



HAL
open science

Non-Pharmacological Interventions : Terminology Acquisition and Visualization

The Loc Nguyen

► **To cite this version:**

The Loc Nguyen. Non-Pharmacological Interventions : Terminology Acquisition and Visualization. Other [cs.OH]. Université Montpellier, 2018. English. NNT : 2018MONT090 . tel-02181583

HAL Id: tel-02181583

<https://theses.hal.science/tel-02181583>

Submitted on 12 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Informatique

École doctorale 166 I2S

Unité de recherche LIRMM

NON-PHARMACOLOGICAL INTERVENTIONS: TERMINOLOGY ACQUISITION AND VISUALIZATION

Présentée par The Loc NGUYEN
Le 12 juin 2018

Sous la direction de Anne LAURENT
et Grégory NINOT

Devant le jury composé de

Thérèse LIBOUREL, Professeur émérite, Université de Montpellier

Maria RIFQI, Professeur, Université Paris 2

Florence SEDES, Professeur, Université de Toulouse

Samuel SZONIECKY, MCF, Université Paris 8

Raphael TROUILLET, MCF HDR, Université Paul Valéry

Sylvie RAPIOR, Professeur, Université de Montpellier

Anne LAURENT, Professeur, Université de Montpellier

Grégory NINOT, Professeur, Université de Montpellier

Président du jury

Rapporteur

Rapporteur

Examineur

Examineur

Examineur

Co-directrice

Co-directeur



UNIVERSITÉ
DE MONTPELLIER

Dedication

This thesis is lovingly dedicated

to my beloved wife Hang “Meocoi”,

to my son Nhat Nam “Zin” and my little daughter Linh San “Moon”.

Daddy is coming home.

Acknowledgments

Firstly, I dedicate my great thankfulness to my advisors Prof. Anne Laurent and Prof. Grégory Ninot for all their help and guidance that they have given me during my PhD journey.

I would like to thank my thesis jury members for accepting to review my work. Special thanks to Prof. Maria Rifqi and Prof. Florence Sedes for their valuable comments and remarks as reviewers of my dissertation. My sincere thanks also go to Samuel Szoniecky and Raphael Trouillet for their time to act as examiners for my work.

I would also like to express my sincere gratitude to Prof. Thérèse Libourel and Prof. Sylvie Rapior for their precious support and comments. My special thanks go to my friend DuyHoa Ngo for the valuable discussions via Skype between Australia and France.

I wish to thank my colleagues at the Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM) and the Plateforme CEPS for interesting discussions as well as relaxed coffee time.

I would like to thank the Ministry of Education and Training of Vietnam (MOET) for the scholarship and the Hanoi University of Mining and Geology (HUMG), where I work, which helped and gave me an opportunity to study in France.

Also, I am grateful to all my good friends in Montpellier who are not mentioned by names but have always been beside me whenever I needed their assistance and helps in work as well as my life. They really made me less homesick by sport activities (football, badminton, ping-pong, tennis, volleyball) and exciting BBQ parties.

Finally, and most importantly, my love is with my great family, my parents, my elder brother and sister, especially, my beloved wife, my son and daughter for their encouragement and unconditional support.

Montpellier 06/2018,

NGUYEN The Loc

Abstract

The explosion of data on the Internet leads to challenges in working with them. Semantic Web and ontology are required to address those problems. Nowadays, ontology plays more and more an important role as a means in domain knowledge representation.

In recent years, Non-Pharmacological Interventions (NPIs) have attracted a lot of attention in the health care community. NPIs can no longer stop at a professional discipline to describe them (psychotherapy, manual therapy, dietary supplement, adapted physical activity, e-health solution, etc.). It requires access to a more concrete level of description where each NPI can be evaluated by science, monitored by professionals and explained to the patient. To do this, an international and evolutionary classification based on the results of science is necessary. Thus, developing an ontology for NPIs is crucial. This ontology will facilitate bibliographic research, usage statistics and the identification of good practices.

Constructing this ontology manually is time consuming and thus an expensive process. Particularly, the step of collecting the NPI terminology requires much more time than the rest, because of heterogeneous and big resources in the context of NPIs. An automatic or semi-automatic method is thus essential to support NPI experts in this task.

Besides, ontologies are often complex with lots of classes, properties and relationships. They are not easy to understand by domain experts. Therefore, a simple and friendly visualization of the ontology for NPI experts needs to be considered. The visualization does not only help NPI experts to easily understand the ontology but also provides support for the NPI ontology development process.

In this thesis, we propose methodologies to address the aforementioned challenges. The first contribution concerns the semi-automatic process for collecting NPI terms. Two approaches, knowledge-based and corpus-based, are presented to retrieve candidate NPI terms. A new similarity measure for NPI is proposed and evaluated. The second contribution is a new method for ontology visualization based on MindMap. This work aims at providing a simple and friendly tool to visualize an ontology which is used by domain experts. We propose a MindMap-based notation for ontology visualization by transforming ontology components to MindMap elements. A web-based tool is then implemented to convert OWL ontologies to FreeMind documents which can be imported by existing Mind-Mapping applications to make visualizations.

Résumé

Le volume de données disponible croît de manière très importante et ouvre d'importants défis pour les exploiter. Les domaines scientifiques du Web sémantique et des ontologies sont alors une réponse pour aider à traiter les données de manière efficace. Ainsi les ontologies sont actuellement devenues incontournables dans de nombreux domaines d'application pour représenter la connaissance experte.

Le domaine que nous considérons dans nos travaux est celui des Interventions Non Médicamenteuses (INM) nommées Non-Pharmacological Interventions (NPIs) en anglais. Elles sont de plus en plus étudiées sur le plan scientifique. Elles sont liées à divers secteurs : psychologie, thérapies manuelles, nutrition, activités sportives adaptées, solutions e-santé, etc.

Avec l'augmentation de leur usage, il devient de plus en plus nécessaire d'évaluer leur efficacité de manière scientifique, dans une démarche pilotée par des spécialistes et expliquée de manière claire et accessible aux utilisateurs. Pour ce faire, il est essentiel de disposer d'une classification évolutive et consensuelle effectuée au niveau international pour les spécialistes. Dans ce domaine, le développement d'une ontologie est crucial pour faciliter les recherches bibliographiques et mettre en place des bonnes pratiques.

Dans nos travaux, nous nous sommes intéressés à deux enjeux majeurs liés à la construction d'une telle ontologie, d'une part comment effectuer la collecte du vocabulaire et d'autre part comment aider à la compréhension par visualisation.

La construction manuelle de l'ontologie est en effet fastidieuse et longue. En particulier, la collecte des termes liés au domaine des INM nécessite beaucoup d'efforts et de temps tant le champ du vocabulaire est large. Ainsi le terme INM

lui-même est parfois remplacé par d'autres (médecines alternatives, médecines douces, etc). Une méthode automatique ou semi-automatique est alors vue comme une aide importante pour la construction de la représentation de la connaissance.

De plus, les ontologies sont parfois considérées comme difficiles à prendre en main pour les personnes non spécialistes de modélisation, en raison de leur complexité, de leur taille ou des propriétés et relations qu'elles incluent. Ainsi, un outil de visualisation doit être proposé pour les experts des INM. L'outil aura deux buts, d'une part visualiser l'ontologie existante, d'autre part proposer des modifications relatives à la structuration de l'ontologie qui doit se construire de manière collaborative.

Des contributions sont proposées dans cette thèse sur ces deux sujets (construction du vocabulaire et visualisation). Deux approches sont présentées pour la construction, l'une reposant sur la connaissance experte et l'autre sur un corpus. Une mesure de similarité est introduite et évaluée. Pour la visualisation, notre proposition repose sur l'utilisation de cartes conceptuelles. Il s'agit alors de ré-écrire l'ontologie sous ce nouveau format et de proposer des outils permettant de distinguer les différents éléments et liens entre les éléments. Un outil a été implémenté, permettant de transformer les ontologies décrites en OWL pour les visualiser.

Contents

List of Figures	xiii
List of Tables	xv
I Preliminaries	1
1 Introduction	3
1.1 Context	3
1.2 Motivations	10
1.3 Objectives and Contributions	13
1.4 Structure of the thesis	14
2 Background	17
2.1 Semiotic and elements	17
2.2 Knowledge Organization Systems	19
2.2.1 Taxonomy	20
2.2.2 Thesaurus	20
2.2.3 Ontology	21
2.3 Encoding Knowledge Organization Systems	24
2.3.1 Resource Description Framework (RDF)	25
2.3.2 Resource Description Framework Schema (RDFS)	26
2.3.3 Web Ontology Language (OWL)	27

2.3.4	Simple Knowledge Organization System (SKOS)	29
2.4	Visualizing Knowledge Organization Systems	30
2.4.1	MindMap	30
2.4.2	Concept Map	31
II	NPI Terminology Acquisition	35
3	Problem Statement	37
4	Related Works	39
4.1	Biomedical resources	39
4.2	Automatic term extraction	41
4.3	Semantic similarity measures	42
4.3.1	Knowledge-based similarity measures	42
4.3.2	Corpus-based similarity measures	43
4.3.3	Hybrid similarity measures	44
5	Methodology	47
5.1	Term extraction	47
5.1.1	Exploiting knowledge bases	48
5.1.1.1	Extracting terms from WordNet	48
5.1.1.2	Extracting terms from ontologies in BioPortal	49
5.1.2	Exploiting text corpus	51
5.2	Term ranking	53

<i>Contents</i>	xi
5.2.1 Edit-based similarity measures	56
5.2.2 Knowledge-based similarity measures	57
5.2.3 Distributional similarity measures	58
5.2.3.1 Vector representation of words	60
5.2.3.2 Calculating similarity between two words	65
5.2.3.3 The similarity for multi-token terms	66
5.3 Term validation	68
6 Results and Evaluation	69
6.1 Results	69
6.2 Evaluation	73
7 Conclusion	79
III MindMap-based Ontology Visualization	81
8 Problem Statement	83
9 Related Works	85
9.1 Hierarchy-style visualization	85
9.2 Graph-style visualization	86
10 Methodology	93
10.1 MindMap-based visual notation	93
10.2 Transforming OWL into FreeMind XML	93

11 Results and Evaluation	99
11.1 Results	99
11.2 Evaluation	102
12 Conclusion	109
IV General Conclusion and Perspectives	111
13 General Conclusion	113
14 Perspectives	115
14.1 NPI term acquisition improvement	115
14.1.1 Linguistic pattern-based method	115
14.1.2 Rule-based post-filtering step	116
14.2 Term linkage	118
14.3 MindMap-based ontology editor	118
Appendices	119
A NonPhaTex Web-based Application	121
B Retrieved NPI terms	125
Bibliography	137

List of Figures

1.1	NPI taxonomy	7
1.2	Workflows in our customized UPON	12
2.1	The triangle of reference	18
2.2	An example of taxonomy “Animals”	20
2.3	An example of thesaurus “Animals” (taken from [67])	21
2.4	An example of ontology “Computers”	22
2.5	Architecture of the Semantic Web	25
2.6	RDF/XML syntax	28
2.7	RDF/Turtle syntax	29
2.8	A mind map of the NPI taxonomy created by FreeMind	32
2.9	The corresponding FreeMind document of the mind map in Figure 2.8	33
2.10	An example of concept map	34
3.1	NPI term acquisition process	38
5.1	Structure of WordNet	48
5.2	WordNet in RDFS and OWL	49
5.3	Extracting related terms from MeSH	50
5.4	Extracting candidate NPI terms from text corpora	52
5.5	Candidate NPI terms and clusters of seed NPI terms	54

5.6	NPI score calculation	55
5.7	Similarity measure based on the shortest path (taken from [65])	59
5.8	Word vector training process	62
5.9	CBOW and Skip-Gram model [50]	65
5.10	Manual term validation process by NPI experts	68
6.1	Clustering the seed terms with K-means ($k = 3$)	71
6.2	Experimental results	73
6.3	Precision by NPI score threshold	78
9.1	Visualization of General Medical Science ontology by OWLViz .	87
9.2	Visualization of General Medical Science ontology by OntoGraf	88
9.3	Visualization of General Medical Science ontology by SOVA . .	89
9.4	Visualization of General Medical Science ontology by WebVOWL	91
11.1	A part of an OWL document with RDF/XML syntax of Wine ontology	100
11.2	A corresponding part of a FreeMind document	101
11.3	The MindMap visualized by FreeMind tool	103
11.4	The MindMap visualized by Mindomo tool	104
11.5	The multi-colored MindMap visualized by FreeMind tool	105
A.1	NonPhaTex main interface	123

List of Tables

6.1	List of seed terms	70
6.2	Seed terms in the first cluster	71
6.3	Seed terms in the second cluster	72
6.4	Seed terms in the third cluster	72
6.5	Top candidate terms using ACS to calculate NPI score	74
6.6	Top candidate terms using UMBC to calculate NPI score	75
6.7	Top candidate terms using Average Vector to calculate NPI score	76
10.1	MindMap-based visual notation for ontology elements	94
10.2	Mapping ontology elements to FreeMind elements	95
11.1	Visualization capabilities of OWL2MM and other tools	102
11.2	Usability of OWL2MM and other tools	106
14.1	The UMLS semantic types related to NPIs	117
B.1	List of current existing NPI terms	136

Part I

Preliminaries

1	Introduction	3
1.1	Context	3
1.2	Motivations	10
1.3	Objectives and Contributions	13
1.4	Structure of the thesis	14

2	Background	17
2.1	Semiotic and elements	17
2.2	Knowledge Organization Systems	19
2.2.1	Taxonomy	20
2.2.2	Thesaurus	20
2.2.3	Ontology	21
2.3	Encoding Knowledge Organization Systems	24
2.3.1	Resource Description Framework (RDF)	25
2.3.2	Resource Description Framework Schema (RDFS)	26
2.3.3	Web Ontology Language (OWL)	27
2.3.4	Simple Knowledge Organization System (SKOS)	29
2.4	Visualizing Knowledge Organization Systems	30
2.4.1	MindMap	30
2.4.2	Concept Map	31

Introduction

THIS chapter presents relevant information of the thesis such as the definition of research field related to the dissertation as well as an introduction about the context of this work, namely Plateforme CEPS ¹. Then motivations and objectives of the thesis are figured out. Consequently, the thesis contributions are presented for achieving the goals. Finally, the structure of thesis is given.

1.1 Context

“Non Pharmacological Interventions (NPIs) are science-based and non invasive interventions on human health. They aim to prevent, treat, or cure health problems. They may consist in products, methods, programs or services whose contents are known by users. They are linked to biological and/or psychological processes identified in clinical studies. They have a measurable impact on health, quality of life, behavioral and socioeconomic markers. Their implementation requires relational, communicational and ethical skills.” (Plateforme CEPS, 2017).

Some examples of NPIs are: Mindfulness-based stress reduction (MBSR) which is a program using mindfulness to reduce pain from patients; Eye move-

¹<http://www.plateforme-ceps.fr>

ment desensitization and reprocessing (EMDR) which is a form of psychotherapy using eye movements to help trauma victims in processing distressing memories and beliefs; Diet programs such as Cohen diet, Mayo diet.

The Plateforme CEPS [59], founded in 2011, is an Academic Collaborative Platform of methodology experts in clinical non-pharmacological research. Its mission is to foster the monitoring, design, implementation and publication of interventional studies dedicated to the assessment of NPIs' efficacy. In other words, the Plateforme CEPS is involved in improving the quality of NPI studies. This work requires a multidisciplinary approach, from the biological sciences to the human sciences, from technological engineering to mathematics, from economics to political science, from history to philosophy. A single scientific discipline can not capture the complexity of an action of a NPI on human beings. The Plateforme CEPS is operated by the Paul Valéry University in Montpellier, France. Financial support comes from the Contrat de Plan Etat Région - CPER 2015-2020 and Metropole de Montpellier and National Health Agencies.

Although the definition of NPIs has just been defined by the Plateforme CEPS since 2017, other equivalent definitions were mentioned earlier such as Complementary and Alternative Medicine (CAM), Integrative Medicine, Traditional Medicine (TM) [61].

The World Health Organization (WHO) defines traditional medicine as *“the sum total of the knowledge, skill, and practices based on the theories, beliefs, and experiences indigenous to different cultures, whether explicable or not, used in the maintenance of health as well as in the prevention, diagnosis, improvement or treatment of physical and mental illness”* [89].

The WHO also proposes the definition of the terms “complementary medicine” or “alternative medicine” as *“a broad set of health care practices that are not part of that country’s own tradition or conventional medicine and are not fully*

integrated into the dominant health-care system. They are used interchangeably with traditional medicine in some countries” [89]. So-called “CAM” is typically defined as a heterogeneous group of health care practices such as Ayurvedic medicine and other Indian systems of medicine such as Yoga, varieties of Chinese medicine, homeopathy, Swedish massage, Qi Gong [49].

In the United State, the National Center for Complementary and Integrative Medicine (NCCIH) ², formerly known as the National Center for Complementary and Alternative Medicine (NCCAM), defines CAM as *“a group of diverse medical and health care systems, practices, and products that are not presently considered to be part of conventional medicine” [31].*

In Europe, the CAMbrella Project ³ recently defined CAM as follows: *“CAM, as utilised by European citizens, represents a variety of different medical systems and therapies based on the knowledge, skills and practices derived from theories, philosophies and experiences used to maintain and improve health, as well as to prevent, diagnose, relieve or treat physical and mental illnesses. CAM therapies are mainly used outside conventional health care, but in many countries some therapies are being adopted or adapted by conventional health care” [17, 48].*

CAM and TM are now recognized as global health care phenomena (World Health Organization - WHO 2013). They almost occur in all WHO member states, encompassing specific therapies or treatment, healing systems, and health care professions. CAM plays a more and more important role in daily life and medicine. Nearly two-thirds of French and more than 100 million Europeans use it [58]. In the United State of America, more than 30 percent of adults and about 12 percent of children use health care approaches developed outside of mainstream medicine (conventional medicine or West-

²<https://nccih.nih.gov/>

³<https://cambrella.eu/home.php>

ern medicine) [54]. The trend of using CAM in Cancer Care attracts more research [21] and people working with cancer patients [7].

A classification of CAM is very essential. There have also been many attempts to classify and categorize the various types of CAM/TM. One of the most widely used classification, developed by NCCIH, divides CAM therapies into five categories or domains [55]:

- Alternative medical systems: therapy and practice systems that developed separately from conventional medicine, such as traditional Chinese medicine, ayurvedic medicine, homeopathy, and naturopathy;
- Mind-body interventions: techniques based on the human mind that have effect on physical health and symptoms, such as meditation, prayer, and mental healing;
- Biologically-based systems: specialized diets such as those proposed by Drs. Atkins and Ornish, herbal products such as St. John's wort and Ginkgo biloba, and other natural products minerals, hormones, and biologicals;
- Manipulative and body-based methods: therapies related to movement or manipulation of the body, such as chiropractic and massage therapy;
- Energy therapies: the manipulation and application of energy fields to the body such as Qigong, Reiki, and therapeutic touch.

In 2017, the Plateforme CEPS proposed a taxonomy of NPIs including five categories as follows (also illustrated in Figure 1.1):

- Psychological Health Interventions: from prevention programs to psychotherapy interventions (Art Therapy, Health Education, Psychotherapy, Zootherapy);

 Psychological Health Interventions	 Physical Health Interventions	 Nutritional Health Interventions	 Digital Health Interventions	 Other Health NP Interventions
Art Therapy Health Education Psychotherapy Zootherapy	Physical Activity Hortitherapy Physiotherapy Manual Therapy Thermalism	Dietary Supplements Nutritional Therapy	eHealth Devices Therapeutic Games Virtual Reality Therapy	Ergonomic tools Phytotherapy Cosmetic Therapy Wave Therapy Lithotherapy

Figure 1.1 – NPI taxonomy

- Physical Health Interventions: from manual therapy to therapeutic physical activity programs (Physical Activity, Hortitherapy, Physiotherapy, Manual Therapy, Thermalism);
- Nutritional Health Interventions: from supplementary food products to diet interventions (Dietary Supplements, Nutritional Therapy);
- Digital Health Interventions: from health wearable and handheld devices to health coaching programs (eHealth Devices, Therapeutic Games, Virtual Reality Therapy);
- Other Health Interventions: from phytotherapy to aromatherapy (Ergonomic tools, Phytotherapy, Cosmetic Therapy, Wave Therapy, Lithotherapy).

Recently, health care has been an interesting field for data mining researchers. In the framework of health, Non-Pharmacological Interventions (NPIs) are attracting more and more interest. They are indeed an essential complement to curative medicine, especially for improving the quality of life or even life expectancy. However, the data connected to this area have been less explored than the data based on biology, chemistry and medicine. Even

if such data can be taken from similar data sources and treated in a similar manner, no solution is yet available to fully exploit them. This information is often based on textual data. In particular, the vocabularies are not yet shared. Moreover, data crossing, mining, recommending and publishing are still difficult. Exploring the ways to manage and exploit data on NPIs are completely necessary.

Besides, the objectives of the Plateforme CEPS are to contribute to the existing ecosystem by providing researchers with open access resources for the evaluation of NPIs and the systematic review of specific publications [59]. The Plateforme CEPS proposes a metadata search engine named Motrial [60], methodological resources for the NPI research, an international congress, and a research network for NPI community. All of them are freely accessible for researchers and clinicians. More specifically, the objectives of the Plateforme CEPS are ⁴:

1. *to bring together actors in NPI research through the organization of an international congress in Montpellier which has gathered more than 3000 participants from 16 different nationalities in 5 editions since 2011 (iCEPS Conference ⁵);*
2. *to identify all NPIs in an ontology combining professional expertise and artificial intelligence (NPI ontology [61]);*
3. *to improve the identification of quality interventional studies to facilitate systematic reviews and meta-analyzes of NPIs (Motrial search engine);*
4. *to identify researchers and research organizations working on the evaluation of NPIs in France, Europe, and worldwide based on their publications of interventional studies (NIRI system delivered in 2019);*

⁴<https://plateformeceps.www.univ-montp3.fr/fr/page-1/page-11>

⁵<https://plateformeceps.www.univ-montp3.fr/fr/page-2>

5. *to encourage quality studies to better understand the benefits and risks of each NPI through the distribution of slideshows and open source scientific documents (NISHARE system delivered in 2020).*

The objective 3 of Plateforme CEPS aims at building a search engine named Motrial ⁶ to help NPI researchers to identify quickly main publications of NPI trials [60]. It will reduce not only search time but also errors in result of the search process. Besides, this search engine will also help them to identify authors and research team all over the world based on trial publications.

There are various ways to categorize search engines. According to our knowledge, search engines are divided into two groups: traditional search engines and semantic ones.

Traditional search engines work by matching query terms against a keyword-based index. They will fail to match relevant information when the keywords used in a query different from those used in the index, despite having the same meaning. As a result, not all the documents related to the search query can be retrieved. The ambiguity of the query also leads to the retrieval of irrelevant information.

Semantic search engines leverage domain-specific knowledge of ontologies to improve retrieved results. They do not only look for matching keywords, but attempt to identify the intent and deeper meaning of a search based on each of the words used in a query. An important role of ontologies is to expand search queries. It also serves as schemata or intelligent view over information resources. Thus they can be used for indexing, querying, and reference purposes over non-ontological datasets and systems.

The value of search engines is given by quality measures as for example precision and recall [34]. Precision ensures that a result retrieved by a search

⁶<http://www.motrial.fr>

engine is relevant, while recall ensures that all relevant information have been retrieved. Some works have shown that adding domain knowledge, represented by ontologies, increases the quality of search engines [3]. For this reason, the Plateforme CEPS aims at integrating an ontology related to NPI domain into the Motrial. The objective 2, therefore, becomes crucial for the objective 3.

The objective 2 of Plateforme CEPS aims at building a shared ontology that will be considered as an open reference when dealing with NPIs. Ontology is a way of describing and linking data and their interrelations across the globe on the web [1, 16]. An ontology, between other characteristics, defines a standard vocabulary for domain-specific researchers who need to share information and cross data in their fields. It includes definitions of concepts with relations among them for a specific domain. It is often encoded in machine-interpretable format [62]. Nowadays, ontologies is becoming increasingly an important role as backbones in ontology-based search engines. Once the NPI ontology is built, it will be used as a component of the search engine Motrial to expand and enrich queries.

In reality, the objective 2 would be a time-consuming work if it was performed manually by NPI experts. Therefore, an automatic or semi-automatic approach should be considered to help them to reduce time and human efforts in the development of the ontology. This thesis is designed to assist them to achieve this goal.

1.2 Motivations

In order to build an ontology, experts often follow one of ontology development methodologies such as METHONTOLOGY [18], NeOn [19], UPON [14] etc. Each method has different detailed processes. NeOn methodology focuses on the development of ontologies based on the reutilization, modification, restruc-

ture, and the adaptation of already existent ontologies. METHONTOLOGY and UPON propose processes to build ontologies from scratch. In the framework of Plateforme CEPS, we use UPON with some modifications for the NPI ontology development because it is highly scalable and customizable. Basically, our customized UPON is based on five workflows as being shown in Figure 1.2:

1. Requirements: In this workflow, ontology experts and domain ones work together to determine the domain of interest and the scope. The business purpose is also defined in this step via competency questions.
2. Analysis: This workflow concerns acquiring domain resources and building domain lexicon. This is the most time-consuming step in the whole ontology development process. Contributions of domain experts and ontology ones are both required for this workflow.
3. Design: The main goal of this workflow is to give an ontological structure to the lexicon gathered in the previous workflow. Concepts and relationships between them are modeled in this step.
4. Implementation: The purpose of this workflow is to encode the ontology in a formal language, for example, Web Ontology Language (OWL). This step must be performed by ontology engineers.
5. Test: In this final workflow, the consistency and the coverage of ontology are checked. Then the ability of answering competency questions is verified.

In the second workflow (Analysis), acquiring a set of NPI terms to develop the ontology is the most time-consuming task. In the beginning, the Plateforme CEPS had only few NPI terms and the NPI experts have been trying to find out more terms. They have used the traditional method to solve this problem by searching and reading articles, book or web pages on the Internet. As a

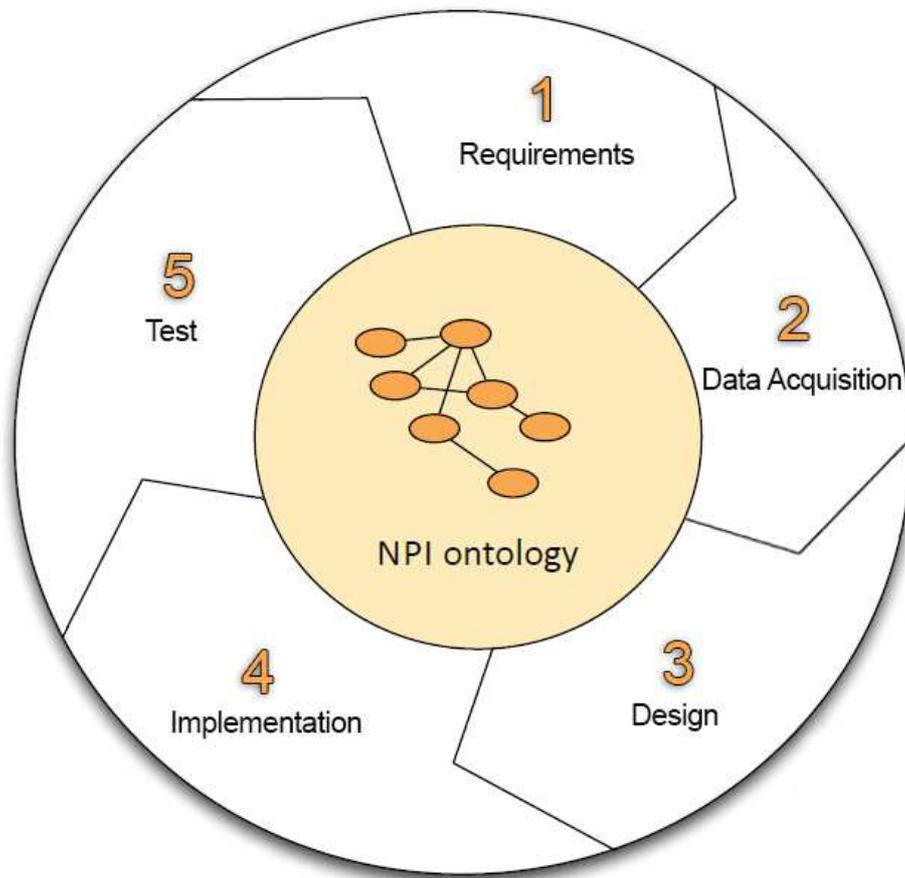


Figure 1.2 – Workflows in our customized UPON

result, they needed a lot of time to figure out several terms. This motivates us to propose a semi-automatic method to collect NPI terms in order to build NPI ontology.

During the ontology development, domain experts need to work and discuss together. They often meet difficulties with some traditional tools such as Microsoft Word, Excel because those tools cannot display ontology-related data efficiently, while dedicated tools like Protege or WebProtege [77, 78] might be complicated for them. A new tool to support domain experts to work effectively in ontology development is essential. It is also our second motivation.

1.3 Objectives and Contributions

This thesis focuses on the objective 2 of Plateforme CEPS (building an ontology of NPIs) which is crucial for the objective 3. We aim at helping the members of Plateforme CEPS to build a shared ontology in a collaborative manner.

We focus on two main goals.

- Initially, the Plateforme CEPS just had some seed NPI terms. One of the goals of this thesis is to acquire more NPI terms based on these seed terms. In reality, NPI terms are spread in many different knowledge bases and corpora. It is hard to distinguish NPI terms from the others. Thus, we propose a method to acquire NPI terms based on semantic similarity from both sources. Given some seed terms, our method not only retrieves similar terms but can also get related terms.
- Ontologies are often complicate to understand by domain experts. Thus efficient visualizations are necessary for them to understand and work on domain knowledge. However, existing visualization tools are often designed for engineering experts. As a result, they are difficult to use by domain experts. Therefore, the remaining goal of this thesis is to develop a new ontology visualization tool for the experts who are working on many different domains, especially NPI domain.

The contributions of my PhD thesis are the following ones:

- NPI terminology acquisition goal:
 - A methodology to acquire new NPI terms from some seed terms;
 - A semantic metric to measure the similarity between two NPI terms;
 - A web-based application to assist NPI experts to validate the NPI term candidates.

- Visualization goal:
 - A MindMap-based notation to visualize elements of an ontology;
 - A methodology, implemented in a Java-based tool, for transforming OWL format to FreeMind format which is imported by existing mind-mapping tools in order to create visualizations.

1.4 Structure of the thesis

This dissertation is organized into 14 chapters which are grouped into 4 parts.

Firstly, in the Part I, Chapter 1 introduces the context of the thesis, then goals and contributions of our work. Next, Chapter 2 summaries background knowledge related to Knowledge Organization Systems (KOS).

Afterwards, chapters 3 to 7, grouped in the Part II, present corresponding work and show the solutions for the NPI terminology acquisition problem. This part starts with the problem statement (Chapter 3) followed by a related works on biomedical resources, term extraction and semantic similarity measures (Chapter 4). From the limitations of current works in literature, a solution based on a distributional similarity measure is proposed to deal with the challenges (Chapter 5). In consequence, Chapter 6 presents some results and evaluations of the proposed methodology. This part is concluded by a summary in the Chapter 7.

From Chapter 8 to Chapter 12, organized in the Part III, we introduce the work on MindMap-based ontology visualization. The same structure as the previous part is repeated with contents for this problem. Chapter 8 introduces the problem and challenges on ontology visualization that we have to address. Chapter 9 presents literature review on hierarchy-style and graph-style visualization for ontologies. In Chapter 10, we propose a visual notation based on

MindMap for ontology visualization. Results and evaluation are presented in Chapter 11 in order to estimate our method. Chapter 12 close this part by a summary.

Finally, in the Part IV, a general conclusion summarizes the whole thesis in Chapter 13. Perspective works are mentioned in Chapter 14.

Background

IN the framework of Plateforme CEPS, researchers in various domains work together. Therefore, the need of a reference terminology is essential. A terminology is a kind of Knowledge Organization Systems such as taxonomy, thesaurus, ontology with different semantic levels [12, 90]. This chapter presents background knowledge related to Knowledge Organization Systems, how to encode and visualize them.

2.1 Semiotic and elements

A word often has more than one definition. According to the Oxford dictionary, there is an average of 28 meanings per word in the 500 most used words in English. As a result, our communication attempts sometimes fail because of misconceptions and ambiguity.

In order to avoid misunderstandings and confusion, communications need to be expressed by the triangle of reference (also known as triangle of meaning, or semiotic triangle) proposed by Ogden and Richard [63], as illustrated in Figure 2.1. The triangle describes a simplified form of relationship between the speaker as subject, a concept as object or referent, and its designation as sign. The “SYMBOL” in Figure 2.1 can be a word, a term or a token. Words have no exact or clear meaning, and meaning depends on context. It is essential to define terms or concepts to avoid this ambiguity.

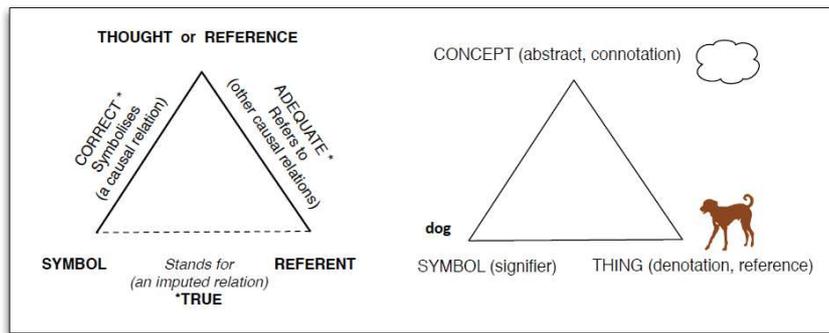


Figure 2.1 – The triangle of reference

For example, when a word “orange” is presented, the corresponding reference may be a fruit or a color. The correct meaning can be well understood if the word is presented in the specific context.

The following parts introduce definitions of some basic elements used in the triangle of reference.

Token: A token is a sequence of continuous characters between two spaces, or between a space and punctuation marks [47]. A token can also be an integer, real, or a number with a colon (for example, 2:00 is a token to indicate time). All other symbols are tokens themselves except apostrophes and quotation marks in a word (with no space), which in many cases symbolize acronyms or citations [8].

Term: A term is a lexical unit that has an unambiguous meaning when used in a text of a specific domain [47]. Terms are the linguistic representation of concepts in a knowledge domain. They are often treated as a single lexical unit even though they might be composed of more than one orthographic word. A set of domain-specific terms constitutes a specialized lexicon or a terminology [68].

Multi-token term: A multi-token term (multi-word term) is a term that is composed of two or more words. The unambiguous semantics of a multi-token term depends on the knowledge area of the concept it describes and cannot be inferred directly from its parts [20, 72].

Controlled vocabulary: A controlled vocabulary is a list of terms that have been enumerated explicitly [67]. This list is controlled by and is available from a controlled vocabulary registration authority. All terms in a controlled vocabulary should have an unambiguous, non-redundant definition [30]. A set of terms is a truly controlled vocabulary if people agree on and have to pick from it whenever they want to refer to the same thing. For example, “Yes, No” is one controlled vocabulary; so is “Mr., Ms., Miss, Mrs., Dr.”; another could be “Cat, Poodle, Mammal, Collie, Dog, Manx, Bulldog”.

Despite being enumerated explicitly, a term of a controlled vocabulary is still misunderstanding sometimes if it lacks a clear description in a specific context. In order to avoid this problem, terms need to be organized into a structured system. The next section presents so-called Knowledge Organization Systems [12, 90].

2.2 Knowledge Organization Systems

The rapidly increasing amount of information in domain knowledges poses challenges to retrieval, integration and reuse of information related to specific contexts. Efficient methods for organizing knowledge systems are crucial and plays more and more important roles. In order to strengthen semantic meaning and give a specific context of words or terms, people often organize them into Knowledge Organization Systems (KOS) such as taxonomy, thesaurus, ontology [12, 90]. Following parts present definitions of those systems.

2.2.1 Taxonomy

A taxonomy is a collection of controlled vocabulary terms organized into a hierarchical structure [67]. Each term in a taxonomy is in one or more parent-child relationships to other terms in the taxonomy. In practice, the parent-child relationships are often “is-a” type. However, different types of parent-child relationships such as “whole-part”, “genus-species”, “type-instance” may be included in a taxonomy. In some taxonomies, a term can have multiple parents. Such taxonomies are called poly-hierarchy [30]. For example, in the example animal controlled vocabulary mentioned above, if the Cat is a broader term for Manx, the Dog is a broader term for Collie and Bulldog, and the Mammal is a broader term for Dog and Cat, then the controlled vocabulary becomes a simple taxonomy (see Figure 2.2).

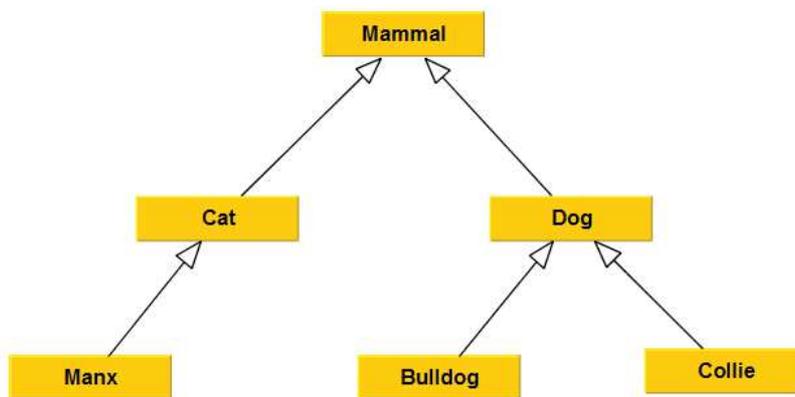


Figure 2.2 – An example of taxonomy “Animals”

2.2.2 Thesaurus

A thesaurus is a networked collection of controlled vocabulary terms. This means that a thesaurus might have associative relationships in addition to parent-child relationships [30]. It stores more metadata than a taxonomy. A thesaurus can be considered as a taxonomy with additional metadata. There-

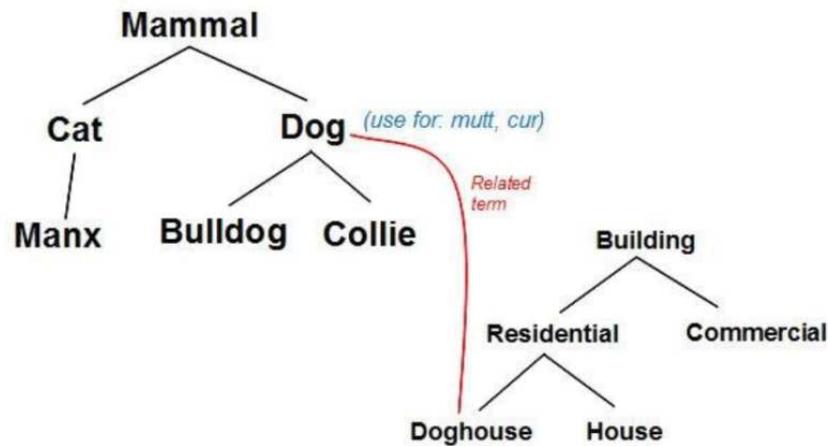


Figure 2.3 – An example of thesaurus “Animals” (taken from [67])

fore, the terms “thesaurus” and “taxonomy” are sometimes used interchangeably [67]. A thesaurus might store some relationships of terms such as “Opposite”, “Use For”. For example, “Yes” is an opposite term of “No”, “Doghouse” is used for “Dog”. These relationships help search systems to redirect to the preferred term of a particular term that is not considered to be the best one. For instance, in Figure 2.3, if someone inputs the term “Dog”, then a search system can output the term “Doghouse” thanks to the advantage of the “Use For” relationship.

2.2.3 Ontology

Ontologies are playing a more and more important role in many fields as efficient knowledge representation. There are many definitions of ontologies, but some definitions proposed by following authors achieved more agreement than the rest. In 1993, Grubber originally defined an ontology as a “*specification of a conceptualization*” [23]. Later, in 1997, Borst introduced a definition in which the conceptualization should be shared view and must be expressed in formal format. Specifically, he defined an ontology as a “*formal specification of a shared conceptualization*” [10]. Guarino then detailed the notions of con-

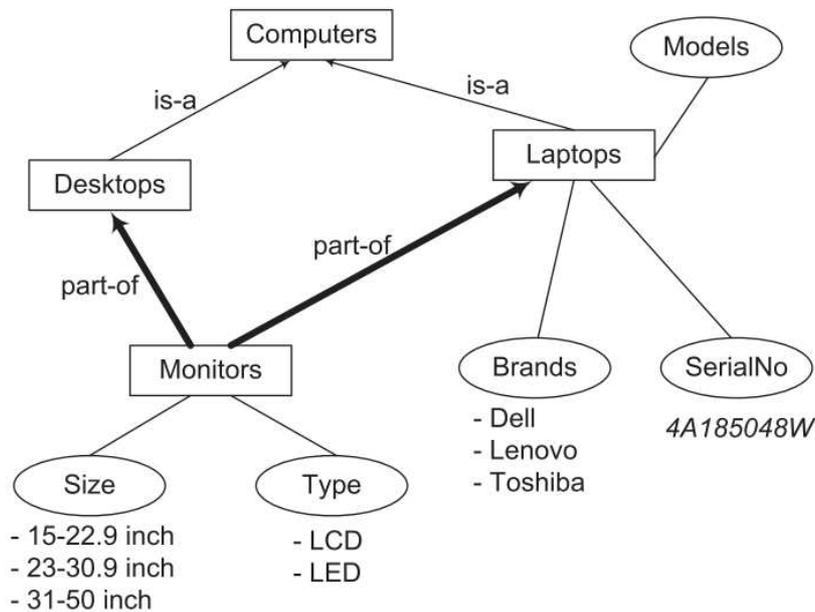


Figure 2.4 – An example of ontology “Computers”

ceptualization and explicit specification. He further discussed the importance of shared explicit specification [25]. In 2009, Grubber proposed a more detail definition in the context of computer and information sciences. He described an ontology as a “*set of representational primitives which is defined to model a domain of knowledge*” [24]. According to him, the representational primitives are typically classes, attributes of classes, and relationships between classes. They include information about their meaning and constraints on their logically consistent application.

Figure 2.4 shows an illustration of the Computers ontology.

As can be seen from the figure, main components of an ontology include:

- **Classes (also called Concepts):** a class describes a set of similar terms of the domain or task. A class can have some subclasses and can also have more than one superclass. For instance, *Computers* could be described as a class including subclasses such as *Desktops* and *Laptops*. In Figure 2.4,

classes are depicted by rectangles.

- Attributes (also called Datatype Properties): describe features of instances of the class, so a class is a set of instances with similar properties. For example, each of *Monitors* has attributes *Size* and *Type*. In Figure 2.4, attributes are displayed by ovals.
- Relations (also called Object Properties): are used to represent connections between elements. In Figure 2.4, they are shown as arrows which point from the subclasses (*Monitors*) to their related superclasses (*Desktops*, *Laptops*).
- Instances (also called Individuals): are occurrences of a specific element. In Figure 2.4, they are displayed in italic strings under their corresponding classes. For example, *4A185048W* is the value of the attribute *SerialNo* and is an instance of the class *Laptops*.
- Taxonomy: is a hierarchy of concepts in which a superclass and a subclass are related to each other by a “is-a” relation.

Another popular example of ontology is about wine knowledge ¹. This ontology has classes for groups of wines (Bordeaux, Merlot, and so on). The Bordeaux class has sub classes like Sauternes, StEmilion. Specific wines such as ChateaudYquemSauterne, ChateauChevalBlancStEmilion are instances of the sub classes, respectively. Each wine has some descriptions like body, color, flavor which are properties of wine classes. The relation between Bordeaux class and its sub classes is the “is-a” relation which creates a taxonomy.

From taxonomy to ontology, the semantic meaning of terms is strengthened. Those Knowledge Organization Systems are useful for people when they want to clearly describe terminology. It is necessary to represent those systems

¹<https://www.w3.org/TR/2003/PR-owl-guide-20031215/wine>

by computer languages in order to make computer understand and to enable information exchange between computer applications. The next section will introduce several models to encode Knowledge Organization Systems.

2.3 Encoding Knowledge Organization Systems

Knowledge Organization Systems can be encoded by traditional technologies such as XML, JSON, even HTML. However, those technologies are often limited for local purposes. Nowadays, in the increasingly demand in globally exchanging and sharing data, Knowledge Organization Systems are often represented by Linked Data technology in the framework of Semantic Web [91].

The Semantic Web [85] is an extension of the current World Wide Web. It aims at organizing the available information and resources in a structure so that computers can easily access and understand. The initial version of the Web was designed in a way that humans could easily understand the content of the pages and navigate among them. Computers, on the other hand, have very limited capability of processing and understanding of unstructured information. The Semantic Web provides a common framework for defining and sharing the explicit semantics of the presented information in a machine readable form.

As can be seen from the Figure 2.5, the Semantic Web is basically built on languages specifically designed for data and semantics such as Extensible Markup Language (XML), Resource Description Framework (RDF), Web Ontology Language (OWL). It use URIs to identify things on the Internet. Unicode is used in layers of the Semantic Web in order to support all natural languages over the world. RDF helps data to be easier linked and shared. OWL enables computers to perform complicated tasks by utilizing domain-defined information with the support of a formal reasoning model.

The following presents in detail some key technologies in Semantic Web

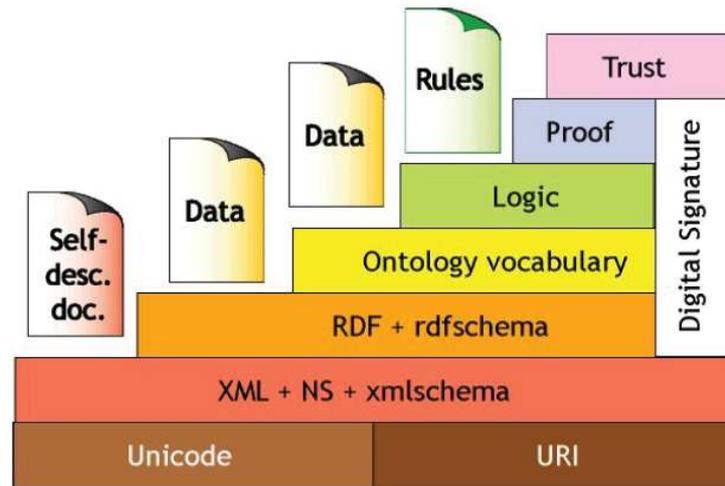


Figure 2.5 – Architecture of the Semantic Web

including RDF, RDFS, OWL.

2.3.1 Resource Description Framework (RDF)

Resource Description Framework (RDF) [84] is a standard model for sharing and exchanging data on the Web. It represents resources (in particularly web resources) as statements in expressions of the form subject-predicate-object which is known as “triples”. A collection of RDF statements forms a directed, labeled graph namely RDF graph. The subject and object are two resources represented by the graph nodes. The predicate is represented by the edge linking the subject and object. URIs are used to name the subject, predicate and object of an RDF statement. In some cases, the subject and object may be a blank node (anonymous resource). The object may be also a Unicode string literal. Using this model, RDF allows structured and semi-structured data to be mixed, exposed, and shared across different applications.

There are several serialization formats to represent RDF data. The following introduces some popular serializations.

- **RDF/XML**: an XML-based syntax defined by the W3C. It serializes an RDF graph as an XML document. It can be read and written by existing tools for XML. RDF/XML is difficult to read and write by human. Despite it is still used today, many RDF users prefer other RDF serializations because of their human-friendly features.
- **Turtle**: a compact and human-friendly format. Its syntax is similar to the syntax used to express RDF queries in the SPARQL. This format is very common in the Semantic Web community.
- **N-Triples**: a very simple and line-based RDF serialization. This format is easy to parse because one triple is represented in one line.
- **JSON-LD** (JavaScript Object Notation for Linked Data): a serialization using JSON to encode Linked Data. It is designed for developers who are familiar with traditional JSON. Its primary goal is to enable developers transform their existing JSON to JSON-LD without much effort.

2.3.2 Resource Description Framework Schema (RDFS)

Resource Description Framework Schema (RDFS) or RDF Schema [83] is a schema language providing a data-modeling vocabulary to add semantics to RDF model. It is used to describe taxonomies of RDF classes and properties. RDFS is written in RDF, but it also extends definitions for some RDF elements (for example, *rdfs:range* and *rdfs:domain* are used to set the range and domain of resources). RDFS provides some classes to describe groups of related resources and the relationships between them. It also allows to describe the meaning of a relationship or a class in human-readable text (for example, *rdfs:label* allows to define human-readable names of resources). RDFS is also used to indicate that a class is a subclass of a more general class (for instance,

rdfs:subClassOf is used to show that “Music Therapy” is a subclass of “Art Therapy”).

2.3.3 Web Ontology Language (OWL)

In computer science and artificial intelligence, ontology languages are formal languages used to describe ontologies. They allow the encoding of knowledge about specific domains and often include reasoning rules that support the processing of that knowledge. Some popular ontology languages are: OWL, CycL, DAML+OIL, F-Logic. Among them, OWL is mostly used by people working on ontologies.

Web Ontology Language (OWL) [87] is a formal language designed by the W3C to represent ontologies in the framework of Semantic Web. As can be seen in the Figure 2.5, OWL is built on RDF and RDFS (RDF Schema). It adds, among other vocabulary of RDF and RDFS, more vocabulary for describing resources on the Internet. For example, vocabulary to express the disjointness (*owl:disjointWith*), equality (*owl:equivalentClass*, *owl:equivalentProperty*, *owl:sameAs*), cardinality (*owl:cardinality*), restriction (*owl:allValuesFrom*, *owl:someValuesFrom*).

The primary purpose of an ontology is to specify and represent things with formal and implicit meanings. This can be realized by using concepts and instances which are called classes and individuals in OWL respectively [37].

- A class in OWL is a group of individuals which share some properties. The built-in class named Thing is the class of all individuals and is the superclass of other OWL classes.
- Individuals are members or instances of OWL classes. Thus individuals inherit semantics and properties defined in the classes to which they belong.

```
<owl:Class rdf:ID="Wine">
  <rdfs:subClassOf rdf:resource="food:PotableLiquid" />
  <rdfs:label xml:lang="en">wine</rdfs:label>
  <rdfs:label xml:lang="fr">vin</rdfs:label>
</owl:Class>
```

Figure 2.6 – RDF/XML syntax

Classes and individuals in OWL can be related together by using properties. There are two types of property in OWL:

- Object properties (*owl:ObjectProperty*) link individuals (instances) of a OWL class to individuals of another OWL class.
- Datatype properties (*owl:DatatypeProperty*) link individuals (instances) of OWL classes to literal (data) values.

OWL adds more semantics to RDF and RDFS. It makes properties and classes of OWL more implicit. OWL can express two instances, in different data sources, are the same. For example, the instance “Barack Obama” on Wikipedia is the same with the instance “Barack Hussein Obama” on BBC News although the labels are different. This is essential to create Linked Data where data represented in different schemas can be joined together.

OWL is based on description logic, so it supports reasoning to infer new information based on available resources. For example, if A is subclass of B, and B is subclass of C are two statements defined in OWL format, then OWL can infer that A is also subclass of C.

As OWL is built on RDF, it is also represented in triples. It supports a variety of syntaxes such as RDF/XML, RDF/Turtle, OWL2 XML, OWL2 Functional, Manchester and others. Figures 2.6 and 2.7 are examples of the wine ontology based on a wine class encoded by OWL with different syntaxes.

```
AnnotationProperty: rdfs:label

vin:Wine rdf:type owl:Class ;
         rdfs:subClassOf food:PotableLiquid ,
         rdfs:label "vin"@fr ,
                   "wine"@en .
```

Figure 2.7 – RDF/Turtle syntax

2.3.4 Simple Knowledge Organization System (SKOS)

The Simple Knowledge Organization System (SKOS) [86] is a Semantic Web vocabulary created by W3C for expressing and publishing semi-formal knowledge organization systems such as thesauri, taxonomies, classification schemes, subject heading lists or any other type of controlled vocabulary (ontologies are the most formal knowledge organization systems) [88]. SKOS aims at defining hierarchical structures which have weak and ambiguous semantics. In other words, it is used for organizing knowledge, while OWL focuses on knowledge representation. SKOS can be used alone or in combination with OWL. The main units of SKOS are concepts, known as SKOS Concepts, with labels and definitions. Those concepts are organized into a hierarchy with broader and narrower relations. Since SKOS is built upon RDF, thus it is also represented in triples and can be serialized in RDF syntaxes [32].

Aforementioned models to encode Knowledge Organization Systems are efficient for computer understanding and interpretation. Whereas, they are complex for people to understand when they look into their syntaxes. Visual representations of Knowledge Organization Systems are thus necessary. The next section will introduce methods to visualize those Knowledge Organization Systems.

2.4 Visualizing Knowledge Organization Systems

2.4.1 MindMap

A MindMap is a diagram used to organize and visualize information in hierarchical structure. It uses images, words in order to represent ideas around a central concept. Those ideas are connected directly to the central concept by edges, and the same principle is repeated for their child ideas. An example of MindMap for the NPI taxonomy is represented by FreeMind tool in Figure 2.8.

FreeMind document is an XML-based file format which is created by FreeMind mind-mapping application to store visual diagrams. It is popular and supported by most existing mind-mapping applications.

The list of elements and their attributes of a FreeMind document as follows:

- **map** (root element): an XML element, with a single required attribute “version”.
- **node** (parent element: node, map): main content element with attributes: id, text, link, folded, color, position (left or right, only for children of the root).
- **edge** (parent element: node): an element to connect nodes. This element has attributes: style, color, width.
- **font** (parent element: node): contains information about the format of node text including: name, size, bold, italic.
- **arrowlink** (parent element: node): an element to show relationship between nodes other than edges. It has attributes: color, destination (id

of the target node), `startarrow` (arrow style), `endarrow` (arrow style).

Figure 2.9 is an example of a FreeMind document associated with the mind map for the NPI taxonomy mentioned previously in Figure 2.8.

2.4.2 Concept Map

A concept map is a diagram that organizes and represents concepts and relationships among them. The concepts are usually enclosed in circles or boxes. The relationships between two concepts are indicated by a connecting line with labels which are referred to as linking words or linking phrases.

Figure 2.10 shows an example ² of concept map.

Discussion

As can be seen from Figure 2.8 and Figure 2.10, a topic in Concept Map may have multiple children and parents. This is useful in modeling complex relationships between topics. However, it is quite confused compared with MindMap in representing simple and concentrated topics. That is the main reason why in the Plateforme CEPS we choose MindMap to visualize the NPI ontology which is a hierarchical and focused knowledge domain.

In this chapter, we presented problems related to terminologies, from unstructured knowledge (word, term, controlled vocabulary) to structured knowledge (taxonomy, thesaurus, ontology). We also introduced some methods to encode and visualize the Knowledge Organization Systems. As we mentioned previously in Chapter 1, the objective of the Plateforme CEPS is to develop an ontology for NPIs. One of the first important steps is NPI term acquisition. The next part presents the challenges and our solution in detail.

²<https://www.lucidchart.com/pages/examples/concept-map>

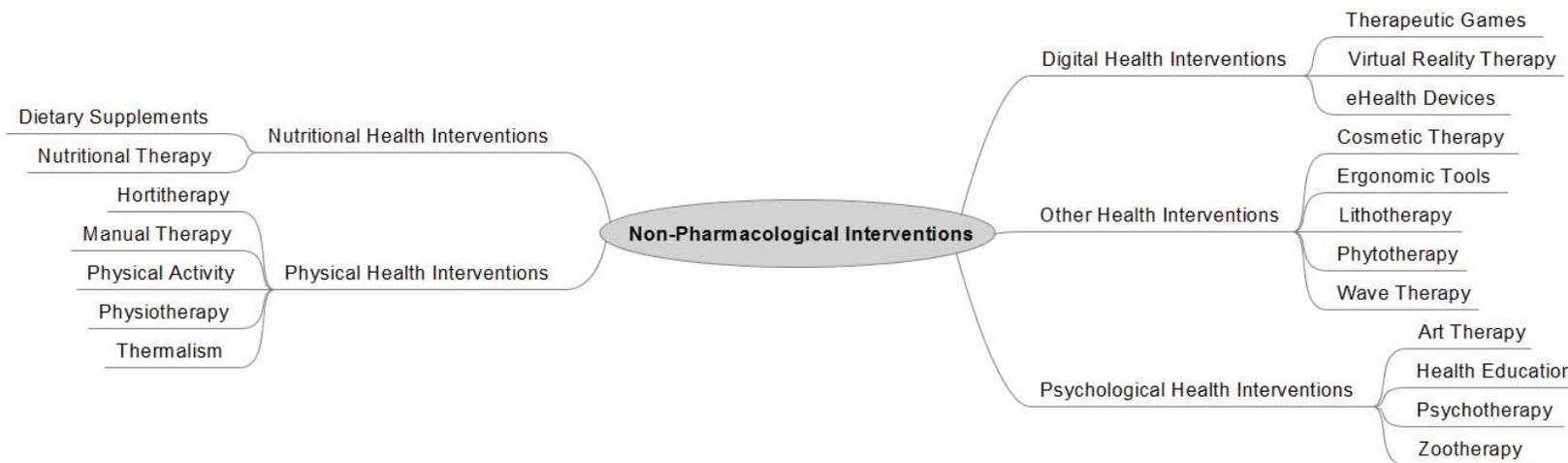


Figure 2.8 – A mind map of the NPI taxonomy created by FreeMind

```

<map version="1.0.1">
<!-- To view this file, download free mind mapping software FreeMind from http://freemind.sourceforge.net -->
<node CREATED="1522327824872" ID="ID_142909177" MODIFIED="1522328645452" TEXT="Non-Pharmacological Interventions">
<font BOLD="true" NAME="SansSerif" SIZE="12"/>
<node CREATED="1522328062391" ID="ID_305698972" MODIFIED="1522328606475" POSITION="right" TEXT="Digital Health Interventions">
<font NAME="SansSerif" SIZE="12"/>
<node CREATED="1522328122198" ID="ID_485128284" MODIFIED="1522328124155" TEXT="Therapeutic Games"/>
<node CREATED="1522328143903" ID="ID_1704785224" MODIFIED="1522328145763" TEXT="Virtual Reality Therapy"/>
<node CREATED="1522328154760" ID="ID_920936542" MODIFIED="1522328156077" TEXT="eHealth Devices"/>
</node>
<node CREATED="1522328067128" ID="ID_412189843" MODIFIED="1522328606475" POSITION="left" TEXT="Nutritional Health Interventions">
<font NAME="SansSerif" SIZE="12"/>
<node CREATED="1522328167581" ID="ID_1613867068" MODIFIED="1522328168492" TEXT="Dietary Supplements"/>
<node CREATED="1522328191320" ID="ID_1608765698" MODIFIED="1522328192711" TEXT="Nutritional Therapy"/>
</node>
<node CREATED="1522328080176" ID="ID_794656949" MODIFIED="1522328606475" POSITION="right" TEXT="Other Health Interventions">
<font NAME="SansSerif" SIZE="12"/>
<node CREATED="1522328298319" ID="ID_1297365388" MODIFIED="1522328299307" TEXT="Cosmetic Therapy"/>
<node CREATED="1522328306780" ID="ID_852601913" MODIFIED="1522328307690" TEXT="Ergonomic Tools"/>
<node CREATED="1522328316855" MODIFIED="1522328316855" TEXT="Lithotherapy"/>
<node CREATED="1522328322086" ID="ID_966017167" MODIFIED="1522328323558" TEXT="Phytotherapy"/>
<node CREATED="1522328328703" ID="ID_1235570833" MODIFIED="1522328330662" TEXT="Wave Therapy"/>
</node>
<node CREATED="1522328090207" ID="ID_1728256888" MODIFIED="1522328606475" POSITION="left" TEXT="Physical Health Interventions">
<font NAME="SansSerif" SIZE="12"/>
<node CREATED="1522328335179" ID="ID_20766561" MODIFIED="1522328336238" TEXT="Hortitherapy"/>
<node CREATED="1522328342372" ID="ID_685154868" MODIFIED="1522328343828" TEXT="Manual Therapy"/>
<node CREATED="1522328350551" ID="ID_248235503" MODIFIED="1522328351911" TEXT="Physical Activity"/>
<node CREATED="1522328356039" ID="ID_1459316352" MODIFIED="1522328356875" TEXT="Physiotherapy"/>
<node CREATED="1522328360610" ID="ID_68770268" MODIFIED="1522328361820" TEXT="Thermalism"/>
</node>
<node CREATED="1522328097247" ID="ID_1086804464" MODIFIED="1522328606475" POSITION="right" TEXT="Psychological Health Interventions">
<font NAME="SansSerif" SIZE="12"/>
<node CREATED="1522328369478" ID="ID_523099783" MODIFIED="1522328370385" TEXT="Art Therapy"/>
<node CREATED="1522328374398" ID="ID_411344516" MODIFIED="1522328375401" TEXT="Health Education"/>
<node CREATED="1522328379198" ID="ID_1783293289" MODIFIED="1522328380415" TEXT="Psychotherapy"/>
<node CREATED="1522328388509" ID="ID_1455680129" MODIFIED="1522328389870" TEXT="Zootherapy"/>
</node>
</node>
</map>

```

Figure 2.9 – The corresponding FreeMind document of the mind map in Figure 2.8

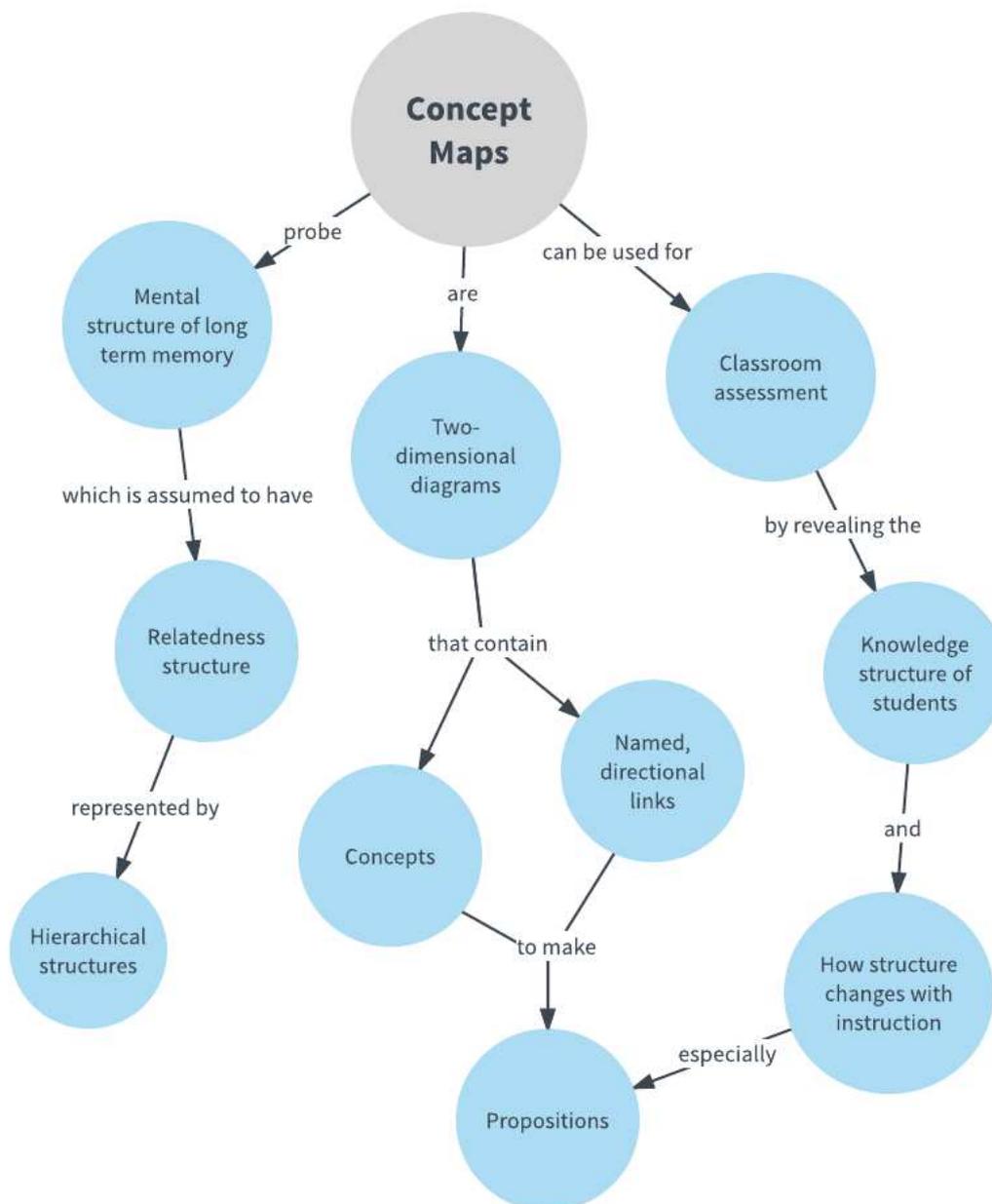


Figure 2.10 – An example of concept map

Part II

NPI Terminology Acquisition

3	Problem Statement	37
4	Related Works	39
4.1	Biomedical resources	39
4.2	Automatic term extraction	41
4.3	Semantic similarity measures	42
4.3.1	Knowledge-based similarity measures	42
4.3.2	Corpus-based similarity measures	43
4.3.3	Hybrid similarity measures	44

5	Methodology	47
5.1	Term extraction	47
5.1.1	Exploiting knowledge bases	48
5.1.2	Exploiting text corpus	51
5.2	Term ranking	53
5.2.1	Edit-based similarity measures	56
5.2.2	Knowledge-based similarity measures	57
5.2.3	Distributional similarity measures	58
5.3	Term validation	68
6	Results and Evaluation	69
6.1	Results	69
6.2	Evaluation	73
7	Conclusion	79

Problem Statement

TERMINOLOGY acquisition is a prerequisite and important step in an ontology development process. Given some seed terms of NPI, the mission of this thesis is to find out as many as possible new NPI terms related to the seed terms. In other words, from the initial set $S = \{s_1, s_2, \dots, s_n\}$ where $s_i (i = 1, 2, \dots, n)$ is a NPI term, we need to build the result set $R = \{r_1, r_2, \dots, r_m\}$ where $r_j (j = 1, 2, \dots, m)$ is also a NPI term.

According to our knowledge, there is no methodology in literature that can work completely automatically in extracting, expanding or enriching a terminology. Normally, it is necessary to have the intervention from domain experts in manual validation. Therefore, we propose a process illustrated in Figure 3.1 including three phases to collect NPI terms as follows:

- Term extraction: a phase to automatically retrieve candidate NPI terms which are synonyms or related terms of given seed NPI terms. The result of this process is a set of candidate NPI terms $C = \{c_1, c_2, \dots, c_t\}$ which is the input for the next step.
- Term ranking: this phase removes bad candidates and ranks the remaining candidate terms by NPI score. The result of this phase is a list of potential NPI terms $P = \{p_1, p_2, \dots, p_h\}$ which is evaluated manually by NPI experts in the next phase.

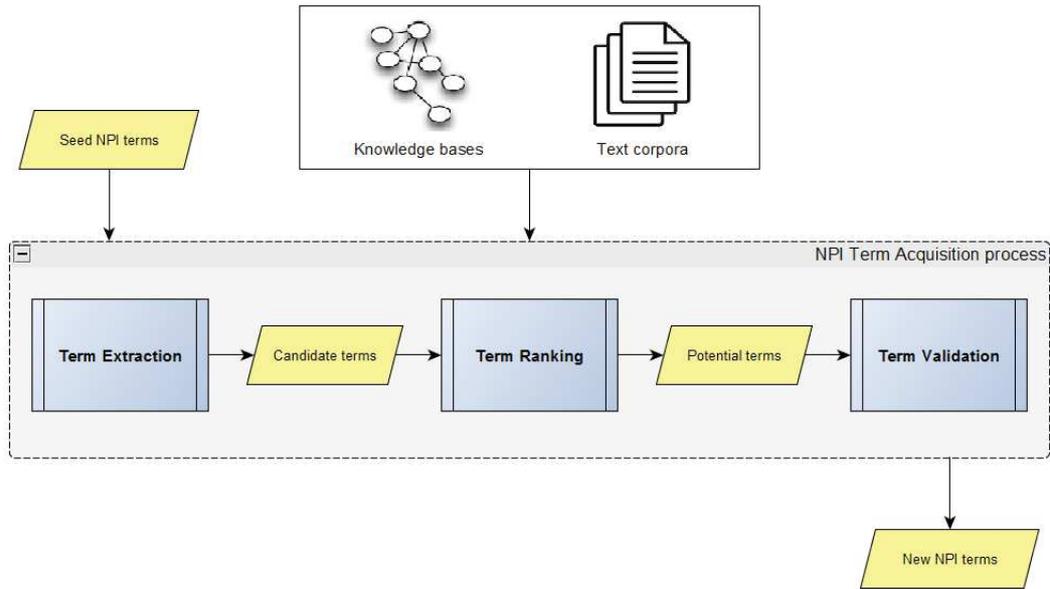


Figure 3.1 – NPI term acquisition process

- Term validation: a manual work of NPI experts to get evaluated NPI terms by choosing the correct NPI terms from the list of potential NPI terms P .

As the third phase is the manual work of NPI experts, it is considered as out of scope of this dissertation. Therefore, the following content of this thesis focuses on the first and second phases. Some methods are brought out to analyse and discuss in the next parts. From the limitations of current solutions in literature, we propose new methods to solve the problems.

Related Works

THIS chapter summarizes related work on semantic similarity measures as well as automatic term extraction. A comparison between the works is also presented in this part. Some useful biomedical resources are also introduced.

For the challenges on NPI terminology acquisition, we consider that the core problem is how to identify which terms are similar to the seed terms. In order to realize that, we need to have a semantic similarity measure for the context of NPI. This motivates us to study related works on semantic similarity literature.

4.1 Biomedical resources

MEDLINE [81] is a journal citation database provided by the National Library of Medicine (NLM). It contains more than 24 million citations from over 5,600 journals published around the world. MEDLINE is a rich source of biomedical text that lends itself well to research on text mining, information extraction, and natural language processing in biomedical domains. PubMed [80] is a large biomedical bibliographic database that is well known to users around the globe. It has more than 28 million references ¹ include the MEDLINE

¹The latest statistics from <https://www.ncbi.nlm.nih.gov/pubmed>

database plus other types of citations. PubMed allows individuals to conduct searches directly by entering search terms on web pages and viewing results, and supports software-based queries across the Internet with programming utilities offered by the NLM. PubMed Central (PMC) [79] is an electronic archive of full-text journal articles, offering free access to its contents. PMC contains more than 4 million articles, most of which have a corresponding entry in PubMed. PubMed does not have citations for certain types of PMC material, such as book reviews, that are considered out of scope for PubMed. These items constitute a small portion of the total PMC collection and there are no current plans to include them in PubMed.

BioPortal [71] is the biggest repository of biomedical ontologies on the Internet. It has more than 700 ontologies up to the time of writing. The ontologies in BioPortal are encoded in OWL, OBO and other formats. A large number of medical terminologies distributed by the U.S. National Library of Medicine are represented in its own proprietary format.

The Unified Medical Language System (UMLS) [9, 36], created in 1986 and maintained by the U.S. National Library of Medicine, is a compendium of many controlled vocabularies in the biomedical sciences. It provides a mapping structure among these vocabularies and thus allows one to translate among the various terminology systems. UMLS may also be viewed as a comprehensive thesaurus and ontology of biomedical concepts. It is intended to be used mainly by developers of systems in medical informatics. The UMLS integrates over 2 million names for some 900 000 concepts from more than 60 families of biomedical vocabularies, as well as 12 million relations among these concepts. Vocabularies integrated in the UMLS Metathesaurus include the NCBI taxonomy, Gene Ontology, the Medical Subject Headings (MeSH), OMIM and the Digital Anatomist Symbolic Knowledge Base [9].

4.2 Automatic term extraction

Automatic Term Extraction (ATE), or Automatic Term Recognition (ATR), is a subtask of information extraction which aims at identifying or extracting automatically technical terminology from unstructured texts. It is a prerequisite step in developing knowledge systems.

The authors in [76] described FlexiTerm which is a method for automatic term extraction from a domain-specific corpus. They proposed two steps for the term recognition process including: linguistic filtering and termhood calculation. In the first step, linguistic filtering is used to select term candidates. Then a termhood calculation, based on a frequency-based measure as evidence, is performed to identify a candidate as a term. In order to improve the quality of termhood calculation, the authors used the bag-of-words approach to manage syntactic variation. The method was evaluated on five biomedical corpora and achieved good results with precision (94.56%), recall (71.31%) and F-measure (81.31%).

The authors in [92] developed a toolkit named Java Automatic Term Extraction (JATE). It is free available as a modular, adaptable and scalable library. The power of this tool comes from 10 algorithms implemented within the Apache Solr framework. It advances existing ATE tools mainly by enabling a significant degree of customization and adaptation thanks to the flexibility under the Solr framework.

The authors in [46] presented a methodology for automatic term extraction in biomedical domain. This methodology offered several measures based on linguistic, statistical, graphic and web aspects [43, 44]. The authors modified some baseline measures (i.e. C-value, TF-IDF, Okapi) and proposed the new hybrid measures F-TFIDF-C and F-OCapi, which combine C-value with TF-IDF and Okapi respectively to obtain better results [82]. Their proposed

measures improve precision of multi-word term extraction in biomedical domain. The results showed that their method outperforms state-of-the-art methods [42]. Finally, they combined their contributions to develop a tool named BioTex [45] to automatically extract biomedical terms from documents. It is available as a web application and as a Java library for using in programming. This system allows users to manually validate extracted terms and export the list of the terms.

4.3 Semantic similarity measures

4.3.1 Knowledge-based similarity measures

Knowledge-based measures rely on formal expressions of knowledge explicitly defining how the compared entities, i.e. concepts or instances, must be understood. They often take advantage of pre-defined dictionaries or ontologies as semantic graphs or semantic networks.

WordNet [53] is a large lexical database of English. It is the most important resource available to researchers in computational linguistics, text analysis, and many related areas. Its design is inspired by current psycholinguistic and computational theories of human lexical memory. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets), each representing one underlying lexicalized concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

There are some algorithms for computing the semantic similarity between two given words or concepts using the knowledge-base WordNet taxonomy. The algorithms are classified in two different groups: Path-based measure and Information content-based measure.

Path-based Measures In the path-based measures, the similarity between two concepts is a function of the length of the path linking the concepts and the position of the concepts in the taxonomy.

Information content-based measure This type of measure assumes that each concept includes much information in WordNet. The IC based measures compute the similarity using the information content of each concept. The more common information two concepts share, the more similar the concepts are.

In the biomedical domain, semantic similarity measures have been evaluated on the Systematized Nomenclature of Medicine-Clinical Terminology (SNOMED CT); Medical Subject Headings (MeSH); and the Unified Medical Language System (UMLS), a collection of biomedical source vocabularies that includes SNOMED CT and MeSH [64].

4.3.2 Corpus-based similarity measures

Corpus-based measures are generally used to compare words, sentences or texts based on NLP techniques. They most often only rely on statistical analysis of word usage in texts, e.g. based on the analysis of word (co-)occurrences and the linguistic contexts in which they occur.

In recent years, approaches based on distributional hypothesis has achieved much success. The general principle is representing words as real number vectors in Vector Space Model which is known as distributed representation for words. This idea was first proposed in [70] and has become a successful paradigm, especially for statistical language modeling [69]. The detailed idea of the approaches is using a sliding window with a specific size through a text corpus and get the probability of words appearing together with the center

words of the window. The words surrounding the center words are called context words. For example, if the window size is 5, then there are 5 context words behind and 5 context words ahead the center words. Two well-known methods can be used in this step to get the probability of the context words: count-based method by using LSA [26], Glove [66] techniques and learn-based method by using word2vec algorithm [52]. After words are represented as vectors, the similarities between them are calculated by measuring the angles between their corresponding vectors. The well-known Cosine measure is often used for this purpose [75].

The world-wide-web is the largest database on earth, and the context information entered by millions of independent users averages out to provide automatic semantics of useful quality. In [11] the authors presented a new theory of similarity between words and phrases based on information distance and Kolmogorov complexity. They used the world-wide-web as database, and Google as search engine. Then they constructed a method to automatically measure the similarity between words and phrases by using Google page counts. They used WordNet as an objective baseline to compare the performance of their method. The method was turned out to agree well with the WordNet semantic concordance made by human experts. The mean of the accuracies of agreements is 0.8725. Besides, the authors also showed that their method could be a good translation tool with the accuracy even reaching to 100% on their test set between English and Spanish. The method is also applicable to other search engines and databases.

4.3.3 Hybrid similarity measures

Statistical word similarity measures have limitations that related words can have similarity scores only as high as their context overlap. Also, word similarity is typically low for synonyms having many word senses since information

about different senses are mixed together. Knowledge-based measures, otherwise, have high accuracy as the knowledge domains are manually constructed by human. However, their vocabularies are often limited. As a result, many words cannot be found in the dictionary. A solution to reduce the above issues is using additional information from WordNet. Hybrid measures are combinations of knowledge-based and corpus-based approaches to take advantage of both methods.

The authors in [26] developed a metric which measures the semantics of a word without considering its lexical category. For example, the noun “husband” is semantically similar to the verb “marry” in the UMBC Phrase Similarity Service. The metric gives highest scores to similar words and lowest scores to dissimilar words. In order to realize the idea, they constructed a model by combining LSA word similarity and WordNet knowledge. However, their word similarity model is restricted by the size of its predefined vocabulary. That means for some words being out of the vocabulary, it is impossible to calculate the similarity among these words. In order to solve this problem, the authors presented two approaches relied on lexical features and semantic features [33].

Discussion

According to our observations and experiments, the knowledge-based similarity measures often have high accuracy. Because knowledge bases are manually created and validated by experts. They are also organized in good structures. However, the knowledge bases cannot cover all domains. WordNet is a general dictionary, so it does not include many terms in biomedical domain. Even UMLS is a big biomedical resources, it does not include all NPI terms. As a result, when two terms are not found in their vocabularies, it is impossible to calculate the similarity between them.

The similarity measures based on corpus have good coverage as text corpus include tremendous documents. Google News and Freebase are two of the biggest text corpora on the Internet. Unfortunately, a corpus-based similarity is dependent on quality of a corresponding text corpus. Therefore, existing similarity measures such as UMBC and Google News word2vec do not return good accuracy for computing the similarity between NPI terms. This is our motivation to propose a semantic similarity measure for NPI domain. Our methodology is detailed in the next chapter.

Methodology

THIS chapter presents our methodology to automatically collect NPI terms for the NPI ontology development. Two approaches for term extraction are introduced with limitations of each approach. Besides, some similarity measures are discussed in this part along with a comparison between them.

The main idea of our methodology is an expansion of the existing seed NPI terms. Starting from the seed terms taking from the NPI taxonomy (Figure 1.1), we find new terms that are semantically related to them. This process is then repeated for the retrieved new NPI terms until meeting requirements from NPI experts. The challenge here is how to know if a new term is semantically similar to the seed terms in the NPI context. Following parts will present approaches to address this challenge.

5.1 Term extraction

In this section, we introduce two approaches to extract terms from two kind of sources: knowledge bases and text corpora. Those terms are considered as candidate NPI terms.

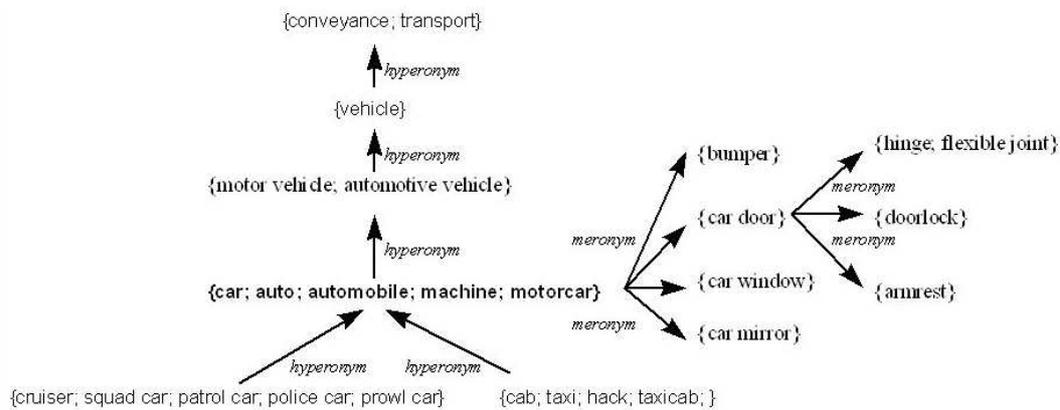


Figure 5.1 – Structure of WordNet

5.1.1 Exploiting knowledge bases

This approach exploits existing external knowledge resources such as dictionaries, lexicons or thesauri. We mainly rely on the semantic relationships in the hierarchy to explore the new terms. In the scope of this thesis, we focus on exploiting terminologies of WordNet and existing ontologies from BioPortal.

5.1.1.1 Extracting terms from WordNet

In this method, we search WordNet item that match with the seed term. After that, we get every synsets of that matched term. Since NPI terms are nouns, we keep noun synsets only and remove the rest. From the noun synsets, we get synonyms, hyponyms and hypernyms of the associated term.

For example, starting with the seed term “Physiotherapy”, we get a noun synset. From this synset, we can extract the synonyms including “physical therapy”, “physiatrics” and the direct hyponym “rehabilitation”.

These above works can be done with WordNet API for the traditional version (Figure 5.1) and with SPARQL or Apache Jena for WordNet RDF version (Figure 5.2).

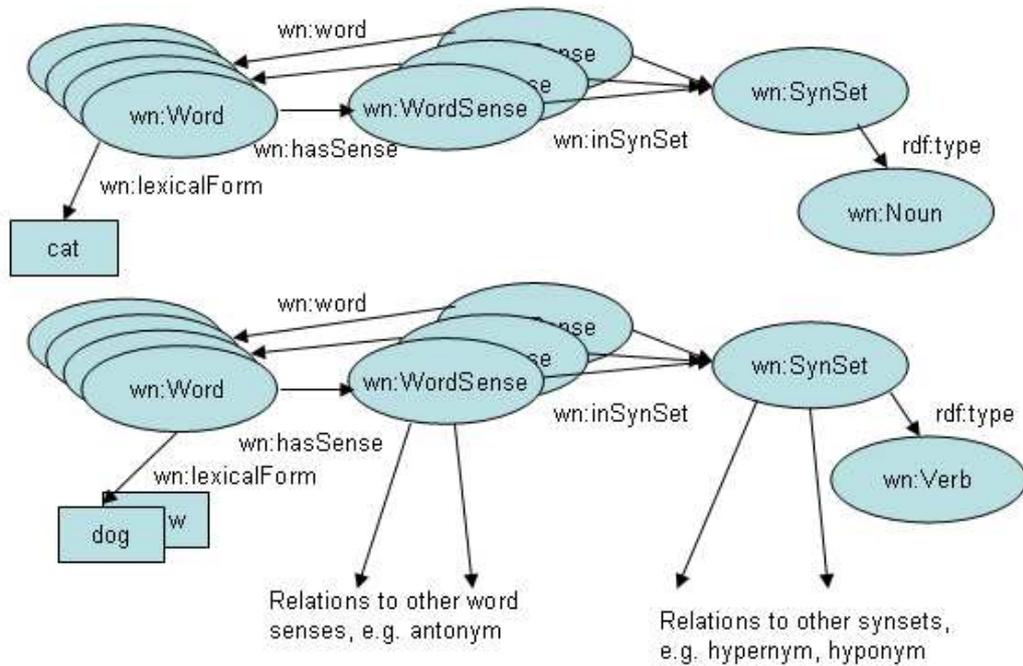


Figure 5.2 – WordNet in RDFS and OWL

5.1.1.2 Extracting terms from ontologies in BioPortal

To the best of our knowledge, BioPortal is the largest biomedical ontology repository. This method directly gets terms which are available in existing ontologies in BioPortal. We present the process with steps as follows.

Step 1 (Get ontology classes related to the initial terms): This step uses “Term Search” endpoint of BioPortal API to search entire terms and properties of BioPortal repository with parameter “q” being the initial term. The retrieved result is a JSON string containing a list of ontology classes which includes information such as ontology id, preferred label, alternative labels (synonyms), parent classes, child classes. The step is repeated with synonyms of the initial term as the “q” parameter.

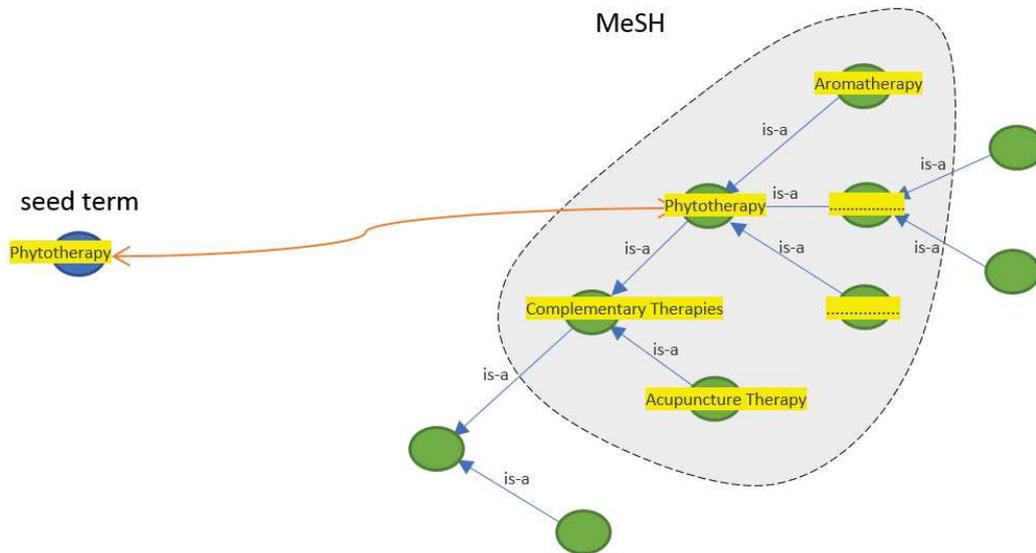


Figure 5.3 – Extracting related terms from MeSH

Step 2 (Extract labels and synonyms of the relevant classes): In this step, the result retrieved in the first step is processed by using JSON parser in order to extract labels and their synonyms if available. The parent classes and child classes of retrieved classes can be exploited by following links in the JSON result. The process of extracting labels and synonyms then is applied again for parent and child classes.

For example, starting from the seed term “Phytotherapy”, we get some classes such as “Phytotherapy” of the Medical Subject Headings (MeSH), “Botanical Therapy” of the National Cancer Institute Thesaurus. From the former, we get the synonyms terms “Herb Therapy” and “Herbal Therapy”. For the latter, we retrieve terms including “Herbal Medicine”, “Botanical Therapy”, “Herbal Therapy”, “Herbalism”. This example is illustrated in Figure 5.3.

In reality, labels of the same classes may be highly different because of various conventions in naming process. There is no standard rule for labeling concepts in an ontology, the same concepts in the same domain of knowledge might be assigned with different labels in diverse ontologies. For example, in

BioPortal, some labels of classes are named as “medicine, alternative”, or “alternative_medicine”, instead of “alternative medicine”. In other ontologies, labels can be named with brackets such as “medicine complementary (alternative)”.

In some ontologies, we found terms that are indicated specific languages. Therefore, the process of querying BioPortal to get synonyms can obtain terms in other languages than English.

Limitations

We did some experiments with the two knowledge-based approaches, and we found two limitations of them:

- Limited coverage: the input words cannot be found in the dictionary. For instance, the terms “Zootherapy”, “Virtual Reality Therapy” do not exist in WordNet vocabulary. While “Hortitherapy”, “Thermalism”, “Lithotherapy” cannot be even found in both WordNet and BioPortal.
- Different context: each ontologies are designed by different users with various point of view. That is why even when an input term is matched in the knowledge base, its relevant terms are not NPI terms.

Hence, we need to use other approaches to avoid the aforementioned limitations. An approach based on text corpora will be discussed in the next part.

5.1.2 Exploiting text corpus

According to NPI experts at the Plateforme CEPS, many NPI terms exist in text documents on the Internet, for example in Wikipedia, PubMed. Among such databases, PubMed is a reference and has more NPI terms compared with

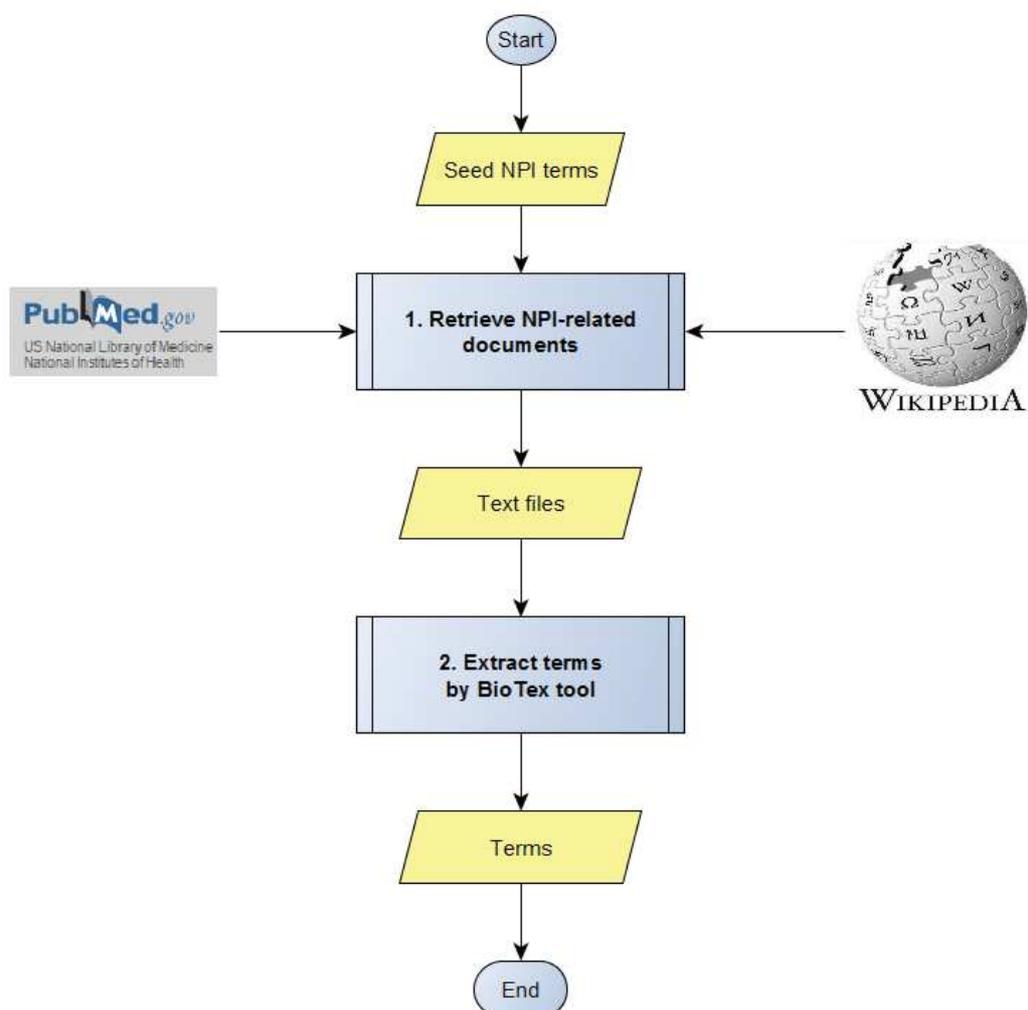


Figure 5.4 – Extracting candidate NPI terms from text corpora

the remaining ones. In each PubMed abstract, many NPI terms are mentioned. This motivates us to propose an automatic method in order to extract the candidate NPI terms. The process, illustrated in Figure 5.4, includes two steps as follows:

Step 1 (Retrieve NPI-related documents): The purpose of this step is to create a text corpus including text documents related to the given NPI terms (the seed NPI terms). We focus on getting results from PubMed and

Wikipedia. Furthermore, we enrich the set of retrieved documents with web pages returned by Google search engine. Finally, the retrieved documents are stored as text files which are input for the next step.

Step 2 (Extract terms by BioTex tool): In this step, we use BioTex tool [42] to extract biomedical terms from the texts retrieved in the previous step. We choose this tool instead of other tools such as FlexiTerm [76], JATE [92], because BioTex is optimized for working with biomedical text. It takes texts as input and return a list of potential biomedical terms which is sorted by score in descending order. The higher score a term has, the greater probability it is a biomedical term.

The resulted list of this approach is supplemented by related terms which are extracted from knowledge bases (BioPortal, WordNet). The combined list then becomes input for the following phase.

5.2 Term ranking

In this phase, we perform two tasks corresponding to two steps as follows:

Step 1 (Pre-filtering): This step aims at removing bad candidate NPI terms in order to reduce computational space for the next step. Firstly, we cluster the seed NPI terms into some clusters. In order to do that, we need to represent terms as vectors in Vector Space Model (the detail is presented in 5.2.3.1). Then a bad candidate NPI term is defined as a point outside the clusters. Our idea is illustrated in Figure 5.5.

Besides, a term is also a bad candidate NPI term if it is existed in the blacklisted term list. The blacklist includes every terms in Disease Ontology,

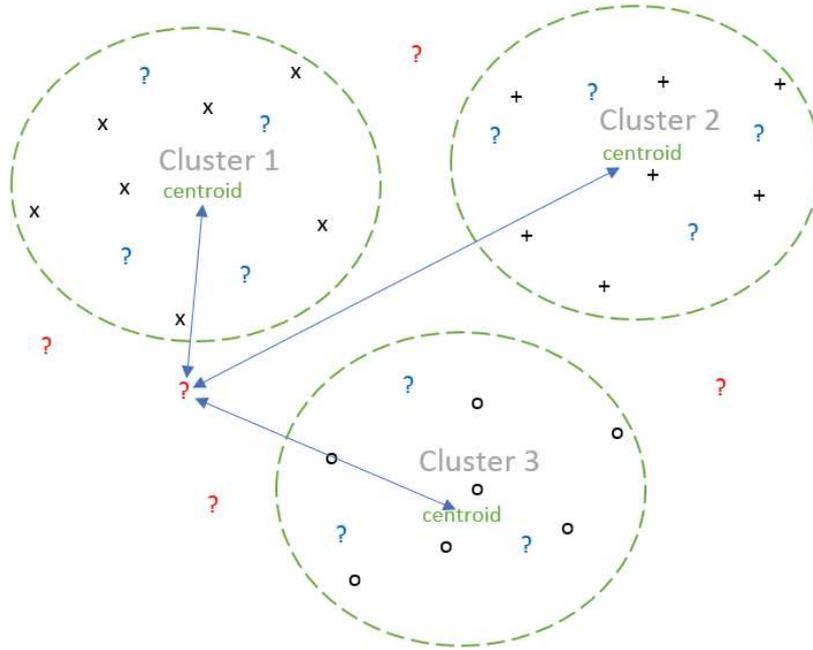


Figure 5.5 – Candidate NPI terms and clusters of seed NPI terms

the terms marked as non-NPI and the existing NPI terms (including the NPI terms found and initial ones).

After this step, we have a shorter list of candidate terms which is the input for the next step.

Step 2 (Ranking): In this step, the remaining terms in the space of the clusters (and not in the blacklist) are calculated NPI score by the equation 5.1 and then sorted descending by the score. The candidate terms with high NPI score are considered as potential NPI terms which are finally evaluated by NPI experts to find out new NPI terms.

$$NPI_score(c) = \max_{1 \leq i \leq k} \left(sim(c, s_i) \right) \quad (5.1)$$

where:

c : a candidate term

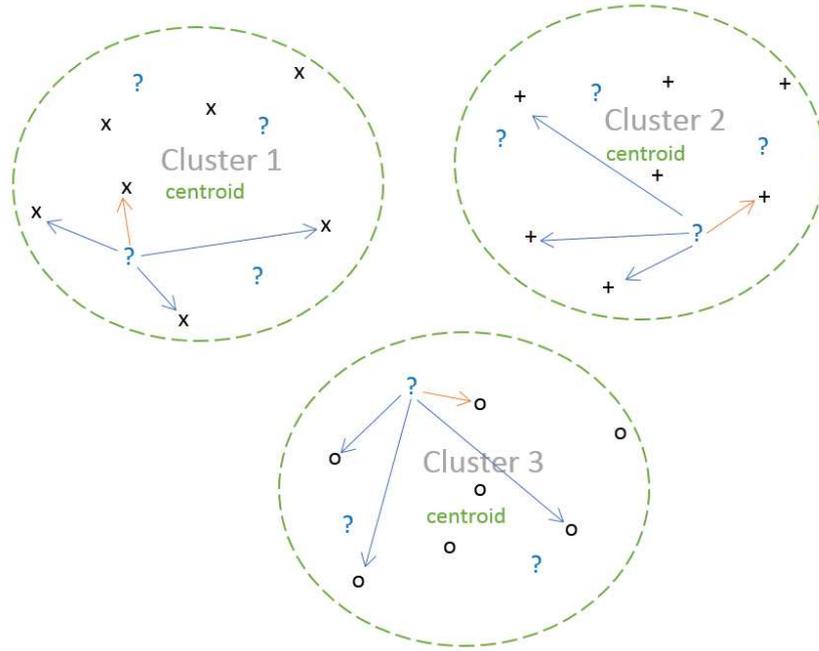


Figure 5.6 – NPI score calculation

s_i : seed term in the same cluster with c

k : number of seed terms in the cluster

The underlying idea of the NPI score calculation (illustrated in Figure 5.6) is as follows: Given a cluster including n NPI terms $S = \{s_1, s_2, \dots, s_n\}$, a candidate term c which is in the same cluster with all terms of S . Firstly, we need to calculate the similarity for each pair (c, s_i) where $s_i \in S$. Then the NPI score is computed as the maximum value of the similarity of every pairs. This idea is applied for other candidate terms with other clusters.

If we have $NPI_score(c_k, s_i) \geq \delta$ (here δ is a similarity threshold which is determined through later experiments) then c_k has the high probability to be a NPI term. A best position is proposed to assign each potential NPI term c_k to the correct node in the NPI taxonomy.

The challenge here is how to measure the similarity between two terms in the context of NPI? or how to calculate $sim(c, s_i)$ where c is a candidate term

and s_i is a seed NPI term.

In the following, we present some existing similarity measures that can be used to deal with the problem. From their limitations, we propose a new similarity measure based on text corpus for NPI domain.

5.2.1 Edit-based similarity measures

In general, Edit-based similarity compares two strings based on individual characters. Depending on the computation method, some following Edit-based similarity measures are frequently used in practice:

- Hamming distance: Number of positions in which two strings (of equal length) differ. In other words, it measures the minimum number of *substitutions* required to change one string into the other, or the minimum number of *errors* that could have transformed one string into the other. For example, $dist_{Hamming}(karolin, kathrin) = 3$.
- Levenshtein distance: Minimum number of character *insertions*, *deletions*, and *substitutions* necessary to transform one string into the other. For example, $dist_{Levenshtein}(kitten, sitting) = 3$. Levenshtein distance between two strings of the same length is strictly less than the Hamming distance. For example, $dist_{Levenshtein}(flaw, lawn) = 2$ (delete “f” from the front; insert “n” at the end), while $dist_{Hamming}(flaw, lawn) = 4$.

The above distance values is commonly transformed to their similarity value by counting the compensation of its normalized value. For example, the Levenshtein similarity is calculated by the formula:

$$sim_{Levenshtein}(s_1, s_2) = 1 - \frac{dist_{Levenshtein}(s_1, s_2)}{\max(|s_1|, |s_2|)} \quad (5.2)$$

As can be seen from the examples above, the Edit-based measures only relies on the sequences of characters of the words and their ordering. The meanings of the words are not taken into account. As a result, these measures cannot calculate semantic similarity of two words. Indeed, according to the computational method of these measures, the similarity score of two words “*dog*” and “*animal*” is zero since they do not share any character. This means that they are not related each other despite the fact that they have the closed semantics. In this case, semantic similarity measures are essential to be used. The semantic measures enable to overcome the limitation of such Edit-based measures by comparing the semantics of the words. In practice, semantic measures rely on existing structured knowledge bases (for example, taxonomies, thesauri, ontologies) and text corpora [27]. Following content of this thesis will present two kinds of semantic similarity measures: knowledge-based and corpus-based.

5.2.2 Knowledge-based similarity measures

The similarity based on knowledge is calculated by leveraging the “is a” relation between terms in existing knowledge bases. In this part, we focus on two well-known resources: WordNet and UMLS. Path-based measures are used to compute the similarity between to concepts in those resources.

The general idea of Path-based approach is measuring the distance between concepts (nodes) in in the hierarchical taxonomy (tree). This distance is computed by counting nodes between the shortest “is a” path that connects the two concepts. The similarity score between concept a and b is thus calculated by the formula:

$$path(a, b) = \frac{1}{shortest_is_a_path(a, b)} \quad (5.3)$$

where $shortest_is_a_path(a, b)$ is the shortest path connecting the two concepts (nodes) a and b by the “is a” relation.

Figure 5.7 shows an example of Path-based measure in biomedical knowledge base UMLS. According to the taxonomy, the similarity score between two terms “*bacterialinfections*” and “*yeastinfections*” is 0.25.

Knowledge-based similarity measures using WordNet and UMLS often give high accuracy because they are manually built by experts. However, many terms do not exist in the vocabularies of WordNet and UMLS. In this situation, we cannot compute the similarity for them. While the similarity measures based on syntactic such as Edit-based similarity (Hamming distance, Levenshtein distance), Token-based similarity (morphological and syntactic analysis) do not return good result for term pairs that are different from syntactic structure. We thus propose a new similarity measure for NPI based on a specific subset of PubMed that related to NPI.

5.2.3 Distributional similarity measures

In general, this method relies on the distributional hypothesis (Harris, 1954): “words that appear in the same contexts share semantic meaning”.

Several approaches have been proposed to define the context of a word. An initial approach considers the context of a word as a document of a text corpus in which the word occurs. According to this approach, each document is known as a bag of words. Then, by counting the occurrences of words in the documents through a text corpus, the documents are represented as vectors in which each dimension corresponds to a separate word. If a word occurs in a document, then its value in the corresponding dimension of the vector is non-zero. There are several ways to compute those values, which are known as term weights, such as TF (term frequency), IDF (inverse document

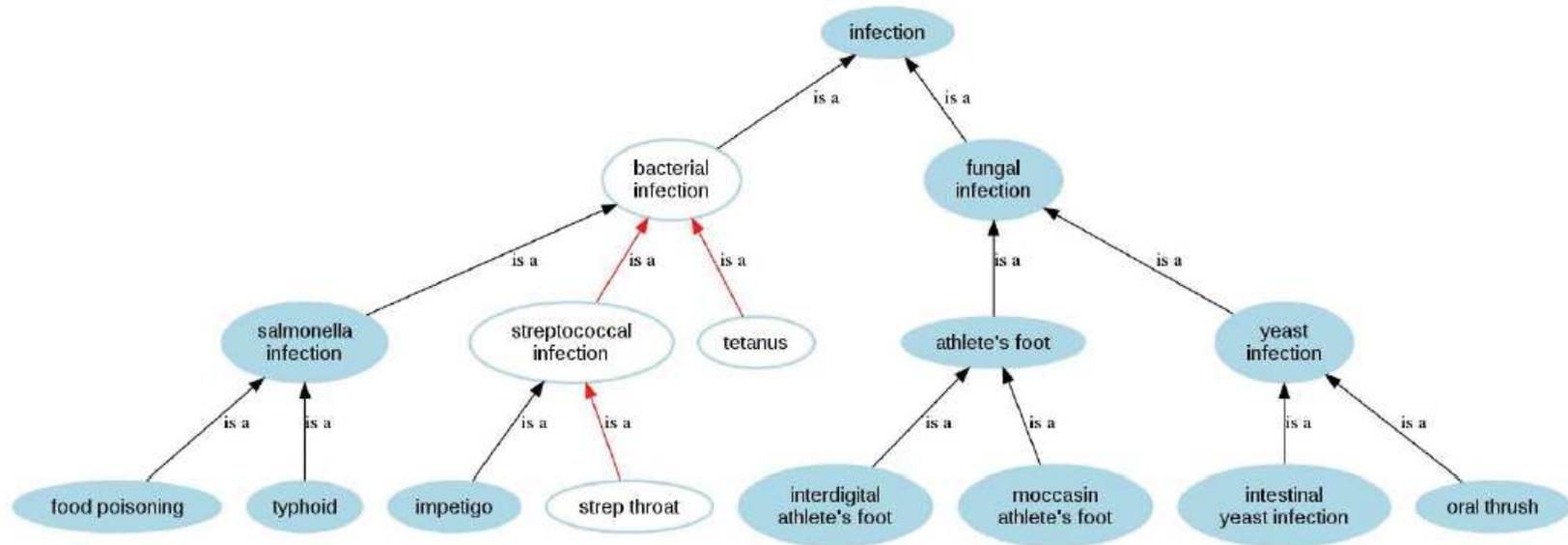


Figure 5.7 – Similarity measure based on the shortest path (taken from [65])

frequency), and TF-IDF (the combination of TF and IDF) [47]. Such vectors, called document vectors, capture the relative importance of the words in the documents. The set of document vectors establishes a document-word matrix which is considered as a semantic model in Vector Space Model (a well-known model which has been widely used in Information Retrieval) [27].

Other approaches consider the appearance of words in paragraph, sentence, word window as the context of words. Recent works achieved much success in using word window as context for representing words as vectors in Vector Space Model [50] and [51]. We follow the idea from the authors and apply it to the domain of NPIs.

5.2.3.1 Vector representation of words

In order to calculate the similarity between words, effective representations of them are necessarily considered. Word vectors are one the most common types of word representation in the current NLP literature nowadays. There are some methods to represent a word in the vector form such as: one-hot encoded vector (or local representation), word embeddings (or distributed representation).

The former, one-hot encoded vector, is a naive and simple method. A vector representation of a word is a vector where 1 stands for the position where the word exists and 0 everywhere else. For example, given a dictionary = ['alternative', 'medicine', 'treatment', 'therapy', 'complementary', 'herbal'], then the vectors representing words 'medicine' and 'therapy' look like [0, 1, 0, 0, 0, 0] and [0, 0, 0, 1, 0, 0] respectively. Obviously, the Cosine measure between these two vectors is zero. Consequently, the similarity between two words 'medicine' and 'therapy' is nothing. In other words, this representation method is not meaningful.

The latter, word embeddings, is an effective method achieving much success

recently. A word is represented by its around context words or, say another way, it is embedded in the space of other words. Each word is represented by a distribution of weights across those context words. As a result, vector representation of words look like [0.65, 0.78, 0.42, 0.82, 0.95]. There are some techniques to perform this representation such as original LSA with context words counted from word-document matrix, LSA word similarity with context words counted word-word co-occurrence matrix, word2vec with context words learned from shallow neural network (CBOW and Skip-gram models). The last technique word2vec outperforms the first two techniques [50].

Hence, we apply the word2vec technique with specific training data for NPI domain. Our process includes three steps as illustrated in Figure 5.8.

Step 1: Creating NPI text corpus

As we mentioned above, there are many NPI terms appearing in PubMed articles. However, only a number of articles in PubMed are free access. The remaining ones allow users to access to their abstracts. Therefore, we create the text corpus of NPI by retrieving a subset of PubMed abstracts. This work is done by using PubMed API (esearch, efetch services).

Starting from the seed NPI terms, we use PubMed API to retrieve XML records corresponding to PubMed abstracts that relate to the seed terms. We then use an XML parser to extract only titles and abstract texts which are saved to text files in order to use in the next step.

In order to have a better quality text corpus in context of NPI, we filter PubMed abstracts so that only articles belonging to species Human are retrieved.

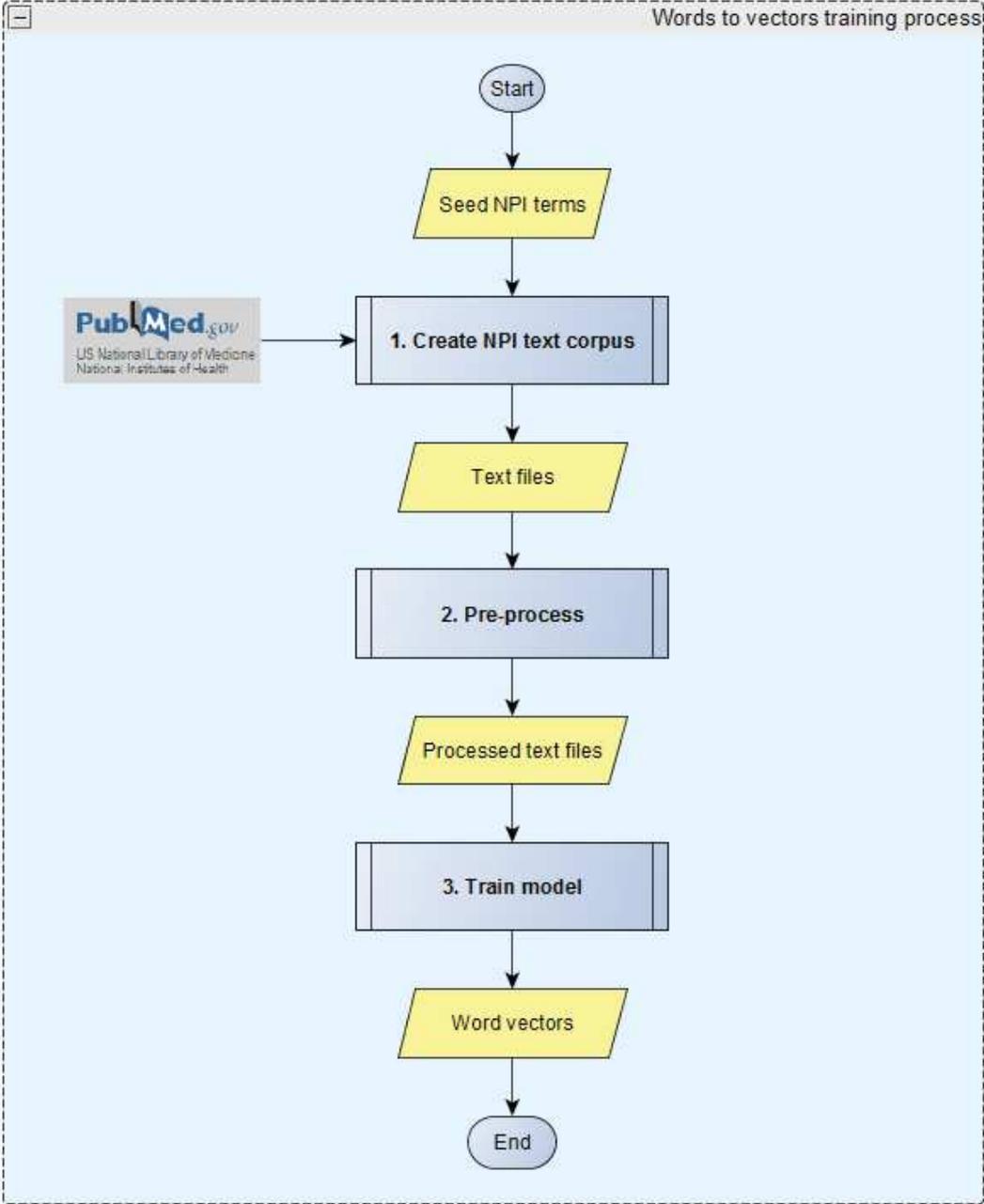


Figure 5.8 – Word vector training process

Step 2: Preprocessing

This is an important step to have valuable input for training model in the next step. We do three following tasks.

Sentence segmentation We segment paragraphs of each text document retrieved above to sentences. The reason is that we want to get context words in the scope of each sentence and avoid mixing contexts between different sentences.

Tokenization This task is to separate tokens (words, symbols etc.) by space in order to get words like “medicine” instead of “medicine.” (with period), “medicine,” (with comma) or “medicine;” (with semi-colon) etc.

Term modeling The purpose of this task is to link tokens of multi-token terms by underscore. For example, we want to have “alternative_medicine” (with underscore linking the two tokens) instead of “alternative medicine” (which is tokenized as separated tokens “alternative” and “medicine”). In order to do this, we use compound terms in MeSH to detect and link separate tokens as multi-token terms. With the remaining terms that do not exist in the vocabularies of MeSH, we use phrase modeling approach to learn combinations of tokens [51]. The main idea of this approach is looking for tokens that co-occur (i.e., appear one after another) together much more frequently than a threshold. The formula 5.4 proposed by [51] is used to determine whether two tokens A and B constitute a phrase.

$$\frac{count(AB) - count_{min}}{count(A) * count(B)} * N > threshold \quad (5.4)$$

where:

- $count(A)$: the number of times that token A appears in the text corpus;
- $count(B)$: the number of times that token B appears in the text corpus;
- $count(AB)$: the number of times that the tokens A and B appear together in the text corpus in order;
- N : the total size of the vocabulary of the text corpus;
- $count_{min}$: a user-defined parameter to ensure that accepted phrases occur at least the number of times;
- $threshold$: a user-defined parameter to indicate the score in the left of the formula must be greater than a threshold in order to tokens A and B constitute a phrase.

Note that in this step we do not need to remove stop words from the text corpus because we want to keep origin contexts of each word in the documents. Besides, the most common words in the corpus that do not carry necessary information are automatically discarded by a simple subsampling approach mentioned in [51]. The underlying idea of the approach is to counter the imbalance between rare and frequent words in the training set (for example, “in”, “the”, and “a” provide less information value than rare words). Each word w_i in the training set is discarded with probability $P(w_i)$ which is calculated by the formula:

$$P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}} \quad (5.5)$$

where $f(w_i)$ is the frequency of word w_i and t is a chosen threshold, typically around 10^{-5} as mentioned in [51].

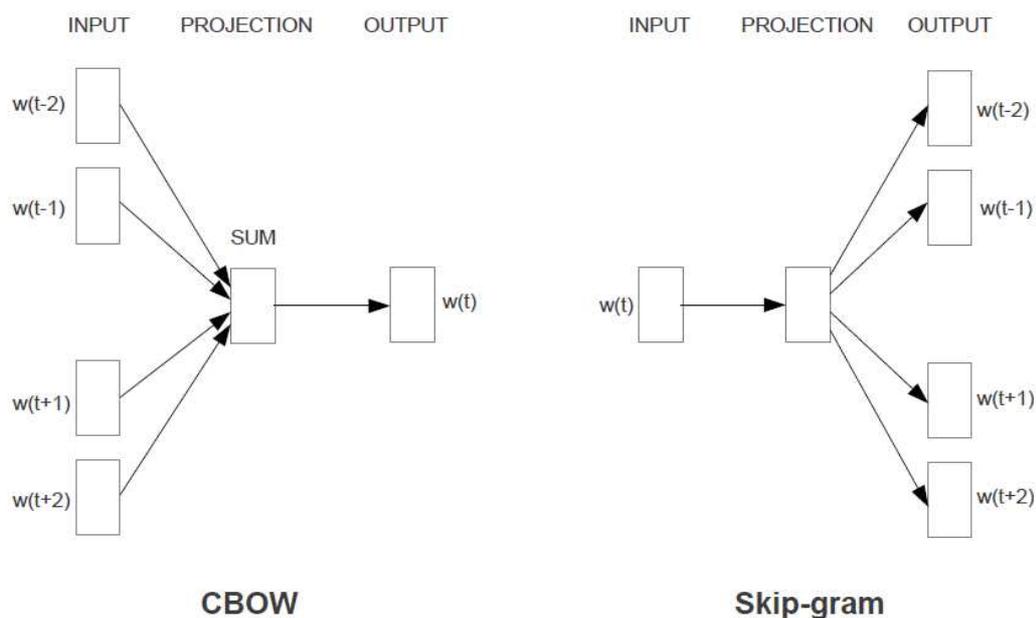


Figure 5.9 – CBOW and Skip-Gram model [50]

Step 3: Training the model

We choose context of each word being a group of words which is determined by a window with the size specified at 5 (5 context words behind and 5 context words ahead the center words). The window is sliced through every sentence of the NPI text corpus to identify the context words. Using CBOW and Skip-gram models [50] to predict the context words, we have word vectors that represent each word in the space of its context words. We choose Skip-gram model because it is better than CBOW model for semantic similarity and relatedness of biomedical terms [93] which is mainly considered in our work.

5.2.3.2 Calculating similarity between two words

After having vector representation of words, we can easily compute the similarity between two words by Cosine measure of two vectors corresponding to the two words. Let $x(x_1, x_2, \dots, x_n)$ and $y(y_1, y_2, \dots, y_n)$ are vectors of word w_1

and word w_2 respectively, then we have the similarity $sim(w_1, w_2)$ which is calculated by the equation:

$$sim(w_1, w_2) = \cos(x, y) = \frac{x \cdot y}{|x|_2 |y|_2} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}} \quad (5.6)$$

5.2.3.3 The similarity for multi-token terms

Although multi-token terms were established in the ‘‘Term modeling’’ taks of the preprocessing step, many other ones cannot be recognized. Therefore, there are not corresponding vectors for those terms. As a result, the similarity between them cannot be calculated directly. Some approximate similarities are presented in the following content.

Average vector

Given two multi-token terms A and B:

- $A = A_1 A_2 \dots A_n$ (A_i is a token of term A)
- $B = B_1 B_2 \dots B_m$ (B_j is a token of term B)
- \vec{X}_i and \vec{Y}_j are vectors of A_i and B_j

Then vectors of A and B are \vec{X} and \vec{Y} respectively, which are calculated as follows:

$$\vec{X} = \frac{\sum_{i=1}^n \vec{X}_i}{n} \quad \vec{Y} = \frac{\sum_{j=1}^m \vec{Y}_j}{m}$$

Finally, $sim(A, B)$ is computed as for two words (single terms).

Average Cosine Similarity

In fact, multi-token terms can be split into a set of tokens. Therefore, the similarity between them can be computed by using approximate methods. The idea of computing similarity value between two multi-token terms is inherited from the token-based similarity measures.

General idea of token-based similarity is separating compound terms into tokens using some separators (space, hyphen, punctuation, special characters) and also converting to lowercase. Then, the similarity score between two original terms is computed by measuring the degree of similarity between their collections of tokens. It means that the similarity score between two strings are calculated by comparing their tokens rather than the whole term themselves.

For example, if we want to compute the similarity between two terms “alternative medicine” and “complementary therapy”, we calculate the similarity of each pair (“alternative”, “complementary”), (“alternative”, “therapy”), (“medicine”, “complementary”) and (“medicine”, “therapy”).

Now we can apply a similarity measure for single words in order to compute the similarity score between two tokens. Having the similarity scores between every pair of tokens, the two widely used aggregation methods, ExtendedJaccard and Monge-Elkan, are used to calculate the similarity between two multi-token terms. In this thesis, we use a simple method called average cosine similarity (ACS) for the same purpose. It is calculated by the formula:

$$ACS(A, B) = \frac{1}{nm} \left(\sum_{i=1}^n \sum_{j=1}^m sim(A_i, B_j) \right) \quad (5.7)$$

where A_i and B_j are tokens of multi-token terms A and B respectively.

5.3 Term validation

This phase, illustrated in Figure 5.10, is manual work of NPI experts on the list of the potential NPI terms retrieved from the automatic phases above. For each potential NPI term, NPI experts discuss together to decide whether it is a NPI or non-NPI term. Finally, the new NPI term is positioned into the correct category of NPI taxonomy. The decision is based on knowledge and experience of NPI experts. Therefore, it is out of this thesis' scope. However, we also build a web-based application to aid the experts working on the validation (see Appendix A for more detail).

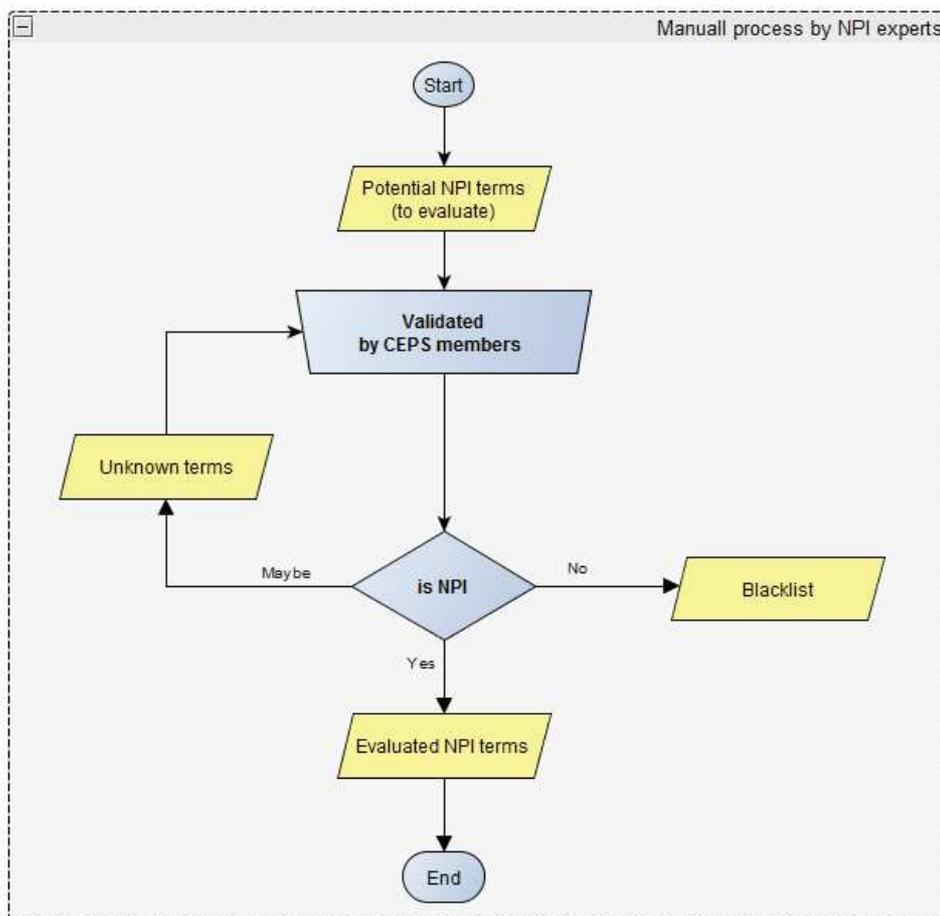


Figure 5.10 – Manual term validation process by NPI experts

Results and Evaluation

THIS CHAPTER presents experiments and results of the NPI term acquisition process followed by a discussion. The evaluation is included to estimate between the NPI similarity measure with the others.

6.1 Results

Initially, we had the seed NPI terms as presented in the Table 6.1.

Firstly, we collected 439,244 PubMed abstracts related to the seed NPI terms and 504,044 compound MeSH terms for training data. This work was done by our tool with the support of the PubMed API and Java. The retrieved documents were preprocessed before being put into the Gensim word2vec tool ¹ to produce word vectors.

In term extraction phase, by exploiting Wikipedia, we collected wiki pages, related to the seed NPI terms, which are then extracted to get title and content only. The retrieved texts were processed by BioTex tool ² to extract terms which were supplemented with the terms extracted from BioPortal. As a result we achieved a list of 3,686 candidate terms which is the input for the

¹<https://radimrehurek.com/gensim/models/word2vec.html>

²<http://tubo.lirmm.fr/biotex/>

Table 6.1 – List of seed terms

#	Term	#	Term
1	Non-Pharmacological Interventions	2	Psychological Health Interventions
3	Art Therapy	4	Health Education
5	Psychotherapy	6	Zootherapy
7	Physical Health Interventions	8	Physical Activity
9	Physiotherapy	10	Manual Therapy
11	Thermalism	12	Nutritional Health Interventions
13	Dietary Supplements	14	Nutritional Therapy
15	Digital Health Interventions	16	eHealth Devices
17	Therapeutic Games	18	Virtual Reality Therapy
19	Other Health Interventions	20	Ergonomic Tools
21	Phytotherapy	22	Cosmetic Therapy
23	Wave Therapy	24	Lithotherapy
25	Alternative Medicine	25	Complementary Medicine
27	Natural Medicine	28	Traditional Medicine
29	Integrative Medicine	30	Behavioral Medicine
31	Alternative Therapy	32	Complementary Therapy
33	Complementary Treatment		

term ranking phase. This work was done by our tool with the support of the BioPortal REST API and wikipedia library in Python.

In term ranking phase, we used K-means algorithm implementation in Python to perform the clustering task on the set of seed terms. Through our experiments and observations on the visualization of the clustering result (Figure 6.1), we found that the clusters were the best with parameter $k = 3$ (k is number of clusters). The seed terms thus were clustered into 3 clusters as listed in Tables 6.2, 6.3 and 6.4. Then we calculated and compared the distances from each candidate term to the centroids of their clusters. If the distances are greater than the maximum distance of items to the centroid of their cluster, then those candidate terms are removed.

In the last step of this phase, our proposed method was used to compute the NPI score of candidate terms. Some achieved results are illustrated in

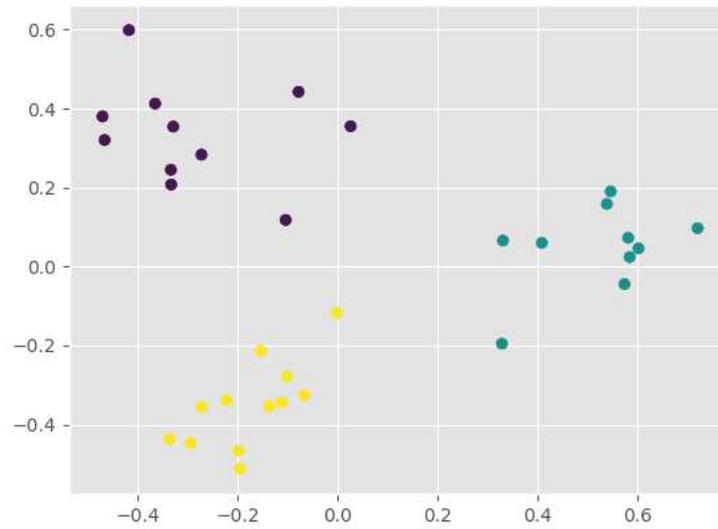
Figure 6.1 – Clustering the seed terms with K-means ($k = 3$)

Table 6.2 – Seed terms in the first cluster

#	Term	Cluster label
1	Non-Pharmacological Interventions	0
2	Psychological Health Interventions	0
3	Health Education	0
4	Physical Health Interventions	0
5	Physical Activity	0
6	Nutritional Health Interventions	0
7	Dietary Supplements	0
8	Digital Health Interventions	0
9	eHealth Devices	0
10	Other Health Interventions	0
11	Ergonomic Tools	0

Table 6.3 – Seed terms in the second cluster

#	Term	Cluster label
1	Art Therapy	1
2	Psychotherapy	1
3	Physiotherapy	1
4	Manual Therapy	1
5	Nutritional Therapy	1
6	Therapeutic Games	1
7	Virtual Reality Therapy	1
8	Cosmetic Therapy	1
9	Wave Therapy	1
10	Alternative Therapy	1
11	Complementary Therapy	1
12	Complementary Treatment	1

Table 6.4 – Seed terms in the third cluster

#	Term	Cluster label
1	Zootherapy	2
2	Thermalism	2
3	Phytotherapy	2
4	Lithotherapy	2
5	Alternative Medicine	2
6	Complementary Medicine	2
7	Natural Medicine	2
8	Traditional Medicine	2
9	Integrative Medicine	2
10	Behavioral Medicine	2

Tables 6.5 and 6.7. Based on the results, NPI experts will make final decisions.

After this second phase (term ranking), the list was pre-filtered and ranked to become a shorter list with 2,767 potential NPI terms which was the input for the term validation phase.

Finally, the list of potential NPI terms was validated by NPI experts to get 389 new NPI terms. The summary of the experiment is illustrated in Figure 6.2.

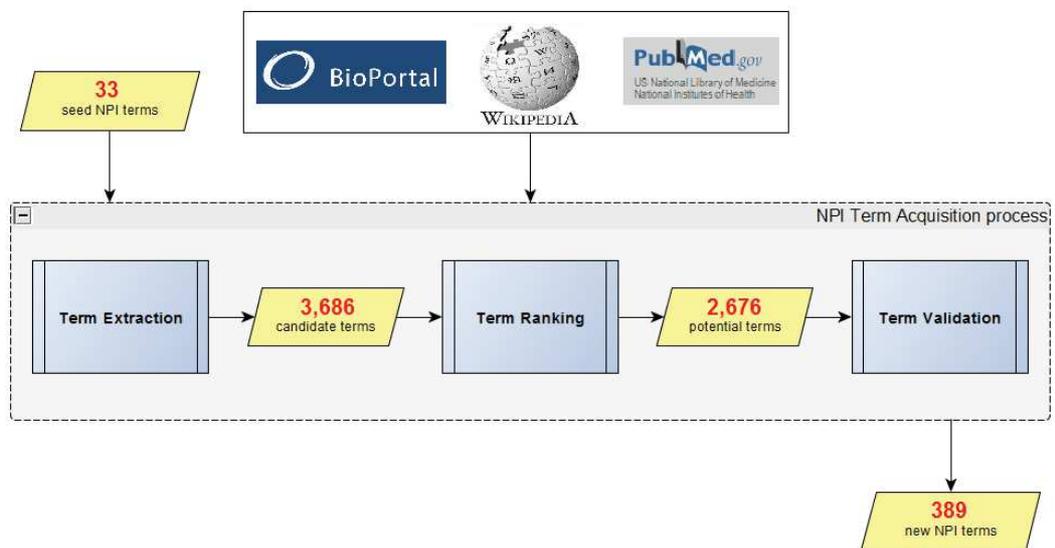


Figure 6.2 – Experimental results

6.2 Evaluation

In order to estimate our similarity measure for NPI, we prepared test data including 1373 terms (376 NPI terms and 997 Non-NPI terms) which are hand-labeled by NPI experts at Plateforme CEPS. Assuming that c is a potential NPI term, $NPI_score(c)$ is a function to compute the NPI score, δ is a score threshold to identify a term being a NPI term or not, we propose the function

Table 6.5 – Top candidate terms using ACS to calculate NPI score

#	Candidate term	NPI score	Closest seed term	Is NPI
1	psychodynamic psychotherapy	0.909100472	psychotherapy	Yes
2	eclectic psychotherapy	0.827152254	psychotherapy	Yes
3	massage physiotherapy	0.819917806	physiotherapy	Yes
4	pt physiotherapy	0.81054997	physiotherapy	Yes
5	immobilisation physiotherapy	0.793248715	physiotherapy	Yes
6	interpersonal psychotherapy	0.781985492	psychotherapy	Yes
7	art psychotherapy	0.777390757	psychotherapy	Yes
8	couple psychotherapy	0.77106033	psychotherapy	Yes
9	supportive psychotherapy	0.76110063	psychotherapy	Yes
10	brief psychotherapy	0.756509859	psychotherapy	Yes
11	rehabilitation	0.756469097	physiotherapy	Yes
12	physiotherapy manipulation	0.753760891	physiotherapy	Yes
13	ayurveda	0.746957455	phytotherapy	Yes
14	psychodynamic therapy	0.738788117	psychotherapy	Yes
15	ethnomedicine	0.734458289	zootherapy	Yes
16	parent-infant psychotherapy	0.729624925	psychotherapy	Yes
17	expressive psychotherapy	0.725312732	psychotherapy	Yes
18	individual psychotherapy	0.717097224	psychotherapy	Yes
19	homeopathy	0.717019641	phytotherapy	Yes
20	immobilization physiotherapy	0.716897153	physiotherapy	Yes
21	short-term psychotherapy	0.715192848	psychotherapy	Yes
22	rational psychotherapy	0.714876664	psychotherapy	Yes
23	integrative psychotherapy	0.714842356	psychotherapy	Yes
24	cbt	0.713947944	psychotherapy	Yes
25	medicine	0.707301836	complementary medicine	No
26	otorhinolaryngology	0.703660769	thermalism	No
27	rational-emotive psychotherapy	0.699626188	psychotherapy	Yes
28	psychoanalytic therapy	0.697056604	psychotherapy	Yes
29	folklore	0.690546695	zootherapy	No
30	shamanism	0.689552871	zootherapy	No

Table 6.6 – Top candidate terms using UMBC to calculate NPI score

#	Candidate term	NPI score	Closest seed term	Is NPI
1	nutrition therapy	1	nutritional therapy	Yes
2	education health	1	health education	Yes
3	complementary therapy no	1	complementary therapy	No
4	physical therapy physiotherapy	1	physiotherapy	No
5	he health education	1	health education	Yes
6	pt - physiotherapy	1	physiotherapy	Yes
7	he - health education	1	health education	Yes
8	school health	0.9660684	health education	No
9	health education profession	0.9348058	health education	No
10	alternative treatment	0.9226736	alternative therapy	Yes
11	alternative medicine therapy	0.9028952	alternative therapy	Yes
12	nutrition intervention	0.89406794	nutritional health intervention	Yes
13	health education and awareness	0.89172745	health education	No
14	medical nutrition therapy	0.88933265	nutritional therapy	Yes
15	nutrition disorder therapy	0.8890289	nutritional therapy	Yes
16	mental health education	0.8876382	health education	Yes
17	dance therapy	0.88399255	art therapy	Yes
18	community health education	0.88018155	health education	No
19	public health education	0.8789342	health education	No
20	dental health education	0.87686723	health education	No
21	health education resource	0.87598723	health education	No
22	art psychotherapy	0.8737393	art therapy	Yes
23	marketed as dietary supplement	0.8733861	dietary supplement	No
24	drugs - health education	0.8692732	health education	No
25	virtual reality immersion therapy	0.86924887	virtual reality therapy	Yes
26	nutritional intervention	0.8691811	nutritional health intervention	Yes
27	health education study	0.868764	health education	No
28	national health education	0.8685779	health education	No
29	combination dietary supplement	0.86734766	dietary supplement	Yes
30	health education specialist	0.8608667	health education	No

Table 6.7 – Top candidate terms using Average Vector to calculate NPI score

#	Candidate term	NPI score	Closest seed term	Is NPI
1	education health	1	health education	Yes
2	stimulative psychotherapy	0.999446988	psychotherapy	Yes
3	health education credentialing	0.996650994	health education	No
4	mongolian traditional medicine	0.9880777	traditional medicine	Yes
5	parent-infant psychotherapy	0.984883845	psychotherapy	Yes
6	eclectic psychotherapy	0.982328773	psychotherapy	Yes
7	rational-emotive psychotherapy	0.981239617	psychotherapy	Yes
8	virtual reality immersion therapy	0.978054225	virtual reality therapy	Yes
9	certified health education	0.977644742	health education	No
10	immobilisation physiotherapy	0.97006011	physiotherapy	Yes
11	folk medicine	0.968921602	traditional medicine	Yes
12	health education code	0.965599775	health education	No
13	psychodynamic psychotherapy	0.964828908	psychotherapy	Yes
14	community health education	0.964672446	health education	No
15	health education resource	0.962411761	health education	No
16	health occupation education	0.959998906	health education	Yes
17	dental health education	0.959929764	health education	No
18	health education general	0.957329571	health education	No
19	manual arts therapy	0.957079649	manual therapy	Yes
20	explorative psychotherapy	0.955182254	psychotherapy	Yes
21	health education specialist	0.95434624	health education	No
22	health education profession	0.953878522	health education	No
23	public health education	0.953245342	health education	No
24	traditional chinese medicine	0.947161973	traditional medicine	Yes
25	couple psychotherapy	0.947116375	psychotherapy	Yes
26	health promotion and education	0.945357919	health education	No
27	expressive psychotherapy	0.945218801	psychotherapy	Yes
28	irrigation therapy	0.942307174	cosmetic therapy	No
29	health education and awareness	0.940779686	health education	No
30	pt physiotherapy	0.94032532	physiotherapy	Yes

$IsNPI(c)$ to predict new NPI terms based on their NPI scores as follows:

$$IsNPI(c) = \begin{cases} True & \text{if } NPI_score(c) \geq \delta \\ False & \text{if } NPI_score(c) < \delta \end{cases} \quad (6.1)$$

By assigning different values for δ , we can calculate the precision with each specific value of δ by formulas 6.2.

$$Precision_{\delta} = \frac{tp_{\delta}}{tp_{\delta} + fp_{\delta}} \quad (6.2)$$

where:

- tp_{δ} (true positive): number of NPI terms which are predicted correctly;
- fp_{δ} (false positive): number of Non-NPI terms which are predicted as NPI terms.

We did experiments with different thresholds δ for our method and UMBC measure [14] to make a comparison between them. Then we got the results as illustrated in the Figure 6.3.

There are two reasons that helps our method got better result than the UMBC measure:

- The first reason is because we have trained the model on a subset of PubMed documents related to NPIs. That means we put the NPI context to the calculation process. Where as, UMBC has been trained on many text copora that are not related to NPIs. As a result, it has many noises when it is used for NPI domain.
- The second reason is because we have used the Skip-gram model with the word2vec algorithm which outperforms the LSA model used in UMBC.

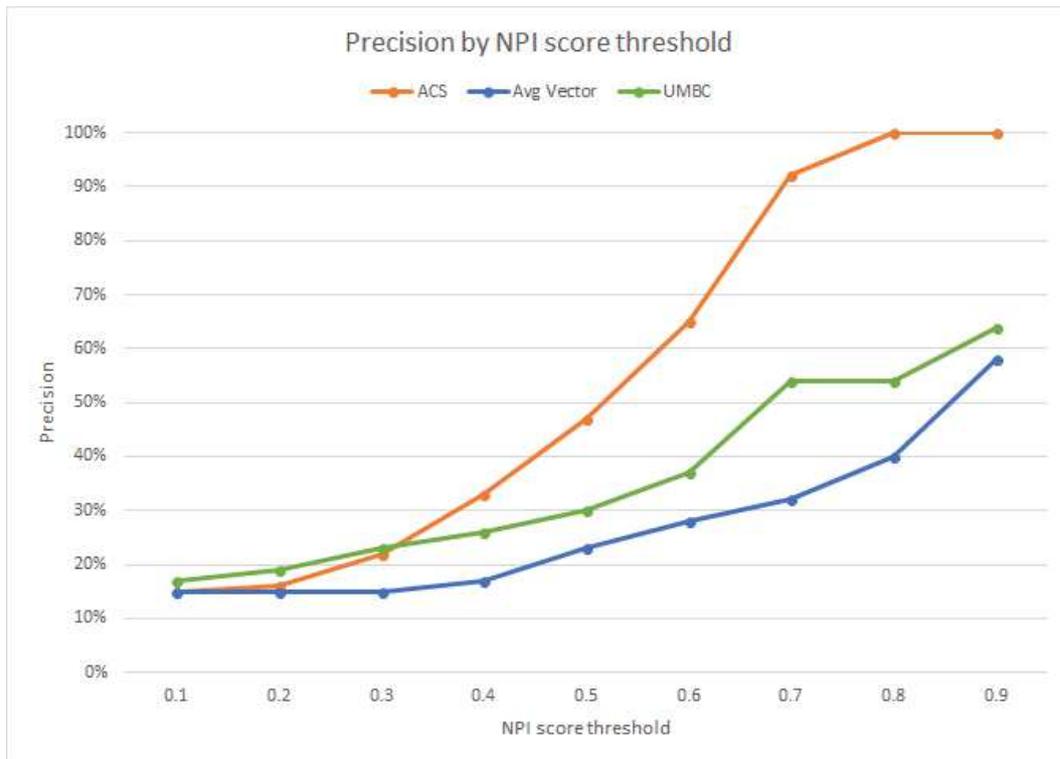


Figure 6.3 – Precision by NPI score threshold

The achieved result is significant in comparison with UMBC measures because our measure is trained on a text corpus of more than 400 thousand documents, while UMBC measure is trained on a text corpus up to 100 million web pages.

Conclusion

IN this chapter, we have presented a solution to automatically collect NPI terms. The underlying idea of the method is an expansion of the seed NPI terms by collecting relevant terms from various sources.

Firstly, a text corpus for NPI was created by gathering documents related to the seed NPI terms. We focused on exploiting some big and well-known resources: PubMed abstracts, ontologies in BioPortal, Wikipedia pages, Google search result pages.

Secondly, a list of candidate terms was extracted from the unstructured documents. Here, we used BioTex which is a state-of-the-art biomedical term extraction tool to deal with this problem. The list were then filtered to remove poor candidates. To achieve this purpose, we compared the distance from each candidate term to the centroids of clusters which are established by clustering the seed terms with K-means algorithm. In order to use K-means, we first had to represent the terms as feature vectors in the Vector Space Model. This task was done with word vectors trained by the word2vec model (Skip-gram model) on the NPI text corpus.

Finally, a ranking process was performed to figure out the potential terms which was then manually evaluated by NPI experts. The best positions in the NPI taxonomy were also suggested for the potential terms. These tasks have done with our proposed similarity measure for NPIs which outperformed UMBC similarity measure as a baseline.

As mentioned previously, our work limits on supporting NPI experts to collect NPI terms. After having a list of terms, experts in the Plateforme CEPS put the terms into WebProtege to make an OWL ontology for NPIs. This ontology was then imported to BioPortal ¹ in order to share and get feedback from community [57]. A full ontology for NPIs is the goal of the Plateforme CEPS. However, in the first steps, they just built a lightweight ontology [13] including a taxonomy and synonyms, related terms [22]. Developing a domain ontology is a collaborative and iterative work, therefore the experts often discuss together to make decisions. A visualization of the resulted ontology on a big screen is thus crucial for them at the time of discussion. This motivates us to propose a method to visualize the ontology. Our method does not only limit to this ontology, but it can also use for other ontologies. The next part will detail our idea.

¹<http://bioportal.bioontology.org/ontologies/NPI>

Part III

MindMap-based Ontology Visualization

8	Problem Statement	83
9	Related Works	85
9.1	Hierarchy-style visualization	85
9.2	Graph-style visualization	86
10	Methodology	93
10.1	MindMap-based visual notation	93
10.2	Transforming OWL into FreeMind XML	93
11	Results and Evaluation	99
11.1	Results	99
11.2	Evaluation	102
12	Conclusion	109

Problem Statement

ONTOLOGIES include complex concepts and relations. Therefore, an efficient visualization of ontology is essential to end-users for understanding and working on domain knowledge. However, existing visualization tools are often difficult to use by domain experts, for instance for focusing on subparts of the domain. In this chapter, we introduce a new method allowing to visualize ontologies as MindMaps that are known as efficient visualization tools. Our tool, OWL2MM, automatically transforms OWL ontologies into FreeMind¹ documents which can then be imported in any mind-mapping application to visualize ontologies.

Although several ontology visualization tools have been introduced in the literature, most of them are not complete, as they are not able to visualize all elements of a large ontology. Some of tools such as OWLViz [29], OntoGraf [74] and TGViz [2] only focus on each part of an ontology, not the whole. Thus, users do not have an overview of ontology structure. Besides, several tools like WebVOWL [41], OWLGrEd [6] and SOVA [35] are powerful and can visualize most of key elements of ontologies. But they display a lot of ontology elements on screen. As a result, this makes the visualization confused in case of large ontology.

A feature that users, especially domain experts, would expect from ontology visualization tools is the ability to customize colors for each sub domain

¹http://freemind.sourceforge.net/wiki/index.php/Main_Page

of ontologies. As far as our knowledge, there does not exist any ontology visualization tool that meets this requirement. Most of them focus on putting different colors for corresponding components of ontologies.

In this chapter, we present an approach using elements of MindMap to visualize ontology elements. According to [73], there are some similarities between MindMaps and ontologies. This motivates us to do this research. Another reason why we pursue this approach is that MindMap visualization is more familiar and accepted by domain experts. Besides, MindMaps can collapse nodes which can help users to customize the view of ontology visualization, especially the visualization of large ontologies. Our tool, OWL2MM, transforms any OWL ontology (OWL file) into a FreeMind document (MM file) which is then imported by existing mind-mapping tools to create a corresponding visualization as a MindMap. OWL2MM allows users to put multiple colors for content of classes or for class types (internal, external, deprecated classes).

Related Works

VISUALIZATION of ontologies is not an easy task because ontologies are often complicated and have large amount of data. An ontology is not only a hierarchy of concepts but also includes relations among concepts and various properties related to each concept. Furthermore, each concept may have many instances attached to it. Hence, it is not simple to create a visualization that will effectively display all this information on a limited screen.

There are two common categories of ontology visualization tools:

- Hierarchy-style visualization mainly focuses on basic structure of an ontology and organizes it as a tree. Those tools, therefore, depict merely classes, “is-a” relations and sometimes individuals of an ontology. They bring a simple and friendly visualization to end-users.
- Graph-style visualization allows to visualize an ontology as a graph with nodes and edges with labels. This kind of visualization is close to the nature of ontologies. It is preferred by ontology developers.

9.1 Hierarchy-style visualization

OWLviz [29] is a mapping visualization plugin designed for Protégé. It allows the user to view an ontology as a concept map. One of the primary require-

ments in our research was the ability to create mind maps and topic maps. Therefore, this functionality within Protégé would significantly raise its stock. OWLViz is one of the solutions to this dilemma. However, OWLViz does not illustrate the relationships between each object, nor does it allow the user to create or edit the ontology within this view. It can only display the relation of Super-class and Sub-class (“is-a” relation).

OntoGraf [74], also a plugin of Protégé, gives support for interactively navigating the relationships of an OWL ontologies. OntoGraf can show other relations in addition to the ‘is-a’ relationship. It provides automatically organized structure with interactive relationships among classes. The domain and range for each property are shown with colorful arcs. Various layouts are supported for automatically organizing the structure of the ontology. Different relationships are supported: subclass, individual, domain/range object properties, and equivalence. Relationships and node types can be filtered to help users create the view they desire. OntoGraf represents various types of property relations, but do not show datatype properties and property characteristics required to fully understand the information modeled in ontologies.

9.2 Graph-style visualization

SOVA (Simple Ontology Visualization API) [35] is a Protégé plugin for full ontology visualization. It can show all ontology’s elements: classes, individuals, properties, anonymous classes and relations between these object. SOVA supports 3 automatic visualization types: Force directed tree layout, Node tree layout, Radial tree layout. Furthermore, it has an option allowing users to choose what displays as label of classes (class ID, class label, class IRI).

The authors in [56] presented a visual notation for the Web Ontology Language (OWL) providing an integrated view on the classes and individuals of

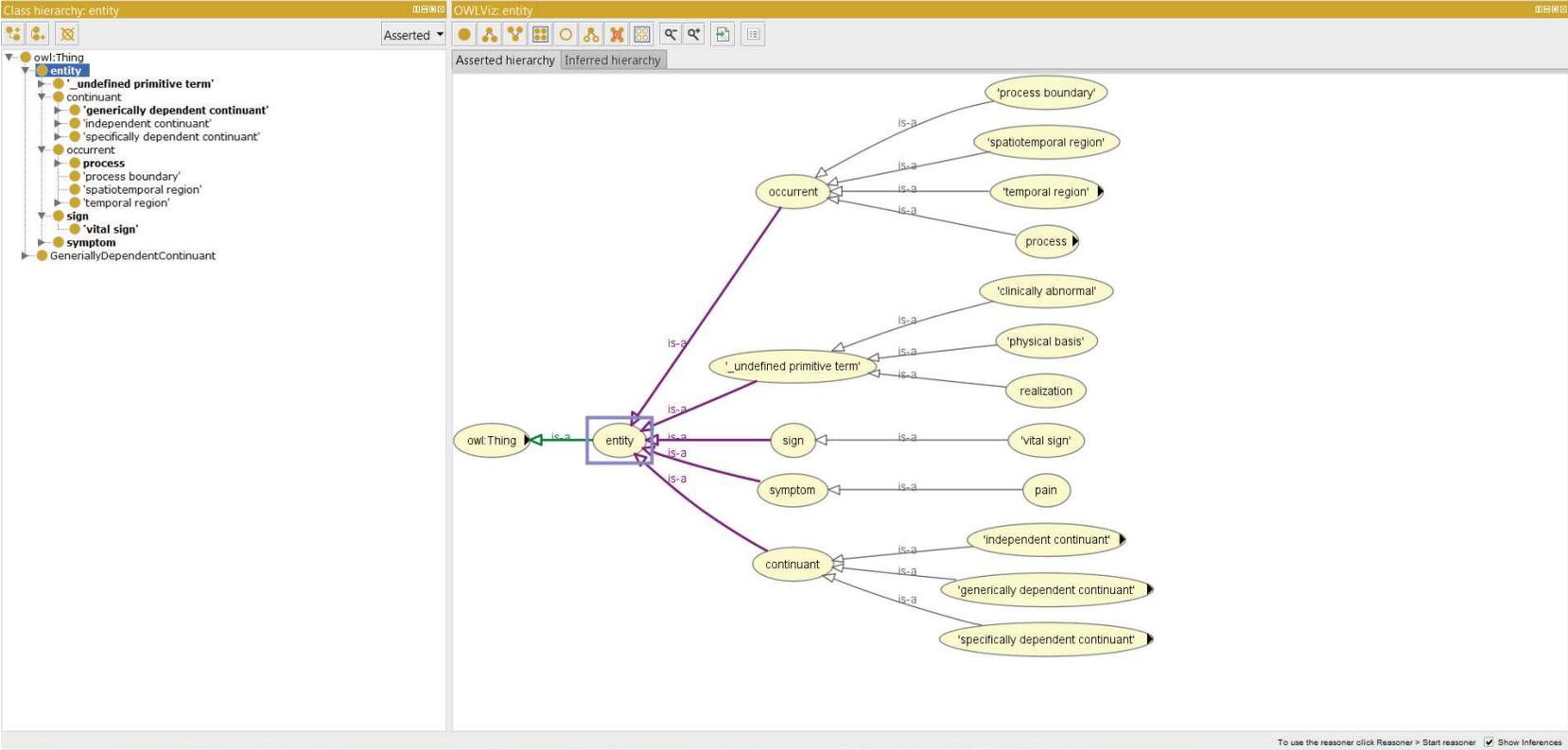


Figure 9.1 – Visualization of General Medical Science ontology by OWLViz

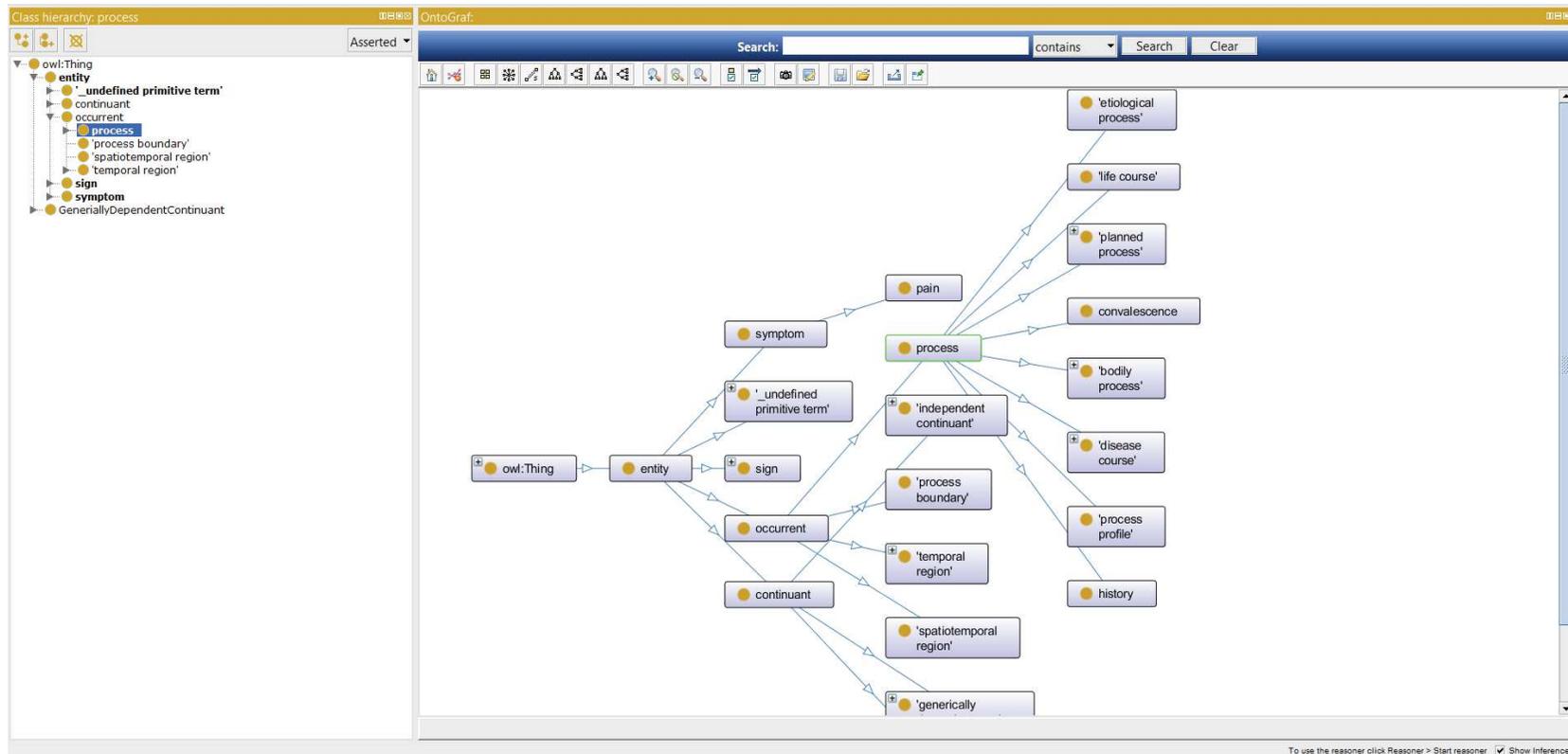


Figure 9.2 – Visualization of General Medical Science ontology by OntoGraf

ontologies. The classes are displayed as circles, with the size of each circle representing its connectivity in the ontology. The individuals are represented as sections in the circles so that it is immediately clear from the visualization how many individuals the classes contain. VOWL [41], the aforementioned Visual Notation for OWL Ontologies, was developed as a means to both obtain a structural overview of OWL ontologies and recognize various attributes of ontology elements at a glance. It has been implemented in two different tools, a plugin called ProtégéVOWL [38] for Protégé and a responsive web application named WebVOWL [40].

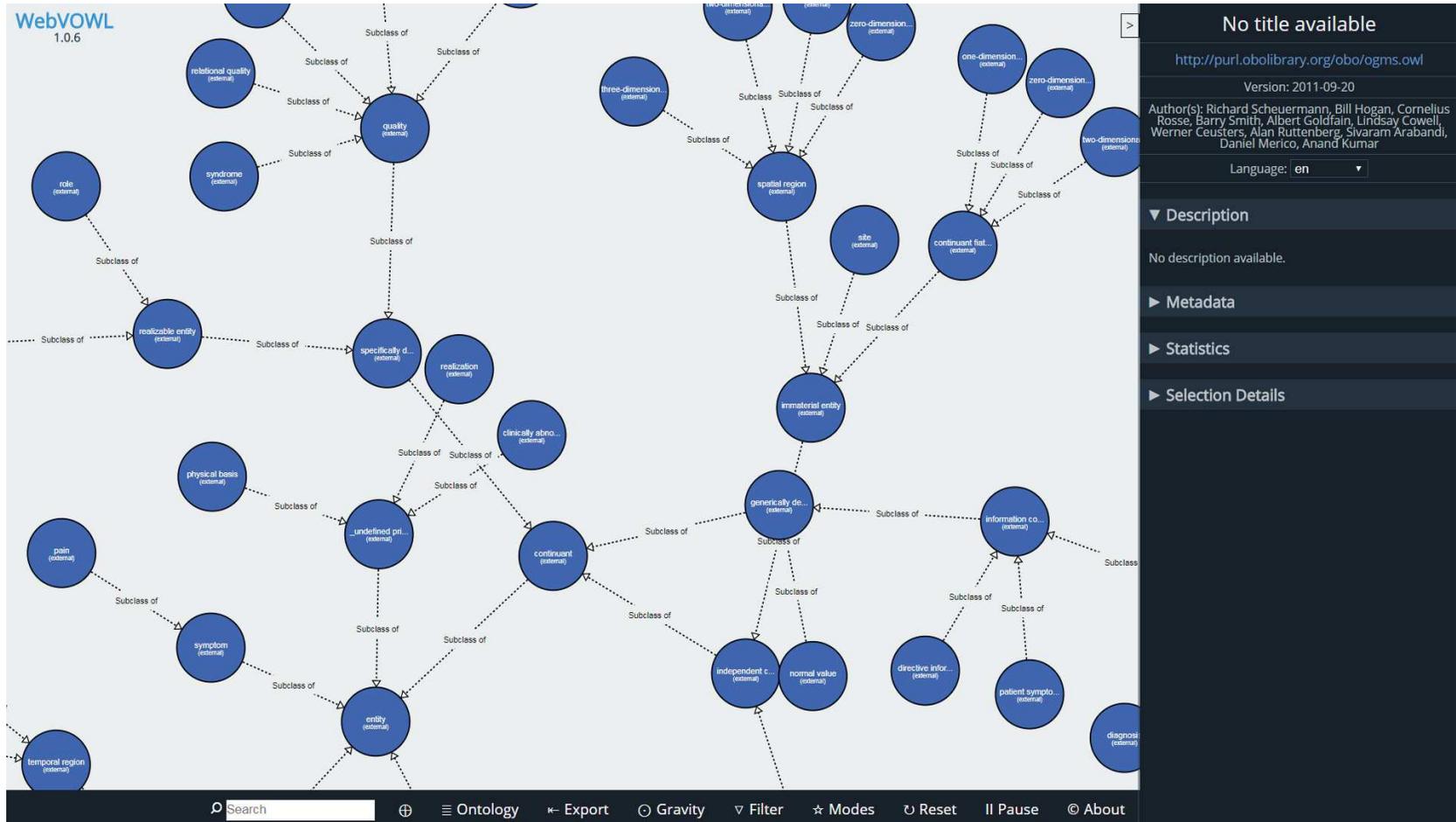


Figure 9.4 – Visualization of General Medical Science ontology by WebVOWL

Methodology

THIS chapter presents the work on ontology visualization problem, especially the visualization for domain experts. A new method is introduced to allow visualizing ontologies as MindMaps that are known as efficient visualization tools. Finally, a tool named OW2MM is built in order to automatically transform OWL ontologies into FreeMind documents which can then be imported in any mind-mapping application to visualize ontologies.

10.1 MindMap-based visual notation

In this section, we define a visual notation for the user-oriented representation of ontologies. It provides graphical depictions for elements of the OWL ontology by combining existing MindMap elements. In order to have a unity between visualization tools, we use some colors of the color scheme which is proposed in [39] for class types of ontologies. Our proposal is detailed in Table 10.1.

10.2 Transforming OWL into FreeMind XML

The idea of our transformation is based on mappings between elements of OWL ontologies to elements of FreeMind documents in Table 10.2.

Table 10.1 – MindMap-based visual notation for ontology elements

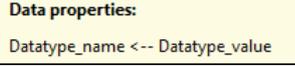
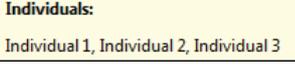
OWL element	FreeMind representation	Description
owl:Thing		A bubble node of MindMap with black label “Thing”, white background and light black border
owl:Class		A bubble node of MindMap with black label, light blue background and light black border
owl:Class (External class, different base URI)		A bubble node of MindMap with white label, dark blue background and light black border
owl:DeprecatedClass		A bubble node of MindMap with black label, light grey background and light black border
rdfs:subClassOf		A fork node with text “is-a”, transparent background and black label, to express A is a subclass of B.
owl:ObjectProperty		A fork node and an arrow link to describe a relation from class C (domain) to class D (range).
owl:DatatypeProperty		A tooltip or a popup window (stored in note element of MindMap) with a black bold text as the heading and text lines including a pair of name and value.
Individuals		A tooltip or a popup window (stored in note element of MindMap) with a black bold text as the heading and a text line including individuals (instances) separated by a comma.

Table 10.2 – Mapping ontology elements to FreeMind elements

OWL element	FreeMind elements
owl:Thing	<node ID="1" BACKGROUND_COLOR="#ffffff" COLOR="#000000" STYLE="bubble" TEXT="Thing"> </node>
owl:Class	<node ID="2" BACKGROUND_COLOR="#aacdff" COLOR="#000000" STYLE="bubble" TEXT="Class"> </node>
owl:Class (External class, different base URI)	<node ID="3" BACKGROUND_COLOR="#3366cc" COLOR="#ffffff" STYLE="bubble" TEXT="Class"> </node>
owl:DeprecatedClass	<node ID="4" BACKGROUND_COLOR="#cccccc" COLOR="#000000" STYLE="bubble" TEXT="Class"> </node>
rdfs:subClassOf	<node ID="5" STYLE="fork" TEXT="is-a"> </node>
owl:ObjectProperty	<node ID="6" STYLE="fork" TEXT="owl:ObjectProperty"> <arrowlink COLOR="#b0b0b0" DESTINATION="nodeID" ENDARROW="Default" ENDINCLINATION="143;0;" ID="Arrow1" STARTARROW="None" STARTINCLINATION="143;0;"/> </node>
owl:DatatypeProperty	<richcontent TYPE="NOTE"> <html> <head> </head> <body> <h4>Data properties: </h4> <p> Datatype name: Datatype value </p> </body> </html> </richcontent>
Individuals	<richcontent TYPE="NOTE"> <html> <head> </head> <body> <h4>Individuals: </h4> <p> Individual 1, Individual 2, Individual 3 </p> </body> </html> </richcontent>

Assuming that `mm-Thing`, `mm-Class`, `mm-ClassExternal`, `mm-DeprecatedClass`, `mm-subClassOf`, `mm-ObjectProperty`, `mm-DatatypeProperty`, `mm-Individual` are names of FreeMind elements corresponding to OWL ontology elements, respectively in Table 10.2. Then we have functions:

- *fmThing()*: returns `mm-Thing`.
- *fmClass(text)*: returns `mm-Class` with `TEXT` replaced by the parameter “text”.
- *fmClassExternal(text)*: returns `mm-ClassExternal` `TEXT` replaced by the parameters “text”.
- *fmDeprecatedClass(text)*: returns `mm-DeprecatedClass` with `TEXT` replaced by the parameter “text”.
- *fmSubClassOf()*: returns `mm-subClassOf`.
- *fmObjectProperty(id1, text, id2)*: returns `mm-ObjectProperty` with `ID`, `TEXT` and `DESTINATION` replaced by the parameters “id1”, “text” and “id2”.
- *fmDatatypeProperty(id, name, value)*: returns `mm-ObjectProperty` with “id”, “name” and “value” from parameters.
- *fmIndividual(names)*: returns `mm-Individual` with “names” being individuals.

The transformation is formulated in the Algorithm 1.

`VisitClass(C, Lnode)` is a recursive function to visit an OWL class in order to read its properties and relations connected to it. Then this function creates corresponding elements of FreeMind document as formulated in Table 10.2. Its algorithm is described in Algorithm 2.

Algorithm 1: Transforming OWL document to FreeMind document

```

input :  $D_{owl}$  = OWL document
output:  $D_{freemind}$  = FreeMind document

1  $D_{freemind} \leftarrow \emptyset$ ;
2  $MapElement \leftarrow$  XML element <map>;
3 Append  $MapElement$  to  $D_{freemind}$ ;
4  $RootNode \leftarrow fmThing()$  ;           // <node> corresponding to
   owl:Thing
5 Append  $RootNode$  to  $D_{freemind}$ ;
6  $L_{class} \leftarrow$  List of first level classes of  $D_{owl}$  ;           // Subclasses of
   owl:Thing
7  $L_{node} \leftarrow \emptyset$ ;           // List of FreeMind elements <node>
8 foreach class  $C \in L_{class}$  do
9   VisitClass( $C, L_{node}$ );
10  Append  $L_{node}$  to  $D_{freemind}$ ;
11   $L_{node} \leftarrow \emptyset$ ;
12 end
13  $L_{attr} \leftarrow$  List of properties of  $D_{owl}$ ;
14  $AttrNode \leftarrow \emptyset$ ;
15 foreach property  $P \in L_{attr}$  do
16   if  $P$  is ObjectProperty then
17      $id1 \leftarrow$  ID of domain;
18      $id2 \leftarrow$  ID of range;
19      $AttrNode \leftarrow fmObjectProperty(id1, P_{label}, id2)$  ; //  $P_{label}$ :
   label of  $P$ 
20   else if  $P$  is DatatypeProperty then
21      $id \leftarrow$  ID of domain;
22      $AttrNode \leftarrow fmDatatypeProperty(id, P_{domain}, P_{range})$  ;
   //  $P_{domain}$ : label of domain,  $P_{range}$ : label of range,
23   Add  $AttrNode$  to  $L_{node}$ ;
24 end
25 Append  $L_{node}$  to  $D_{freemind}$ ;

```

Algorithm 2: Visiting an OWL class and create corresponding Free-Mind elements

```

input :  $C$  = OWL class
output:  $L_{node}$  = List of FreeMind elements <node>

1  $IsaNode \leftarrow fmSubClassOf()$ ; // FreeMind element <node>
2 Add  $IsaNode$  to  $L_{node}$ ;
3  $SubNode \leftarrow \emptyset$ ; // FreeMind element <node>
4  $E_{text} \leftarrow \emptyset$ ; // labels of equivalent classes of  $C$ 
5  $L_{equiv} \leftarrow$  List of equivalent classes of  $C$ ;
6 foreach class  $E \in L_{equiv}$  do
7 |  $E_{text} \leftarrow E_{text} + E_{label}$ ; //  $E_{label}$ : label of  $E$ 
8 end
9  $C_{label} \leftarrow C_{label} + E_{text}$ ; // join  $C_{label}$  and  $E_{text}$  into a string
10 if  $C$  is external class then
11 |  $SubNode \leftarrow fmExternalClass(C_{label})$ ;
12 else if  $C$  is deprecated class then
13 |  $SubNode \leftarrow fmDeprecatedClass(C_{label})$ ;
14 else
15 |  $SubNode \leftarrow fmClass(C_{label})$ ;
16 end
17 Add  $SubNode$  to  $L_{node}$ ;
18  $L_{indi} \leftarrow$  List of individuals of  $C$ ;
19  $IndiNode \leftarrow \emptyset$ ; // FreeMind element <richcontent>
20 foreach individual  $I \in L_{indi}$  do
21 |  $IndiNode \leftarrow fmIndividual(I_{label})$ ; //  $I_{label}$ : label of  $I$ 
22 end
23 Add  $IndiNode$  to  $L_{node}$ ;
24 if  $C$  has subclasses then
25 |  $L_{sub} \leftarrow$  List of subclasses of  $C$ ;
26 | foreach class  $S \in L_{sub}$  do
27 | | VisitClass( $S, L_{node}$ );
28 | end
29 end

```

Results and Evaluation

IN this chapter, we present results and evaluation of the MindMap-based ontology visualization method. A comparison is made to show advantages and disadvantages of our method and the others.

11.1 Results

We use Apache Jena ¹, a free and open source Java framework, to build the tool named OWL2MM. The tool has an OWL file as input and an FreeMind file as output. It has some options to allows users to choose: displaying elements with label (rdfs:label) or local name; multiple colors for class types or sub domains. The tool also allows users to use predefined colors or customize with their own colors.

Figure 11.1 is a part of an OWL ontology “Wine” with RDF/XML syntax which is input data of our tool. A corresponding part of a FreeMind document as an output of our tool (with the option “customize colors by class types”) is illustrated in Figure 11.2.

In order to display the result as a visual MindMap, users can use FreeMind tool or other mind-mapping tools because, according to our knowledge, FreeMind file is the most popular MindMap file so it can be supported by many

¹<https://jena.apache.org>

```
<owl:Class rdf:ID="Region" />

<owl:ObjectProperty rdf:ID="locatedIn">
  <rdf:type rdf:resource="&owl;TransitiveProperty" />
  <rdfs:domain rdf:resource="http://www.w3.org/2002/07/owl#Thing" />
  <rdfs:range rdf:resource="#Region" />
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="adjacentRegion">
  <rdf:type rdf:resource="&owl;SymmetricProperty" />
  <rdfs:domain rdf:resource="#Region" />
  <rdfs:range rdf:resource="#Region" />
</owl:ObjectProperty>

<owl:Class rdf:ID="VintageYear" />

<owl:DatatypeProperty rdf:ID="yearValue">
  <rdfs:domain rdf:resource="#VintageYear" />
  <rdfs:range rdf:resource="&xsd;positiveInteger" />
</owl:DatatypeProperty>
```

Figure 11.1 – A part of an OWL document with RDF/XML syntax of Wine ontology

```

<node BACKGROUND_COLOR="#aaccff" COLOR="#000000" ID="289" STYLE="bubble" TEXT="Region">
  <font NAME="SansSerif" SIZE="12"/>
  <linktarget COLOR="#CC0000" DESTINATION="289" ID="AR_4" SOURCE="OP_4"/>
  <linktarget COLOR="#CC0000" DESTINATION="289" ID="AR_8" SOURCE="OP_8"/>
  <node COLOR="#CC0000" ID="OP_8" POSITION="left" STYLE="fork" TEXT="adjacentRegion">
    <font NAME="SansSerif" SIZE="8"/>
    <arrowlink COLOR="#CC0000" DESTINATION="289" ENDARROW="Default" ID="AR_8" STARTARROW="None"/>
  </node>
</node>
<node COLOR="#CC0000" ID="OP_4" POSITION="left" STYLE="fork" TEXT="locatedIn">
  <font NAME="SansSerif" SIZE="8"/>
  <arrowlink COLOR="#CC0000" DESTINATION="289" ENDARROW="Default" ID="AR_4" STARTARROW="None"/>
</node>
<node BACKGROUND_COLOR="#aaccff" COLOR="#000000" ID="293" STYLE="bubble" TEXT="VintageYear">
  <font NAME="SansSerif" SIZE="12"/>
  <linktarget COLOR="#CC0000" DESTINATION="293" ID="AR_3" SOURCE="OP_3"/>
</node>

```

Figure 11.2 – A corresponding part of a FreeMind document

current mind-mapping tools. Figure 11.3 and 11.4 are MindMaps which are automatically created from the resulting file by FreeMind and Mindomo tools, respectively.

When end-users choose the option to make different colors for each sub domain instead of each class type of ontologies, the result is showed in FreeMind mind-mapping tool as in Figure 11.5.

11.2 Evaluation

We chose some well-known ontology visualization tools and made comparisons between OWL2MM and them. The first comparison, Table 11.1, is based on elements of an ontology that tools are able to visualize.

Table 11.1 – Visualization capabilities of OWL2MM and other tools

	OWL2MM	OWLviz	OntoGraf	SOVA	OWLGrEd	TGViz	WebVOWL
Classes	x	x	x	x	x	x	x
equivalentClass	x	x	x	x	x	x	x
subClassOf	x	x	x	x	x	x	x
Object properties	x		x	x	x	x	x
Datatype properties	x		x	x	x		x
Individuals	x		x	x	x	x	x
disjointWith				x	x		x
Intersection				x	x		x
Union				x	x		x
Complement				x	x		x

It can be seen from Table 11.1 that WebVOWL, OWLGrEd and SOVA are powerful. However, WebVOWL and SOVA have overlooked visualizations of large ontologies. OWLGrEd uses UML-style notation that is difficult to read and requires a training for domain experts. Our tool, OWL2MM, aims at a

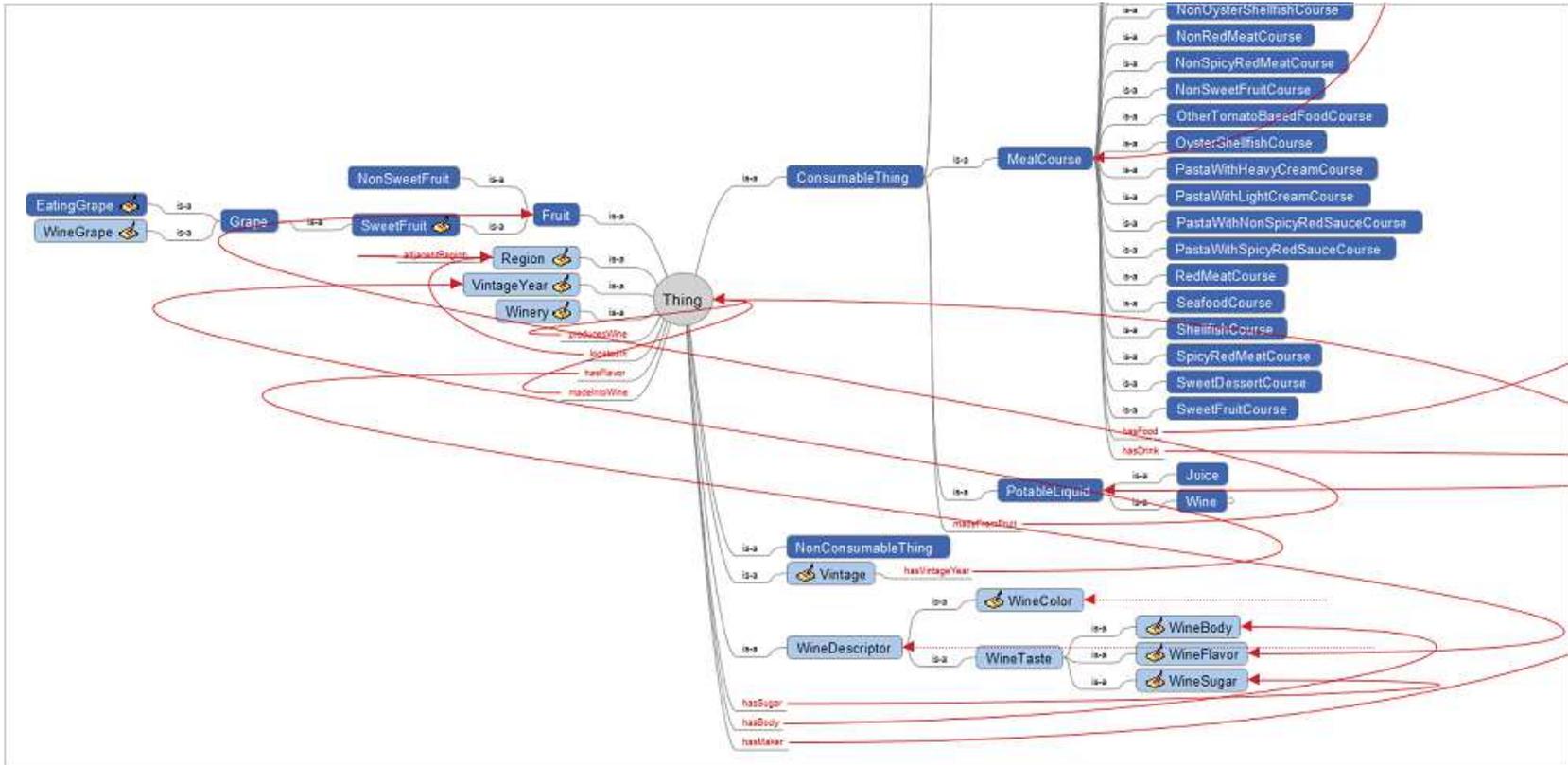


Figure 11.3 – The MindMap visualized by FreeMind tool

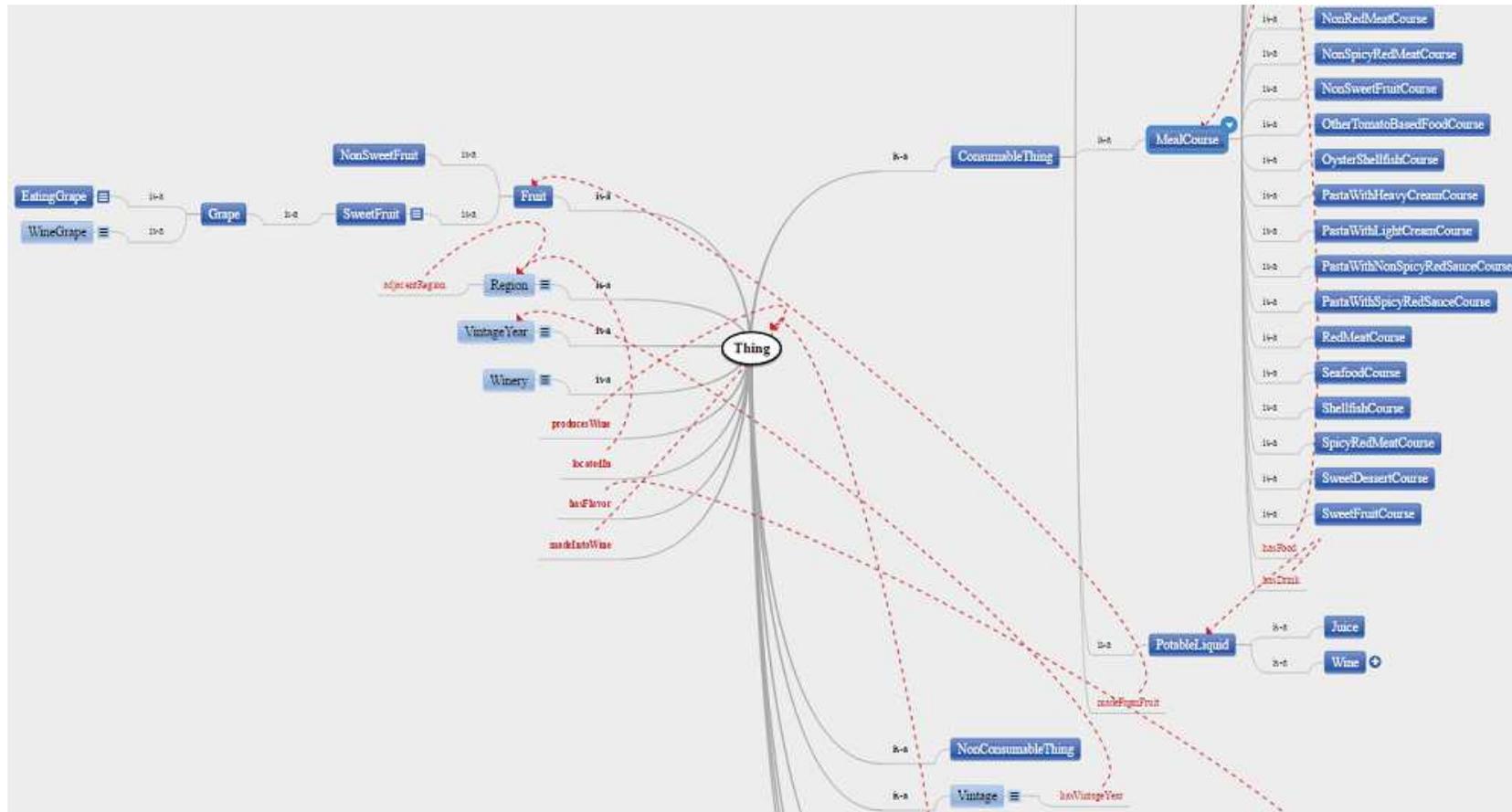


Figure 11.4 – The MindMap visualized by Mindomo tool

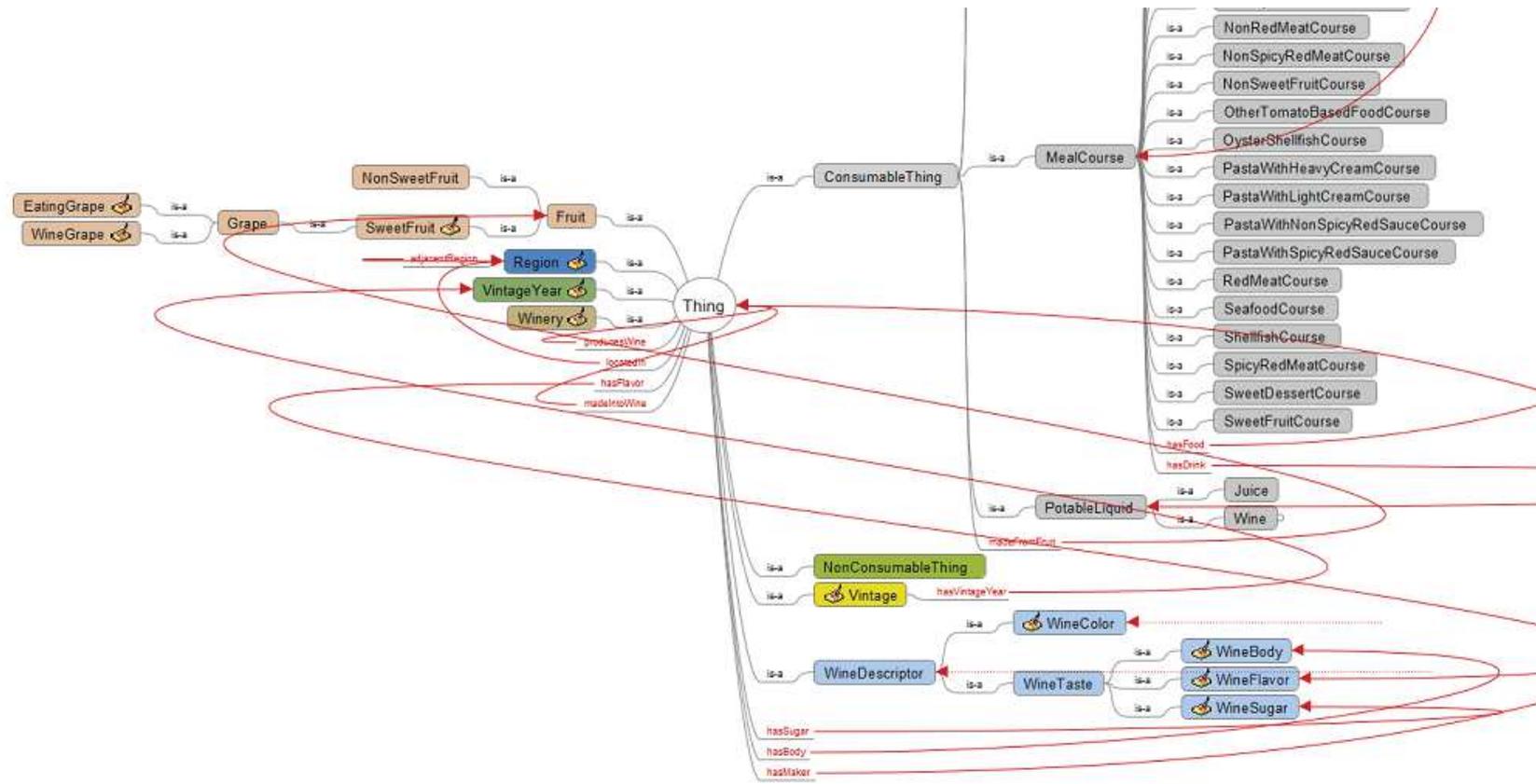


Figure 11.5 – The multi-colored MindMap visualized by FreeMind tool

Table 11.2 – Usability of OWL2MM and other tools

	OWL2MM	OWLviz	OntoGraf	SOVA	OWLGrEd	TGViz	WebVOWL
Colors for class types	x						x
Colors for sub domains	x						
Collapse and expand	x		x				
Drag and drop	x	x	x		x		x
In/Out zoom	x	x	x	x	x	x	x
Search	x		x	x		x	x
Custom layout	x	x	x	x			
In-line editing	x			x	x		
State saving	x			x			

simple visualization for casual users without much knowledge about ontologies. That is why the tool focuses on visualizing class hierarchy, but not complicated constraints as disjointWith, Intersection, Union and so on. OWL2MM also centres on representing classes and object properties on screen, instead of many elements of ontologies. Data properties and individuals are displayed in popup or tooltips of MindMap nodes. They just appear when users hover mouse or click on corresponding nodes. In contrast, OWLGrEd, SOVA and OntoGraf display individuals on screen. This makes the visualization more complicated and confused for the ontologies that have a big number of individuals.

Both OWL2MM and WebVOWL fill different colors for each types of ontology class (internal, external, deprecated). Users can easily realize a class belongs to the ontology (internal class) or is reused from another ontology (external class). Besides, our tool makes different colors based on predefined color list for each sub domain of ontologies. End-users can also choose specific colors for each sub domain as they want.

Another advantage of OWL2MM versus the others is that it allows end-users to show/hide sub classes of a specific class. Because OWL2MM reuse

expand/collapse abilities of existing mind-mapping tools. With this feature, users can easily work with large visualizations.

Conclusion

IN this part, we present a solution to visualize ontologies as MindMaps by transforming OWL documents to FreeMind documents. We also build a tool based on Apache Jena to assist this job. The tool reads OWL files and then converts ontology elements to FreeMind elements which are saved as FreeMind files. The resulting files are imported by most existing mind-mapping tools to create corresponding visualizations as MindMaps.

One of the strengths of ontology visualizations by OWL2MM is that users can customize views of the visualizations by expanding or collapsing ontology parts. This feature is useful when users work with large ontologies. Besides, our tool allows users to choose specific colors for each sub domain of ontologies in order to easily observe each sub domain knowledge. Moreover, users can customize the visualizations by existing functions of mind-mapping tools such as changing layout, editing content and so on.

Re-using features of existing mind-mapping tools helps us to create the visualization of ontologies without building new graphical tools which requires a lot of time and human effort. The ontology visualizations based on MindMaps can be understood by users without much expertise in ontologies.

The ontology used in the experiment of this work is moderate size. However, there is no limitation of the transformation from OWL ontologies to FreeMind documents. The visualization is thus limited by the maximum number of nodes that mind-mapping tools support.

Although OWL2MM mostly meets requirements of an ontology visualization, our tool still has some limitations. It does not visualize complex axioms such as `disjointWith`, `Intersection`, `Union` and so on. We, therefore, continue to improve our tool next time in order to have a stronger ontology visualization tool for end-users, especially domain experts.

Part IV

General Conclusion and Perspectives

13 General Conclusion	113
14 Perspectives	115
14.1 NPI term acquisition improvement	115
14.1.1 Linguistic pattern-based method	115
14.1.2 Rule-based post-filtering step	116
14.2 Term linkage	118
14.3 MindMap-based ontology editor	118

General Conclusion

THE explosion of data on the Internet leads to challenges in working with them. Semantic Web and ontology are required to address those problems. Health care, especially NPI, is receiving more and more consideration from researchers and community. The missions of Plateforme CEPS are thus crucial to contribute in improving quality of life. This thesis contributes a part to support Plateforme CEPS realizing the missions.

As we mentioned in the Part I, NPIs are involved in multidisciplinary domain. For this reason, NPI terms are often confused and not general. Many NPI terms cannot be found in general knowledge base such as WordNet which is a large lexical database of English. Therefore, we proposed two approaches to acquire NPI terms from ontologies in BioPortal and from unstructured texts such as abstracts in PubMed, web pages in Wikipedia and Google. The main idea of those two approaches is an expansion of the seed NPI terms step by step. In other words, we collect new terms that have semantically similarity with the seed NPI terms.

BioPortal is the most comprehensive repository of biomedical ontologies over the world. There are 694 ontologies with 8,841,123 classes stored in BioPortal at the time of this writing. Although there are not any ontologies on NPI domain, but many NPI terms are found in classes of some ontologies in BioPortal. This motivated us to extract candidate NPI terms by searching classes that include the seed NPI terms. From the retrieved classes, we ex-

tracted their synonyms and parent classes as well as their child classes. The synonyms of their parents and children are also extracted as candidate NPI terms.

PubMed comprises more than 28 million citations for biomedical literature from MEDLINE, life science journals, and online books. Most of NPI terms can be found in the PubMed database. In order to collect NPI terms from PubMed, we must address two tasks: first, how to extract candidate terms from unstructured texts in PubMed abstracts; second, how to get only NPI terms and eliminate non-NPI terms. For the first task, we used a state-of-the-art tool BioTex to get a list of biomedical terms. Those candidates were filtered by comparing to the seed NPI terms using a semantic similarity measure in the second task. Three types of similarity measure including edit-based, knowledge-based and distributional ones were introduced and compared. Among them, the distributional similarity measure gained more good results over the remaining in context of NPI.

The candidate NPI terms retrieved from the above approach were manually validated by NPI experts in the final step to get the official list of NPI terms. This list is be used to construct an ontology for NPI.

In the part II, we introduced a method to visualize an ontology as a MindMap. This method provides a simple view of ontology to domain experts who do not have much knowledge on computer science as well as ontology engineering. To fulfill the objectives of this part, we first proposed a notation based on MindMap for ontology visualization. Each component of an ontology is represent by an element or a combination of elements of a MindMap. Finally, we built a tool to transform an OWL ontology to a FreeMind document which is then imported by existing Mind-Mapping tools to make a visualization of the ontology as a MindMap.

Perspectives

IN the ontology development process, ontology engineers and domain experts need to have a collaborative framework to work together. Therefore, a global approach which integrate different phases in ontology development is essential. It should include modules for preparing data, designing and implementing as well as visualizing the ontology being built. Discussion and version management functions need to be taken into account because ontology development is often a collaborative work between different experts. It is a motivation for our future work to develop a tool integrating the modules into a framework to help experts in building ontologies.

Besides, we find that our contributions have some limitations. We thus present some works that remains to be done in the following.

14.1 NPI term acquisition improvement

14.1.1 Linguistic pattern-based method

Our current method highly depends on the quality of input documents. However, we cannot know which documents contain rich NPI term candidates. In theory, taking the whole NPI text corpus as input might solve this problem. But in practice, it is not possible because this will take much time for extracting and ranking term steps. Therefore, we plan to use a linguistic patterns

to directly acquire NPI terms from the text corpus. The intuitive idea can be realized in the following steps:

- Step 1 (POS tagging text corpus): For a start, a NPI relevant text corpus is gathered. Then we perform POS tag the whole corpus.
- Step 2 (Noun phrase recognition): A simple chunking technique is used in this step to combine adjectives and nouns in order to make noun phrases.
- Step 3 (Candidate extraction): This step aims at extracting every pair (NP_1, NP_2) that appears together in the text corpus same as Hearst patterns [28] such as: NP_1 also known as NP_2 ; NP_1 aka NP_2 ; NP_1 also called NP_2 ; NP_1 which is called NP_2 ; NP_1 which is named NP_2 ; NP_1 or NP_2 .
- Step 4 (Term retrieval): Given a term NP_1 , which is a noun phrase, its semantically related term is NP_2 if NP_1 and NP_2 appears together more than n times in the candidate pairs retrieved in the previous step. Here n is a threshold determined by users.

14.1.2 Rule-based post-filtering step

According to the NPI experts at the Platforme CEPS, if NPI terms are found in the UMLS metathesaurus, they often have the UMLS semantic types related to NPIs as listed in Table 14.1. Therefore, we plan to use simple rules to enhance the accuracy of our current method. This proposal is used along with the linguistic pattern-based approach above and our current method as a post-filtering step. The underlying idea can be expressed as follows:

- Firstly, candidate NPI terms are retrieved by our current method or the linguistic pattern-based approach;

- Next, we check every candidate NPI term in the UMLS metathesaurus. If a matching term is found and it has one of the UMLS semantic types as in Table 14.1, then a weight α is added to the existing similarity score of the corresponding candidate NPI term. This means that those candidate terms have higher capability to be NPI terms. The weight α is determined by users through experiments.

Table 14.1 – The UMLS semantic types related to NPIs

#	Semantic Type	#	Semantic Type
1	Activity	23	Mental or Behavioral Dysfunction
2	Behavior	24	Natural Phenomenon or Process
3	Biologic Function	25	Occupation or Discipline
4	Biologically Active Substance	26	Occupational Activity
5	Biomedical Occupation or Discipline	27	Organism
6	Body System	28	Organism Attribute
7	Daily or Recreational Activity	29	Organism Function
8	Disease or Syndrome	30	Organization
9	Educational Activity	31	Phenomenon or Process
10	Environmental Effect of Humans	32	Physical Object
11	Food	33	Physiologic Function
12	Fungus	34	Plant
13	Health Care Activity	35	Professional or Occupational Group
14	Health Care Related Organization	36	Qualitative Concept
15	Human	37	Quantitative Concept
16	Human-caused Phenomenon or Process	38	Research Activity
17	Idea or Concept	39	Research Device
18	Individual Behavior	40	Self-help or Relief Organization
19	Intellectual Product	41	Sign or Symptom
20	Manufactured Object	42	Social Behavior
21	Medical Device	43	Therapeutic or Preventive Procedure
22	Mental Process	44	Vitamin

14.2 Term linkage

Although in our current method, the potential positions are suggested for new NPI terms, however, the suggestions are not correct in some cases. For example, in case there are more than one position having the equivalent similarity score. Therefore, the work to assign a new NPI term to a correct position in the NPI taxonomy need to be improved. We plan to propose a method that considers the context of target terms. More specifically, we will rely on the similarity between the candidate term and not only the target term but also its child nodes and parent nodes in the NPI taxonomy. Furthermore, we want to enrich the NPI taxonomy with relationship and axiom layers in order to make a full ontology for NPIs.

14.3 MindMap-based ontology editor

Our current MindMap-based visualization tool for ontologies does not support users to save their modification on the visualizations to OWL format. The changes of ontologies must be performed in ontology editors. Therefore, an ontology visualization tool should be upgraded to become an ontology visualization editor allowing users to modify and save the visualization to OWL file. As our knowledge, OWLGrEd desktop application has already supported visualizing and editing ontologies by visual graphical interaction [4, 5]. However, this tool uses UML style to visualize ontology elements, as a result it is complex for domain experts. Our future work aims at developing a simple and friendly MindMap-based ontology editor for domain experts.

Appendices

NonPhaTex Web-based Application

In order to help NPI experts in the Plateforme CEPS in validating potential NPI terms, we built a web-based application named NonPhaTex as illustrated in Figure A.1. The web application is developed by Java, Bootstrap ¹ and JQuery ², while its database is stored in MySQL. This tool integrates our works on term acquisition and ontology visualization presented in the main content of this thesis. It brings to users a convenient visual interface with some functions:

1. The first function, on the left of the application interface, shows the NPI taxonomy. It allows users to collapse and expand nodes to upper or lower levels. Besides, it provides a search function to help users to quickly navigate to specific terms as they want.
2. The second function, in the center of the interface, lists existing terms related to the current selected term on the left.
3. The third function, on the right of the interface, displays potential NPI terms associated to the current selected terms on the left. Those candi-

¹<https://getbootstrap.com/>

²<https://jquery.com/>

dates will be evaluated by NPI experts by clicking on Yes, No, Maybe buttons.

4. The fourth function, above the first function, allows users to export the current taxonomy along with the existing related terms to OWL and XML file in order to continue working with the terms on WebProtege and Excel applications.

NPI taxonomy Export as **4** **Existing related terms** **Candidate terms**

Type to search...

- Non-Pharmacological Intervention
 - Psychological Health Intervention
 - Art Therapy
 - Health Education
 - Psychotherapy
 - Zootherapy
 - Physical Health Interventions
 - Physical Activity
 - Hortitherapy
 - 1** **Physiotherapy**
 - Manual Therapy
 - Thermalism
 - Nutritional Health Interventions
 - Dietary Supplements
 - Nutritional Therapy
 - Digital Health Interventions
 - eHealth Devices
 - Therapeutic Games
 - Virtual Reality Therapy
 - Other Health Interventions

botanical therapy			
cardiac rehabilitation			
comprehensive rehabilitation			
medicinal herb			
physical rehabilitation			
2 phytotherapy or herbalism			
plant extract			
pulmonary rehabilitation			
rehabilitation intervention			
rehabilitation program			
rehabilitation programmes			

speech therapy	Yes	No	Maybe
podiatry	Yes	No	Maybe
chiroprody	Yes	No	Maybe
radiography	Yes	No	Maybe
dietetics	Yes	No	Maybe
social work	Yes	No	Maybe
nursing	Yes	No	Maybe
3 midwifery	Yes	No	Maybe
optometry	Yes	No	Maybe
hospital treatment	Yes	No	Maybe
dentistry	Yes	No	Maybe
visiting	Yes	No	Maybe
pharmacy	Yes	No	Maybe
therapy	Yes	No	Maybe
clinical psychology	Yes	No	Maybe
counselling	Yes	No	Maybe
radiology	Yes	No	Maybe
support services	Yes	No	Maybe

Figure A.1 – NonPhaTex main interface

Retrieved NPI terms

Up to this time, we achieved the list of NPI terms as follows.

#	NPI term	#	NPI term
1	Acai	325	La'au lapa'au
2	Acai berry	326	Lactoferrin
3	Acceptance and commitment therapy	327	Laetrile
4	ACT	328	L-Arginine
5	Actimarch program	329	Lavender
6	Actinotherapy	330	Lemon BalmLemon
7	Acupressure	331	Licorice root
8	Acupuncture	332	Lifestyle intervention
9	Acupuncture appointments	333	Light therapy
10	Acupuncture treatment	334	Limitation diet
11	Adapted physical activity	335	Lithotherapy
12	Adjunct treatment	336	Lobelia
13	Adjunctive intervention	337	L-Tryptophan
14	Adjuvant intervention	338	Lutein
15	Adjuvant therapy	339	Lycium
16	Advitha kriya-dhanwantraï prakriya	340	Lycopene
17	Aerobic activity	341	Macrobiotic lifestyle
18	Aerobic exercise	342	Magnesium
19	Aerobic resistance	343	Magnet therapy

20	Aetherolea	344	Magnetic healing
21	Alexander technique	345	Magnets
22	Alfalfa	346	Management health intervention
23	Aloe vera	347	Manganese
24	Alpha lipoic acid	348	Manipulative therapy
25	Alpha-linolenic Acid	349	Manual lymphatic drainage
26	Alternative care	350	Manual therapy
27	Alternative medicine	351	Marijuana
28	Alternative medicine therapy	352	Marijuana plant
29	Alternative therapeutic	353	Marijuana use
30	Alternative therapy	354	Massage
31	Alternative treatment	355	Massage therapy
32	Amino acids	356	Mayo diet
33	Amputee rehabilitation	357	MBSR
34	Amygdalin	358	Medical acupuncture
35	Antidepressant therapy	359	Medical care
36	Antidepressant treatment	360	Medical devices
37	Antioxidants	361	Medicinal herb
38	Anti-Wrinkle Treatment	362	Medicinal mushroom
39	APAD program	363	Medicinal mushrooms
40	Apitherapy	364	Medicinal plant
41	Aristolochic Acid	365	Meditation
42	Aroma therapy	366	Mediterranean diet
43	Aromatherapy	367	Mega-vitamin therapy
44	Art therapy	368	Megavitamins
45	Arts-based health approach	369	Melaleuca Oil
46	Ascorbic acid/vitamin C	370	Melatonin
47	Ashtanga vinyasa yoga	371	Memory training
48	Ashtanga yoga	372	Methylsulfonylmethane

49	Atkins diet	373	Meziere method
50	Attachment therapy	374	Mhealth
51	Auricular acupuncture	375	Milk thistle
52	Auriculotherapy	376	Mind and body practice
53	Autogenic training	377	Mind body therapy
54	Autosuggestion	378	Mind-body intervention
55	Avocado-soybean unsaponifiables	379	Mind-body medicine
56	Ayurveda	380	Mind-body practice
57	Ayurvedic medicine	381	Mind-body therapy
58	Bach flower therapy	382	Mindfulness meditation
59	Balneotherapy	383	Mindfulness-Based Stress Reduction
60	Bates method	384	Mineral
61	Behavior conditioning therapy	385	Mistletoe
62	Behavior therapy	386	Mobile health
63	Behavioral approach	387	MorindaOmega-3 Fatty Acids
64	Behavioral intervention	388	Motivational intervention
65	Behavioral medicine	389	Motivational intervention
66	Behavioral psychotherapy	390	MSM
67	Behavioral therapy	391	Multi-component intervention
68	Behavioral treatment	392	Multimodal therapy
69	behaviour conditioning therapy	393	Multimodal treatment
70	Behaviour therapy	394	Multimodality therapy
71	Behavioural intervention	395	Multimodality treatment
72	Behavioural support	396	Multivitamins
73	Behavioural therapy	397	Musculoskeletal manipulations
74	Belladonna	398	Music intervention
75	Bikram yoga	399	Music therapy
76	Bilberry	400	Nasal irrigation

77	Biodanza	401	Natural antimicrobial
78	Biofeedback procedure	402	Natural health
79	Biofeedback technique	403	Natural medicine
80	Biofeedback therapy	404	Natural therapy
81	Bioresonance therapy	405	Naturopathic medicine
82	Biotin	406	Naturopathic therapy
83	Bitter orange	407	Naturopathy
84	Black cohosh	408	Naturopathy therapy
85	Black psyllium	409	Neurofeedback
86	Black tea	410	Neuro-linguistic programming
87	Bladderwrack	411	Neuropsychotherapeutic intervention
88	Blepharocalyxin	412	Niacin Vitamin B3
89	Blessed thistle	413	Non pharmacological intervention
90	Blond psyllium	414	Non-conventional therapy
91	Blood irradiation therapies	415	Noni
92	Blueberry	416	Nonpharm intervention
93	Blue-Green Algae	417	Non-pharm intervention
94	Body intervention	418	Non-pharmacologic intervention
95	Body work	419	Nonpharmacological approach
96	Body-based manipulative therapy	420	Non-pharmacological approach
97	Bodybuilding	421	Nonpharmacological intervention
98	Boron	422	Non-pharmacological intervention
99	Boswellia	423	Nonpharmacological method
100	Botanical therapy	424	Non-pharmacological strategy
101	Botanicals	425	Nonpharmacological therapy
102	Breathing exercise	426	Non-pharmacological therapy
103	Bromelain	427	Nonpharmacological treatment
104	Butterbur	428	Non-pharmacological treatment

105	Calcium	429	Nutritional healing
106	Calendula	430	Nutritional health intervention
107	CAM	431	Nutritional supplements
108	Cannabinoids	432	Nutritional therapy
109	Cardiac rehabilitation	433	Occupational health
110	Care medicine	434	Occupational therapy
111	Carnitine	435	Omega-6 Fatty Acids
112	Carotenoids	436	Oral hygiene
113	CBT	437	Orgonomy
114	Cessation method	438	Orthomolecular medicine
115	Cessation program	439	Osteomyology
116	Chamomile	440	Osteopathic manipulative treatment
117	Change management	441	Osteopathy
118	Chasteberry	442	Otago program
119	Chi exercise	443	Other NPIs
120	Chinese food therapy	444	Palliative care
121	Chinese herb	445	Palliative therapy
122	Chinese herbal medicine	446	Palliative treatment
123	Chinese martial arts	447	Palmitic acid
124	Chinese medicine	448	Panax ginseng
125	Chiropractic	449	Pantothenic acid
126	Chondroitin	450	Passionflower
127	Chromium	451	PEM-ES Program
128	Chromium diet	452	Pennyroyal
129	Chromium picolinate	453	Peppermint
130	Chromotherapy	454	Peppermint oil
131	Cinnamon	455	Pet therapy
132	Clove	456	Pharmaceutical plant

133	Cocoa	457	Phosphate Salts
134	Cognitive approach	458	Photopheresis
135	Cognitive assessment	459	Photoradiation therapy
136	Cognitive behavior modification	460	Phototherapy
137	Cognitive behavior therapy	461	Physical activity
138	Cognitive behavioral therapy	462	Physical conditioning
139	Cognitive behaviour therapy	463	Physical exercise
140	Cognitive behavioural therapy	464	Physical health intervention
141	Cognitive intervention	465	Physical rehabilitation
142	Cognitive psychotherapy	466	Physical therapy
143	Cognitive rehabilitation	467	Physical therapy procedure
144	Cognitive stimulation therapy	468	Physiotherapy
145	Cognitive therapy	469	Physiotherapy procedure
146	Cognitive therapy approach	470	Phytoestrogens
147	Cognitive-behavioral therapy	471	Phytotherapy
148	Cognitive-behavioral therapy	472	Pilates
149	Cognitive-behavioural therapy	473	Placebo product
150	Cohen diet	474	Plant estrogens
151	Colloidal silver	475	Plant extract
152	Colloidal silver therapy	476	Play therapy
153	Colon hydrotherapy	477	Polarity therapy
154	Color therapy	478	Policosonal
155	Combination therapy	479	Polyphenols
156	Combination treatment	480	Pomegranate
157	combined modality treatment	481	Power yoga
158	combined treatment	482	Pranic healing
159	combined treatment modalities	483	Prayer therapy
160	Comfort care	484	Prenatal care
161	Comfort measure	485	Pressure therapy

162	Comfort therapy	486	Prevention health strategy
163	Complementary and alternative medicine	487	Primary prevention action
164	Complementary approach	488	Probiotics
165	Complementary food	489	Problem-adaptation therapy
166	Complementary medicine	490	Progressive relaxation
167	Complementary therapy	491	Prophylaxis care
168	Complementary treatment	492	Propolis
169	Comprehensive rehabilitation	493	Protein Supplements
170	Computer-delivered intervention	494	Psychoanalysis cure
171	Concurrent therapy	495	Psychoeducative health intervention
172	Conditioning therapy	496	Psychological health intervention
173	Control diet	497	Psychological therapy
174	Cosmeceutical Peels	498	Psychology training
175	Cosmetic therapy	499	Psychosocial interventions
176	Cranberry	500	Psychotherapeutic technique
177	Craniosacral therapy	501	Psychotherapy
178	Creatine	502	Pulmonary rehabilitation
179	Crystal healing	503	Pycnogenol
180	Cupping	504	Qi Gong
181	Curcumin	505	Qigong
182	Cybertherapy	506	Quitting smoking
183	Dance movement therapy	507	Radiation therapy
184	Dance therapy	508	Radionics
185	Dandelion	509	Rational emotive therapy
186	Decision support	510	Rebirthing
187	Decision-making approach	511	Red clover
188	Deep Breathing	512	Red yeast rice
189	Detoxification	513	Reflexology

190	Device therapy	514	Reflexology intervention
191	Devil's Claw	515	Regulation therapy
192	Diet	516	Rehabilitation care
193	Dietary supplement	517	Rehabilitation exercise
194	Dietary supplementation	518	Rehabilitation intervention
195	Digital health Interventions	519	Rehabilitation program
196	Dimethylamylamine	520	Reiki
197	Distilled oil	521	Reinforcement training
198	DMAA	522	Relaxation technique
199	DMSO	523	Relaxation therapy
200	Double Chin Treatment	524	Relaxation training
201	Dowsing	525	Reminiscence therapy
202	Dukan Diet	526	Renshen
203	Echinacea	527	Replacement therapy
204	Echium Oil	528	Resistance exercise
205	E-cigarette	529	Restriction diet
206	Ecopsychosocial interventions	530	Resveratrol
207	Educ health	531	Rhodiola
208	Ehealth	532	Riboflavin Vitamin B2
209	E-health	533	S-Adenosyl-L-methionine
210	Ehealth device	534	SAMe
211	E-health device	535	Saw palmetto
212	Ehealth intervention	536	Schisandra
213	Elderberry	537	Secondary prevention action
214	Electroacupuncture	538	Seitai
215	Electrohomeopathy	539	Selenium
216	Electromagnetic therapy	540	Self-hypnosis
217	Electronic cigarette	541	Self-management health intervention

218	EMDR	542	Self-management health strategy
219	End-of-life care	543	Senna
220	Energy drinks	544	Serious game
221	Energy medicine	545	Shiatsu
222	Energy therapt	546	Shingles
223	Ephedra	547	Siddha medicine
224	Ephedrine Alkaloids	548	Siddha yoga
225	Ergonomic tool	549	Silibinin
226	Essential oil	550	Silver
227	Ethereal oil	551	Silver Acetate
228	Eucalyptus	552	Silymarin
229	European Elder	553	Sivananda yoga
230	European Mistletoe	554	Skin healing
231	Evening Primrose oil	555	Skin Needling
232	Exercise	556	Smoking cessation
233	Exercise intervention	557	Smoking cessation method
234	Exercise program	558	Smoking cessation Ottawa program
235	Exercise therapy	559	Sonopuncture
236	Exercise training	560	Sophrology
237	Eye movement desensitization and reprocessing	561	Sound therapy
238	Fasting	562	Soy
239	Feldenkrais method	563	Spinal manipulation
240	Fenugreek	564	Spiritual mind treatment
241	Feverfew	565	St. John's Wort
242	Fish oil	566	Steroids
243	Flaxseed oil	567	Stimulation therapy
244	Flower essence therapy	568	Structural Integration
245	FolateGarlic	569	Support groups

246	Folic acid	570	Supportive care
247	Gait therapy	571	Swedish Massage
248	Gamma linolenic acid	572	Tai Chi
249	Gamma-tocopherol	573	Tai-chi exercise
250	Garden therapy	574	Talk therapy
251	Gdm prevention	575	Tantric yoga
252	German new medicine	576	Taurine
253	Ginger	577	Tea
254	Ginkgo biloba	578	Tea tree oil
255	Ginseng	579	Technological aid
256	Ginseng intervention	580	Telecare
257	Ginseng treatment	581	Telehealth
258	Glucosamine	582	Telemedicine
259	Glycyrrhizin	583	Tertiary prevention action
260	Goldenseal	584	Thai massage
261	Gotu Kola	585	Thalassotherapy
262	Grape seed extract	586	Therapeutic exercise
263	Grapefruit	587	Therapeutic game
264	Green tea	588	Therapeutic horseback riding
265	Group psychotherapy	589	Therapeutic touch
266	Group therapy	590	Therapy support
267	Gua sha	591	Thermalism
268	Guided imagery	592	Thermalism care
269	Handling device	593	Thiamine
270	Hatha Yoga	594	Thunder god vine
271	Hawaiian massage	595	Thymus Extract
272	Healing interventions	596	Tobacco cessation
273	Healing plant	597	Tobacco use cessation
274	Health care	598	Traditional Chinese Medicine

275	Health claim	599	Traditional healing
276	Health education	600	Traditional Japanese medicine
277	Health prevention	601	Traditional Korean medicine
278	Health prevention action	602	Traditional medicine
279	Healthcare	603	Traditional Mongolian medicine
280	Herb therapy	604	Traditional Tibetan medicine
281	Herbal medicine	605	Traditional Vietnamese medicine
282	Herbal tea	606	Trager approach
283	Herbal therapy	607	Transcendental Meditation
284	Herbal viagra	608	Trigger point
285	Herbalism	609	Tui na
286	Herbology	610	Turmeric
287	Herbs	611	Unani medicine
288	High-calorie diet	612	Unconventional medicine
289	Hippotherapy	613	Urine therapy
290	Holistic medicine	614	Valerian
291	Home remedies	615	Video-music therapy
292	Homeopathy	616	Viniyoga
293	Honey	617	Vinyasa yoga
294	Hoodia	618	Vipassana
295	Hops	619	Virtual reality exposure therapy
296	Hormone replacement therapy	620	Virtual reality immersion therapy
297	Horse therapy	621	Virtual reality therapy
298	Hortitherapy	622	Visualization
299	humor therapy	623	Vitamin A
300	Hydrotherapy	624	Vitamin B
301	Hygiene care	625	Vitamin B1
302	Hyperbaric oxygen therapy	626	Vitamin B12
303	Hypericum	627	Vitamin B5

304	Hypnosis	628	Vitamin B6
305	Hypnotherapy	629	Vitamin D
306	Impairment intervention	630	Vitamin E
307	Integrated intervention	631	Vitamin K
308	Integrative and comprehensive care	632	Vitex
309	Integrative and comprehensive health intervention	633	Volatile oil
310	Integrative medicine	634	Water cure
311	Intensive lifestyle intervention	635	Wave therapy
312	Iodine	636	Weight Control
313	Iron	637	Weight management
314	Isometric exercise	638	White Tea
315	Isopathy	639	Wild Yam
316	Iyengar yoga	640	Yoga
317	Jen shen	641	Yoga intervention
318	Kampo	642	Yoga meditation
319	Kava	643	Yoga session
320	Kg diet	644	Yoga therapy
321	Kinesitherapy	645	Yohimbe
322	Knee rehabilitation	646	Zeaxanthin
323	Kundalini yoga	647	Zinc
324	Laau lapaau	648	Zootherapy

Table B.1 – List of current existing NPI terms

Bibliography

- [1] Zeeshan Ahmed, Detlef Gerhard, A Zeeshan, and D Gerhard. Role of Ontology in Semantic Web Development. volume 11, pages 2007–15, 2007.
- [2] Harith Alani. TGVizTab: An Ontology Visualisation Extension for Protégé. In *Knowledge Capture (K-Cap'03), Workshop on Visualization Information in Knowledge Engineering*, pages 3–9, 2003.
- [3] R Aravindhan and Mano Chitra M. A Review on Ontology Based Search Engine. *International Journal of Advanced Research in Computer and Communication Engineering*, 3(10):8232–8236, 2014.
- [4] Janis Barzdins, Guntis Barzdins, Karlis Cerans, Renars Liepins, and Arturs Sprogis. UML Style Graphical Notation and Editor for OWL 2. In *CEUR Workshop Proceedings*, volume 596, pages 102–114. 2010. ISBN 3642161006. doi: 10.1007/978-3-642-16101-8_9.
- [5] Janis Barzdins, Guntis Barzdins, Karlis Cerans, Renars Liepins, Arturs Sprogis, and Lumilv RenarsLiepins. OWLGrEd: a UML Style Graphical Editor for OWL. In *Ontology Repositories and Editors for the Semantic Web*, 2010.
- [6] Janis Barzdins, Karlis Cerans, Renars Liepins, and Arturs Sprogis. Advanced ontology visualization with OWLGrEd. In *CEUR Workshop Proceedings*, volume 796, 2011. doi: 10.3233/978-1-61499-161-8-41.
- [7] Eran Ben-Arye, Bella Shulman, Yael Eilon, Rachel Weitiz, Victoria Cherniak, Ilanit Shalom Sharabi, Osnat Sher, Hiba Reches, Yfat Katz, Michal Arad, Elad Schiff, Noah Samuels, Ofer Caspi, Shahr Lev-Ari, Moshe Frenkel, Abed Agbarya, and Hana Admi. Attitudes Among Nurses Toward the Integration of Complementary Medicine Into Supportive Cancer

- Care. *Oncology Nursing Forum*, 44(4):428–434, 2017. ISSN 0190-535X. doi: 10.1188/17.ONF.428-434.
- [8] Yaakov HaCohen-KernerAsaf ApplebaumJacob Bitterman. Experiments with Language Models for Word Completion and Prediction in Hebrew. *Advances in Natural Language Processing*, NA(SEPTEMBER 2014):450–462, 2014. ISSN 16113349. doi: 10.1007/978-3-319-10888-9_44.
- [9] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(90001):267D–270, jan 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh061.
- [10] Willem Nico Borst. *Construction of engineering ontologies for knowledge sharing and reuse*. PhD thesis, sep 1997.
- [11] R.L. Cilibrasi and P.M.B. Vitanyi. The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, mar 2007. ISSN 1041-4347. doi: 10.1109/TKDE.2007.48.
- [12] CLIR. Knowledge Organization Systems: An Overview. URL <https://www.clir.org/pubs/reports/pub91/1knowledge/>.
- [13] John Davies. Lightweight Ontologies. In *Theory and Applications of Ontology: Computer Applications*, pages 197–229. Springer Netherlands, Dordrecht, 2010. ISBN 9789048188468. doi: 10.1007/978-90-481-8847-5_9.
- [14] Antonio De Nicola, Michele Missikoff, and Roberto Navigli. A software engineering approach to ontology building. *Information Systems*, 34(2): 258–275, 2009. ISSN 03064379. doi: 10.1016/j.is.2008.07.002.
- [15] Deborah L. McGuinness and Frank van Harmelen. OWL Web Ontology Language Overview, 2004. URL <https://www.w3.org/TR/owl-features/>.

- [16] Biswanath Dutta. Examining the interrelatedness between ontologies and Linked Data. *Library Hi Tech*, 35(2):312–331, jun 2017. ISSN 0737-8831. doi: 10.1108/LHT-10-2016-0107.
- [17] EuroCAM. What is CAM? URL <http://www.cam-europe.eu/cam-definition.php>.
- [18] M Fernández-López, A Gómez-Pérez, and Natalia Juristo. METHONTOLOGY: From Ontological Art Towards Ontological Engineering. *AAAI-97 Spring Symposium Series*, SS-97-06:33–40, 1997. doi: 10.1109/AXMEDIS.2007.19.
- [19] Asuncion Gomez-perez Figueroa and Mari Carmen Suarez. Neon methodology for building ontology networks: a Scenario- Based Methodology. *Demetra EOOD*, (February):1–18, 2009. ISSN 01692046. doi: 10.1016/j.landurbplan.2011.04.007.
- [20] Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130, aug 2000. ISSN 1432-5012. doi: 10.1007/s007999900023.
- [21] Moshe Frenkel, Eran Ben-Arye, and Lorenzo Cohen. Communication in Cancer Care: Discussing Complementary and Alternative Medicine. *Integrative Cancer Therapies*, 9(2):177–185, 2010. ISSN 15347354. doi: 10.1177/1534735410363706.
- [22] A. Gérazine, T.-L. Nguyen, F. Carbonnel, E. Guerdoux-Ninot, and G. Ninot. Ontologie des interventions non médicamenteuses. *Revue d'Épidémiologie et de Santé Publique*, 66:S42, mar 2018. ISSN 03987620. doi: 10.1016/j.respe.2018.01.093.

- [23] Thomas R Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, jun 1993. ISSN 10428143. doi: 10.1006/knac.1993.1008.
- [24] Thomas R Gruber. Ontology. In *Encyclopedia of Database Systems*, pages 1963–1965. Springer US, Boston, MA, 2009. doi: 10.1007/978-0-387-39940-9_1318.
- [25] Nicola Guarino and Daniel Oberle. What is An Ontology. In *Handbook on Ontologies*, pages 1–17. 2009. ISBN 978-3-540-70999-2. doi: 10.1007/978-3-540-92673-3.
- [26] Lushan Han, Abhay Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. UMBC_EBIQUITY-CORE: Semantic Textual Similarity Systems. *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*, 1:44–52, 2013.
- [27] Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. *Semantic Similarity from Natural Language and Ontology Analysis*, volume 8. may 2015. ISBN 9781627054461. doi: 10.2200/S00639ED1V01Y201504HLT027.
- [28] Marti A Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics* -, volume 2, page 539, Morristown, NJ, USA, 1992. Association for Computational Linguistics. doi: 10.3115/992133.992154.
- [29] Matthew Horridge. OWLViz. URL <https://protegewiki.stanford.edu/wiki/OWLViz>.
- [30] InfoGrid. What are the differences between a vocabulary, a taxonomy, a thesaurus, an ontology, and a meta-model? URL <http://infogrid.org/trac/wiki/Reference/PidcockArticle>.

- [31] Institute of Medicine of the National Academies. *Complementary and Alternative Medicine In the United States*. 2005. ISBN 0309092701.
- [32] Simon Jupp. Simple Knowledge Organisation System (SKOS), jan 2010. URL <http://ontogenesis.knowledgeblog.org/240>.
- [33] Abhay Kashyap, Lushan Han, Roberto Yus, Jennifer Sleeman, Taneeya Satyapanich, Sunil Gandhi, and Tim Finin. Robust semantic text similarity using LSA, machine learning, and linguistic resources. *Language Resources and Evaluation*, 50(1):125–161, mar 2016. ISSN 1574-020X. doi: 10.1007/s10579-015-9319-2.
- [34] Rakesh Kumar, P K Suri, and R K Chauhan. Search Engines Evaluation. *DESIDOC Bulletin of Information Technology*, 25(2):3–10, 2005. ISSN 09714383.
- [35] Piotr Kunowski and Tomasz Boinski. SOVA. URL <https://protegewiki.stanford.edu/wiki/SOVA>.
- [36] D A Lindberg, B L Humphreys, and A T McCray. The Unified Medical Language System. *Methods Archive*, 32:281–291, 1993. ISSN 0026-1270.
- [37] Linked Data Tools. Introducing RDFS & OWL. URL <http://www.linkeddatatools.com/introducing-rdfs-owl>.
- [38] Steffen Lohmann, Stefan Negru, and David Bold. The ProtégéVOWL Plugin: Ontology Visualization for Everyone. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8798, pages 395–400. 2014. ISBN 9783319119540. doi: 10.1007/978-3-319-11955-7_55.
- [39] Steffen Lohmann, Stefan Negru, Florian Haag, and Thomas Ertl. VOWL 2: User-Oriented Visualization of Ontologies. In *Knowledge Engineering and Knowledge Management - 19th International Conference*,

- {EKAW} 2014, Linköping, Sweden, November 24-28, 2014. *Proceedings*, volume 8876, pages 266–281. 2014. ISBN 978-3-319-13703-2. doi: 10.1007/978-3-319-13704-9_21.
- [40] Steffen Lohmann, Vincent Link, Eduard Marbach, and Stefan Negru. WebVOWL: Web-based Visualization of Ontologies. pages 154–158. 2015. doi: 10.1007/978-3-319-17966-7_21.
- [41] Steffen Lohmann, Stefan Negru, Florian Haag, and Thomas Ertl. Visualizing ontologies with VOWL. *Semantic Web*, 7(4):399–419, may 2016. ISSN 22104968. doi: 10.3233/SW-150200.
- [42] Juan Antonio Lossio-Ventura, Clement Jonquet, and Mathieu Roche. Yet Another Ranking Function for Automatic Multiword Term Extraction. In *LNCS*, number 8686, pages 52–64, 2014. ISBN 9783319108889. doi: 10.1007/978-3-319-10888-9.
- [43] Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. Integration of linguistic and web information to improve biomedical terminology extraction. In *Proceedings of the 18th International Database Engineering & Applications Symposium on - IDEAS '14*, pages 265–269, New York, New York, USA, 2014. ACM Press. ISBN 9781450326278. doi: 10.1145/2628194.2628208.
- [44] Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. Biomedical Terminology Extraction: A new combination of Statistical and Web Mining Approaches. In *12th International Workshop on Statistical Analysis of Textual Data, JADT'14*, number i, pages 421–432, 2014.
- [45] Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. BioTex: A system for biomedical terminology ex-

- traction, ranking, and validation. In *CEUR Workshop Proceedings*, volume 1272, pages 157–160, 2014.
- [46] Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal*, 19(1-2):59–99, apr 2016. ISSN 1386-4564. doi: 10.1007/s10791-015-9262-2.
- [47] Christopher D. Manning, Prabhakar Ragahvan, and Hinrich Schutze. *An Introduction to Information Retrieval*. Number c. 2009. ISBN 0521865719. doi: 10.1109/LPT.2009.2020494.
- [48] Michael McIntyre. The Regulation of Complementary and Alternative Medicine (CAM) in the EU. URL <https://www.srab.dk/media/1130/cam-regulation-in-europe.pdf>.
- [49] J. Melorose, R. Perroy, and S. Careas. CAM 2020. *Statewide Agricultural Land Use Baseline 2015*, 1, 2015. ISSN 1098-6596. doi: 10.1017/CBO9781107415324.004.
- [50] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *Arxiv*, (October): 1–12, 2013. ISSN 15324435. doi: 10.1162/153244303322533223.
- [51] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Nips*, pages 1–9, 2013. ISBN 2150-8097. doi: 10.1162/jmlr.2003.3.4-5.951.
- [52] Tomas Mikolov, Wen-Tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, number June, pages 746–751, 2013. ISBN 9781937284473.
- [53] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An On-line Lexical

- Database *. *International Journal of Lexicography*, 3(4):235–244, 1990. ISSN 0950-3846. doi: 10.1093/ijl/3.4.235.
- [54] National Center for Complementary and Integrative Health. Complementary, Alternative, or Integrative Health: What’s In a Name?, 2015. URL <https://nccih.nih.gov/health/integrative-health>.
- [55] NCCAM. Expanding Horizons of Healthcare: Five-Year Strategic Plan 2001-2005, 2000. URL <https://nccih.nih.gov/sites/nccam.nih.gov/files/about/plans/fiveyear/fiveyear.pdf>.
- [56] Stefan Negru and Steffen Lohmann. A Visual Notation for the Integrated Representation of OWL Ontologies. In *Proceedings of the 9th International Conference on Web Information Systems and Technologies*, pages 308–315. SciTePress - Science and Technology Publications, 2013. ISBN 978-989-8565-54-9. doi: 10.5220/0004373003080315.
- [57] The Loc Nguyen, Anne Laurent, Sylvie Rapior, François Carbonnel, Raphaël Trouillet, Gérard Bourrel, and Gregory Ninot. Defining a Collaborative Ontology for Non-Pharmacological Interventions. In *IBTN 2016*, 2016.
- [58] Grégory Ninot. Intervention non-médicamenteuse I NM : un concept pour lever les ambiguïtés sur les médecines douces et complémentaires. doi: 10.4267/2042/65110.
- [59] Grégory Ninot, Sylvain Agier, Simon Bacon, Claudine Berr, Isabelle Boulze, Gérard Bourrel, François Carbonnel, Valérie Clément, Michel David, Aurélie Gerazime, Adeline Gomez, Estelle Guerdoux-Ninot, Anne Laurent, Kim Lavoie, Thérèse Libourel, Béatrice Lognos, Francis Maffre, Jérôme Maitre, Sophie Martin, Grégory Mercier, Bertrand Nalpas, The Loc, Nguyen, Agnes Oude Engberink, Jean-louis Pujol, Xavier Quantin,

- Sylvie Rapior, Pierre Senesse, Anne Stoebner-Delbarre, and Raphaël Trouillet. La Plateforme CEPS : Une structure universitaire de réflexion sur l'évaluation des interventions non médicamenteuses (INM). *HEGEL - HEpato-GastroEntérologie Libérale*, 7(1):53–56, feb 2017. ISSN 2115-452X. doi: 10.4267/2042/62022.
- [60] Grégory Ninot, Fabienne Amadori, Jérôme Maître, Sylvie Rapior, Loric Rivière, Raphaël Trouillet, and François Carbonnel. Motrial, le premier méta-moteur de recherche des études cliniques sur les interventions non médicamenteuses (INM). *HEGEL - HEpato-GastroEntérologie Libérale*, (1), feb 2018. ISSN 2115-452X. doi: 10.4267/2042/65113.
- [61] Grégory Ninot, Isabelle Boulze-Launay, Gérard Bourrel, Aurélie Gérazime, Estelle Guerdoux-Ninot, Béatrice Lognos, Thérèse Libourel, Grégoire Mercier, Agnès Oude Engberink, Sylvie Rapior, Pierre Senesse, Raphaël Trouillet, and François Carbonnel. De la définition des interventions non médicamenteuses à leur ontologie. *HEGEL - HEpato-GastroEntérologie Libérale*, (1), feb 2018. ISSN 2115-452X. doi: 10.4267/2042/65114.
- [62] Natalya F Noy and Deborah L Mcguinness. *Ontology Development 101: A Guide to Creating Your First Ontology*. Technical report, Stanford Knowledge Systems Laboratory, 2001.
- [63] Charles Kay Ogden and I. A. Richards. The Meaning of Meaning: a Study of the Influence of Language upon Thought and of the Science of Symbolism. *Nature*, 111(2791):566–566, apr 1923. ISSN 0028-0836. doi: 10.1038/111566b0.
- [64] Ted Pedersen, Serguei V S Pakhomov, Siddharth Patwardhan, and Christopher G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299, 2007. ISSN 15320464. doi: 10.1016/j.jbi.2006.06.004.

- [65] Ted Pedersen, Serguei Pakhomov, Bridget McInnes, and Ying Liu. Measuring the Similarity and Relatedness of Concepts in the Medical Domain : IHI 2012 Tutorial. Technical report, 2012.
- [66] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Stroudsburg, PA, USA, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162.
- [67] Lewis Philip and Wysocki Eversole. Controlled vocabularies taxonomies and ontologies. *Top Braid*, pages 1–5, 2013.
- [68] Carlos Eduardo Ramisch. Multi-word terminology extraction for domain-specific documents. Master’s thesis, 2009.
- [69] Czech Republic and Tomas Mikolov. *Statistical Language Models Based on Neural Networks*. PhD thesis, 2012.
- [70] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, oct 1986. ISSN 0028-0836. doi: 10.1038/323533a0.
- [71] Manuel Salvadores, Paul R Alexander, Mark A Musen, and Natalya F Noy. BioPortal as a Dataset of Linked Biomedical Ontologies and Terminologies in RDF. *Semantic Web*, 4(3):277–284, 2013. ISSN 15700844. doi: 10.3233/SW-2012-0086.
- [72] Eric Sanjuan, James Dowdall, Fidelia Ibekwe-Sanjuan, and Fabio Rinaldi. A symbolic approach to automatic multiword term structuring. *Computer Speech and Language*, 19(4):524–542, 2005. ISSN 08852308. doi: 10.1016/j.csl.2005.02.002.

- [73] Biplab K. Sarker, Peter Wallace, and Will Gill. Some observations on mind map and ontology building tools for knowledge management. *Ubiquity*, 2008(March):1–9, mar 2008. ISSN 15302180. doi: 10.1145/1366313.1353570.
- [74] Sean Falconer. OntoGraf. URL <https://protegewiki.stanford.edu/wiki/OntoGraf>.
- [75] Amit Singhal. Modern Information Retrieval: A Brief Overview. *Bulletin of the Ieee Computer Society Technical Committee on Data Engineering*, 24(4):1–9, 2001. ISSN 00218979. doi: 10.1.1.117.7676.
- [76] Irena Spasić, Mark Greenwood, Alun Preece, Nick Francis, and Glyn Elwyn. FlexiTerm: a flexible term recognition method. *Journal of Biomedical Semantics*, 4(1):27, oct 2013. ISSN 2041-1480. doi: 10.1186/2041-1480-4-27.
- [77] Tania Tudorache and Natasha Noy. Collaborative Ontology Development with Protégé. Technical report, 2009.
- [78] Tania Tudorache, Csongor Nyulas, Natalya F. Noy, and Mark A. Musen. WebProtégé: A Collaborative Ontology Editor and Knowledge Acquisition Tool for the Web. *Semantic web*, 4(1):89–99, jan 2013. ISSN 1570-0844. doi: 10.3233/SW-2012-0057.
- [79] U.S. National Library of Medicine. PMC Overview, . URL <https://www.ncbi.nlm.nih.gov/pmc/about/intro/>.
- [80] U.S. National Library of Medicine. PubMed®: MEDLINE® Retrieval on the World Wide Web, . URL <https://www.nlm.nih.gov/pubs/factsheets/pubmed.html>.
- [81] U.S. National Library of Medicine. MEDLINE®, . URL <https://www.nlm.nih.gov/pubs/factsheets/medline.html>.

- [82] Juan Antonio Lossio Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. Towards a Mixed Approach to Extract Biomedical Terms from Text Corpus. *International Journal of Knowledge Discovery in Bioinformatics*, 4(1):1–15, jan 2014. ISSN 1947-9115. doi: 10.4018/ijkdb.2014010101.
- [83] W3C. RDF Vocabulary Description Language 1.0: RDF Schema (RDFS), . URL <https://www.w3.org/2001/sw/wiki/RDFS>.
- [84] W3C. Resource Description Framework (RDF), . URL <https://www.w3.org/RDF/>.
- [85] W3C. Semantic Web, . URL <https://www.w3.org/standards/semanticweb/>.
- [86] W3C. Simple Knowledge Organization System (SKOS), . URL <https://www.w3.org/2001/sw/wiki/SKOS>.
- [87] W3C. Web Ontology Language (OWL), . URL <https://www.w3.org/OWL/>.
- [88] W3C. SKOS Simple Knowledge Organization System Reference, . URL <https://www.w3.org/TR/skos-reference/>.
- [89] World Health Organization. Traditional, complementary and integrative medicine, 2017. URL <http://www.who.int/traditional-complementary-integrative-medicine/about/en/>.
- [90] Marcia Lei Zeng. Knowledge Organization Systems (KOS). *Knowledge Organization*, 35(2):160–182, 2008. ISSN 0943-7444, 0943-7444. doi: 10.1002/meet.145044019.
- [91] Marcia Lei Zeng and Philipp Mayr. Knowledge Organization Systems (KOS) in the Semantic Web: A Multi-Dimensional Review. 2018.

- [92] Ziqi Zhang, Jie Gao, and Fabio Ciravegna. JATE 2.0 : Java Automatic Term Extraction with Apache Solr. In *10th International Conference on Language Resources and Evaluation (LREC'16)*, Portorôz, Slovenia, number May, pages 2262–2269, 2016.
- [93] Yongjun Zhu, Erjia Yan, and Fei Wang. Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. *BMC Medical Informatics and Decision Making*, 17(1):95, dec 2017. ISSN 1472-6947. doi: 10.1186/s12911-017-0498-1.