entity \sqsubseteq immaterial-anatomical-continuant
 anatomical-space
 anatomical-space \sqsubseteq immaterial-anatomical-continuant
 anatomical-structure
 anatomical-structure \sqsubseteq Object
 anatomical-structure $\sqsubseteq \neg$ medical-device
 anatomical-structure $\sqsubseteq \neg$ pharmacological-substance
 anatomical-structure $\sqsubseteq \neg$ organism
 anatomical-structure $\sqsubseteq \neg$ computer-system
 anatomical-surface
 anatomical-surface \sqsubseteq anatomical-boundary-entity
 bodily-feature
 bodily-feature \equiv organismal-continuant \sqcup Quality bodily-feature \sqsubseteq Entity
 clinical-act
 clinical-act \sqsubseteq ProcessualEntity
 clinical-act $\sqsubseteq \exists$ has_agent (\exists hasRole healthcare-professional-role)
 clinical-act $\sqsubseteq \exists$ has_participant patient
 clinical-administration-act
 clinical-administration-act \sqsubseteq clinical-act
 clinical-analysis-act
 clinical-analysis-act \sqsubseteq clinical-investigation-act
 clinical-artifact
 clinical-artifact \sqsubseteq representational-artifact
 clinical-artifact $\sqsubseteq \exists$ subjectOfDescription person
 clinical-artifact $\sqsubseteq \exists$ composedBy person
 clinical-diagnosis
 clinical-diagnosis $\sqsubseteq \exists$ hypothesizedProblem pathological-disposition
 clinical-diagnosis $\sqsubseteq \exists$ outputOf clinical-analysis-act
 clinical-diagnosis \sqsubseteq clinical-finding
 clinical-diagnosis $\sqsubseteq \neg$ sign-recording
 clinical-diagnosis $\sqsubseteq \neg$ laboratory-test-finding
 clinical-diagnosis $\sqsubseteq \neg$ anamnesis
 clinical-examination
 clinical-examination \sqsubseteq screening-act
 clinical-finding
 clinical-finding $\sqsubseteq \exists$ outputOf clinical-act
 clinical-finding $\sqsubseteq \exists$ composedBy (\exists hasRole clinician-role)
 clinical-finding \sqsubseteq clinical-artifact
 clinical-finding $\sqsubseteq \exists$ representationOf bodily-feature
 clinical-finding $\sqsubseteq \neg$ symptom-recording
 clinical-finding $\sqsubseteq \neg$ recorded-clinical-situation
 clinical-finding $\sqsubseteq \neg$ patient-record
 clinical-investigation-act
 clinical-investigation-act $\sqsubseteq \exists$ investigates (etiologic-agent $\sqcup \exists$ hasIndication therapeutic-act)
 clinical-investigation-act \sqsubseteq clinical-act
 clinical-investigation-act $\sqsubseteq \exists$ hasOutput clinical-artifact
 clinical-phenotype
 clinical-phenotype \sqsubseteq aggregate-bodily-feature
 clinician-role

clinician-role \sqsubseteq healthcare-professional-role
 computer-system
 computer-system \sqsubseteq Object
 computer-system $\sqsubseteq \neg$ pharmacological-substance
 computer-system $\sqsubseteq \neg$ organism
 computer-system $\sqsubseteq \neg$ anatomical-structure
 computer-system $\sqsubseteq \neg$ medical-device
 ogcp-entities
 ogcp-entities \equiv patient-record \sqcup recorded-clinical-situation \sqcup sign-recording \sqcup symptom-recording $\sqcup \forall$ representationOf vital-sign
 data-entity
 diagnostic-image
 diagnostic-image \sqsubseteq image
 diagnostic-image $\sqsubseteq \exists$ outputOf clinical-investigation-act
 diagnostic-procedure
 diagnostic-procedure \sqsubseteq clinical-investigation-act
 etiologic-agent
 etiologic-agent \sqsubseteq Continuant
 etiologic-agent $\sqsubseteq \exists$ hasConsequence pathological-disposition
 excised-anatomy
 excised-anatomy \sqsubseteq anatomical-structure
 excised-anatomy $\sqsubseteq \exists$ actedUponBy (procedure $\sqcap \exists$ hasMethod removal)
 extra-organismal-continuant
 extra-organismal-continuant \sqsubseteq ObjectAggregate
 genetic-abnormality
 genetic-abnormality $\sqsubseteq \exists$ disruptsPhysiology
 genetic-abnormality \sqsubseteq etiologic-agent
 genetic-disease
 genetic-disease \sqsubseteq pathological-disposition
 genetic-disease $\sqsubseteq \exists$ isConsequenceOf genetic-abnormality
 genetic-disease $\sqsubseteq \neg$ idiopathic-disease
 genetic-disease $\sqsubseteq \neg$ infectious-disease
 healthcare-professional-role
 healthcare-professional-role \sqsubseteq Role
 idiopathic-disease
 idiopathic-disease \sqsubseteq pathological-disposition
 idiopathic-disease $\sqsubseteq \neg$ genetic-disease
 idiopathic-disease $\sqsubseteq \neg$ infectious-disease
 image
 image \sqsubseteq representational-artifact
 immaterial-anatomical-continuant
 immaterial-anatomical-continuant \sqsubseteq DependentContinuant
 immaterial-anatomical-continuant $\sqsubseteq \neg$ state
 immaterial-anatomical-continuant $\sqsubseteq \neg$ immaterial-pathological-continuant
 immaterial-pathological-continuant
 immaterial-pathological-continuant \sqsubseteq DependentContinuant
 immaterial-pathological-continuant $\sqsubseteq \neg$ immaterial-anatomical-continuant
 immaterial-pathological-continuant $\sqsubseteq \neg$ state
 infectious-disease
 infectious-disease $\sqsubseteq \exists$ isConsequenceOf pathogen
 infectious-disease \sqsubseteq pathological-disposition
 infectious-disease $\sqsubseteq \neg$ genetic-disease
 infectious-disease $\sqsubseteq \neg$ idiopathic-disease
 laboratory-test
 laboratory-test \sqsubseteq clinical-investigation-act
 laboratory-test $\sqsubseteq \forall$ hasOutput laboratory-test-finding
 laboratory-test-finding
 laboratory-test-finding \equiv clinical-finding $\sqcap \forall$ outputOf laboratory-test
 laboratory-test-finding $\sqsubseteq \neg$ clinical-diagnosis
 laboratory-test-finding $\sqsubseteq \neg$ sign-recording
 laboratory-test-finding $\sqsubseteq \neg$ anamnesis
 linguistic-construct
 longitudinal-patient-medical-history
 longitudinal-patient-medical-history \sqsubseteq ProcessAggregate
 material-pathological-entity

material-pathological-entity \sqsubseteq MaterialEntity
 material-pathological-entity $\sqsubseteq \exists$ derives_from anatomical-structure $\sqcup \exists$ transformation_of anatomical-structure
 material-pathological-entity $\sqsubseteq \exists$ participates_in morphologic-alteration medical-device
 medical-device \sqsubseteq Object
 medical-device $\sqsubseteq \neg$ organism
 medical-device $\sqsubseteq \neg$ anatomical-structure
 medical-device $\sqsubseteq \neg$ pharmacological-substance
 medical-device $\sqsubseteq \neg$ computer-system
 medical-history-screening-act
 medical-history-screening-act \sqsubseteq screening-act
 medical-problem
 medical-problem \equiv etiologic-agent \sqcup pathological-disposition \sqcup pathological-process $\sqcup \exists$ representedBy sign-recording $\sqcup \exists$ representedBy symptom-recording
 medical-problem \sqsubseteq Entity
 medical-therapy
 medical-therapy \sqsubseteq therapeutic-act
 medication
 medication \sqsubseteq pharmacological-substance
 metadata-entity
 morphologic-alteration
 morphologic-alteration $\sqsubseteq \forall$ has_participant physical-anatomical-entity
 morphologic-alteration $\sqsubseteq \forall$ has_agent pathological-disposition
 morphologic-alteration $\sqsubseteq \forall$ part_of pathological-process
 organism
 organism \sqsubseteq Object
 organism $\sqsubseteq \neg$ medical-device
 organism $\sqsubseteq \neg$ pharmacological-substance
 organism $\sqsubseteq \neg$ computer-system
 organism $\sqsubseteq \neg$ anatomical-structure
 organismal-continuant
 organismal-continuant \equiv anatomical-structure \sqcup extra-organismal-continuant \sqcup immaterial-anatomical-continuant \sqcup material-pathological-entity \sqcup organism
 organismal-continuant \sqsubseteq Continuant
 organismal-process-aggregate
 organismal-process-aggregate \sqsubseteq ProcessAggregate
 pathogen
 pathogen \equiv etiologic-agent \sqcap organism
 pathological-disposition
 pathological-disposition \sqsubseteq Disposition
 pathological-disposition $\sqsubseteq \forall$ located_in (patient $\sqcup \forall$ part_of patient)
 pathological-disposition $\sqsubseteq \exists$ isConsequenceOf etiologic-agent $\sqcup \exists$ participates_in (pathological-process $\sqcap \exists$ has_participant etiologic-agent)
 pathological-disposition $\sqsubseteq \exists$ agent_in (pathological-process $\sqcap \exists$ has_part morphologic-alteration)
 pathological-disposition $\sqsubseteq \neg$ physiological-disposition
 pathological-process
 pathological-process $\sqsubseteq \exists$ has_agent pathological-disposition
 pathological-role
 pathological-role \sqsubseteq Role
 patient
 patient \equiv person $\sqcap \exists$ hasRole patient-role $\sqcap \exists$ participates_in clinical-act
 patient-record
 patient-record $\sqsubseteq \exists$ has_proper_part representational-artifact
 patient-record $\sqsubseteq \exists$ representationOf longitudinal-patient-medical-history
 patient-record \sqsubseteq clinical-artifact
 patient-record $\sqsubseteq \neg$ symptom-recording
 patient-record $\sqsubseteq \neg$ recorded-clinical-situation
 patient-record $\sqsubseteq \neg$ clinical-finding
 patient-role
 patient-role \sqsubseteq Role
 person
 person \sqsubseteq organism
 pharmacological-substance
 pharmacological-substance \sqsubseteq Object
 pharmacological-substance $\sqsubseteq \neg$ organism

pharmacological-substance $\sqsubseteq \neg$ computer-system
 pharmacological-substance $\sqsubseteq \neg$ medical-device
 pharmacological-substance $\sqsubseteq \neg$ anatomical-structure
 physical-anatomical-entity
 physical-anatomical-entity \sqsubseteq organismal-continuant
 physical-therapy
 physical-therapy \sqsubseteq therapeutic-act
 physician
 physician \equiv person $\sqcap \exists$ hasRole physician-role $\sqcap \exists$ participates_in (clinical-investigation-act \sqcup therapeutic-act)
 physician-role
 physician-role \sqsubseteq clinician-role
 physiological-disposition
 physiological-disposition \sqsubseteq Disposition
 physiological-disposition $\sqsubseteq \neg$ pathological-disposition
 physiological-role
 physiological-role \sqsubseteq Role
 procedure
 procedure \equiv diagnostic-procedure \sqcup therapeutic-procedure
 procedure $\sqsubseteq \exists$ actsOn organismal-continuant
 procedure \sqsubseteq clinical-act
 procedure $\sqsubseteq \exists$ approachSite immaterial-anatomical-continuant
 psychological-therapy
 psychological-therapy \sqsubseteq therapeutic-act
 recorded-clinical-situation
 recorded-clinical-situation \equiv clinical-artifact $\sqcap \forall$ includes clinical-artifact
 recorded-clinical-situation \sqsubseteq clinical-artifact
 recorded-clinical-situation $\sqsubseteq \neg$ patient-record
 recorded-clinical-situation $\sqsubseteq \neg$ symptom-recording
 recorded-clinical-situation $\sqsubseteq \neg$ clinical-finding
 removal
 removal \sqsubseteq SurgicalMethod
 representational-artifact
 representational-artifact $\sqsubseteq \exists$ representationOf Entity
 representational-artifact \sqsubseteq GenericallyDependentContinuant
 screening-act
 screening-act \sqsubseteq clinical-investigation-act
 self-examination
 self-examination $\sqsubseteq \exists$ has_agent patient
 self-examination \sqsubseteq ProcessualEntity
 sequela
 sequela \equiv bodily-feature $\sqcap \exists$ isConsequenceOf pathological-process
 sequela $\sqsubseteq \neg$ vital-sign
 sign-recording
 sign-recording \sqsubseteq clinical-finding $\sqcap \exists$ outputOf clinical-examination
 sign-recording $\sqsubseteq \neg$ clinical-diagnosis
 sign-recording $\sqsubseteq \neg$ anamnesis
 sign-recording $\sqsubseteq \neg$ laboratory-test-finding
 state
 state \sqsubseteq DependentContinuant
 state $\sqsubseteq \neg$ immaterial-anatomical-continuant
 state $\sqsubseteq \neg$ immaterial-pathological-continuant
 substance-administration
 substance-administration $\sqsubseteq \exists$ has_participant medication
 substance-administration \sqsubseteq medical-therapy
 symptom-recording
 symptom-recording $\sqsubseteq \exists$ representationOf bodily-feature
 symptom-recording $\sqsubseteq \exists$ outputOf (medical-history-screening-act \sqcup self-examination)
 symptom-recording $\sqsubseteq \exists$ composedBy patient
 symptom-recording \sqsubseteq clinical-artifact
 symptom-recording $\sqsubseteq \neg$ patient-record
 symptom-recording $\sqsubseteq \neg$ clinical-finding
 symptom-recording $\sqsubseteq \neg$ recorded-clinical-situation
 syndrome
 syndrome \sqsubseteq idiopathic-disease

therapeutic-act
 therapeutic-act \sqsubseteq clinical-act
 therapeutic-procedure
 therapeutic-procedure $\sqsubseteq \exists$ hasMethod SurgicalMethod
 therapeutic-procedure \sqsubseteq therapeutic-act
 vital-sign
 vital-sign \sqsubseteq bodily-feature
 vital-sign $\sqsubseteq \neg$ sequela

Object properties

part_of
 part_of \equiv part_of
 actedUponBy
 $\langle \text{http://purl.org/ogcp/actedUponBy} \rangle \equiv \langle \text{http://purl.org/ogcp/actsOn} \rangle \top \sqsubseteq \forall$ actedUponBy procedure
 actsOn
 $\langle \text{http://purl.org/ogcp/actedUponBy} \rangle \equiv \langle \text{http://purl.org/ogcp/actsOn} \rangle \exists$ actsOn Thing \sqsubseteq Process
 $\top \sqsubseteq \forall$ actsOn (organismal-continuant \sqcup physical-anatomical-entity)
 agent_in
 annotatedFunction
 approachSite
 \exists approachSite Thing \sqsubseteq procedure
 $\top \sqsubseteq \forall$ approachSite immaterial-anatomical-continuant
 composedBy
 \exists composedBy Thing \sqsubseteq representational-artifact
 $\top \sqsubseteq \forall$ composedBy person
 derives_from
 disruptsPhysiology
 \exists disruptsPhysiology Thing \sqsubseteq genetic-abnormality
 findingSite
 \sqsubseteq located_in
 \exists findingSite Thing \sqsubseteq clinical-finding
 $\top \sqsubseteq \forall$ findingSite physical-anatomical-entity
 hasConsequence
 $\langle \text{http://purl.org/ogcp/hasConsequence} \rangle \equiv \langle \text{http://purl.org/ogcp/isConsequenceOf} \rangle$
 hasContraindication
 $\langle \text{http://purl.org/ogcp/isContraindicationFor} \rangle \equiv \langle \text{http://purl.org/ogcp/hasContraindication} \rangle$
 \exists hasContraindication Thing \sqsubseteq therapeutic-act
 hasIndication
 $\langle \text{http://purl.org/ogcp/hasIndication} \rangle \equiv \langle \text{http://purl.org/ogcp/isIndicationFor} \rangle \exists$ hasIndication Thing \sqsubseteq therapeutic-act
 hasInput
 \exists hasInput Thing \sqsubseteq Process
 $\top \sqsubseteq \forall$ hasInput representational-artifact
 hasMethod
 $\top \sqsubseteq \forall$ hasMethod SurgicalMethod
 hasOutput
 $\langle \text{http://purl.org/ogcp/outputOf} \rangle \equiv \langle \text{http://purl.org/ogcp/hasOutput} \rangle \exists$ hasOutput Thing \sqsubseteq Process
 $\top \sqsubseteq \forall$ hasOutput representational-artifact
 hasRole
 \exists hasRole Thing \sqsubseteq IndependentContinuant
 $\top \sqsubseteq \forall$ hasRole Role
 has_agent
 has_part
 $\langle \text{http://www.obofoundry.org/ro/ro.owl\#has_part} \rangle \equiv \langle \text{http://www.obofoundry.org/ro/ro.owl\#part_of} \rangle$
 has_participant
 $\langle \text{http://www.obofoundry.org/ro/ro.owl\#participates_in} \rangle \equiv \langle \text{http://www.obofoundry.org/ro/ro.owl\#has_participant} \rangle$
 has_proper_part
 hypothesizedProblem \sqsubseteq representationOf \exists hypothesizedProblem Thing \sqsubseteq clinical-diagnosis
 $\top \sqsubseteq \forall$ hypothesizedProblem pathological-disposition
 includes
 $\top \sqsubseteq \forall$ includes Thing
 investigates
 isConsequenceOf

$\langle \text{http://purl.org/ogcp/hasConsequence} \rangle \equiv \langle \text{http://purl.org/ogcp/isConsequenceOf} \rangle$
 $\text{isContraindicationFor}$
 $\langle \text{http://purl.org/ogcp/isContraindicationFor} \rangle \equiv \langle \text{http://purl.org/ogcp/hasContraindication} \rangle$
 isIndicationFor
 $\langle \text{http://purl.org/ogcp/hasIndication} \rangle \equiv \langle \text{http://purl.org/ogcp/isIndicationFor} \rangle \sqsupseteq \forall \text{ isIndicationFor procedure}$
 located_in
 measurementOf
 $\sqsubseteq \text{representationOf} \sqsupseteq \forall \text{ measurementOf QualityValue}$
 outputOf
 $\langle \text{http://purl.org/ogcp/outputOf} \rangle \equiv \langle \text{http://purl.org/ogcp/hasOutput} \rangle \sqsupseteq \forall \text{ outputOf ProcessualEntity}$
 part_of
 $\text{part_of} \equiv \text{part_of}$
 $\langle \text{http://www.obofoundry.org/ro/ro.owl\#has_part} \rangle \equiv \langle \text{http://www.obofoundry.org/ro/ro.owl\#part_of} \rangle \text{ participates_in}$
 $\langle \text{http://www.obofoundry.org/ro/ro.owl\#participates_in} \rangle \equiv \langle \text{http://www.obofoundry.org/ro/ro.owl\#has_participant} \rangle \text{ rep-}$
 $\text{resentationOf} \langle \text{http://purl.org/ogcp/representedBy} \rangle \equiv \langle \text{http://purl.org/ogcp/representationOf} \rangle \text{ representedBy}$
 $\langle \text{http://purl.org/ogcp/representedBy} \rangle \equiv \langle \text{http://purl.org/ogcp/representationOf} \rangle$
 $\text{subjectOfDescription}$
 transformation_of

Data properties

comment
 definition
 editorialNote
 occursDuring
 startsNoEarlierThan
 startsNoLaterThan
 stopsNoEarlierThan
 stopsNoLaterThan

Individuals

SurgicalMethod
 aggregate-bodily-feature
 annotatedFunction
 clinical-artifact
 clinical-finding
 clinical-investigation-act
 etiologic-agent
 genetic-abnormality
 genetic-disease
 hasConsequence
 hasContraindication
 hasInput
 hasMethod
 hasOutput
 healthcare-professional-role
 idiopathic-disease
 includes
 investigates
 measurementOf
 medical-problem
 morphologic-alteration

occursDuring
organism
organismal-continuant
pathogen
procedure
recorded-clinical-situation
sequela
sign-recording
state
therapeutic-procedure
vital-sign

Index

- Clinical Knowledge (CK)
 - Representation, 37
- Computer Semantic Interoperability (CSI), 51
- Open Biological and Biomedical Ontologies (OBO)
 - Foundry principles, 53
- Ontology for General Clinical Practice (OGCP)
 - proposal, 47
- Web Ontology Language (OWL) verbalization, 95
- Translation Memories workflow with CP-ESB, 98
- Automatic Ontology Learning, 68
- Basic Formal Ontology, 54
- Top-Domain Ontology for the Life Sciences, 59
- Clinical Controlled Language translation, 72
- Computer Based Patient Record Ontology, 60
- Co-Reference Resolution, 74
- Cardiovascular Disease Ontology, 62
- Discourse Based Enhancement, 92
- Discourse Controller used for enrichment control, 80
- Disease Ontology, 62
- Extraction of Attributes and Values, 75
- Foundational Model of Anatomy, 59
- Knowledge Acquisition
 - through NLP, 72
- Knowledge Base Controlled Natural Language in-terrogation, 95
- Knowledge Base instance creation, 82
- Named Entity Recognition, 74
- Ontology Driven Expanded Semantic Annotation, 76
- Ontology for General Medical Science, 61
- Ontological Realism, 38
 - applied to OGCP, 53
- Part Of Speech tagging, 74
- Relations Ontology, 55
- Symptom Ontology, 60
- Translation Memory Manager tools, 66
- Translation Memory as a controlled technical jar-gon repository, 66
- Vital Signs Ontology, 60
- Word Sense Disambiguation, 74
- CNL
 - generation, 95
- NLP
 - Pragmatic Interpretation in, 78
- OBO Foundry ontologies, 52
- OBO foundry
 - ontologies integration, 54
- OGCP
 - Classes, 136
 - clinical-administration-act, 137
 - Data properties, 142
 - Individuals, 142
 - Object properties, 141
 - presuppositions, 47
 - Role, 136
- OGCP ontologies alignment, 52
- SNOMED-CT, 49
- UMLS CORE, 51
- Appendix
 - OGCP Description Logics (DL), 136
 - Symbols and terminology, 134
- Clinical concept guidance, 92

- Clinical Knowledge
 - Acquisition, 43
 - Using the OGCP for CCL building, 44
- Clinical Practice
 - Ontological Relations, 38
- Clinical reasoning, 91
- Clinical text available sources, 39
 - Cardiology and ICU in Portalegre district, 42
 - Structure, scope, adequacy, 42
- Comparable high standard formal results evaluation, 116
- Conclusions, 113
 - Extension of the concept to different clinical specialties, 115
 - Extension of the concept to wider geographical coverage, 115
 - Future Work, 115
 - Resource access limitations, 114
 - State of The Art Restrictions, 113
 - OWL-Time, 114
 - SO, 114
 - Foundational ontology weaknesses, 114
 - Time or other type of constraints, 114
- Current on-going controlled results, 110
- Domain experts validation, 111
- Future Work
 - Software module to complement the AOL, 115
- Introduction
 - Research context and motivation, 1
 - Research questions, 2
 - Scientific Innovation, 3
- Knowledge Acquisition
 - Text interpretation, 83
 - DRS rewriting methodology, 88
 - OGCP enhancements in order to represent healthcare practice episodes, 86
 - Ontology structure considerations, 84
 - Preliminary considerations, 84
- OWL-Time Ontology, 55
- Quality indicators, 96
- Reasoning with effective logics, 82
- Reasoning with the Discourse Structure, 80
- Results
 - OGCP Population examples, 100
- Round Trip Debug and Repair, 80
- Smart instance creation, 77
- State of the Art
 - Web Ontology Language (OWL) Reasoners, 14
 - Controlled Natural Language, 21
 - Knowledge Acquisition through Information Extraction from text NLP, 19
 - Ontology Learning, 18
 - OWL Reasoning, 12
 - RDF data, 8
 - Biomedical Knowledge Representation, 18
 - Biomedical Resources, 17
 - Consequence driven reasoning, 15
 - Ethical and legal issues, 20
 - Generic and Biomedical Ontologies, 17
 - Identity tracking, 19
 - Privacy and De-identification, 20
 - Reasoning about Knowledge, 11
 - Kinds of , 11
 - Uses for Reasoning, 11
 - Reasoning Support for OWL, 13
 - Semantic Web, 7
 - Temporal Information Retrieval (IR), 19
- Supervised tutoring, 64
- System Architecture, 97
- Text interpretation, 83
- The ELK reasoner, 15
- Tokenization, 73
- Tools and Technologies
 - Apache Jena, 35
 - Description Logics, 9
 - Description Logics and Web Ontology Language v.2, 8
 - Protégé, 32
 - Protégé Ace View Plugin, 34
 - Protégé OWL API, 34
 - Apache OpenNLP, 28
 - Apache Tika, 29
 - Apache UIMA, 29
 - Apache cTAKES, 31

OWL, 10

UIMA Ruta Workbench, 30

ACE

 Attempto Controlled English, 22

 ACE to OWL limitations, 27

 Anaphoric references, 26

 Constraining ambiguities, 25

 From ACE to OWL, 24

Tools for Ontology manipulation, 32

Tools for specialized NLP, 27