

University of Southampton Research Repository

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.

Thesis: Author (Year of Submission) "Full thesis title", University of Southampton, name of the University Faculty or School or Department, PhD Thesis, pagination.

Data: Author (Year) Title. URI [dataset]

University of Southampton

Faculty of Engineering and Physical Sciences

School of Electronics and Computer Science

**Towards an Understanding of the Semantic Web
Opportunities, Usages, Affordances and Challenges in
Health Research**

by

Mona Mohammed AlmoFarreh

Thesis for the degree of PhD Computer Science

March 2021

University of Southampton

Abstract

Faculty of Faculty of Engineering and Physical Sciences

School of Electronics and Computer Science

Thesis for the degree of PhD Computer Science

**Towards an Understanding of the Semantic Web Opportunities, Usages,
Affordances and Challenges in Health Research**

by

Mona Mohammed AlmoFarreh

The semantic web (SW) offers tools for supporting data integration and sharing across disparate resources in the web. Meanwhile, health research needs an efficient approach for handling heterogeneous data integration for the massive amounts of available health-related data to help discovering new scientific breakthroughs. In this thesis, the current and potential relationships between the semantic web and health research are aimed to be understood and identified through systematically reviewing the literature and examining the SW features in a proof-of-concept health-related demonstrator.

Firstly, a systematic literature review of 447 articles addressing health questions and using the SW standards was conducted to map the literature and identify any gaps or opportunities. The results of the review were analysed in a mixed approach of quantitative and qualitative methods producing two taxonomies: 1) the health aims and 2) the SW features taxonomies. The review revealed the most and least addressed health questions as well as the used SW features in the literature.

Secondly, a semantic web-based demonstrator was developed to represent the NHS dispensed prescriptions topic and examine some of the identified SW features. The prescriptions demonstrator consists of three interlinked OWL ontologies: the BNF, NHS and prescriptions ontologies along with their converted RDF instances. Moreover, two health questions, inspired from the traditional health literature and suggested by health experts in a focus group, were translated into SPARQL queries and ran across the ontologies to test more of the SW features.

It has been learned that the SW has a potential in supporting health research and accelerating research findings in the areas of: data representation, data integration and knowledge discovery. However, there are some challenges need resolving for a better result such as: data accessibility, security, quality, heterogeneity and lack of user-friendly tools.

Table of Contents

Table of Contents	i
Table of Tables	xi
Table of Figures	xiii
Research Thesis: Declaration of Authorship	xvii
Acknowledgements	xix
Chapter 1 Introduction	1
1.1 Research Questions	2
1.2 Thesis Structure.....	3
1.2.1 Chapter One	3
1.2.2 Chapter Two	3
1.2.3 Chapter Three.....	4
1.2.4 Chapter Four.....	4
1.2.5 Chapter Five.....	4
1.2.6 Chapter Six	5
1.2.7 Chapter Seven	5
1.2.8 Chapter Eight.....	6
Chapter 2 Background	7
2.1 The Semantic Web Vision	7
2.2 Linked Data	8
2.3 The SW Architecture.....	11
2.3.1 The Low-level Components (Standards)	11
2.3.2 The Higher-level Components.....	12
2.3.3 An Example: The Drug Ontology (DrOn)	13
2.4 The Uses of the Semantic Web in Health Data	15
2.4.1 Overview	15
2.4.1.1 Data Publishing	15
2.4.1.2 Data Exchange.....	15
2.4.1.3 Data Integration	16

2.4.1.4	Data Retrieving	16
2.4.1.5	Data Monitoring.....	16
2.4.2	Conclusion.....	17
2.5	Linking Data Traditionally and Semantically.....	17
2.5.1	Overview.....	17
2.5.2	Data Types.....	19
2.5.3	The Linking Technique	21
2.5.4	Strengths and Challenges	22
2.5.5	Conclusion.....	23
2.6	Summary.....	25
Chapter 3	Methodology	27
3.1	Systematic Literature Review	28
3.1.1	Identification	29
3.1.2	Screening.....	30
3.1.3	Eligibility.....	30
3.1.4	Inclusion.....	31
3.1.5	Qualitative Analysis	31
3.1.6	Quantitative Analysis.....	31
3.2	Building a Demonstrator	31
3.2.1	Choosing a Topic	32
3.2.2	Choosing Data Resources	33
3.2.3	Extracting Concepts from the Data	33
3.2.4	Modelling Ontologies	33
3.2.5	Converting Data into RDF	33
3.2.6	Uploading Data to Triplestore.....	34
3.2.7	Testing the Demonstrator.....	34
3.3	Health Use Cases.....	34
3.3.1	Choosing the Health Question	35
3.3.2	Finding Data Sources	35
3.3.3	Translating the Question	36

3.3.4	Executing the Query	36
3.3.5	Analysing the Results	36
3.4	Summary	37
Chapter 4 The Semantic Web's Uses in Health Research: a Systematic Review		39
4.1	Introduction	39
4.1.1	Similar Reviews	40
4.1.2	The Study's Contribution	41
4.2	Methods	42
4.2.1	Systematic Review Phases	42
4.2.1.1	Identification	43
4.2.1.2	Screening	43
4.2.1.3	Eligibility	44
4.2.1.4	Inclusion	44
4.2.2	Analysis Methods	44
4.2.2.1	Qualitative Analysis	44
4.2.2.2	Quantitative Analysis	45
4.2.3	Limitations	46
4.3	Results	46
4.3.1	Health Aims Taxonomy	47
4.3.1.1	Definitions of Aims	48
4.3.1.2	Usage Analysis	53
4.3.2	Semantic Web Features Taxonomy	59
4.3.2.1	Definitions of the SW Taxonomy's Features	60
4.3.2.2	Usage Analysis	65
4.3.3	Variance in Using the SW's Features across Health Aims	71
4.3.3.1	Usage Analysis	71
4.3.3.2	Examples	76
4.4	Discussion	86

4.4.1	Addressed Health Questions.....	86
4.4.2	The Semantic Web's Features.....	88
4.4.3	The Semantic Web's Affordances in Health Research	90
4.4.4	Challenges and Recommendations	91
4.5	Conclusion	93
4.6	Summary.....	95
Chapter 5	Semantic Web Demonstrator for Health Research.....	97
5.1	Prescriptions in England.....	97
5.1.1	Why Prescriptions?.....	98
5.1.2	About the NHS.....	98
5.1.3	About the BNF	99
5.1.4	Examples of Prescription Projects.....	100
5.1.4.1	Mapping English GP prescribing data: a tool for monitoring health- service inequalities.....	100
5.1.4.2	OpenPrescribing.....	101
5.1.4.3	iView.....	101
5.1.4.4	Information Services Portal	102
5.1.4.5	Prescribing Analytics	102
5.1.4.6	Medicine statistics: GP prescribing by constituency, 2015.....	102
5.2	Demonstrator Architecture	103
5.2.1	Data Access Component	104
5.2.1.1	Data Sources	104
5.2.1.2	Prescription data.....	105
5.2.1.3	Practice information and NHS structure data	105
5.2.1.4	Medication data.....	106
5.2.2	The Mapper.....	107
5.2.2.1	Extract	107
5.2.2.2	Transform	114
5.2.2.3	Load.....	119
5.2.3	Querying Interface.....	120

5.2.3.1	Testing the NHS Ontology.....	121
5.2.3.2	Testing the BNF Ontology.....	124
5.2.3.3	Testing the Prescriptions Ontology.....	127
5.2.3.4	Testing the Integration of the Three Ontologies.....	130
5.2.3.5	Testing Inference.....	133
5.2.3.6	Testing Aggregation.....	136
5.3	Analysis.....	139
5.3.1	Data Representation.....	139
5.3.2	Data Integration.....	140
5.3.3	Knowledge Discovery.....	141
5.3.4	Limitations.....	142
5.4	Summary.....	143
Chapter 6	Health Research: Two Use Cases.....	145
6.1	Case One: Prescribing Inequalities for Diabetic Medications.....	145
6.1.1	Case Description.....	145
6.1.2	The Query.....	146
	Step 1: Identifying metformin hydrochloride products and presentations.....	149
	Step 2: Identifying the linked prescriptions with the metformin hydrochloride's presentations.....	151
	Step 3: Calculating the total NIC per practice.....	152
	Step 4: Finding the total number of registered patients.....	153
	Step 5: Calculating the cost per person rate.....	154
6.1.3	Results.....	154
6.2	Case Two: The Effect of Living in a Coastal City or Town upon the Prescribing of Antidepressants.....	157
6.2.1	Focus Group Design.....	157
6.2.2	The Focus Group's Results.....	158
6.2.3	Case Description.....	160
6.2.4	The Query.....	160
	Step 1: Identifying cities/towns and towns located in England from Wikidata.....	164

Step 2: Identifying cities/towns' populations from Wikidata	164
Step 3: Defining coastal cities/towns according to Wikidata	165
Step 4: Identifying the list of NHS practices in the selected coastal cities/towns	166
Step 5: Selecting all BNF presentations for antidepressants	167
Step 6: Calculating the antidepressants prescribing rate per each city/town .	168
6.2.5 Results.....	169
6.3 Analysis.....	172
6.4 Summary.....	176
Chapter 7 Discussion.....	177
7.1 Types of Addressed Health Questions	177
7.2 The SW's Uses in Health Research.....	179
7.2.1 Data Representation.....	179
7.2.2 Data Integration	181
7.2.3 Knowledge Discovery.....	182
7.2.4 Updating Knowledge	184
7.2.5 Data Sharing	184
7.3 The Affordances and Challenges of Using the Semantic Web in Health Research	185
7.3.1 The Semantic Web's Affordances for Health Research.....	185
7.3.1.1 Representing various topics and questions	186
7.3.1.2 Simplicity and flexibility in representing data	186
7.3.1.3 Representing logical conditions.....	186
7.3.1.4 Heterogenous data integration	186
7.3.1.5 Local and remote data integration	186
7.3.1.6 Incorporation between public and private datasets	187
7.3.1.7 Flexibility in data linking.....	187
7.3.1.8 Exploring linked data.....	187
7.3.1.9 Checking for logical inconsistencies.....	187
7.3.1.10 Inferring new information	188
7.3.2 More Affordances.....	188

7.3.2.1	Ability to handle big data.....	188
7.3.2.2	Ability to integrate multiple data at a time	188
7.3.2.3	Aggregating data based on certain condition.....	189
7.3.2.4	Supporting arithmetic functions	189
7.3.3	Challenges Facing Employing the Semantic Web in Health Research.....	189
7.3.3.1	Data accessibility	189
7.3.3.2	Data security	190
7.3.3.3	Data quality.....	190
7.3.3.4	Data heterogeneity	191
7.3.4	More Challenges.....	192
7.3.4.1	Data availability.....	192
7.3.4.2	Complexity of building ontologies	192
7.3.4.3	Lack of efficient user-friendly tools.....	193
Chapter 8	Conclusion.....	195
8.1	Contributions	197
8.1.1	Health Aims Taxonomy	197
8.1.2	Semantic Web Features Taxonomy	197
8.1.3	The Prescriptions Demonstrator	198
8.1.4	List of the Affordances of Employing the Semantic Web in Health Research 198	
8.1.5	List of the Faced Challenges of Employing the Semantic Web in Health Research.....	199
8.2	Recommendations.....	199
8.3	Future Work	200
8.3.1	Reasoning over rules in health topics addressing decision-based questions	200
8.3.2	Standardising the process of publishing linked data on the web	200
8.3.3	Improving user-friendly tools for accessing linked data.....	200
8.3.4	Supporting security-related studies especially in the case of incorporation between private and public data.....	200
8.3.5	Improving federated querying technology	201

8.4	Ending Statement.....	201
Bibliography.....		203
Appendix A The Uses of the Semantic Web in Health Data Matrices		221
A.1	General Information.....	221
A.2	Domain Information.....	225
A.3	Aims / Goals.....	226
Appendix B The Traditional Linked Data Matrices.....		229
B.1	General Information.....	229
B.2	Domain Information.....	231
B.3	Aims / Goals.....	231
B.4	Methods / Tools.....	232
B.5	Data Resources	232
B.6	Challenges.....	234
B.7	Strengths.....	234
Appendix C The Semantic Linked Data Matrices		235
C.1	General Information.....	235
C.2	Domain Information.....	236
C.3	Aims / Goals.....	236
C.4	Methods / Tools.....	237
C.5	Data Resources	237
C.6	Technologies.....	239
C.7	Weaknesses	239
C.8	Strengths.....	239
Appendix D The Thematic Matrix of the Systematic Review for the Semantic Web's Uses in Health Research.....		240
Appendix E Converting CSV to RDF Script Code.....		253
Appendix F Focus Group Inputs and Outputs		255
F.1	Information Sheet	255
F.2	Consent Form.....	259
F.3	The Prescriptions Demonstrator.....	260

F.4	Suggested Datasets	261
F.5	The Resulted Questions	270

Table of Tables

Table 1: A summary of the research questions, methods and expected outputs	38
Table 2: The usage rates of the health aims taxonomy's categories	53
Table 3: The usage rates of the categories in the sw features taxonomy.....	65
Table 4: variance in the use of the semantic web's features across health aims.....	72
Table 5: The representations of the Panadol tab 500mg BNF presentation code	100
Table 6: The chosen open datasets for the prescriptions demonstrator	104
Table 7: Extracting the prescriptions ontology information	108
Table 8: Extracting the BNF ontology information	109
Table 9: Extracting the NHS ontology information	111
Table 10: Part of the region office - area team CSV file	116
Table 11: Details of the splitting and converting steps for the prescriptions dataset	118
Table 12: An example of the prescriptions' division into groups	153
Table 13: A snapshot of the colour coded results for the cost-per-person query.....	156
Table 14: The list of suggested questions in the focus group with their exclusion reasons	158
Table 15: The results for the antidepressants prescribing rates for coastal and non-coastal cities/towns	169
Table 16: The results of the manually added coastal cities/towns for the antidepressants query	170
Table 17: The results of the manually added non-coastal cities/towns for the antidepressants query	171

Table of Figures

Figure 1: The Linked Open Data (LOD) diagram in March 2019 (Insight Centre for Data Analytics, 2019)	10
Figure 2: The W3C technology stack	11
Figure 3: The Drug ontology (DrOn) (Hanna et al., 2013)	13
Figure 4: An example of a triple in DrOn	14
Figure 5: The methodology steps.....	27
Figure 6: Systematic review steps.....	28
Figure 7: Steps to building a Semantic Web-based demonstrator	32
Figure 8: Steps for applying health use cases	35
Figure 9: The four PRISMA steps followed in the systematic review.....	42
Figure 10: The health aims taxonomy	47
Figure 11: The division of the literature into four main health aims	54
Figure 12: The division of the literature according to the health sub-aims (addressed questions).....	55
Figure 13: Taxonomy of the semantic web's features	59
Figure 14: The average usage of the semantic web's main features.....	66
Figure 15: The usage's rate of the semantic web's sub-features	67
Figure 16: An Example for 'Diagnosis and decision support'	77
Figure 17: An Example for 'Monitoring patients via sensor devices'	77
Figure 18: An Example for 'Treatment and drugs recommendation'	77
Figure 19: An Example for 'Examination using medical tools or devices'	77
Figure 20: An Example for ' Finding causes of diseases'	77
Figure 21: An Example for 'Promoting public awareness'.....	80
Figure 22: An Example for 'Epidemiology and environment surveillance'	80

Figure 23: An Example for 'Supporting assistive technologies for special-needs people'	80
Figure 24: An Example for 'Understanding social behaviours'	80
Figure 25: An Example for 'Supporting clinical trials and secondary use of EHRs'	82
Figure 26: An Example for 'Clinical pathways and patients care plans'	82
Figure 27: An Example for 'Clinical guidelines and policies'	82
Figure 28: An Example for 'Training staff and supporting learning'	82
Figure 29: An Example for 'Workflow, communication and business processes'	82
Figure 30: An Example for 'Finding adverse drug events'	84
Figure 31: An Example for 'Drugs discovery'	84
Figure 32: An Example for 'Testing drugs'	84
Figure 33: The linkage between the datasets in the 'Mapping English GP prescribing data: a tool for monitoring health-service inequalities' (Rowlingson et al., 2013)	100
Figure 34: The prescription demonstrator architecture	103
Figure 35: The prescriptions ontology	108
Figure 36: The BNF ontology	110
Figure 37: The NHS ontology	113
Figure 38: The integration of the NHS, BNF and prescriptions ontologies	114
Figure 39: Transforming the CSV files into RDF files for NHS, BNF and prescriptions ontologies	115
Figure 40: An example of a mapping rule using OpenRefine	116
Figure 41: The PrescriptionRepository information in GraphDB	119
Figure 42: A SPARQL Query for Testing the NHS Ontology	121
Figure 43: The elements used in testing the NHS ontology	122
Figure 44: A snapshot of the results of the NHS ontology testing use case	123
Figure 45: The implemented query for the BNF ontology testing use case	124
Figure 46: The data used in the BNF ontology testing query	125

Figure 47: Snapshot of the results of the BNF ontology testing use case	126
Figure 48: The implemented query for the prescriptions ontology test case	127
Figure 49: The used data in the prescriptions ontology test query	128
Figure 50: A snapshot of the results of the prescriptions ontology test case	129
Figure 51: The implemented query for testing the integration of all three ontologies.....	130
Figure 52: The data used to test the integration of all three ontologies	131
Figure 53: A snapshot of the results of the integration of all three ontologies	132
Figure 54: The query that was implemented to test for inference.....	133
Figure 55: The data used in testing the inference query	134
Figure 56: A snapshot of the results of the inference test case	135
Figure 57: The implemented query for testing aggregation	136
Figure 58: The data used in testing aggregation query.....	137
Figure 59: A snapshot of the results of testing aggregation use case.....	137
Figure 60: The SW features taxonomy for the prescriptions demonstrator	139
Figure 61: The SPARQL query for finding the cost-per-person rate	147
Figure 62: The data concepts used in the five steps of the cost-per-person query	148
Figure 63: Step 1 of the cost-per-person query.....	149
Figure 64: the links between the chemical substance coded '0601022b0' and its related products and presentations	150
Figure 65: Step 2 of the cost-per-person query.....	151
Figure 66: Step 3 of the cost-per-person query.....	152
Figure 67: Step 4 of the cost-per-person query.....	153
Figure 68: Step 5 of the cost-per-person query.....	154
Figure 69: Cost-per-person for Metformin spending in London (Rowlingson <i>et al.</i> , 2013)	155
Figure 70: The percentages of the five colour coded cost-per-person rates	156

Figure 71: The ‘depression rates in coastal cities/towns’ query	162
Figure 72: The data concepts used in the six steps of the ‘depression rates in coastal cities/towns’ query.....	163
Figure 73: Step 1 of the ‘depression rates in coastal cities/towns’ query.....	164
Figure 74: Step 2 of the ‘depression rates in coastal cities/towns’ query.....	164
Figure 75: Step 3 of the ‘depression rates in coastal cities/towns’ query.....	165
Figure 76: Step 4 of the ‘depression rates in coastal cities/towns’ query.....	166
Figure 77: Step 5 of the ‘depression rates in coastal cities/towns’ query.....	167
Figure 78: Step 6 of the ‘depression rates in coastal cities/towns’ query.....	168
Figure 79: The results for the coastal cities/towns query.....	169
Figure 80: The results for the non-coastal cities/towns	170
Figure 81: The thesis's summary figure	196

Research Thesis: Declaration of Authorship

Print name: Mona Mohammed AlmoFarreh

Title of thesis: Towards an Understanding of the Semantic Web Opportunities, Usages, Affordances and Challenges in Health Research

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission.

Signature:



Date: 22/03/2021

Acknowledgements

To my family... the joy of my life...

First and foremost, I am grateful and thankful for almighty God for giving me the strength and patience to finish this work. I am also grateful for my sponsors, King Saud University and the Saudi government, for a full financial support for this research.

I am sincerely grateful for Dr. Mark Weal for supporting, encouraging and advising me from the beginning of this journey till the end. I also want to thank Prof. Susan Halford for her wise comments and supervision during the early stages of the PhD.

Finally, I am extremely grateful for all members of my family, who were there for me in every step of the way, my father and mother, my sister and brothers, and my in-law's family.

Meaili and Bushra, my lovely children, thank you for being so patient with me through this journey.

Ahmad, my supporting and loving husband, you've been there for me from day one till the last moment. Thank you my love.

Chapter 1 Introduction

Health research is essential for overcoming human's health challenges by monitoring, diagnosing and responding to the rise of diseases worldwide (World Health Organization (WHO), no date). Health research refers to the scientific studies that is interested in improving knowledge about human health by addressing and answering questions in order to find better ways for treating people (Cancer Council Australia, 2018).

Health research needs to address questions involving the integration of heterogeneous data sources, as well as analysing data, in order to discover new knowledge (Cheung *et al.*, 2009). Interdisciplinary approaches to facilitate 'remote' collaboration between health researchers are gaining more attention, especially in life science studies, in attempts to find new scientific breakthroughs (Sagotsky *et al.*, 2008). The need for data integration in health research is becoming increasingly significant in attempts to expose hidden patterns in the huge available amounts of health data. The integrating and sharing of health data, together with discovering new knowledge and breakthroughs, will lead to improving both healthcare practice and products (Zenuni *et al.*, 2015).

In addition, there is a tremendous amount of health-related data produced every day from various health applications and devices (Zaveri and Ertaylan, 2017). For example, electronic health records (EHRs) in health institutions are frequently updated with different patients' data which are then stored in private data warehouses. Other sources of health data are governmental and organisational reports and registries that can be either private, with limited access, or even open for the public to use. Data generated from medical/health devices and sensors like Fitbit has also captured the attention of the academic community lately.

The semantic web (SW) offers tools and standards for data integration and sharing to inform heterogeneous resources on the web (Machado *et al.*, 2013). The SW is based on the idea of structuring and linking data concepts on the web in a machine-friendly format to permit the usage of enormous amounts of linked data in computer processing (Berners-Lee, Hendler and Lassila, 2001). There is a need for a set of practices and standards to handle structuring, publishing and linking data to apply the SW vision on the web of documents today. As a solution, Tim Berners-Lee back in 2006 (Berners-Lee, 2006) proposed the concept of linked data (LD). LD offers exchanging data on a big-scale in an interoperable re-usable manner; a process which could be the answer to resolving the heterogeneous data integration challenges. The interlinked data in the huge LD world is represented in the form of ontologies. Each ontology is designed to serve a certain purpose or topic, and is supported by several SW technologies for data representing, integrating

and discovering. This situation is an opportunity for health researchers to benefit from what the SW can offer to support knowledge discovery and data integration.

This thesis aims to understand the current and potential relationships between the SW and health research. By studying the existing health projects employing SW technologies in the literature, the types of addressed health questions can be identified. Studying the different types of addressed health questions leads to a better understanding of the potential opportunities and existing gaps in the literature. In order to understand how the SW is employed in health research, the main used features of the SW evident in the health research literature are identified. By studying and analysing the literature, a better understanding of the faced challenges in applying the SW in health research as well as the gained affordances can be achieved.

In addition to understanding and analysing the current SW and health research relationships from the published literature perspective, a further analysis from a practical point of view helps in expanding the understanding of the SW's future in health research. A proof-of-concept model for demonstrating the use of the SW standards in representing a health-related topic is implemented to demonstrate the SW features in addressing health-related questions. This implementation helps in analysing the affordances, as well as the challenges in the process, from a practical perspective.

1.1 Research Questions

To achieve the research's aim of understanding the relationships between the SW and health research currently and potentially, three main research questions are addressed in this work.

What are the main health questions being addressed in health research employing semantic web technologies?

The first research question focuses on understanding the current health topics and questions being discussed and addressed in the literature. It will be answered by systematically reviewing the health literature using the SW and identifying the main topics and questions being discussed and attracting attention. By identifying the main health questions addressed in the SW accessed literature, a better understanding of the opportunities for, and existing gaps in the application of this technology can be achieved.

How are the semantic web features being used in health research?

The second research question is concerned with the usage of the SW features in the literature of health research. It will be answered by systematically reviewing and identifying the main SW features used in the literature. By looking at the usage rates of each identified feature across different health aims, a better understanding of the research trends and gaps can be reached.

What are the affordances and challenges in employing the semantic web for health research?

The third research question discusses the identified affordances and faced challenges of employing SW technologies in health research. This question will be answered from a literature and practical perspectives. The health literature using the SW approach will be reviewed and analysed for any mentioned affordances or challenges. More affordances and challenges will be identified from a practical point of view. A demonstrator based on the SW technologies representing a health topic will be developed along with real health use cases to be addressed using the demonstrator. This experiment will help in understanding any potential opportunities for this technology when employed in health research.

1.2 Thesis Structure

This thesis aims to understand the current and potential relationships between the SW and health research. The opportunities, uses, affordances and challenges facing attempts to apply SW technology to health research are discussed from the literature and practice points of view across the eight chapters of this thesis.

1.2.1 Chapter One

The first chapter addresses the research problem in integrating heterogeneous health data and then introduces the study's main aim and research questions. The aim of this research is to understand the relationships between the SW and health research currently and potentially by studying: a) addressed health questions, b) SW features, c) SW affordances and d) challenges facing SW in health research.

1.2.2 Chapter Two

The second chapter reviews the main concepts of the SW and LD to understand whether the SW vision was fulfilled in the literature or not. Also, the traditional approach of linking health data is reviewed in opposed to the SW approach. The traditional linking process was mainly performed

on a personal level by independent data centres, while the semantic linking was performed at different levels, mostly linking with open biomedical and clinical ontologies.

1.2.3 Chapter Three

The third chapter discusses the methodology adopted in this thesis. The methodology consists of three main parts: 1) a systematic review for the addressed health questions and used SW features in the literature, 2) a proof-of-concept demonstrator for the SW features representing a chosen health topic (NHS prescriptions) and 3) two health use cases applied to the demonstrator; one case comes from the traditional health literature and the other is suggested by health experts in a focus group.

1.2.4 Chapter Four

The fourth chapter systematically reviews the interdisciplinary literature relating to the use of SW for carrying out health research. In the systematic review, the literature is mapped from a health and technical perspectives. The first research question is answered in this chapter by identifying the main health questions that have been addressed in the literature. Four main health aims were found: i) medical, ii) public health, iii) health management and iv) pharmaceutical. Represented as a health aims taxonomy, 17 specific types of addressed health questions are categorised under the four main aims cited immediately above.

The second contribution in this chapter is the SW features taxonomy. The taxonomy includes five main semantic web features: i) data representation, ii) knowledge discovery, iii) data integration, iv) data sharing and v) updating knowledge, as well as 12 sub-features. From analysing the usage of the identified SW features across the health aims, both of opportunities and gaps in the literature were identified.

Finally, the affordances and challenges mentioned in the literature were analysed. The affordances were mainly around three areas: i) representing knowledge, ii) integrating data and iii) discovering knowledge. Challenges regarding data accessibility, security, quality and heterogeneity were also identified.

1.2.5 Chapter Five

The fifth chapter describes the proof-of-concept SW demonstrator modelling of a particular health management topic: *the dispensed NHS prescriptions in England*. The topic was chosen because it has not been discussed before, according to the results of the reviewed topics

presented in chapter four. Moreover, the prescriptions topic is a good candidate for demonstrating data integration between two different domains, health management and pharmaceutical domains.

The purpose of building the demonstrator was to test the identified SW features in practice, so as to analyse any affordances or challenges that may face any users of this technology. In chapter five, the architecture of the demonstrator including: a) the datasets used, b) the mapping process and c) testing queries are illustrated. Three interlinked ontologies were produced: i) NHS ontology, ii) BNF ontology and iii) prescriptions ontology, along with the converted RDF instances related to them.

The prescriptions demonstrator showed an ability to represent the prescriptions topic that links the pharmaceutical and health management domains. SPARQL queries succeeded in retrieving and exploring the linked data in the demonstrator. Few technical issues were noticed in the process of implementing the SW demonstrator, such as handling big data for graph displaying or converting to RDF.

1.2.6 Chapter Six

The sixth chapter describes two health-related cases applied to the prescriptions demonstrator introduced in the previous chapter. The aim for addressing these ‘use cases’ is to test the ability of the SW to address real health case scenarios and to analyse the process for any challenges or affordances. The first case is extracted from the health literature and discusses ‘prescribing inequalities for diabetic medicines in England’. The second case was suggested by health experts and discusses ‘the effect of living in a coastal city and the rate of prescribing antidepressants’. Both cases were translated to SPARQL queries and successfully achieved results from the demonstrator. However, issues in data availability and quality were faced that hindered the process of analysing feasible results, especially in the second use case. Some of the noticed affordances in this experiment were the ability to explore, link and retrieve big amounts of data per query. Another interesting SW affordance was the ability to invoke a remote data source (Wikidata) successfully and easily.

1.2.7 Chapter Seven

The seventh chapter discusses the findings from the previous chapters in order to analyse the affordances and challenges for applying the SW approach in health research. There were 14 identified affordances of the SW from both literature and practice perspectives. Regarding the challenges, there were seven found challenges in literature and practice.

1.2.8 Chapter Eight

Chapter eight concludes the thesis by listing the main contributions of this work along with the recommended future work in this field. This thesis concluded that the SW has the ability to support and accelerate health research findings by: a) representing different domains of knowledge, b) integrating heterogeneous data, c) discovering knowledge by exploring linked data and reasoning over rules. However, the process faces some challenges regarding: 1) data accessibility, 2) data security, 3) data quality, 4) data heterogeneity, 5) data availability, 6) ontologies complexities and 7) lack of SW tools.

Chapter 2 Background

In the previous chapter the thesis's aim was introduced: to understand the current and potential relationships between the SW and health research. This chapter elaborates on the main theoretical concepts about the SW in the literature, before attempting to answer the research questions of this thesis. In addition to reviewing the SW's concepts, the uses of the SW in health research are introduced, with a focus on the use of linked data (LD) traditionally and semantically.

2.1 The Semantic Web Vision

The term 'semantic web' (SW) was first used by Tim Berners-Lee in his article 'The Semantic Web' in 2001 (Berners-Lee, Hendler and Lassila, 2001). In this article, Berners-Lee introduced the idea of structuring data published on the web to allow machines applying intelligent processing to that data. Concepts and terms within a document need first to be defined in a language that is able to be understood by a machine, in order to allow for more complex processing involving artificial intelligence (AI) methods to take place. Berners-Lee and his colleagues had a vision of the SW as a global data repository that is combined with the semantics of interconnected data concepts, where some interesting inferences and reasoning processes can be applied on a big scale (Shadbolt, Hall and Berners-Lee, 2006).

One of the things that is promised from the SW is the ability to reach information easily by following a series of links to more data concepts (Berners-Lee, 2006). The SW, also called the web of data, is based on the idea of large-scale structured data linking. A standard format is used to structure the data in the form of triples. Triples are modelled as a simple sentence that consists of three related concepts: subject, predicate and object. Triples in the SW are formed by using the resource description framework (RDF) standard (Gandon and Schreiber, 2014) and RDF schema (RDFS) (Brickley and Guha, 2014). To identify these data concepts globally, they are assigned a unique identifier called the Uniform Resource Identifier (URI) to represent each resource (Shadbolt, Hall and Berners-Lee, 2006). The collection of these data concepts, represented by URIs and connected as triples, is called ontology (Berners-Lee, Hendler and Lassila, 2001). The term ontology has been defined by (Gruber, 1995) as "an explicit specification of a conceptualization", which enables both the representation and exchanging of knowledge in some domain. Thus, another promised benefit of the SW is to facilitate data interoperability and the exchanging of knowledge between systems via the use of standards and ontologies.

Besides defining data concepts in an ontology and classifying them in a taxonomical matter, ontologies also define associative relationships between the concepts' definitions and instances, or even mappings to other related concepts in different ontologies (Shaban-Nejad et al., 2016). One of the recommended SW standards in developing ontologies is the web ontology language (OWL) that also supports inference and consistency checking (McGuinness and Harmelen, 2004).

Ontologies also support the use of inference rules using reasoning based on logic. The semantic web rule language (SWRL) is based on OWL and allows the defining of if-then types of rules to infer new knowledge about OWL individuals (World Wide Web Consortium (W3C), 2004).

Automated reasoning in the SW is inherited from knowledge representation systems in artificial intelligence (AI). Knowledge representation systems in AI were subject to centralised control and difficult to manage with the increase in the amount of information they were handling. However, the SW has potential in representing data globally on a big scale, at the price of allowing some false answers or false inferring. The SW vision in automating reasoning for intelligent agents is 'to be able to answer complicated questions involving data from diverse sources' (Berners-Lee, Hendler and Lassila, 2001).

Another improvement that the SW promised to offer is in searching for knowledge processes. The search process nowadays mostly depends on keyword searching, which can result in unexpected answers sometimes. This outcome is because searching engines take the searched words as a solid block, without realising what are the semantics underneath these series of characters (Antoniou and Harmelen, 2008). For example, if a user typed the word 'orange' into a searching engine, the results could be about the colour orange or the fruit. However, if the search supported data semantics, the meaning noun or adjective of the word 'orange' can be determined, and therefore the search results will be more accurate.

2.2 Linked Data

The practice of linking datasets using SW standards is called linked data (LD). Linked data has been defined as "the set of best practices for publishing and connecting structured data on the Web" (Bizer, Heath and Berners-Lee, 2009). The LD term was first introduced in Berners-Lee's preliminary report 'Linked Data' (Berners-Lee, 2006), where he defined four rules for publishing data on the web. These four rules have become known as the 'Linked Data Principles':

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).

4. Include links to other URIs, so that they can discover more things.

These principles explain the ideas of linking data in the SW by: i) using a standard global naming system, ii) defining mappings between existing resources, iii) allowing data sharing and accessing and iv) finally re-using others' resources to share the effort. By following these rules, Berners-Lee sought to follow his vision in enabling machines to explore the web of data.

Ontologies are used to allow machines to: a) process data within documents, b) share data between systems and c) integrate heterogeneous datasets (Helfin, 2004). One of the main benefits of LD is the ability to integrate data between different domains under controlled universal standards and policies. Health research systems are very good examples for the need of data integration. For instance, translational research requires the integration of private clinical data with proprietary pharmaceutical data (Machado et al., 2013), while health economic research tends to link the outcomes of: a) randomised controlled trials (RCT), b) observational studies and c) patients' routine data (Husain et al., 2012). To achieve feasible data sharing between any two systems, those systems must be interoperable with each other. In a semantic context, two systems are not interoperable if they use different terms to describe similar concepts, or use identical terms to mean different things (Miller, 2000). One example of this problem could be using the terms 'patient record' or 'medical record' or 'health record' in three different systems to describe the same concept.

Although developing ontologies is a complex process that involves cooperation between domain experts and computer scientists, it is an idea gaining increasing attention in several different fields. The number of interlinked public ontologies available on the cloud or linked open data (LOD) (Heath and Bizer, 2011) is growing every year. In March 2019, there were 1,239 datasets with 16,147 links according to (the Insight Centre for Data Analytics, 2019), see figure 1. In 2007, there were only 12 registered datasets in the cloud, a total that jumped to 570 in 2014.

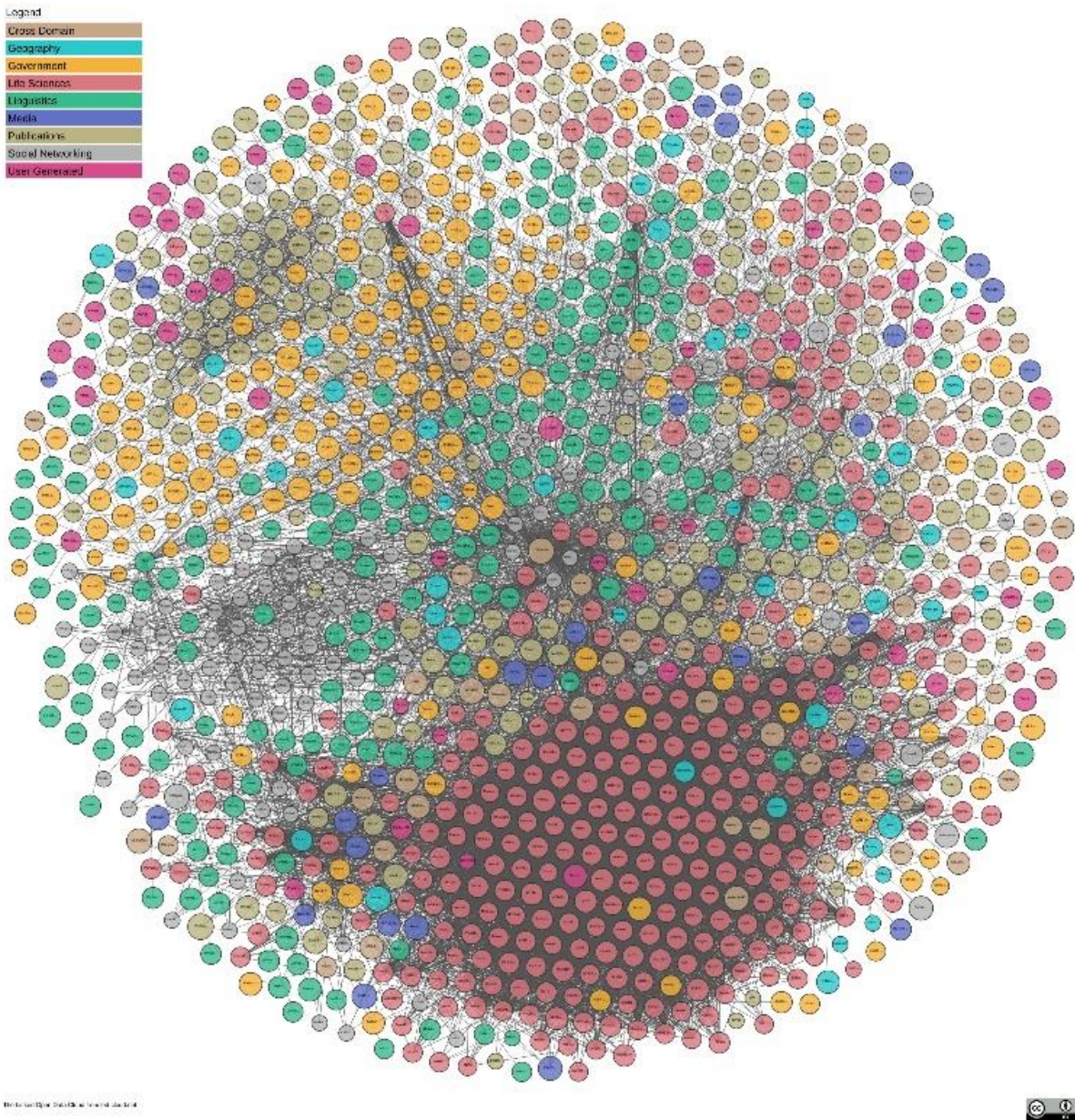


Figure 1: The Linked Open Data (LOD) diagram in March 2019 (Insight Centre for Data Analytics, 2019)

2.3 The SW Architecture

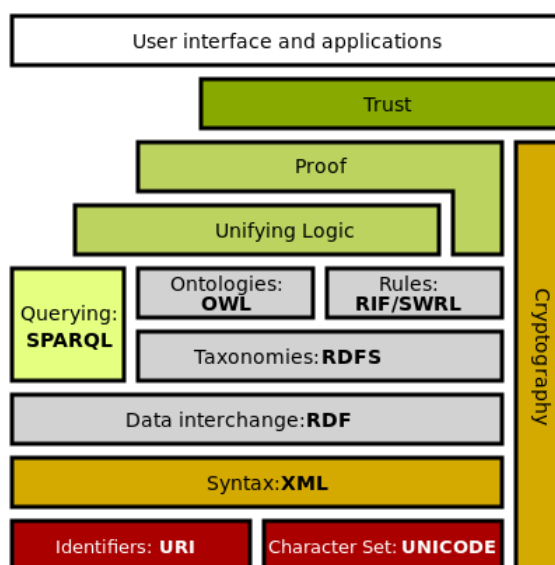


Figure 2: The W3C technology stack

The SW architecture is represented in a layered stack of technologies and standards that are aiming to ultimately be compatible with each other. Berners-Lee illustrated the semantic web architecture in a model named a 'semantic layer cake' in 2001, since then the model has experienced several updates (Antoniou and Harmelen, 2008). Figure 2 shows one version of the layer cake model that consists of high and low-level components. The lower layers, from the bottom up to the OWL layer, are the low-level technologies, which have been standardised (RDF, OWL and SPARQL), while the top layers of trust, proof and logic are still evolving and therefore not yet fully realised. It is worth mentioning that the layer cake model has been frequently debated and no final universally accepted version exists at this stage.

2.3.1 The Low-level Components (Standards)

Starting from the lowest level of the SW stack, resource identifiers take the form of URI or IRI. Berners-Lee defines the URI as "a compact sequence of characters that identifies an abstract or physical resource" (Berners-Lee, Fielding and Masinter, 2005). URIs have a special syntax to uniquely identify each resource universally, which usually starts with 'http://'. They can be considered as the main building block of the SW, as they can be used to identify any concept or resource globally. By this way, the machines will create a better interpretation of the data concepts within the documents (Shadbolt, Hall and Berners-Lee, 2006).

On top of the resource identifiers reside three layers representing data: XML, RDF and RDFS. The eXtensible markup language (XML) is a syntax language that is used for structuring text documents (Bray et al., 2008). RDF and RDFS are based on XML syntax. RDF is an abstract graph model that uses triples to link resources with each other. Each triple contains a subject and an object, representing the two data concepts that are related to each other by an association, which is represented by a predicate. Moreover, each subject, object and predicate are assigned unique URIs to represent them (Shadbolt, Hall and Berners-Lee, 2006). RDF schema (RDFS) language provides a basic vocabulary to describe the ontologies that structure RDF resources (Brickley and Guha, 2014).

Another ontological language is the web ontology language (OWL). OWL is based on description logic (DL), and for this reason it is more data expressive than RDFS (Shadbolt, Hall and Berners-Lee, 2006). One of OWL's major features is its ability to more adequately describe properties and classes, such as in the use of disjointedness, cardinality or enumerated classes (Helfin, 2004). OWL also relies on the triples concept, where a network of triples builds an ontology that enables: a) knowledge representation, b) logical inference and c) data interoperability (Farinelli, Barcellos De Almeida and Linhares De Souza, 2014).

Another standard that is compatible with OWL is the rule interchange format (RIF), which is responsible for exchanging rules between web rule-based systems, such as ontologies written in OWL (World Wide Web Consortium (W3C), 2013a). Moreover, the semantic web rule language (SWRL) is also built upon OWL axioms. SWRL enable rules in the form of antecedent (body) and consequent (head) to be combined with OWL ontologies (World Wide Web Consortium (W3C), 2004).

The last standard in the low-level components is the query language SPARQL. SPARQL is an RDF query language that is designed to retrieve data by the use of the triples concept. This means that a query consists of three sections in order to represent: a) subject, b) predicate and c) object. Variables can replace any of these sections to represent the aimed-for data that is to be retrieved (World Wide Web Consortium (W3C), no date b). SPARQL endpoints are web services used for querying data sources (DuCharme, 2011).

2.3.2 The Higher-level Components

Researchers are currently working on realising the rest of the components of the semantic web model, as many of the developed technologies are still evolving. In the semantic layer cake model, each layer is relying on the one beneath it. For example, the logic layer is an enhancement for ontological languages like OWL and RDFS. Proof and cryptography are essential to achieving trust

in the SW technology. Proof relies on a deductive process to validate the resulted knowledge, while cryptography involves the use of digital signature and encryption (Antoniou and Harmelen, 2008).

On the top level of the semantic stack stands the user interface concept, which enables humans through the use of different applications to employ the semantic web. However, there is a crucial factor that controls this human-technology interaction, which is trust. The idea of 'trust' in computer science often refers to verifying the identity of the information source; a task which can be successfully achieved with the help of the 'cryptography' and 'proof' layers. Trust in the semantic web also involves agents and automated 'reasoners' choosing alternative information resources (Artz and Gil, 2007). Trust is a very important element in the stack; it does not matter how good the technology is if there was zero trust from the human using it.

2.3.3 An Example: The Drug Ontology (DrOn)

For a better understanding of the SW standards, an example of the drug ontology (DrOn) is discussed in this section. DrOn contains drug products and information about their ingredients for the use in semantic web applications (Hanna et al., 2013). Figure 3 shows the main entities and relationships in DrOn. Briefly, each drug is identified by a national drug code (NDC) and is represented by a clinical drug form, which contains some ingredient details.

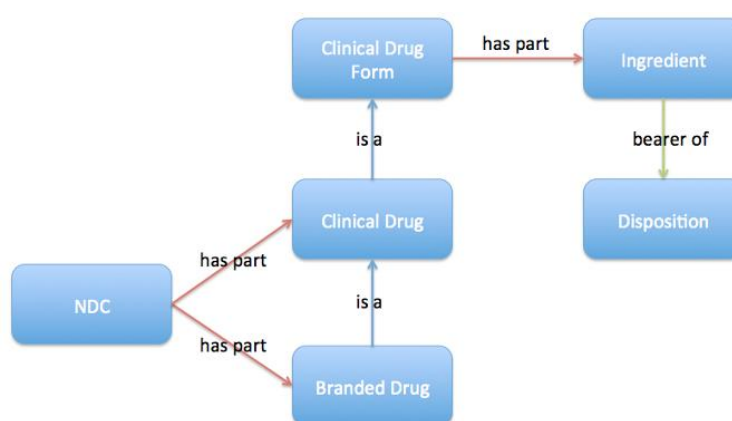


Figure 3: The Drug ontology (DrOn) (Hanna et al., 2013)

The taxonomy of DrOn has been developed by RDFS, while OWL2 has been used to represent the higher characteristics of the ontology. In the ontology graph, the concept of a triple is represented by two boxes, with an arc connecting them. The graph shows the subject for the first box, the

object for the second box and the predicate for the interconnecting arc, see Figure 4 (a). The schema or taxonomy of the domain is represented by the triples concepts using a language like RDFS or OWL, and this is what is shown in Figure 4 (b) as well. Finally, the classes' instances, the data itself, is represented, also by using the triples concept informed by the RDF standard, see Figure 4 (c). For example, 'Panadol' is an instance of the 'branded drug' class, and is connected by the 'is a' relationship to 'Paracetamol', which is an instance of the 'clinical drug' class.

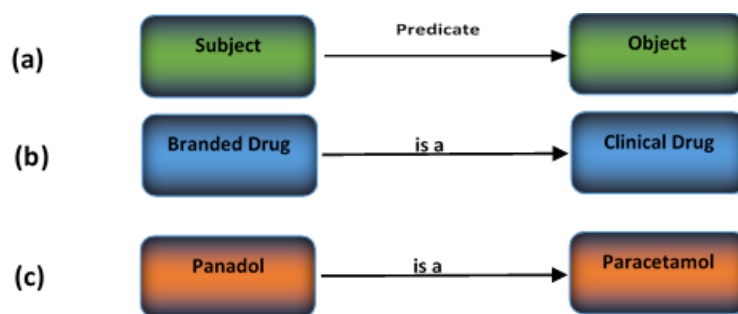


Figure 4: An example of a triple in DrOn

In the semantic web world, each resource is identified by a URI identifier. For instance, the 'Paracetamol' instance is identified by 'http://purl.obolibrary.org/obo/CHEBI_46195', while 'Panadol' is assigned a different URI 'http://purl.obolibrary.org/obo/DRON_00086554'.

For discovering knowledge in the SW world, there are two approaches: a) exploratory and b) inferential. The former is to explore linked resources by tracking the connections between them. By referring to the DrOns example in Figure 3, the query 'What is the clinical drug form for paracetamol?' can be answered by tracing the link between 'clinical drug' and 'clinical drug form' in a straightforward manner. However, there are other kinds of queries that can pose, and then answer, more interesting questions by traversing various linked ontologies among different disciplines. For example, it is possible an application needs some answers about drug interactions with paracetamol so as to trace the mappings between the drug ontology and the drug-drug interactions ontology (DDI) to obtain the data in order to answer the inquiry.

The second approach of discovering knowledge is the inference approach. Inference is more complex than exploration because it involves revealing knowledge that does not have an explicit representation in the ontology (Machado et al., 2013). Descriptive logic principles are the base of the inference approach. Referring again to the drugs example in Figure 3, if the 'has part' property, linking 'clinical drug' with 'ingredient', has an inverse property called 'is part of', the following knowledge could be inferred: 'ingredient' 'is part of' 'clinical drug form'. Knowing that

the ontology explicitly mentions that 'clinical drug form' 'has part' 'ingredient' but not vice versa, the reasoner will be able to infer the inverse relationship and consequently add it to the knowledge base.

2.4 The Uses of the Semantic Web in Health Data

Improvements in technologies help to increase the quantity of available health data on the web day by day. Following this study's aim to explore the SW's relationships with health research, this section will focus on reviewing some of the SW's uses in managing health data.

2.4.1 Overview

From narratively reviewing the literature, five main technical topics or papers' themes were identified: i) data integration, ii) data publishing, iii) data retrieving, iv) data exchange and v) data monitoring. See Appendix A for more details about the collected information.

2.4.1.1 Data Publishing

The data publishing category discusses publishing data online, either as private data with limited access or as open data. Some projects aim to publish governmental health data as linked open data (LOD) such as (Bukhari and Baker, 2013; Jovanovik, Najdenov and Trajanov, 2013; Rinciog and Posea, 2015). For instance Bukhari and Baker (2013) aim to re-publish the open Canadian health census data as LOD, as part of the open government movement. Another paper focused on building emergency healthcare ontology with limited access for authorised users, also linking it with LOD (Poulymenopoulou, Malamateniou and Vassilacopoulos, 2015).

2.4.1.2 Data Exchange

The data exchange category discusses several topics related to sharing data between different systems, such as: a) data interoperability and b) data accessibility. Liu and Wang (2016) presented an access control model for the Health Care and Life Sciences Research Group (HCLS), while others looked into the areas of data interoperability and managing data exchange between health systems (Hussain et al., 2012; Sonsilphong, Arch-Int and Arch-Int, 2012; Khan et al., 2013; Lim, Kim and Lee, 2013; Nie et al., 2013; Fareedi and Hassan, 2014; Cruz, Espinoza and Vidal, 2015; Menezes, Cook and Cavalini, 2016). For example Hussain *et al.* (2012) proposed a semantic interoperability system for reusing EHRs to improve clinical trials.

2.4.1.3 Data Integration

Data integration involves linking data from different domains or systems. In this category, there were some papers distinguished from the others regarding their aims. Their aims were motivated by a health background rather than a technical one, unlike the rest of the papers. In other words, these papers were attempting to answer health-related research questions (Pathak *et al.*, 2012; Pathak, Richard C Kiefer and Chute, 2012; Pathak, Richard C. Kiefer and Chute, 2012; Jesualdo Tomás Fernández-Breis *et al.*, 2013; Jyotishman Pathak, R. C. Kiefer and Chute, 2013; Jyotishman Pathak, R. Kiefer and Chute, 2013; Beyan *et al.*, 2014; Shah *et al.*, 2014; Odgers and Dumontier, 2015). An example of this type of papers is Pathak, R. Kiefer and Chute (2013) who concentrated on discovering some drug-drug interactions, while Pathak *et al.* (2012) aimed to diagnose patients with a specific disease.

The rest of the papers had technical aims and used some health examples as exemplar cases, but the authors were not trying to answer health question *per se*. For instance Sreekanth and Biswas (2014) aimed to develop a web service that would be able to compare various health insurance services to help health insurance seekers to find the most suitable policy by taking into consideration the issue of information asymmetry.

2.4.1.4 Data Retrieving

In this category, authors aimed to study the process of searching for and querying data. In some of the literature the researchers investigated different techniques for improving the data retrieving process by using natural language processing (Ostankov, Rohrbein and Waltinger, 2014; Al-Nazer, Albukhitan and Helmy, 2016). Other researchers chose to vary that aim by concentrating on search techniques. For instance Miyazaki *et al.* (2015) aimed to explore the content information in TV health programmes by developing an information hub using LOD.

2.4.1.5 Data Monitoring

What is meant by data monitoring is the process of analysing and evaluating the consumption and production of linked data. New technologies usually motivate academics to be interested in analysing the technologies' growth and also evaluating it. For example Tilahun *et al.* (2014) evaluated linked data technologies for use in the healthcare domain by developing a LOD-based health information system. Fotopoulou *et al.* (2016) analysed linked data by testing it in an urban environment study.

2.4.2 Conclusion

In conclusion, the learnt lesson from this preliminary review is that the literature in this field of interest involves topics from the: a) publishing, b) retrieving, c) exchanging, d) integrating and e) monitoring of data. The interesting point here is that most of the authors used health examples in their work but were primarily motivated by and interested in technical issues or topics. This situation can be seen as computer scientists trying to test and improve different technologies, which implies that this area is still growing. On the other hand, the few papers that were informed by a health perspective are also interesting, as they are an indication that the SW is starting to capture the attention of health researchers. The semantic web is now being used as a tool in conducting their research and for addressing health questions. In the rest of this thesis, this type of literature will be focused on; the type that has health motivation and aims driving it.

2.5 Linking Data Traditionally and Semantically

Health research faces many challenges due to its complex nature and the increase of multi-morbidity (Kotwal *et al.*, 2016). Specifically, data integration in health research faces challenges in operating the data collections due to their size and complexity, in addition to issues regarding the privacy and confidentiality of participants (Christen and Churches, 2006). Moreover, the lack of: i) data interoperability, ii) unified data representations and iii) data standards hinder the possibility of data being smoothly shared (Sagotsky *et al.*, 2008).

In this section, the challenges associated with efforts to link health data will be discussed. Two approaches to linking health research data will be reviewed. First is the traditional approach where SW technologies are not involved, while the second approach uses the SW as an integration tool to answer health questions. The main reviewed information in both approaches is made up of the used data types and linking methods, as well as some of the strengths and challenges mentioned in the literature.

2.5.1 Overview

The traditional linked data approach is defined as the process that involves aggregating data about a certain person from different resources while at all times maintaining his/her privacy (Kotwal *et al.*, 2016). To achieve this outcome means that the linking is usually based on a personal identifier and not performed with, or informed by, any other identifiers. Due to the sensitivity of the integrated information, the linking process is managed by independent data centres that provide customised linked datasets with encrypted identifiers for the research use.

The semantic linked data approach is the same type of literature that has been discussed earlier in the previous section of this chapter. It was found that one of the SW users with health data was answering health motivated questions. This type of research, that uses the SW as a tool for linking health data, will be discussed here and compared against the traditional linking data approach as both types of literature are trying to answer health questions, but by using different methods.

For reviewing the traditional approach, 11 key papers were chosen from different known health journals, such as BioMed Central (BMC) and British Medical Journal (BMJ), whereas the previously reviewed papers in the last section were chosen for the semantic approach, more details can be found in Appendix B. The traditional papers discussed different topics from the: a) clinical, b) pharmacological, c) epidemiological and d) management fields.

The topic of predicting the risk of developing a disease was discussed by several researchers like (Husain et al., 2012; Buckley et al., 2016; Clark *et al.*, 2016). Husain et al. (2012) and Buckley et al. (2016) tried to achieve risk prediction by aggregating available patients' data to produce a more comprehensive health history to analyse. Clark *et al.* (2016) tried to find associations between two different concepts; those of heart failure and cancer treatments. The researchers concluded that patients receiving cancer treatment had a high risk of developing heart failures and such treatment might cause mortality in certain cases. Cornish et al. (2015) aimed to find inter-dependent conditions between the duration of breastfeeding and IQ at the age of 15 by analysing linked data. In the pharmacology domain Juurlink et al. (2009) aimed to study the drugs interaction between the linked data of patients samples used both of proton pump inhibitors and clopidogrel. Falster et al. (2015) aimed to identify any factors contributing to positive early childhood development in Aboriginal children in Australia by finding associations between various administrative information and the children's developmental outcomes. In the health management field Falster, Jorm and Leyland (2016) aimed to identify and predict health services usage patterns to evaluate healthcare outcomes, while Ellis *et al.* (2013) tried to predict the healthcare expenditure for each individual by finding possible associations between costs and some patients' characteristics. The research team found that a patient's age and lifestyle were highly associated with his/her healthcare expenditure. Finally, Robertson et al. (2012) tried to investigate the burden of heart failure admissions on health services by finding associations between health burdens and patients' characteristics.

Regarding the semantic approach, there were nine reviewed papers around similar health fields, as in the traditional approach, see Appendix C for more details. Some of the discussed topics in this section were: a) predicting the risk of developing a disease, b) finding inter-dependent

conditions between health domains, c) discovering drugs interactions and d) epidemiological planning.

In the clinical field several researchers worked on identifying genotype-phenotype mappings to predict the possibility of developing a disease in a cohort (Pathak et al., 2012; Pathak, Richard C Kiefer and Chute, 2012). Moreover, Fernández-Breis et al. (2013) proposed a semantic linked model for predicting the risk of developing colorectal cancer. Shah et al. (2014) tried to find inter-dependent relationships between clinical and oral health domains by finding specific associations between the aggregated data from both fields. Other researchers tried to find patterns in patient's records to find drugs interactions (Pathak, Richard C. Kiefer and Chute, 2012; Jyotishman Pathak, R. C. Kiefer and Chute, 2013; Jyotishman Pathak, R. Kiefer and Chute, 2013). In the area of epidemiological planning Beyan et al. (2014) linked several historical records for individuals in order to analyse maternal and infants mortality rates.

From reviewing the chosen papers in the two approaches, it can be concluded that the questions asked regarding the different health topics are similar in aiming to identify patterns and correlations between the different concepts. However, it is important to note that the semantic linked data approach is still in its early stages. All the reviewed papers in this approach had a joint aim in testing the ability of the SW to answer their proposed questions; they were proof-of-concept studies rather than health studies. Therefore, it would be expected that the research literature at this stage will try to imitate or replicate some of the traditional health questions for testing purposes.

2.5.2 Data Types

Part of the reviewed information in both approaches was looking into the different data types and sources that were used. The data types in the traditional linked data papers were from three different broad sources: i) hospital-originated data, ii) government-originated data and iii) organisation-originated data.

The hospital data was used either in its electronic form as electronic health records (EHRs); or in other cases its paper-based form. Patients' records are cumulative, as they often consist of a range of reports such as: a) diagnosis, b) symptoms, c) treatments or d) laboratory results, as well as some personal demographic information. For example Denaxas *et al.* (2012) was profiling data resources for cardiovascular disease research and used many other longitudinal primary care data, such as: i) diagnoses, ii) symptoms, iii) prescriptions, iv) vaccinations and v) blood test results. Another type of hospital data was administrative data that was collected during hospital care processes. For instance Husain *et al.* (2012) used appointments information to count the

number of visits for each patient, while Falster, Jorm and Leyland (2016) evaluated health outcomes and performances by using: a) health services admissions, b) discharges and c) transfers information. Hospital financial records were another type of data used to relate to the costs, expenditures and burdens of providing health services (Ellis *et al.*, 2013). Hospital data seems to be the most accessed information type, as all the reviewed papers used some information from hospital records.

Another popular type of data was governmental registries with open access. An example of this source is births' and deaths' registries. Such census sources are one of the most common & popular types used in health research; all are cited by the following list of researchers (Denaxas *et al.*, 2012; Robertson *et al.*, 2012; Falster *et al.*, 2015; Clark, R. A. *et al.*, 2016; Falster, Jorm and Leyland, 2016). There was also a more specific type of census that was also used albeit less frequently, for example cancer registries (Buckley *et al.*, 2016; Clark, R. A. *et al.*, 2016).

Many of the clinical researchers use international known standards and terminologies in their linked data; such as the International Statistical Classification of Diseases and Related Problems (ICD) or the Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT). For example Clark *et al.* (2016) analysed the issue of heart failure in cancer patients after those patients were exposed to chemotherapy. All the diagnostic and clinical terms were coded in ICD terms.

Similarly, the semantic linked data approach used the same type of data, but in a structured format to ease and enable the integration process with other ontologies. In addition to the hospital, government and organisation datasets, the use of public ontologies from the LOD was very common. For example, DrugBank (Wishart *et al.*, 2006) and DailyMed (DailyMed, no date) are public pharmacological ontologies that contain a comprehensive amount of information about drugs, chemicals and proteins. The two sources were used in several clinical and pharmacology projects (Pathak, Richard C Kiefer and Chute, 2012; Jyotishman Pathak, R. C. Kiefer and Chute, 2013). Translational Medicine Ontology (TMO) (Dumontier *et al.*, 2010); Sequence Ontology (SO) (Eilbeck *et al.*, 2005); and Diseasesome (Goh *et al.*, 2007) are also some of the many online published repositories that contain significant biomedical information and knowledge. These biomedical ontologies played a major role in clinical and pharmacology papers such as that from Pathak *et al.* (2012); where TMO and SO were used in defining the mappings between phenotype-genotype associations.

Similar to the traditional approach, there was an extensive use of biomedical terminologies and clinical terms but in its structured semantic format. Some of the main terminologies that were used are Systematised Nomenclature of Medicine Standards (SNOMED CT) (Benson, 2012); the National Cancer Institute Thesaurus (NCI Thesaurus) (National Cancer Institute, no date) and

Logical Observation Identifiers Names and Codes (LOINC) (McDonald *et al.*, 1995). Terminologies and ontologies have a strong relationship, particularly where a terminology represents the name of a concept in an ontology (Noy and McGuinness, 2001). In addition to the biomedical ontologies, the LOD is full of ontologies from various domains that can be helpful for a health researcher. For example, one of the significant existing ontologies is DBpedia, which is responsible for presenting all the information on Wikipedia. All these ontologies can open many doors for different research and applications ideas.

2.5.3 The Linking Technique

The linking process in the traditional linked data approach is performed by aggregating personal data as well as protecting the privacy of the person. Hence, the linking process is managed by an independent data centre that generates an encrypted version of the linked data for the researcher's use. The linkage process in these data centres can be performed in two ways: a deterministic or a probabilistic approach. The linking in the deterministic approach is performed by using a person's identifier that can uniquely identify one individual from another. Denaxas *et al.* (2012) used this type of linking when integrating several personal records based on the individual's NHS number.

The second linking approach is by using probabilistic algorithms, where a mixture of personal identifiers, demographic markers and specific personal information can be used to identify an individual (Winkler, 2006). There is available matching software which supports probabilistic link making, like ChoiceMaker. An example of the use of this approach is the work of Falster *et al.* (2015), where the researchers used the name, date of birth, sex and address identifiers to probabilistically link records. On some occasions, a mixture of deterministic and probabilistic approaches can be used, as demonstrated by Husain *et al.* (2012).

The third party data centres play an important role in preparing the linked data for researchers. The data centres offer ready to use linked datasets for immediate access to the researcher. In other cases, the centres can offer to prepare, clean and link the data specifically for a certain project; for example: The Centre for Health Record Linkage (CHeReL) (The Centre for Health Record Linkage (CHeReL), no date). By using the linked data offered by the linking data centres, the researchers can benefit from having access to multiple administrative datasets and registries linked by personal identifier, without having to worry about the privacy issues associated with researchers trying to access personal data (Kotwal *et al.*, 2016). This benefit is because the linkage process is performed firstly by linking the different records per person either deterministically or probabilistically in an inside operation, to which the researcher has no access.

Secondly, an encrypted identifier (different than the original one) that represents each person (patient) is generated for the linked data. The reason to generate these encrypted identifiers is to maintain anonymity and to have a unified identifier for the multiple datasets (The Centre for Health Record Linkage (CHeReL), no date).

While the linking process in the traditional approach was performed on a personal level, only by third party data centres, the linking process in the SW approach is different. Firstly, the linkage is not just limited to personal identifiers. It can be performed on any declared entity that was assigned a URI. The possibilities and flexibility of linking heterogeneous datasets are constantly expanding.

Moreover, the linking process in this approach is performed by building up a knowledge domain that consists of different integrated resources. The integrated resources do not need to be fully implemented by the researcher, as many of the used resources in the reviewed systems were reused public ontologies that had been published in the LOD by others. Another property of the linking process in the SW approach is the ability to deal with private as well as public data, although it should be noted that access to the private datasets is restricted. For example, Pathak *et al.* (2012) managed to integrate public data from the linked open data source (LOD) with private patient records from the Mayo clinic's EHR systems. Another example comes from Odgers and Dumontier (2015) who linked private data with public data to be used in a translational research initiative. The private data was collected from the central repository for EHR data at the Lucile Packard Children's Hospital and the Stanford Hospital and Clinics (STRIDE); the public data was taken from different online biomedical linked datasets like the Systematised Nomenclature of Medicine (SNOMED-CT).

2.5.4 Strengths and Challenges

Health researchers used the traditional linked data approach because it improved the quality of the research by allowing for a breadth and depth overview of the tested sample as agreed by the following researchers (Juurink *et al.*, 2009; Denaxas *et al.*, 2012; Ellis *et al.*, 2013; Cornish *et al.*, 2015; Falster *et al.*, 2015; Clark, R. A. *et al.*, 2016; Falster, Jorm and Leyland, 2016; Hilder *et al.*, 2016). Having a wider overview of the linked datasets helped in identifying population-level patterns and trends (Robertson *et al.*, 2012; Ellis *et al.*, 2013; Falster, Jorm and Leyland, 2016). Moreover, an in-depth overview of the aggregated data gave a 'higher resolution more detailed picture' in a time and cost-effective manner (Denaxas *et al.*, 2012; Husain *et al.*, 2012; Robertson *et al.*, 2012; Falster *et al.*, 2015; Clark, R. A. *et al.*, 2016; Hilder *et al.*, 2016). Furthermore,

traditional linked data opened opportunities to assess real-world healthcare services and evaluate their outcomes (Robertson *et al.*, 2012; Falster *et al.*, 2015; Falster, Jorm and Leyland, 2016).

However, some researchers argued that the routine data used in this approach lacked quality, as it was collected for non-research purposes by non-specialists. For example, some data was missed out in the collection phase (Denaxas *et al.*, 2012; Robertson *et al.*, 2012; Buckley *et al.*, 2016; Clark, R. A. *et al.*, 2016), while other data was ambiguous (Denaxas *et al.*, 2012; Robertson *et al.*, 2012; Buckley *et al.*, 2016). Other researchers had some issues with the linkage process. As a result of these issues, there was some mismatching of links resulting in conflicting data (Denaxas *et al.*, 2012; Robertson *et al.*, 2012; Clark, R. A. *et al.*, 2016; Hilder *et al.*, 2016).

While many researchers had very positive opinions about using the traditional approach, the SW linking option seems to have received the same reactions. Some researchers predicted that by using the SW technology the process of research findings can be accelerated: a) due to effective approaches for knowledge discovery evident when examining biomedical data (Odgers and Dumontier, 2015) or b) due to the integration between public data sources from the LOD cloud and private institute-specific patient data (Pathak, Richard C. Kiefer and Chute, 2012; Jyotishman Pathak, R. C. Kiefer and Chute, 2013; Jyotishman Pathak, R. Kiefer and Chute, 2013). The simplicity and flexibility of domain representation has attracted many health researchers such as (Pathak *et al.*, 2012; Jesualdo Tomás Fernández-Breis *et al.*, 2013; Jyotishman Pathak, R. C. Kiefer and Chute, 2013; Odgers and Dumontier, 2015). However, the prime advantage of the SW is the availability of many open datasets in RDF containing different biomedical information relating to such issues as: a) genes, b) diseases, and c) drugs, which make it a very powerful tool for knowledge acquisition as agreed by the following citations (Pathak *et al.*, 2012; Pathak, Richard C Kiefer and Chute, 2012; Pathak, Richard C. Kiefer and Chute, 2012; Jyotishman Pathak, R. C. Kiefer and Chute, 2013; Jyotishman Pathak, R. Kiefer and Chute, 2013; Odgers and Dumontier, 2015).

Challenges to the SW approach have been minimal. One of the few issues raised was a technical challenge regarding the latency in responses to queries (Pathak *et al.*, 2012). Odgers and Dumontier (2015) faced other issues including: i) URI mismatching and ii) dirty clinical data.

2.5.5 Conclusion

This section reviewed some characteristics of two approaches for linking health data: traditional and semantic linked. The traditional linked data approach was performed by independent data centres that link several health-related datasets based on personal identifiers. The privacy of the participants in the linked data was protected by encrypting their identifiable information and producing new identifiers for research use. On the other hand, the semantic linked data approach

used the SW standards in defining and linking entities in a domain using the triples concept and URIs as global identifiers. In this approach the linking can be performed by anyone with some background knowledge of the SW tools and the linkage can be performed any defined entities with URIs.

The data types used in both approaches were similar in concept but with different formats; for example, both approaches used hospital generated datasets. However, in the case of the SW approach the data was provided from the institution supporting the researchers, unlike the traditional approach where the patients' private data is provided by a third party data centre. Another difference between approaches is the format of the linked data, where all the used datasets in the semantic approach were structured as RDF data. Although the semantic linked data approach cannot offer the ready-to-use personal linked datasets available with the traditional model, the SW technologies and capabilities are very promising due to the amount of available public linked data in the cloud. The LOD includes different types of ontologies from pharmacological, biomedical and medical ontologies to more general ones like DBpedia. The idea of leveraging the public data by linking it to patients' private data from hospitals' warehouses was popular in some projects. Pathak *et al.* (2012) linked private data from the Mayo clinic with public biomedical ontologies. It is important to note that the linking here is not necessarily performed on a personal identifier-level, as any defined resource in an ontology is represented by a unique URI that can be used for linking.

The traditional approach attracted many researchers in a range of health topics due to the efficiency of the sample size, when compared to stand-alone datasets. This advantage allowed the researcher to have a bigger picture and higher resolution perspective of reality, which can ease the process of identifying population-level trends. Moreover, by using the data centre's linked data, the researcher can have access to some encrypted private data that is difficult to be accessed due to privacy issues.

The semantic linked data approach is a new method that has come available to the health research community; however, it succeeded in gaining some positive attention. Some of the reasons mentioned in the literature for adopting the SW approach were: i) the simplicity and ii) the flexibility of the data model, as well as the accessibility to public data; an asset which can lead to better incorporation between the public and private datasets.

2.6 Summary

This chapter reviewed the main concepts of the semantic web (SW) and linked data (LD) principles in order to have a better understanding of the SW vision and how it may affect health research. The SW vision is concerned with structuring and linking the data on the web in order to make the web not just human-friendly but also machine-friendly. The main building block of the SW is the ontology which is a collection of connected and defined data entities representing a specific domain. LD is a concept introduced in the SW context referring to “the set of best practices for publishing and connecting structured data on the Web” (Bizer, Heath and Berners-Lee, 2009). The interconnected data (ontologies) in the cloud is called linked open data (LOD). Building ontologies needs several SW standards such as: URI, RDFS, RDF, OWL, SWRL and SPARQL.

In a preliminary review for the SW uses in health data, it was found that the SW is being used for different reasons such as to: a) publish, b) exchange, c) integrate, d) retrieve and e) monitor data. Another observation is that most of the available literature is technically-oriented. Health related issues were only chosen as examples as the researchers were not motivated by a health issue or question. The smaller part of the literature, where authors were motivated by a health-related aim, will be the focus of this thesis from now on.

Another interesting area was identified in regards to the techniques used for linking health data in the SW approach or ‘traditionally’. A traditional linked data approach was used in health research by linking at a personal level using a third party data centre. The semantic linking was performed at different levels mostly linked to open biomedical and clinical ontologies. Some of the strengths in this approach that attracted health researchers are: a) the efficiency of the sample size in comparing with stand-alone datasets and b) the accessibility to encrypted private linked data, which is considered a big challenge in health research. On the other hand, using the SW tools to link health data is gaining more attention for its simplicity, flexibility and accessibility to public data.

Chapter 3 Methodology

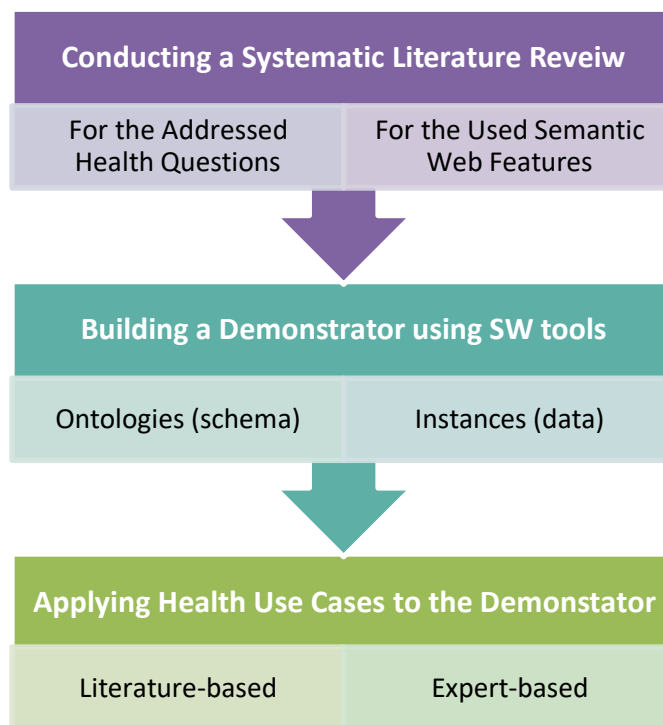


Figure 5: The methodology steps

The main aim of this empirical study is to understand the current and potential relationships between the SW and health research. In order to fulfil this aim and to answer the research questions, three types of methods were used. The first method is for addressing the current relationship between the SW and health research by conducting a systematic literature review. The systematic literature review is used to find: 1) the types of addressed health questions, 2) the main SW features, 3) the affordances and challenges and 4) opportunities and gaps in the literature.

The other two methods are more concerned with analysing the potential of this technology in health domain from a practical point of view. A demonstrator based on the SW technologies and representing a health-related topic was built, which was then used to deal with two health-based cases. The demonstrator, along with the two cases, aimed to answer third research questions; an aim that would be achieved by understanding how SW tools can be used in health research, as well as analysing the affordances and challenges evident when doing so. Figure 5 shows the three methods used in the thesis, while the next sections will discuss each method in detail.

3.1 Systematic Literature Review

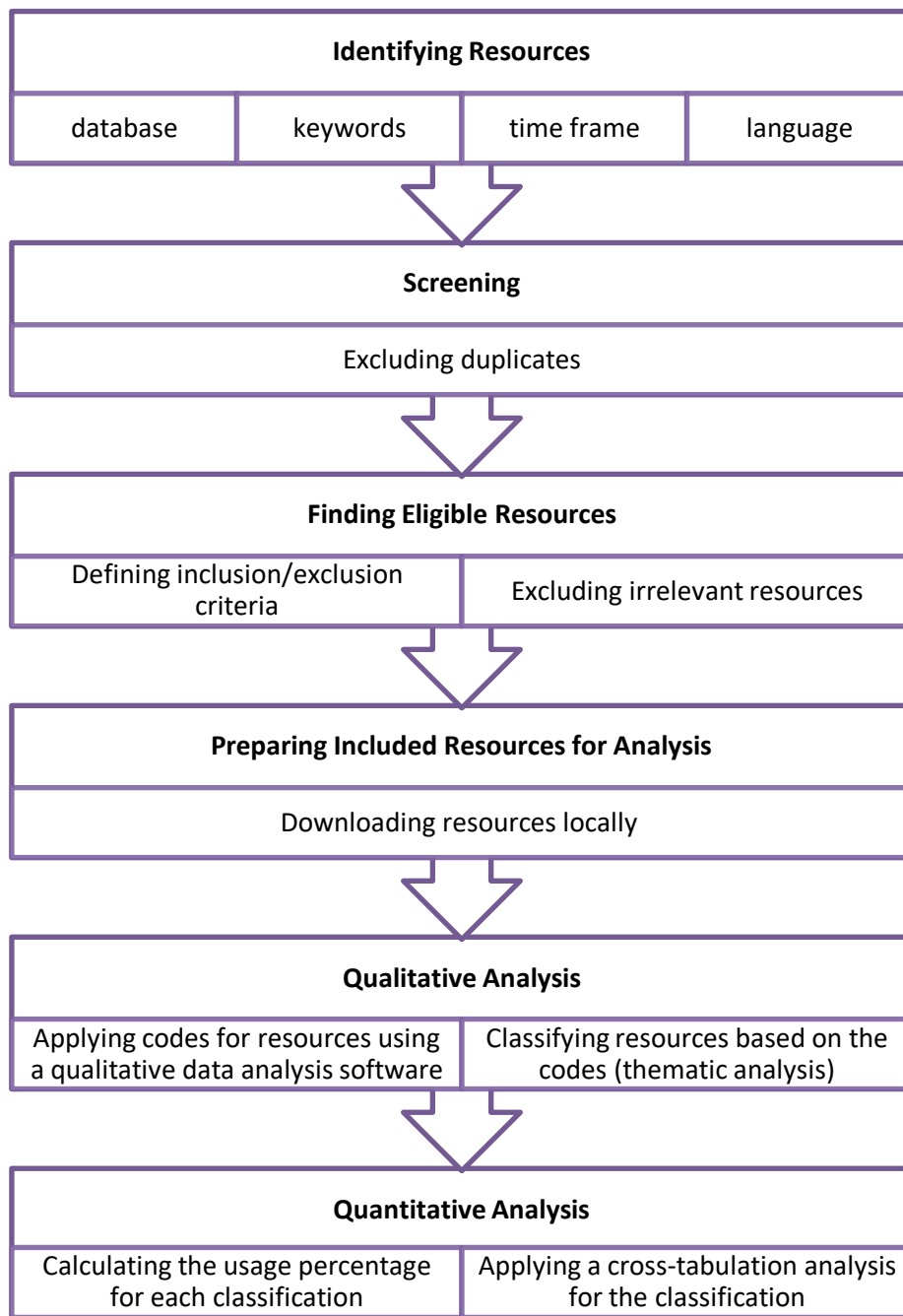


Figure 6: Systematic review steps

Systematic literature reviews are meant to map, classify and interpret available research regarding a specific topic or research area. They are used due to their ability in summarise grounds for a topic or highlight gaps in a research area (Kitchenham, 2004). The health literature exploiting the SW

technologies was systematically reviewed to classify and map the huge amounts of information available, in order to understand the research trends and themes.

The aim of this systematic literature review was to identify the addressed health questions and topics, highlighting the main SW features that were used and then analysing any affordances and challenges in that process. Therefore, the systematic review mapped the literature from two perspectives: the health questions that were addressed and the SW features that were used. The review aimed to answer the first research question in addition to part of the second and third questions.

The systematic review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) steps to guide the reviewing process (Moher *et al.*, 2009). In the PRISMA statement, there were four main phases in conducting a systematic review: i) identification, ii) screening, iii) eligibility and iv) inclusion. After identifying the included studies in the review, the information within these studies was analysed both qualitatively and quantitatively.

Figure 6 shows the steps followed in conducting the systematic review and the next sections explain the steps in greater detail.

3.1.1 Identification

The first step was to choose the searching database for the review. In a longitudinal and cross-disciplinary study conducted by Harzing and Alakangas (2016), a comparison across three popular searching databases namely Scopus, the Web of Science and Google Scholar was performed. The authors concluded that the three databases were sufficient and showed stable growth in the number of publications and citation. However, Scopus showed slightly higher average growth rate in the number of papers per academic at 2.7%, while Google Scholar and the Web of Science registered 2.5% and 2.2% respectively.

Moreover, Google Scholar was tested earlier in a pre-review of the uses of the SW in health data, section 2.4. Although Google Scholar might give a greater number of hits per search, the quality of retrieved results is lower than Scopus. In this preliminary experiment, there were many irrelevant results included. Harzing and Alakangas (2016) also support this idea by stating that Google Scholar's approach is to trace any information available on academic related websites and other low-quality resources such as blogs. Although the Google Scholar's approach expands the searching circle, it also lowers the accuracy and quality of the retrieved resources. In addition, Google Scholar does not

provide any filtering options for the sources' types, while Scopus provides many filtering options including specifying the documents' types.

The second step in identifying the included resources for the systematic review is choosing the searching keywords. In the pre-review mentioned in section 2.4, a simple query including the terms "semantic web" and "linked data" as well as "health" and "healthcare" was used. Many of the needed literature was excluded by using such simple query. The reason was that in some articles the semantic web or linked data was not mentioned explicitly in the article, but they were expressed implicitly by mentioning one or more of the SW standards such as "RDF" or the term "ontology". Therefore, the query was modified into including more semantic web keywords. The search keywords to describe the SW technology were: "semantic web", "linked data", "ontology", "RDF", "OWL" and "SPARQL", while the keywords to describe health subjects were "health" and "healthcare". All these keywords were put together in the following query:

```
TITLE-ABS-KEY ( ( "semantic web" OR "linked data" OR ontology OR rdf OR owl OR sparql ) AND
( health OR healthcare ) ) AND DOCTYPE ( ar OR cp ) AND PUBYEAR > 2000 AND LANGUAGE
( "English" )
```

The third step was to limit the search using some of the advanced filtering options provided by Scopus. Firstly, the searching time frame was limited from 2001, which was the year Berners-Lee wrote his first paper about the SW (Berners-Lee, Hendler and Lassila, 2001), to the date of the search day in July 2018 by using the *PUBYEAR* option. Secondly, the search was limited to articles written in English by using the keyword *LANGUAGE*. Moreover, the results were filtered based on the document type. The chosen included document types for this review were academic journals and professionally relevant conferences by using the term *DOCTYPE (ar OR cp)*. Finally, the search process was limited into looking into the titles, abstracts and keywords using the filtering option *TITLE-ABS-KEY*.

3.1.2 Screening

The screening phase included the included studies from the previous step by visually inspecting the titles, abstracts and keywords. Irrelevant articles were excluded, as were any duplicated resources.

3.1.3 Eligibility

The eligible resources for the systematic review were the ones that followed the exclusion criteria. The criteria included two points:

- 1) Exclude any article that is out of the SW context, such as articles that use traditional linked data, as discussed in the previous chapter.
- 2) Exclude any article that has a purely technical aim rather than a health-related one, an issue also discussed in the previous chapter.

3.1.4 Inclusion

After applying the inclusion/exclusion criteria, all included studies were downloaded locally to be prepared for analysis.

3.1.5 Qualitative Analysis

The qualitative analysis was done by using NVivo, a data analysis software (QSR International Pty Ltd., 2019). For each paper, several codes were added to identify: i) the main aim or ii) the question(s) addressed in the paper, iii) specific topic(s) discussed, iv) the SW tools used and v) SW features. After coding all the downloaded resources, a thematic analysis using a classification matrix was performed using the produced codes. The classification was done by visually analysing and aggregating similar aims and concepts together in order to produce primary and secondary research themes.

3.1.6 Quantitative Analysis

The quantitative analysis took place informed by the themes produced from the previous step. By counting the number of resources in each classification theme, and calculating the usage percentage, the data was represented as numerical tables and charts. In a second analytical step, a multi-dimensional cross tabulation analysis for the different classification themes was applied. The purpose of this tabulation analysis was to calculate the percentage of usage for each of the 'crossed' themes.

3.2 Building a Demonstrator

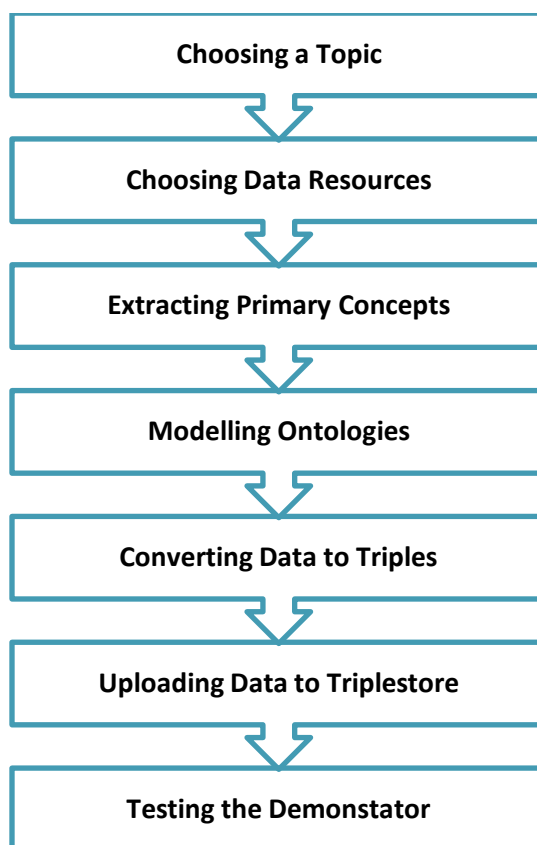


Figure 7: Steps to building a Semantic Web-based demonstrator

This section describes the methodology employed in designing and implementing the semantic web-based demonstrator (proof-of-concept model). The aim of building the demonstrator was to illustrate how the SW tools were used to represent a health-related topic, as well as to analyse any affordances and challenges faced in the process.

Figure 7 shows the steps that were followed in order to build the demonstrator and the next sections explain the steps in more details.

3.2.1 Choosing a Topic

Based on the outcomes of the systematic review and the identified gaps in the literature, a health topic was chosen to be demonstrated. The topic needs to be able to demonstrate the SW standards and features as well as to be accessible to open data on the web.

3.2.2 Choosing Data Resources

When choosing suitable open data from the previous step, it was also important to choose reliable data sources, as well as choosing datasets that were capable to be linked afterwards. The chosen data was then downloaded into a local machine and saved as a 'comma separated values' (CSV) file, if it was not RDF. All CSV files were prepared and cleaned to be converted into RDF files.

3.2.3 Extracting Concepts from the Data

The extract, transform and load (ETL) process was followed in building the main body of the demonstrator. The ETL process is responsible for extracting data from various sources, cleaning and customising the data to prepare it to be uploaded into a data warehouse (Skoutas and Simitsis, 2006).

This step was the beginning of the ontology design process. Firstly, the downloaded datasets were analysed and the primary concepts were extracted. The extracted concepts were meant to be defined as classes or data properties in the designed ontology. The other type of information to be extracted from the datasets was the relationships between concepts to be represented later, as either data or object properties.

3.2.4 Modelling Ontologies

Protégé, a software for designing ontologies developed at Stanford University (Musen, 2015), was used for developing the ontologies employed in this research. The chosen SW standard for building the ontologies was the web ontology language (OWL). The ontologies were based on the extracted concepts and the relationships between them. It is important to note that the aim for building the demonstrator is to prove the feasibility of using the SW standards to represent a health topic. Therefore, the ontology's design was primitive and only meant to represent the basic concepts in the domain.

3.2.5 Converting Data into RDF

After designing the ontologies, the next step is to add instances, in the form of the actual data, to these ontologies. Using OpenRefine (previously Google Refine) combined with RDF extension (Google, 2010), the tabular CSV data was converted into structured RDF files. OpenRefine converts the data depending on and informed by user-defined mapping rules. These rules were defined based on the ontology's design to produce compatible instances with the defined schema.

3.2.6 Uploading Data to Triplestore

The chosen triplestore was the open version of GraphDB from Ontotext (Ontotext, 2017). GraphDB was chosen due to its capability to handle large amounts of triples. For example, 17 billion triples were handled in the UNIPORT study. GraphDB was combined with a user-friendly interface that allows for uploading the triples' repository with OWL or RDF files. The OWL files for the ontologies were uploaded first; then the converted RDF files were uploaded. The repository, where all the uploaded files were saved, was located locally: in the researcher's computer.

3.2.7 Testing the Demonstrator

GraphDB was also combined with a SPARQL endpoint for querying the available repository. Several testing queries were run across the ontologies to test different SW features.

3.3 Health Use Cases

This section aims to use real case scenarios in the demonstrator to try more of the SW features. By applying two health use cases for the SW demonstrator to deal with, lessons regarding the affordances and challenges in answering health questions using SW tools can be explored and analysed.

The health use cases were chosen via two different approaches. The first one was inspired from the traditional health literature and tried to replicate some of the results. The second case was chosen from health researchers' suggestions in a focus group. Both cases were translated into SPARQL queries and were then run across the implemented demonstrator. The steps followed in both 'use' cases are discussed in the next sections, as shown in figure 8.

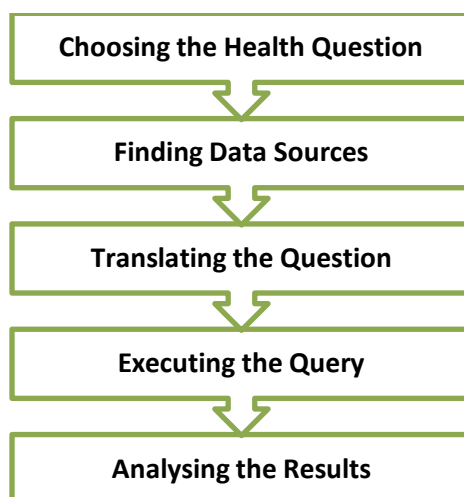


Figure 8: Steps for applying health use cases

3.3.1 Choosing the Health Question

Choosing the health question for the use cases was done differently in each of the two approaches. The chosen topic and question in the first use case was borrowed from the traditional health literature. The needed condition was to select a complex question that involved integrating different heterogeneous datasets in order to answer it. The reasoning behind this approach was to demonstrate as many of the SW technology's abilities as possible.

In the second use case, the health question was taken from suggestions offered by a focus group of health researchers. The aim of the focus group was to allow the participants to brainstorm some interesting health questions from their point of view that can be presented to the newly built demonstrator. The opinions of the health researchers were meant to expand the understanding of possible scenarios where the demonstrator would be useful. The output of the focus group would be a list of suggestions (questions) that were interesting from those experts' perspectives. These questions were filtered to one chosen question that was the most suitable for use with the demonstrator. The filtering criteria were based on the ability of the question to highlight the SW's capabilities, with the possibility of applying it to research issues located in the field of health care.

3.3.2 Finding Data Sources

Because both questions were chosen carefully, the needed data for answering them was mostly available in the built demonstrator. However in other cases, there was a need locate extra data in addition to the open data.

3.3.3 Translating the Question

After preparing the demonstrator for answering the questions and linking it with any needed extra data, the chosen questions were divided into smaller parts to ease the translation process. The questions' parts were then translated into small SPARQL queries nested in a bigger query. Thus, there was one nested query for each question in the use cases.

3.3.4 Executing the Query

When the queries were ready, they were executed in the GraphDB's SPARQL endpoint. The queries' results were saved locally for analysis.

3.3.5 Analysing the Results

The results of the first use case will be analysed in a manner corresponding to the paper that was replicated. However, in the second use case, the health experts will be consulted again for their opinions regarding the results.

3.4 Summary

This chapter discussed the followed methodology evident in this thesis. The three methods used in this thesis were:

1. Conducting a systematic review of the health literature using the SW.
2. Building a demonstrator based on SW technologies representing a health topic.
3. Applying health-related cases across the demonstrator.

The systematic review aimed to map and to classify the literature based on the health questions that were addressed and the SW features that were used. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) steps were followed in order to implement the reviewing process (Moher *et al.*, 2009). The four PRISMA steps were: i) identification, ii) screening, iii) eligibility and finally iv) inclusion. After identifying the included articles in the review, a mixed analysis approach was performed.

The demonstrator was built as a proof-of-concept model to illustrate the use of the SW's standards for researching a health topic and its data and to analyse any affordances or challenges in the process. Choosing a topic for the demonstrator was informed by the topics identified in the systematic review. Building the ontologies and feeding them with instances followed the extract, transform and load (ETL) process (Skoutas and Simitsis, 2006). GraphDB was chosen as the triplestore with which to save and load all needed data (Ontotext, 2017).

Two health use cases were applied to the demonstrator in order to test the SW's features and analyse any affordances or challenges in the process. The first case was inspired by the traditional health literature and tried to replicate some of the results. The second case was chosen from suggestions offered by a focus group of health researchers. Both cases were translated into SPARQL queries and run across the ontologies in the SW demonstrator. The results of the queries were analysed in corresponding to the replicated paper in the first use case and according to the experts opinions in the second case.

Table 1 shows a summary of the methods used in answering the research questions and their outcomes.

Research Aim:	<i>Understanding the relationship between the semantic web and health research currently and potentially</i>		
Research Questions:	<i>What are the main health questions being addressed in health research employing semantic web technologies?</i>	<i>How are the semantic web features being used in health research?</i>	<i>What are the affordances and challenges in employing the semantic web for health research?</i>
Methods:	<ul style="list-style-type: none"> • Systematic review 	<ul style="list-style-type: none"> • Systematic review 	<ul style="list-style-type: none"> • Systematic review • Demonstrator • Use cases
Expected Outputs:	The addressed health aims and questions in the literature	The identified SW features in the literature with their usage rates	The identified affordances and challenges from literature and practice

Table 1: A summary of the research questions, methods and expected outputs

Chapter 4 **The Semantic Web's Uses in Health Research: a Systematic Review**

The previous chapters of this thesis introduced and explained the aim of this thesis and the methods used in achieving this aim. Sections 2.4 and 2.5 in the second chapter started exploring the inter-related literature covering the fields of SW and health research. The SW's uses in health research, from a data cycle perspective, were reviewed as well as the uses of the traditional and semantic LD approach in health research. One of the main findings in those reviews was that using SW and LD tools in health research captured the attention of researchers. The existence of proof-of-concept studies, or papers testing the SW approach in answering a specific health question, was evident. While most of the available literature is of a technical nature, this study focuses on the smaller portion of the literature that addresses health questions and then uses the SW as a tool for answering them; this research is not motivated by a purely technical aim.

Motivated by how the SW tools help health research in managing data, this chapter continues investigating the uses of the SW in health research by proposing a deeper investigation of the literature. The literature is systematically reviewed for the purpose of identifying research gaps, trends, challenges and affordances of the SW in the health domain. The first part of this systematic review aims to address health questions and topics in the literature. The second part of the review attempts to understand the different uses and features of the SW in general and across the different health research themes.

4.1 Introduction

Massive amounts of health data are available publicly and more are produced in private health institutions and by self-monitoring devices every day. However, heterogeneous data sources, incompatible formats and a lack of data interoperability between systems all hinder attempts to exploit and integrate this data (Sagotsky *et al.*, 2008).

It has been suggested that the semantic web (SW) and linked data (LD) can offer solutions relating to: i) data structuring, publishing, integrating and sharing on the web (Berners-Lee, Hendler and Lassila, 2001; Berners-Lee, 2006). The SW offers standards for representing data e.g. RDF and OWL and enables the linking of data by following the LD principles (Berners-Lee, 2006). The 'linked data principles' are:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs, so that they can discover more things

These standards and rules can offer a platform for health researchers to use in managing their data. It is suggested that the SW is gaining popularity in the health field, as mentioned in a state-of-the-art review of the SW in healthcare (Zenuni *et al.*, 2015) and also by the findings in Chapter 2.s

4.1.1 Similar Reviews

Zenuni *et al.* (2015) reviewed the literature from three perspectives: i) ontologies and data repositories, ii) applications and user interfaces and finally iii) data mining and analytic approaches in the SW for the healthcare domain. It was found that the SW is valuable for the health domain. However, there were challenges in mapping from non-semantic formats to ontological concepts, as well as the difficulties presented by ontological maintenance. The paper reviewed the healthcare literature in general, but there were also other reviews focused on specific health domains such as: a) translational medicine or b) health informatics.

An example of focused reviewing is the work of Machado *et al.* (2013), in which 11 translational medicine applications using SW technology were reviewed. The applications covered different cases of medical use in: a) pharmacology, b) neuroscience, and c) cardiovascular diseases. The authors also reviewed the public and private resources used in the applications, as well as the SW standards and knowledge discovery approaches that were used. It was concluded that the SW technologies are fulfilling their role in data integration and exploration, but that those technologies need to be combined with the process of: i) defining mapping between resources and ii) re-using resources.

McArthur (2009) conducted a survey to find opportunities for health information professionals to play a role in developing the SW. This survey was different from others because the targeted readers were health information professionals. The author found five key themes for the health information professional and librarians to support SW development: i) ontological development, ii) knowledge translation, iii) information retrieval, iv) scientific publishing and v) resource classification and indexing.

The last review was the work of Eysenbach (2003) in the field of consumer health informatics. The author reviewed the literature from two perspectives: a) the main building blocks of the SW and b)

the prospects of the SW in the field of consumers' health informatics. The author succeeded in outlining several benefits and challenges of using the SW in the field of knowledge translation for dealing with consumers' health information. Some of the benefits mentioned were: i) improving the abilities of search engines, ii) guiding users to relevant health information and iii) better automatic translations. On the other hand, there were many issues regarding data accessibility, quality and privacy. In conclusion the author suggested that the SW may open more opportunities for health information consumers to find and aggregate information. However, there is a fear that this possibility may lead to overlaying on the web as a source of health information and the role of physicians may be negatively affected in some ways.

4.1.2 The Study's Contribution

This systematic review aims to identify research gaps, trends, challenges and affordances within the field of SW literature that addresses health questions. Firstly, a 'health aims' taxonomy is presented that includes two levels of classification for the literature: i) the general health theme and ii) the specifically addressed health questions. This taxonomy can help researchers to understand the types of questions that the SW could help to deal with. Accompanying the health taxonomy is a quantitative analysis focused on the usage percentages for each classified theme and topic which gives an overview of the research trends and gaps in this specific area of the literature.

The second part of the systematic review aims to understand the different uses and features of the SW in general and across the different health research themes. A second taxonomy for the literature is presented that classifies the literature according to the particular SW features discussed. This taxonomy is also supported with a quantitative analysis that calculated the percentages of use for each SW feature.

In addition to the two taxonomies, a cross-analysis for the use rates between the SW features employed across health themes is produced. This analysis helps researchers to understand which SW features are popular in each health theme and which are not. Moreover, the analysis highlights the impact of the characteristics of addressed health questions in raising or lowering the rates of usage for specific SW features. Finally, this systematic review offers an understanding of the affordances or challenges facing the SW in health research. This understanding is based upon the variance in uses of the specific SW features across different health-related questions addressed that are addressed in the literature. Thus, this systematic review is different from the previously mentioned reviews: a) in terms of the breadth of the health topics discussed and b) the specificity of the discussed affordances and challenges according to the health questions addressed in the literature.

4.2 Methods

The approach followed for systematically reviewing the health literature that focused on the SW was the 'preferred reporting items for systematic reviews and meta-analysis' (PRISMA) (Moher *et al.*, 2009). In the PRISMA statement, there were four main phases in conducting a systematic review: i) identification, ii) screening, iii) eligibility and finally iv) inclusion. Section 4.2.1 discusses the details of each phase, while section 4.2.2 discusses the data analysis process.

4.2.1 Systematic Review Phases

Following PRISMA recommendations, four main steps were followed in choosing the articles for reviewing. Figure 9 shows the details of the four steps.

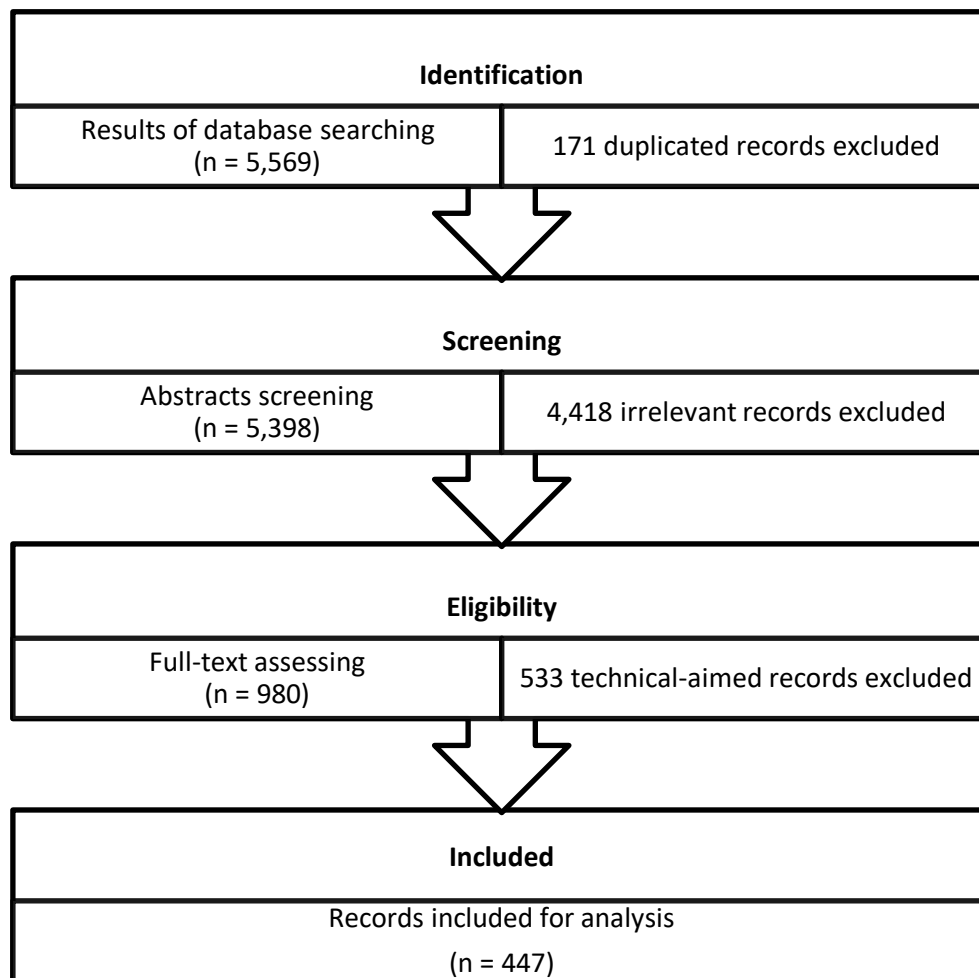


Figure 9: The four PRISMA steps followed in the systematic review

4.2.1.1 Identification

The first step in planning for the systematic review was to choose the searching database. In this case, Scopus was used because of its multidisciplinary nature that is essential in the SW and interdisciplinary health topics. Moreover, Scopus showed slightly higher consistent growth for both publications and citations in comparison to the Web of Science and Google Scholar (Harzing and Alakangas, 2016).

In Scopus, there were several filtering options for the search process. The main step in searching was to choose a combination of keywords that describe the intended literature. The chosen keywords were: "Semantic Web", "Linked Data", "ontology", "RDF", "OWL", "SPARQL", "health" and "healthcare". All these keywords were put together on the following query:

```
TITLE-ABS-KEY ( ( "semantic web" OR "linked data" OR ontology OR rdf OR owl OR sparql ) AND ( health OR healthcare ) ) AND DOCTYPE ( ar OR cp ) AND PUBYEAR > 2000 AND LANGUAGE ( "English" )
```

The second chosen searching option was limiting the searching time frame from 2001 until the date of the searched day, in July 2018. The beginning year of the time frame, 2001, was chosen because it was the year when Berners-Lee wrote his first paper about the SW (Berners-Lee, Hendler and Lassila, 2001).

Finally, the search process was limited to titles, abstracts and keywords for articles published in journals or conferences in English. Any other languages or types of resource such as books or posters were excluded.

After refining these filtering options on Scopus, the total number of resulted hits was 5,569. These hits / records were downloaded into EndNote (Thomson Reuters, 2016), a reference management software, for initial sorting and filtering. 171 records out of the 5,569 were excluded due to duplication, which left 5,398 records ready for the next step.

4.2.1.2 Screening

For the 5,398 records, a visual screening of the titles, abstracts and keywords was performed to exclude any irrelevant articles. The inclusion and exclusion criteria were based on the meaning of the search terms in the article. Some of the used keywords in the search query had other meanings not in the SW context. For example, many articles were excluded as they discussed 'Linked Data' as a tool for integrating data; **not** the Linked Data movement as in the SW context. Other records were

excludes because the keyword 'ontology' did not refer to the SW tool. After applying the criteria 4,418 records were excluded, leaving 980 articles ready for the next step.

4.2.1.3 Eligibility

In this phase, a deeper visual reviewing of the articles was performed to find the eligible ones. The second condition in the inclusion/exclusion criteria was to exclude any articles that have a purely technical aim, rather than a health-related one. For example, the problem of data interoperability between biomedical resources was considered as a technical problem. However, if the problem discussed interoperability between drugs resources in order to discover new drug interactions, then it was considered to have a definite health-related pharmaceutical aim and was included. The number of records excluded after this step was 533 out of 980.

4.2.1.4 Inclusion

The final result after all exclusions was 447 qualified articles that discussed the use of the SW for a health-related aim. A comprehensive review of these papers was conducted in order to find the main health questions that were addressed and the main SW features that were used.

4.2.2 Analysis Methods

The analysis process followed a mixed approach between quantitative and qualitative; details of which are in the following sections.

4.2.2.1 Qualitative Analysis

For the qualitative analysis, the records were uploaded into NVivo (QSR International Pty Ltd., 2019), a qualitative data analysis software. NVivo was used for extracting and coding the papers' contents. For each paper, several codes were added relating to the main aim, the question addressed, specific topic discussed and the SW tools and features that were used. After collecting all these data, a spreadsheet was employed to register the data as a matrix. The matrix can be found in Appendix D.

Through a repetitive visual analysing and aggregating of similar aims and concepts, the researcher managed to provide a classification for the papers. The papers were classified by the discussed health aim and also via the health question being addressed. A two-level hierarchical classification was provided in the form of a two-level taxonomy. For example, the work of (Natsiavas *et al.*, 2018) aimed to detect adverse drug events by building an ontology for representing and sharing information in the pharmacology field. Moreover, Pathak, Kiefer and Chute (2013) aimed to find

potential adverse drug events , specifically drug-drug interactions, for prescribed cardiovascular and gastroenterology drugs in the Mayo Clinic database. Both papers discussed the Adverse Drug Events (ADEs) topic under the pharmacology umbrella. Thus, these two papers along with other similar ones were grouped together under the 'Finding adverse drug events' category.

Following the same methodology, another taxonomy was provided for the SW features used in the literature. These features were either mentioned explicitly in the papers or implicitly deduced from the context.

It is important to note that both taxonomies were built by the researcher and then checked and verified by another independent researcher in order to check their consistency. Moreover, the nature of the categories in both taxonomies is not mutually exclusive, meaning that an article may fit in more than one category. For the health-aim taxonomy, each paper was classified under the best-fit classification, with no repetition being allowed. However, the taxonomy of the SW features had joint sets and so repetition was allowed. In fact, it was the common case to have many features used in one paper.

4.2.2.2 Quantitative Analysis

For the health-aims classification, a quantitative analysis was performed to identify the research trends and any possible gaps in the literature. Two main steps were performed over the chosen papers. Firstly, the number of papers in both of the main categories and sub-categories were identified and then the percentages were calculated. Moreover, two pie charts showing the division of literature by the health aims and addressed questions were provided.

Similarly, the second analysis was performed by identifying the most common SW features used in a given paper. The percentages of the features' usage were calculated along with two bar charts showing the usage rates. The bar charts were chosen to represent the usage rates because the categories here were joint sets and repetition was allowed.

Finally, a two-dimensional cross-tabulation analysis was used to show the rate / frequency of employing the SW features across the health aims. Firstly, the SW features mentioned in each health aim were counted. Then the percentage of a feature's use in each health aim was calculated. These percentages were classified into four usage rates or range values for simplicity. The aim of this analysis was not to focus on the exact number of papers using a specific feature, but to provide an overview of the research trends in this field of literature.

4.2.3 Limitations

The methodology of the study was limited by the defined inclusion/exclusion criteria, as well as the available articles identified via the search database (Scopus) that was used. Thus, the analysis and results achieved were limited by the collected sample data which did not necessarily cover the whole interdisciplinary field of health research and the SW. Moreover, the analyses, especially the qualitative one, was human error prone as the categorisation process was performed under, and informed by, the best-fit policy as mentioned earlier. Regarding SW features that were mentioned in the literature, occasionally they were not identified explicitly by the author(s). When this omission was the case, the targeted feature was discarded in the analysis, which meant that the analysis showed the minimum number of the used features and not necessarily the actual number. The purpose of the usage rate was to give a fair approximation of the research trends.

4.3 Results

The results of systematically reviewing and analysing the joint literature of health topics using the SW technology are discussed in the next sections. The first achieved result is the health aims taxonomy that maps and classifies the literature based on the health questions that were addressed in the reviewed papers. The taxonomy chart, definitions, examples and usage analysis are discussed in section 4.3.1. The second achieved result is the SW features taxonomy that maps the main used SW features in the literature. The taxonomy along with the definitions and usage analysis are discussed in section 4.3.2. Finally, a cross-analysis matrix, which shows the variance and similarities of SW feature usage rates across health aims, is shown in section 4.3.3 accompanied by a series of examples from the literature for mapped papers on the SW features' taxonomy.

4.3.1 Health Aims Taxonomy

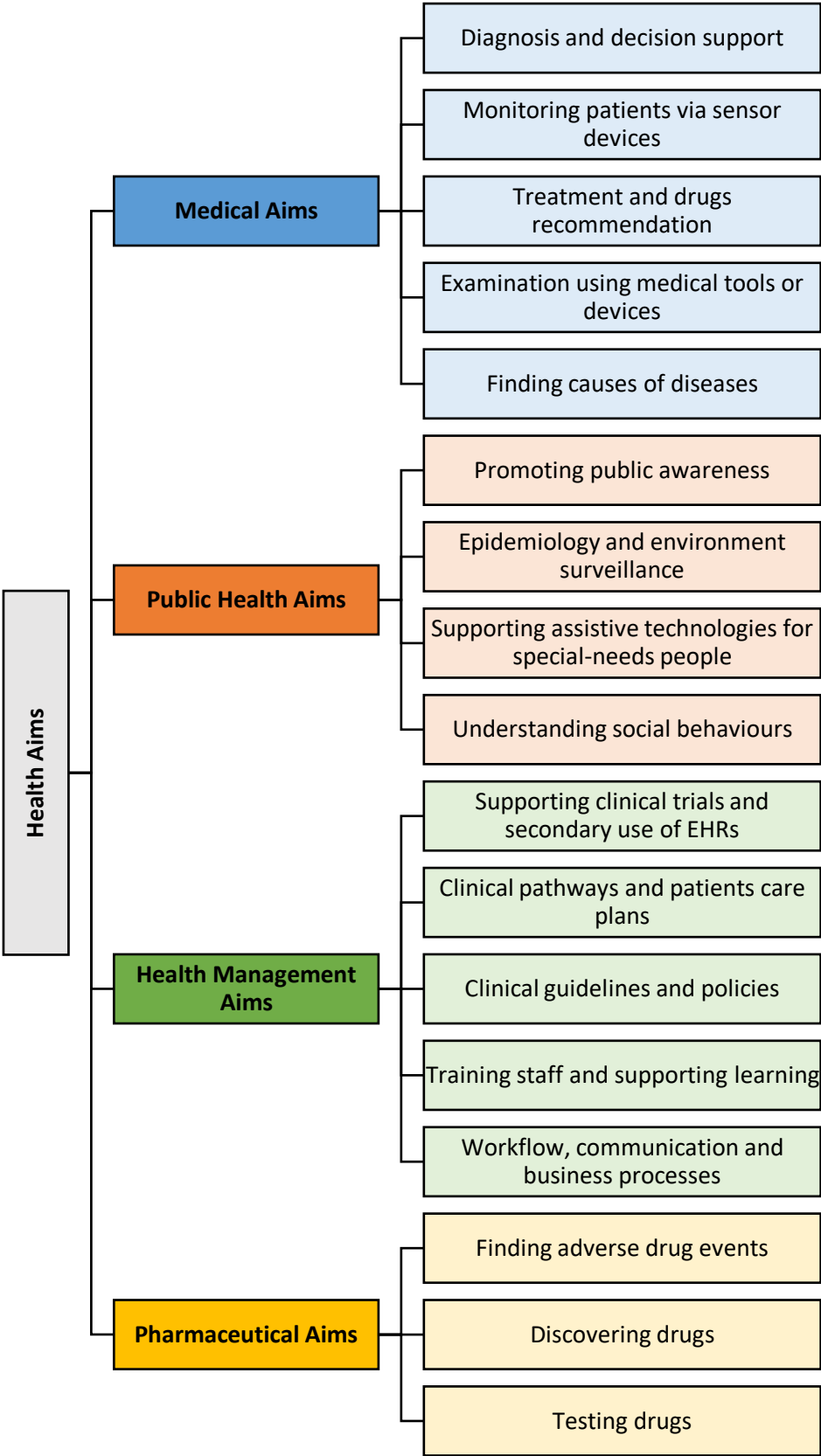


Figure 10: The health aims taxonomy

The health aims taxonomy was built by relying on the health questions that were addressed in the reviewed literature. Figure 10 shows the health taxonomy that represents the reviewed literature. The literature has been divided into four main health aims and 17 addressed health questions under those four aims. The following section will discuss the four main aims and addressed questions.

4.3.1.1 Definitions of Aims

Medical Aims

This aim is related to medical science, which is according to the Oxford Dictionary is defined as “The science or practice of the diagnosis, treatment, and prevention of disease” (Oxford University Press, 01/05/2019). The papers that discussed medical aims were categorised into five sub-aims: i) diagnosis, ii) monitoring, iii) treatment, iv) tools and v) causes of diseases.

Diagnosis and decision support

Diagnosis is one of the main activities in the medical field. It relies on logic to decide the patient’s condition based on known symptoms and signs. This topic was discussed by a big number of papers in the literature. For example Pathak *et al.* (2012) aimed to identify genotype-phenotype associations to predict the possibility of developing type 2 diabetes and hypothyroidism in a cohort.

Monitoring patients via sensor devices

This topic relies on the advances of technology employing sensors that are used to monitor patients in the medical field. Puustjarvi and Puustjarvi (2015) is an example of a paper discussing the monitoring of patients. This paper discussed monitoring health data generated from sensors at home. Another example is Stavropoulos *et al.* (2016) where the authors report their aim to build a monitoring system to support the elderly who are suffering from dementia. This initiative would allow such patients to have an independent life by providing their care givers with feedback via their measurement data being generated from wearable sensors.

Treatment and drug recommendations

After diagnosing a patient’s health issue, their suggested treatment is the next step. There are some papers that discussed the idea of determining the best treatment plan for a patient (Rung Ching Chen *et al.*, 2012). The authors built a recommendation system that helps doctors to decide which anti-diabetic drug is most suitable for each patient. Another example is Ullah *et al.* (2017), where the authors built a system that recommends medicine based on symptoms collected from sensor devices.

Examination using medical tools or devices

One of the initial clinical processes with a patient is examining the symptoms they have to determine their cause. Maragoudakis, Maglogiannis and Lymberopoulos (2008) discussed this idea when examining skin lesion images. The authors developed an ontology for skin lesion images to be used in decision support systems. Another example of an initiative is from Singh *et al.* (2013), where the authors implemented a retrieval system for patients' records based on extracting features from medical images.

Finding causes of diseases

Finding causes for health defects is one of the most interesting biomedical research aims. Gudivada *et al.* (2008) studied the causes of genetic diseases. The authors aimed to find disease-causal genes whose mutations cause health defects.

Public Health Aims

According to the World Health Organisation (WHO), the term 'public health' is defined as "the art and science of preventing disease, prolonging life and promoting health through the organized efforts of society" (World Health Organization, 01/05/2019). This aim discusses public health related topics, such as promoting awareness, epidemiology, assistive technologies and acceptable social behaviours.

Promoting public awareness

Improving the public's health knowledge is one of the most popular questions discussed in the public health aim. One example of promoting awareness is encouraging people to have healthy diets; a point made by Wang *et al.* (2010). This paper offers an example of a health information system that answers users' queries regarding their diary and health status, which eventually can lead to a better healthier lifestyle and therefore disease prevention. Another example in this 'awareness field' was from Velmurugan and Ravi (2016), who aimed to improve people's education on allergies by developing an allergy information ontology.

Epidemiology and environment surveillance

This topic included papers that focused on studying and analysing the distribution of diseases. Sometimes, the study would take place in a closed environment, such as a hospital, rather than in a

broader public scope. For example Shaban-Nejad *et al.*(2012) studied how advances in semantic technology would improve the analysis and detection of ‘hospital-acquired infections’ (HAI).

Supporting assistive technologies for special-needs people

Assistive technologies is a growing interdisciplinary field that is interested in employing technology to help special-needs people, such as the elderly or individuals with disabilities. For example, Baldassini *et al.*(2017) aimed to support the elderly in maintaining a healthier lifestyle in their own living environment by designing a virtual reality system for exercising. The system was based on SW technologies that provide users with tailored physical exercises based on their health conditions.

Understanding social behaviours

This question includes studies that investigate public opinions and behaviours about a certain issue. For example, Birjali, Beni-Hssane and Erritali (2017) analysed semantic suicidal sentiments expressed in social networks.

Health Management Aims

This aim discusses different business processes in a clinical context. The term ‘healthcare management’ is defined as “supervising the functions of a healthcare organization” (Sifaki-Pistolla *et al.*, 2017). The tasks expected from a health manager are directing and leading healthcare units to provide the best healthcare services available (Sifaki-Pistolla *et al.*, 2017). Five different topics were found under this aim: i) secondary use of EHRs, ii) clinical pathways, iii) clinical guidelines, iv) training staff and v) business processes and workflows.

Supporting clinical trials and secondary use of EHRs

Clinical trials are the type of experiments and observations involving patients in a clinical environment. Many researchers are using patients’ records such as their electronic health records (EHRs) as a source for patients’ information and selection. For example, Chondrogiannis *et al.* (2017) demonstrated a system that automatically selects patients for recruitment to clinical trials by semantically representing the eligibility criteria for the study.

Clinical pathways and patients care plans

Clinical pathways have been defined as ‘the optimal multidisciplinary care process performed by a team of health care professionals for a particular diagnosis or procedure’ (Ye *et al.*, (2008). This

question includes articles discussing finding / identifying optimal clinical pathways or care plans for patients. For example, Alexandrou *et al.*(2012) designed an ontology named SEMPETH (SEMantic PATHways) that conceptualised the domain of clinical pathways. Another example is from Wang *et al.*(2015), where the authors aimed to enhance a treatment's quality by creating personalised clinical pathways. They used the SW to achieve data interoperability between EHRs and their proposed clinical pathway ontology.

Clinical guidelines and policies

This type of questions includes any study discussing clinical policies and regulations. An example is the work of Puustjarvi and Puustjarvi (2016). The authors aimed to keep clinicians informed about relevant clinical guidelines by querying two linked ontologies: clinicians' profiles and clinical guidelines.

Training staff and supporting learning

This type of questions includes studies that discuss clinical learning for physicians or hospital staff. An example of this topic is the work of Bajenaru and Smeureanu (2015) that aimed to train healthcare managers by developing an e-learning system.

Workflow, communication and business processes

This topic is considered together with managing healthcare processes. For example Dang *et al.* (2008, 2009) aimed to enable healthcare users and administrators to create and manage medical workflows and also to personalise them. Another example is Kaddari, Malki and Elmdeghri (2016), where the authors addressed the difficulties in communicating between different healthcare actors. They implemented a global unified system that organises medical workflow between stakeholders.

Pharmaceutical Aims

The meaning of the word 'pharmaceutical' is "Relating to medicinal drugs, or their preparation, use, or sale" (Oxford University Press, 01/05/2019). Most of the papers in the pharmaceutical aim discussed adverse drug events, such as drug interactions. Discovering and testing drugs were other topics that were discussed.

Finding adverse drug events

According to the U.S. Department of Health and Human Services (no date), an adverse drug event (ADE) is an injury caused from medical intervention related to a drug, regardless of whether it was: a) medication error, b) adverse drug reaction, c) allergic reaction, or d) an overdose. An example of this topic is the work of Natsiavas *et al.*(2018), where the authors built up an ontology for the communication and representation of pharmacovigilance signals to help in preventing any possible adverse drug events.

Drugs discovery

This topic focuses on discovering new purposes for existing drugs or discovering new drugs. For instance McCusker *et al.* (2014) aimed to discover new purposes for existing drugs and new links between existing drugs and diseases by building a drug repurposing semantic framework.

Testing drugs

Testing drugs is one of the most primary processes in producing drugs; however, few studies discussed this topic in the reviewed literature. One of those few was Kohonen *et al.*(2013), where the authors developed a data warehouse named ToxBank for the purpose of representing and supporting test replacements for repeated dose systemic toxicity testing on animals.

4.3.1.2 Usage Analysis

After defining the health aims and the different types of addressed questions immediately above, this section focuses on analysing the usage rates for each of the defined aims and questions. The usage rate is calculated by counting the number of papers which were categorised under a certain category. Table 2 shows the quantity and the percentage of the categorised papers in each of the four identified health aims and 17 sub-aims (addressed questions).

Main Aim / Sub-aim	Main Aim N (%)	Sub-aim N (%)
Medical Aims	216 (48%)	
<i>Diagnosis</i>		119 (27%)
<i>Monitoring patients</i>		35 (8%)
<i>Treatment</i>		27 (6%)
<i>Medical tools</i>		20 (4%)
<i>Diseases' causes</i>		15 (3%)
Public Health Aims	112 (25%)	
<i>Public awareness</i>		57 (13%)
<i>Epidemiology and surveillance</i>		29 (6%)
<i>Assistive technology</i>		18 (4%)
<i>Social behaviours</i>		8 (2%)
Healthcare Management Aims	83 (19%)	
<i>Secondary use of EHRs and trials</i>		22 (5%)
<i>Patient care plans</i>		16 (4%)
<i>Guidelines and policies</i>		16 (4%)
<i>Learning and training</i>		15 (3%)
<i>Workflow and communication</i>		14 (3%)
Pharmaceutical Aims	36 (8%)	
<i>Adverse drug events</i>		19 (4%)
<i>Drugs discovery</i>		9 (2%)
<i>Testing drugs</i>		8 (2%)
Final Total	447 (100%)	447 (100%)

Table 2: The usage rates of the health aims taxonomy's categories

Moreover, the literature division is represented by pie charts according to the health aims categorisation. Figure 11 shows the division according to the main aims, while figure 12 shows the types of addressed health questions.

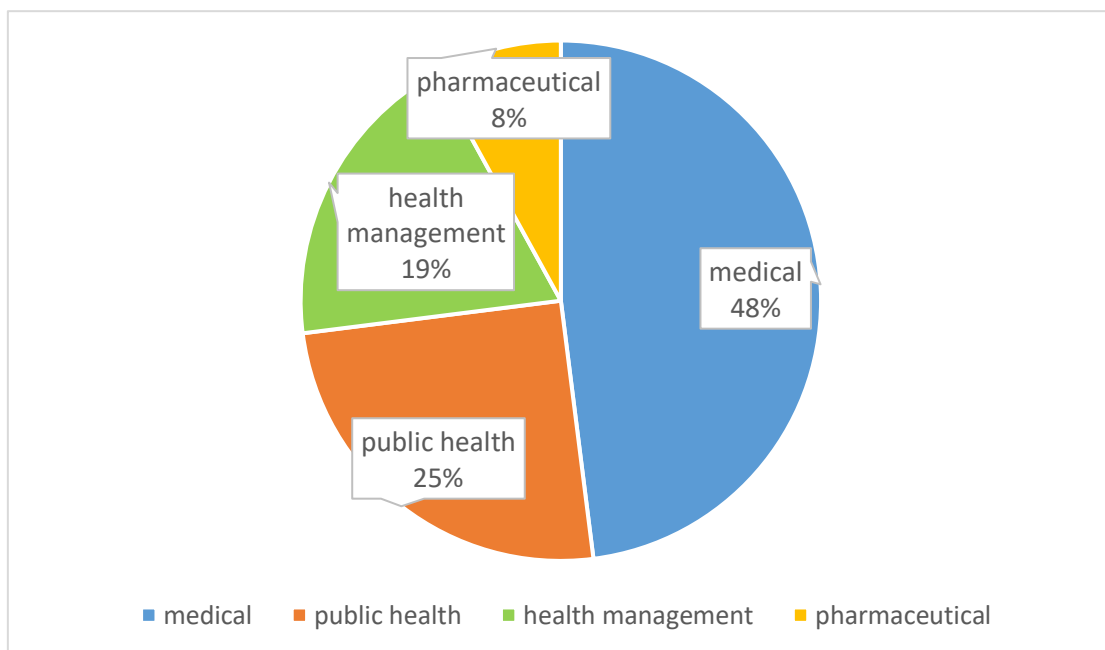


Figure 11: The division of the literature into four main health aims

Figure 11 shows the reviewed literature divided into four main aims: i) medical, ii) public health, iii) health management and finally iv) pharmaceutical. The medical aim has the biggest share of the reviewed papers with almost half of literature addressing medical questions (48%). A quarter of the literature discussed public health related issues questions. The health management aim was 19% of the literature, while pharmaceutical papers represented only 8%. The next figure is a detailed version of the literature division.

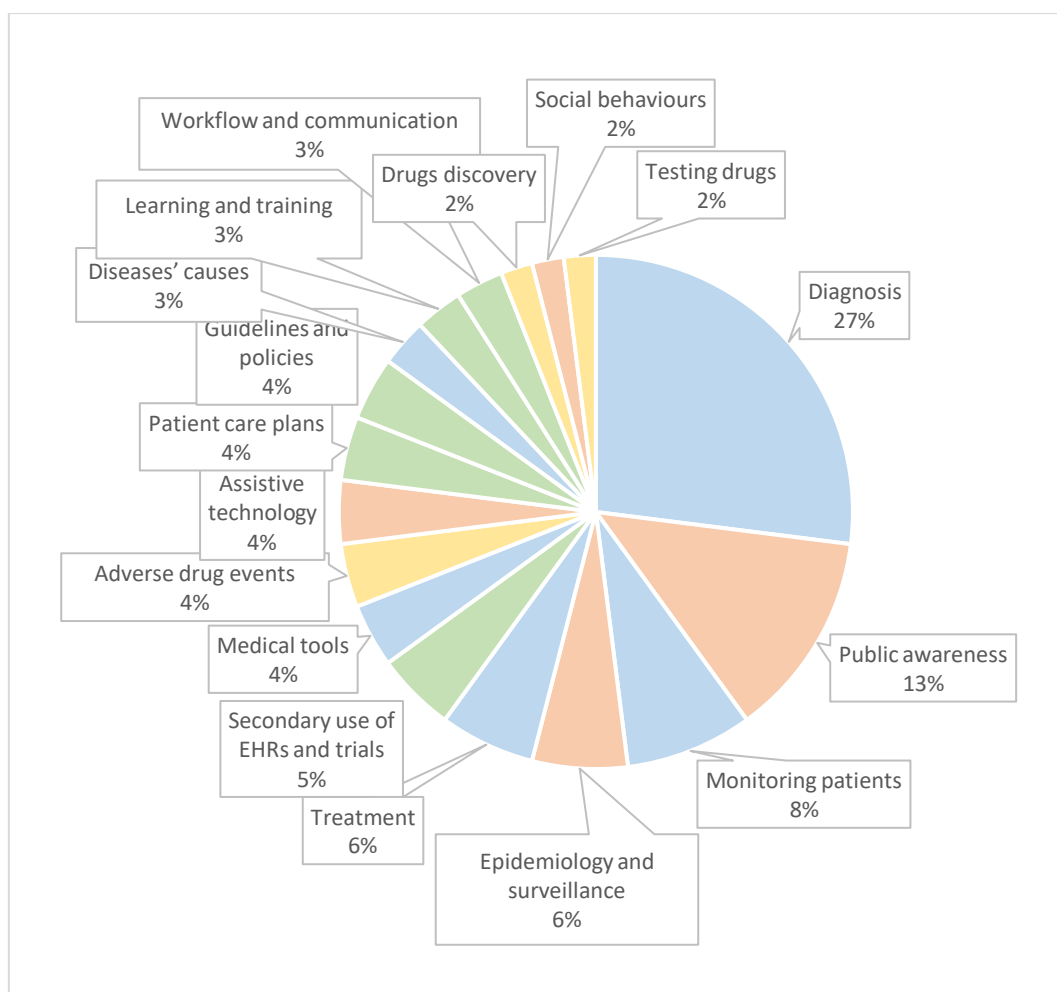


Figure 12: The division of the literature according to the health sub-aims (addressed questions)

Figure 12 shows the division of the reviewed literature according to the specific health questions (sub-aims). The most discussed topic in the literature was 'diagnosis and decision support' which is part of the medical aim; just above quarter of the literature addressed this type of question. The second biggest topic comes from the public health aim which is 'promoting public awareness' with 13% of usage. Other popular topics are: i) 'monitoring patients via sensor devices' with 8%, ii) 'epidemiology and environment surveillance' at 6% and iii) 'treatment and drugs recommendation' also with 6%. All the discussed topics that have more than 5% representation in the literature are either medical or public health topics. The health management and pharmaceutical topics have less share of the literature. For example, 'supporting clinical trials and secondary use of EHRs' has 5% of the literature and this is a health management topic, while 'finding adverse drug events' is part of the pharmaceutical aim with a 4% share of the literature.

In the following sections, the health questions (aims) that were addressed in the literature are discussed according to their popularity in terms of usage. The health aims are divided into four groups. The first one is the most popular questions that featured in the literature, while the last group includes the questions / issues that were the least discussed. Another two groups are for the second and third most discussed topics for papers with usage rates in the range of between 5% and 15%.

The Most Discussed Health Aim

This group is for the aims that have a share of the literature that is greater than 25%, which is only one aim. 'The diagnosis and decision support' aim has a usage percentage of 27%. The diagnosis questions and decision making process in the medical aim rely on the idea of defining rules and conditions. This type of question tries to find new information from the associations of issues that are already known, via inference. For example, to answer a diagnosis area question such as "whether patient X has a specific disease", the associations between different symptoms and diseases need to be examined. Whether the patient has the disease or not will then be inferred based on a defined rule.

In the reviewed literature, one of the studied cases was Mohammadhassanzadeh *et al.* (2017) who used querying and reasoning to find if a specific patient had hepatitis, based on known and inferred information about that patient. These types of questions are frequently used in the medical field. Therefore, these decision-based questions are suitable to be used with the SW as it can offer logic and reasoning features.

The Second Most Discussed Health Aim

This group is for aims that have a usage percentage from 10% to 15%. The only aim belonging to this group is the 'promoting public awareness' aim with 13% use. This aim relies primarily on the huge amounts of health information available online. The role of the SW here is to deliver the right health information to the right user. Therefore, the majority of papers with this aim developed health information systems that recommended personalised health information for the user.

For example, Zaman and Li (2014) proposed a system that suggested health content to social network users, based on similarities of social profiles and behaviours with other users. All this information was mapped to a defined ontology to calculate similarity rates between users of social networks. As a result, 'liked' health-related items by one user were suggested to other similar users.

The Third Most Discussed Health Aims

This group is for aims that have a literature share in the range of 5% to 10 %. There are three aims belonging to this group: i) 'monitoring patients via sensor devices', ii) 'treatment and drugs recommendations' and iii) 'epidemiology and environment surveillance'; the usage percentages were 8%, 6% and 6% respectively.

'Monitoring patients via sensor devices' is one of the aims that relies on the advances in the hardware sector that have opened opportunities for the SW in managing the software controlling these new devices. The SW was chosen in many of these projects for its integration and interoperability capabilities. For instance, Stavropoulos *et al.* (2016) used the SW technologies in integrating heterogeneous sensing data (video and audio) with other physiological and environmental measurements.

The second aim in this group is 'treatment and drug recommendation' aim. The nature of this medical aim is very similar to the 'diagnosis' aim in terms of relying on decisions. For example in Rung Ching Chen *et al.* (2012), the authors used SWRL rules and reasoning capabilities to decide which anti-diabetic drug was most suitable with patients having special conditions.

While the two mentioned aims earlier were from the medical aim, the last aim in this group belonged to the public health aim. Epidemiology and environment surveillance studies took place in both closed and open environments such as hospitals or cities. In epidemiology and surveillance studies, indicators measuring different health aspects were monitored and the relationships between these indicators were studied. The SW was used in these studies to model and link such information. For example, Shaban-Nejad *et al.* (2012) modelled hospital acquired infections to be able to detect risk by understanding the relationships between infection indicators.

The Least Discussed Health Aims

All aims mentioned in the previous groups were either medical or public health aims, while this group has a diverse collection of questions from all the aims. All the health management and pharmaceutical aims are in this group. This group has the biggest number of aims belonging to it: 12 aims with usage rates of less than 5%. The aims that belong to this group are:

i) 'examination using medical tools or devices', ii) 'finding adverse drug events', iii) 'supporting assistive technologies for special-needs pupils', iv) 'clinical Pathways and patients care plans', v) 'clinical Guidelines and Policies', vi) 'finding causes of diseases', vii) 'training staff and supporting

learning', viii) 'workflow, communication and business processes', ix) 'drugs discovery', x) 'understanding social behaviour' and xi) 'testing drugs'.

The papers that addressed these aims used the SW for different reasons, depending on the research aim. For example, the papers that used assistive technologies for supporting special-needs pupils were using the SW for its interoperability and heterogeneous integrating abilities. In Baldassini *et al.* (2017), the authors used the SW tools to integrate heterogeneous information from three different domains: a) information about the user's health condition, b) sensors devices' information in the living environment and c) the user's physical measurements / data recorded by sensory hardware.

Another example from the pharmaceutical aim was in papers aimed at finding adverse drug events. Natsiavas *et al.* (2018) aimed to detect any pharmacovigilance signals by linking different data resources and by inferring new facts from the integrated data.

Another study found the reason for using the SW by health researchers was as a standard for representing data to overcome heterogeneous integration issues. In Chondrogiannis *et al.* (2018), the authors used EHRs to find eligible patients for clinical trials. They used the SW technology here firstly to integrate the heterogeneous EHRs written in different terminologies with the available semantic clinical terminologies on the cloud. Secondly, the SW was used to represent the eligibility criteria and retrieve the records of those eligible only from the integrated contents.

4.3.2 Semantic Web Features Taxonomy

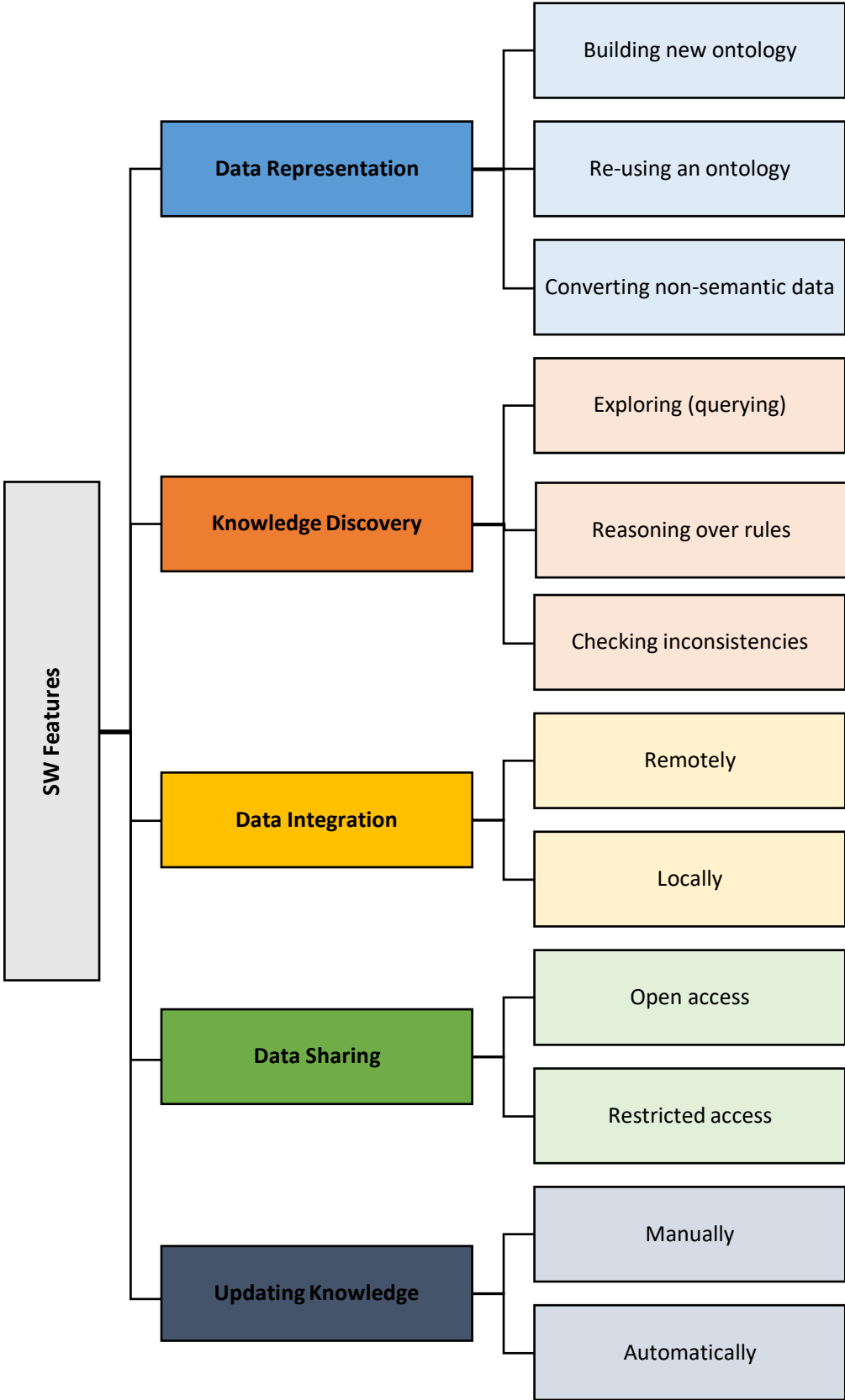


Figure 13: Taxonomy of the semantic web’s features

The taxonomy of SW features was built relying on features of the SW mentioned in the reviewed literature. Figure 13 shows the main and secondary features of the SW taxonomy. Five main features were identified, while the sub-features totalled 12. The following section will discuss the five main features: i) data representation, ii) knowledge discovery, iii) data integration, iv) data sharing and v) updating knowledge. Specific details will be presented under each main category.

4.3.2.1 Definitions of the SW Taxonomy's Features

Data Representation

Data representation is about conceptualising and modelling a domain which can be achieved by using ontologies. Gruber (1995) defined an ontology as “an explicit specification of a conceptualization”. Ontologies concentrate on abstracting a domain and classifying it by defining classes and relationships (World Wide Web Consortium (W3C), 2008). Ontologies can be used for many things such as: a) representing and organising knowledge, b) data integration and c) sharing and discovering new knowledge.

Building a new ontology

Many reviewed papers aimed to represent a certain domain by building an ontology. For example, Dang *et al.* (2008) developed an ontology to describe a healthcare network including hospital resources and processes. Hu *et al.* (2012) built an ontology to represent clinical pathways and treatment procedures. Puustjarvi and Puustjarvi (2016) designed two integrated ontologies to represent clinicians' profiles and clinical guidelines. In the medical field Maragoudakis, Maglogiannis and Lymberopoulos (2008) developed an ontology for skin lesions, while Ciccarese, Wu, Kinoshita, *et al.* (2008) modelled the neuro medicine field by developing a semantic web applications in neuromedicine (SWAN) ontology.

Re-using ontology

The open biomedical ontologies (OBO) foundry is a collaboration involving developers of scientific ontologies to set common principles for ontological development (Smith *et al.*, 2007). One of the OBO foundry's principles is “commitment to collaboration”, an idea which encourages ontology developers to re-use other ontologies in order to avoid duplication and increase interoperability.

In the reviewed literature, some of the reviewed papers followed the principle of re-use. For example, Hogan *et al.* (2016) developed Apollo-SV ontology to model the epidemiological domain of infectious diseases. Around half of the used classes in the ontology are imported from other

ontologies, either by importing the full ontology into Apollo-SV or sometimes by importing partial parts of the target ontology.

Converting non-semantic data

Many of the available data is in non-triple forms. In order to use such exceptional data in a SW system, it must be converted to triple format. Some of the reviewed papers overcame this problem by converting the needed data items and saving them as triples in a triplestore. An example of a paper that followed the conversion approach was by Odgers and Dumontier (2015), where PHP and python scripts were used to transform EHRs into RDF. In other cases, only the needed part of the data was converted rather than the whole database. For example, in the Mayo clinic project for cohort identification (Pathak *et al.*, 2012), the patients' records were kept in a relational data repository. To link this data with the rest of the semantic framework, the data was queried using SQL and then the results were converted to RDF.

Knowledge Discovery

From reviewing the literature, three different approaches of discovering new knowledge were identified.

Querying (exploring)

The querying process means retrieving information from the web of data by using technologies (World Wide Web Consortium (W3C), no date). To answer a query, one or more ontologies are explored and any associations between data elements are traversed. Queries in the semantic web are performed using SPARQL, a query language based on RDF triples. In the reviewed literature Puustjarvi and Puustjarvi (2016) aimed to disseminate clinical guidelines to clinicians. The researchers developed two ontologies; one for the guidelines and the other for the clinicians' profiles. By using SPARQL queries, they were able to retrieve the relevant guidelines for the appropriate clinicians.

Reasoning over rules

Inference or reasoning is explained in the W3C website as an automatic procedure to discover new relationships from known ones based on the data and set of rules (World Wide Web Consortium (W3C), 2008). Alexandrou, Xenikoudakis and Mentzas (2008) designed a semantic recommendation system for personalising treatments. The system contained an adaptive clinical pathway ontology and a set of semantic rules to represent the domain. For instance, one of the defined rules in the

system stated that “if the patient is admitted to the healthcare organization and there is a diagnosis of neurological deficit, then the patient has to be evaluated for thrombolysis eligibility”. Thus, after applying this rule to the available patients’ data, new knowledge and recommendations would be provided by the system.

Checking inconsistencies

Checking inconsistencies in integrated data is possible via inferencing (Dalwadi, Nagar and Makwana, 2012). Chen *et al.* (2012) built an anti-diabetic medicine ontology and used an inference engine (Pellet) to check for any conflicts or contradictions in the ontology’s design.

Data Integration

The main aim of the SW is to link data across the web. Therefore, it would be expected for the data integration’s feature to be mentioned a lot in the literature. Data integration was enabled by defining and then linking ontologies together. For example, there are many ontologies in the medical field that represent diseases and treatment information, along with ontologies from the pharmaceutical field representing drug related information. Linking these ontologies with patients’ private records can open many opportunities for health researchers to address different questions on diagnosing, epidemiology and drug use and many others topics.

Locally

Integrating heterogeneous datasets when developing an ontology was performed widely in the literature to show the strength of the SW in facilitating data interoperability. For example, Shaban-Nejad *et al.* (2016) integrated different datasets such as: a) hospital morbidity and discharge abstracts, b) existing bio-ontologies like infectious disease ontology (IDO), c) medical terminologies and standards like SNOMED-CT and ICD-9 , and d) some other textual resources to develop the integrated hospital acquired infections ontology (HAI).

Remote integration

Integrating data remotely can be done through a special feature in SPARQL, named federated query, which is defined by W3C as “the ability to take a query and provide solutions based on information from many different sources” (World Wide Web Consortium (W3C), 2009). For example, Pathak, R. C. Kiefer and Chute, (2013) used federated query to enquire about drug interactions using data from the public drug data repository, DrugBank, in the LOD in their developed system. Moreover, in the

same paper they performed another remote data integration by performing a federated query into a converted relational database to RDF that contained Mayo clinic electronic medical records.

Data Sharing

According to the W3C, "The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries" (World Wide Web Consortium, 2013) . There are two forms of data sharing in this literature: public and private approaches.

Restricted access

Health data is sensitive, especially if it deals with or involves accessing personal/patient records. Due to this fact, many of the produced ontologies had restricted access. For example, Rung Ching Chen *et al.* (2012) developed a recommendation system to be used by doctors in Taichung's Department of Health. The system's aim was to analyse the real records of diabetic patients and recommend the most appropriate anti-diabetic drug for each case. As seen from this example, this type of sensitive data demands operating as a restricted access model, for authorised persons only.

Open access (LOD)

There are other types of health/biomedical data that represent a specific domain but which do not contain any private personal information. Part of the reviewed literature represented such data as ontologies and shared them publicly in the LOD. An example for this type of sharing was the Apollo ontology (Hogan *et al.*, 2016). The Apollo structured vocabulary (Apollo-SV) ontology represented infectious disease epidemiology and population biology phenomena. Apollo-SV is freely available at the LOD.

Updating Knowledge

One of the SW's advantages is its flexibility in updating knowledge as there is no difference in RDF between updating schema's information and / or the instances (Jyotishman Pathak, R. C. Kiefer and Chute, 2013). In the reviewed literature, a couple of papers mentioned the updating knowledge feature; useful to correct wrong information or to add new information.

Manually

Usually, manual updating of knowledge is performed after consulting domain experts regarding the ontology's design. This happened in the SWAN project (semantic web applications in neuro

medicine), which was an interdisciplinary project aimed to develop a semantically structured framework for biomedical discourse in neuro diseases such as Alzheimer's (Cicarese, Wu, Kinoshita, *et al.*, 2008). The developers of the SWAN ontology updated it manually after consulting experts and scientists about their design.

Automatically

Another form of updating knowledge involves inserting newly inferred knowledge into the knowledge base. Huang *et al.* (2014) presented a semantic rule-based clinical pathway compliance checking system. After reasoning over the compliance rules for patients' data, new inferred facts were added into the system automatically.

4.3.2.2 Usage Analysis

After defining the SW features and giving examples about each one in the previous section, this section focuses on analysing the usage rate(s) for each of the defined features. The usage rate is calculated by counting the number of papers which used a specific feature. For the main categories of the SW features, the average of the usage rate was calculated because repetition was allowed between the sub-feature categories. Table 3 shows the main and sub-features of the SW with the usage percentage of each feature. From the reviewed literature five main SW features and 12 sub-features were identified.

Main/Sub-features	Average of Sub-features N (%)	Sub-features N (%)
Data Representation	250 (56%)	
<i>Building new ontology</i>		415 (93%)
<i>Re-using ontology</i>		232 (52%)
<i>Converting non-semantic data</i>		104 (23%)
Knowledge Discovery	218 (49%)	
<i>Querying (exploring)</i>		344 (77%)
<i>Reasoning over rules</i>		284 (64%)
<i>Checking inconsistencies</i>		26 (6%)
Data Integration	200 (45%)	
<i>Locally</i>		377 (84%)
<i>Remotely</i>		22 (5%)
Data Sharing	126 (28%)	
<i>Restricted access</i>		238 (53%)
<i>Open access (LOD)</i>		14 (3%)
Updating Knowledge	17 (4%)	
<i>Manually</i>		26 (6%)
<i>Automatically</i>		8 (2%)

Table 3: The usage rates of the categories in the sw features taxonomy

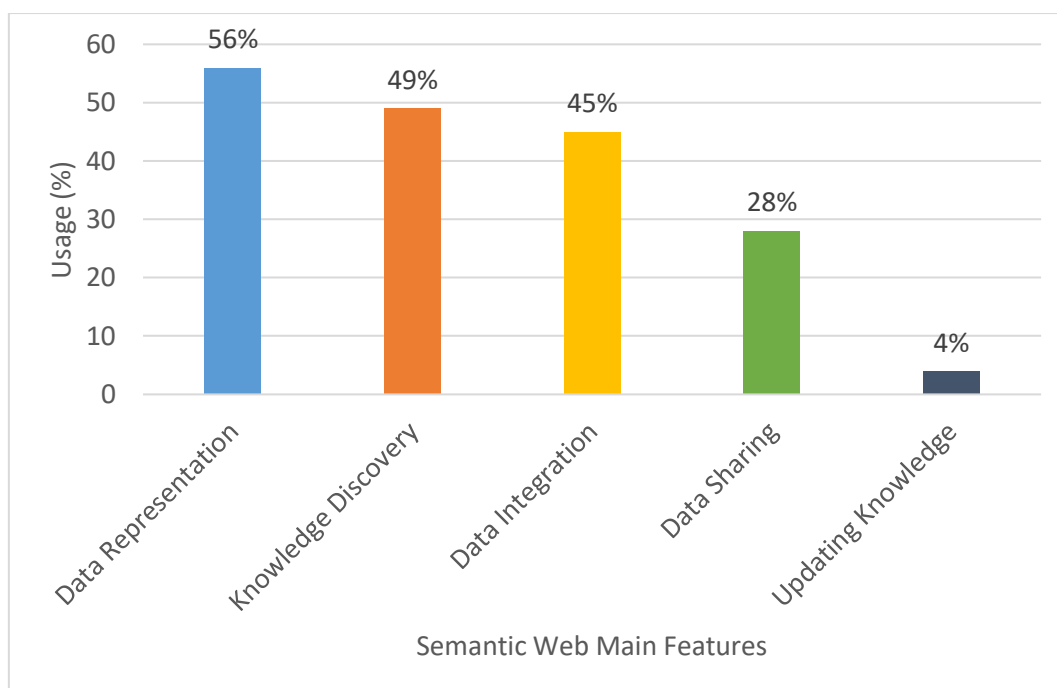


Figure 14: The average usage of the semantic web's main features

Figure 14 shows the main features in ascending order informed by the average number of papers that included them. The average indicator was used rather than the total because it is not possible to find the total for the sub-features' groups, as they are joint sets.

The highest used main feature is 'data representation'; an average of 250 papers used it, which represents around 56% of the literature. 'knowledge discovery' comes second in average of 49% of usage, and 'data integration' comes third with 45%. Data sharing had a smaller portion with 28% of usage, while the lowest usage was for 'updating knowledge' with only 4% of usage. These data show that the most important reasons for using the SW in the health research literature were for: i) representing data, ii) discovering knowledge and iii) integrating data.

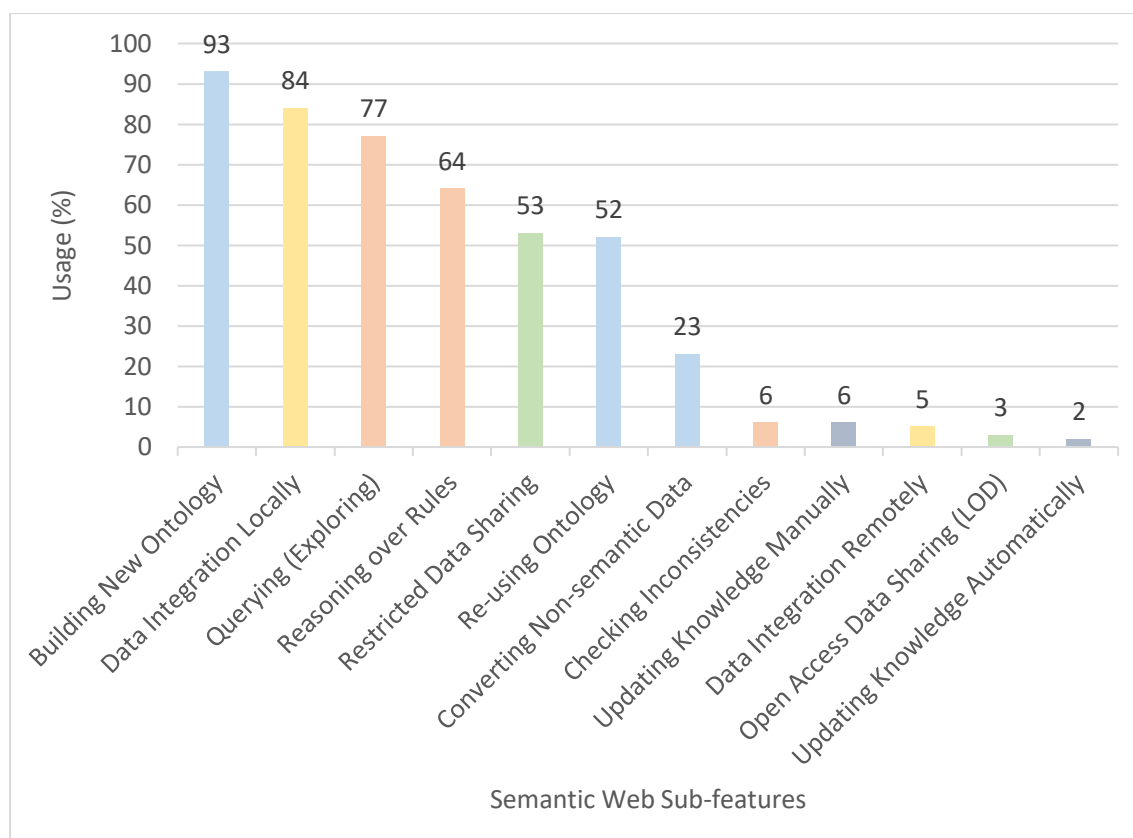


Figure 15: The usage's rate of the semantic web's sub-features

Figure 15 shows the ascending order in the usage of the 12 sub-features of the SW. There is a variant in the usage percentage between the features. Half of the features are used by more than 50% of the literature, while the other half has a very low percentage of usage. The most used feature in the reviewed literature is 'building new ontology' which is part of the data representation feature. The 'building a new ontology' feature achieves a very high usage percentage (93%). On the other hand, the least used feature is 'updating knowledge automatically' with 2% of usage.

Due to the variations in the usage percentages of the features, the SW features are categorised into four groups (quarters). The first group represents the most used features, with a usage percentage greater than 75%. This group involves: i) building a new ontology, ii) local data integration and iii) querying. The second group (50%-74%) also contains three features which are: i) reasoning over rules, ii) restricted data sharing, and iii) reusing ontology. There are no features in the third quarter. However, the fourth quarter housing the least used features totalled six features. Those six features were used less than 25%. In fact, all the features in the fourth section had very low usage percentages; five registered less than 10%; with one 'converting non-semantic data' showing 23% usage. The next sections discuss these groups in detail.

The most used features

The feature that showed the highest usage is 'building new ontology' with 93%. Almost all of the reviewed papers described the building of new ontologies due to the importance of the ontology as the main building block of the SW. Ontologies have been used for different purposes such as representing facts and rules in a domain, integrating data in an interoperable way, and discovering new knowledge. For example, Rung Ching Chen *et al.* (2012) built a system to recommend medications based on the personal condition of each patient. The main building block in the system was two integrated ontologies: medicine and patients' conditions. The system suggested personalised recommendations based on defined rules that associate specific medical conditions with the features of specific medications. By using these defined rules that represented the domain, new knowledge could be discovered as well.

Some of the built ontologies in this reviewed literature were repetitive, with the same domain being the focus of interest in different papers. For example, there were several reports on the building of ontologies for clinical pathways (Hurley *et al.*, 2007; Ye *et al.*, 2009; Alexandrou *et al.*, 2012; Hu *et al.*, 2012; Huang *et al.*, 2014; Wang *et al.*, 2015). However, this repetition was necessary in some cases because each institution or hospital had its own regulations and policies, where each can be slightly different than the others. Thus, it was not always possible to re-use other ontologies.

With a percentage of 84%, local data integration was used in the literature. One of the promises of the SW was to provide an environment in which it would be possible to find and link data globally. However, most of the data integration in the health literature was done at a local level. In many cases, this format is due to the personal and sensitive nature of the integrated health data. Nonetheless, in other projects the integration was performed on both levels locally and remotely (Pathak *et al.*, 2012). In addition to integrating the patients' private data locally, another level of remote linkage was done via federated querying. The remote linkage was performed on general biomedical ontologies in the LOD, such as translational medicine ontology (TMO) and sequence ontology (SO). This type of remotely integrated data was used in only 5% of the reviewed studies. This usage percentage is surprisingly small in comparison with the big promise of linked data and the SW.

The third highest use was for 'querying' at 77%. Although not all the reviewed papers mentioned querying explicitly, it was pointed to via indicators like retrieving RDF data or using SPARQL. On the other hand, there were very few papers that reported the use of other querying languages such as SQL. This type of querying was discarded as it did not involve using SW.

The moderately used features

The moderately used features are those used in between 50% and 75% of the literature. Three features are used in this range. The first one is 'discovering new knowledge via reasoning over defined rules'. This feature showed a 64% of usage. The reason behind the high usage rate is the inferential nature of the discussed health topics. Many of the discussed health topics had a decision nature that was based on fulfilling conditions such as diagnosing. In addition, representing clinical guidelines and policies as defined rules was another aim evident in the reviewed literature that can infer new knowledge by using reasoning techniques.

Restricted data sharing and re-using other ontologies' features came next, mentioned in just over half of the papers. Again for the same reason of data sensitivity, it is not always possible to share data publicly. The data is usually owned by a health institution or hospital and it contains large amounts of personal information about patients, doctors, health situation and finance that is shared within the institution and which should be viewed by authorised people only. Projects that shared their data and published it publicly were never above the 3% level. The published ontologies usually have a biomedical nature and represent a specific biomedical domain, such as the 'semantic web applications in neuromedicine' ontology (SWAN) in the field of neuroscience (Ciccarese, Wu, Kinoshita, *et al.*, 2008) and the ToxBank data warehouse in the pharmaceutical field (Kohonen *et al.*, 2013).

Regarding reusing ontologies, this approach is usually deployed for medical standards and terminologies. Ontologies like the 'systematized nomenclature of medicine - clinical terms' (SNOMED-CT), the 'medical subject headings' (MeSH), and the 'international classification of diseases' (ICD) were used broadly in the literature because many health topics depend on these terminologies to represent them. Using these terminologies in a structured format helps researchers to easily integrate those terminologies into their own work and therefore saves both time and effort. Another type of re-used ontology is biomedical, such as the gene ontology (GO), the infectious disease ontology (IDO), the translational medicine ontology (TMO) and the sequence ontology (SO) which also were used in the literature addressing various health topics.

The least used features

There are six identified features in this category; five with a lower than 10% usage level. The one exception was 'converting non-semantic data' which registered 23% usage. Converting non-semantic data was used to ease the process of populating the semantic triple store with already existed data. Some of the studies explicitly mentioned the conversion step, while most of the literature did not

mention the populating step at all, or failed to explain how the process was achieved. The converted data can be patients' records such as the Mayo clinic's relational data used by Pathak *et al.* (2012). The conversion in this project was performed using the relational to RDF mapping language (R2RML). Another example of converting instances is in the ToxBank data warehouse that consisted of integrated public toxicogenomics data converted into RDF (Kohonen *et al.*, 2013). In this project a tailored tool was developed for the conversion process.

The rest of the features that were rarely used in the literature are: i) checking inconsistencies, ii) updating knowledge manually or automatically, iii) data integration remotely and iv) public data sharing. The last two were mentioned earlier in discussing the 'local data integration' and 'restricted data sharing' features.

Checking for inconsistencies in the designed ontologies was mentioned in some of the literature as a 'checking-for-error' step using a reasoner and the inference feature provided by the SW. The number of papers that explicitly mentioned using this feature is small in comparison with the number of papers that reported using 'reasoning over rules'. While both were using inference capabilities to discover new knowledge, the rate of mentioning error checking was low because this step might be considered as a testing step that was neglected by the authors. However, is important to mention error checking here as it shows one of the SW's capabilities; that of discovering errors and inconsistencies in the data.

Relating to 'checking for inconsistencies', 'updating knowledge' is another feature that was not mentioned a lot in the literature. For the same reasons as in 'checking for inconsistencies' this step might be performed by the majority of projects, but was neglected. This updating feature reflects the advantages of ease and flexibility when using the SW to change or update some of the defined knowledge.

4.3.3 Variance in Using the SW's Features across Health Aims

So far, the results section discussed the health aim's taxonomy and SW's features taxonomy in the previous sections, as well as their definitions and usage analysis. This section combines the results of both taxonomies in a cross-tabulation analysis between the SW's features on the one hand and the main health aims on the other hand. The purpose of this analysis is to show the trends and gaps in usage rates of the features across the health aims. A better understanding of the addressed health questions' impact on the usage rates of the SW's features can be achieved; as well as a better understanding of the challenges faced and potentially available affordances.

The next section shows the tabulated analysis of the SW's relationships with health aims, followed by a discussion of the similarity and disparity of the SW features' usage rates across health aims. Finally, a list of chosen examples from the reviewed literature, mapped on the SW's features taxonomy, is provided at the end of this section to demonstrate how the taxonomies were used in mapping the literature.

4.3.3.1 Usage Analysis

Table 4 shows the usage of the SW's features (rows) across the four main health aims (columns) for the reviewed literature. The usage's rates were categorised into four codes from dark blue to light blue for simplicity, as well as to highlight and demonstrate the differences and similarities in the usages rates across the four health aims' columns.

There are some features that all the aims agreed regarding percentage of use, whereas other features show some differences regarding their usage rates between the four aims. For example, all the aims agreed on using 'building a new ontology' and 'local data integration'. In addition, they also agreed to lesser extents on using: a) 'checking consistencies', b) 'remote data integration', c) 'open access sharing' and with d) 'updating knowledge' either manually or automatically as the least accessed of the main features. However, for the rest of the features there was a disparity in the percentage of usage between the four aims.

The usage of the SW's features across the four health aims is divided into two categories: i) the features that register similar usage percentages across the aims and ii) the features that do not. The following sections discuss the two categories in detail.

SW Features / Health Aims	Medical	Public health	Health Management	Pharmaceutical
Data Representation				
<i>Building New Ontology</i>	4 (93%)	4 (91%)	4 (95%)	4 (92%)
<i>Re-using Ontology</i>	3 (56%)	2 (39%)	3 (57%)	3 (56%)
<i>Converting Non-semantic Data</i>	2 (28%)	1 (10%)	2 (33%)	1 (14%)
Knowledge Discovery				
<i>Querying (Exploring)</i>	4 (86%)	3 (63%)	3 (66%)	4 (89%)
<i>Reasoning over Rules</i>	4 (77%)	2 (42%)	3 (61%)	3 (56%)
<i>Checking Inconsistencies</i>	1 (8%)	1 (4%)	1 (2%)	1 (6%)
Data Integration				
<i>Locally</i>	4 (91%)	4 (77%)	4 (87%)	4 (83%)
<i>Remotely</i>	1 (4%)	1 (5%)	1 (5%)	1 (8%)
Data Sharing				
<i>Restricted Access</i>	3 (50%)	3 (59%)	3 (63%)	2 (33%)
<i>Open Access (LOD)</i>	1 (3%)	1 (2%)	1 (4%)	1 (8%)
Updating Knowledge				
<i>Manually</i>	1 (7%)	1 (4%)	1 (6%)	1 (3%)
<i>Automatically</i>	1 (2%)	1 (2%)	1 (1%)	1 (3%)

4	75% – 100%	Most used
3	50% – 74%	Many used
2	25% – 49%	Little used
1	0% – 24%	Least used

Table 4: variance in the use of the semantic web's features across health aims

The Similar Features' Usage

There are seven features that are agreed as being used at similar rates between the health aims.

Five of these features are in the category of 'least used', while two features are in the 'most used'

category. 'Building a new ontology' and 'integrating data locally' are the most used features in all the aims, while: a) 'checking inconsistencies', b) 'updating knowledge manually', c) 'updating knowledge

automatically', d) 'remote data integration' and e) 'open access sharing' showed the least percent of usage in all the aims.

All the aims agree on building new ontologies, which implies that this feature is a significant one for any research aim. To perform any semantic activity, such as 'querying' or 'reasoning' there is a need to have / create an ontology before any action can be taken. This ontology is the semantic framework that includes any defined elements needed in ensuing semantic activities. Shaban-Nejad *et al.* (2016) said that an ontology is important for knowledge discovery and exchange, while Horridge *et al.* (2014) emphasised that ontologies support data integration and linkage. Dang *et al.* (2008) explained that the importance of using ontologies is for machines to be able to reason about objects in a specified domain, while Ye *et al.* (2009) mentioned that ontologies are 'key technology' due to their affordances in data sharing and reusing in a digital form. Therefore, the reason behind the agreement of all health aims in using and building ontologies is understood as it is the main building block for the employment of SW technology.

'Locally integrating data' is the second used feature where all the aims are agreed on. Integrating heterogeneous datasets is the core idea of the linked data principle. However, the integration levels that have been performed in the majority of the projects were on a local level, not global. This orientation is usually due to the sensitive nature of personal health data. For example, in two cases from the medical field: i) Pathak, Kiefer and Chute (2012) integrated data for one institution: the Mayo clinic and ii) Hu *et al.* (2012) integrated the electronic medical records (EMR) initiative in Miyazaki University hospital. However, that local orientation was not always the case as the public health aim showed a lower percentage of usage in comparison with the others (77%). This lower rate could be related to the fact that some of the epidemiological projects integrated data on a larger geographical scale and not a local one, as in the medical or clinical field. For instance Capiere *et al.* (2014) introduced a distributed medical data source grid network to support large-scale epidemiological analysis. The system used remote data integration via federating different databases from hospitals' and labs' remote servers.

The other five features that showed the lowest usage rates in all the aims are: i) 'checking inconsistencies', ii) 'updating knowledge manually', iii) 'updating knowledge automatically', iv) 'remote data integration' and v) 'open access sharing'. Regarding data integration and sharing, the SW community, at the rise of this technology, described the SW as a global data space where data integration and sharing play an important role (Shadbolt, Hall and Berners-Lee, 2006; Bizer, Heath and Berners-Lee, 2009). However, from reviewing the literature we can see that the researchers used these two features in two different ways that must be acknowledged. 'Data integration' was

used heavily in the literature, but mostly locally. The same applies for 'data sharing'. Usually, the shared data was restricted and did not have open access. This restriction is due to the nature of the used health data, which was sensitive and private. Most of the reviewed projects used patients' records or data from private institutions. For this reason, the ability to perform public data sharing or remote data integration was and still is limited.

In addition to the low percentage of usage for 'remote data integration' and 'public data sharing', 'checking inconsistencies' and 'updating knowledge' also had a low usage rate. This low rate could be because the papers' authors were unaware of these features or, more likely, because these features were used but were neglected to be mentioned by those authors. In all the cases, these two features show the SW's flexibility in finding errors and then fixing them.

The Disparate Features' Usage

There are five features that showed variance in the percentage of their usage among the four main health aims. The first one is 'discovering knowledge via querying'. It is generally the third most frequently used feature in all the aims; however, it showed a high level of usage in some aims whilst only a moderate usage in others. 'public health and health management' showed a lower rate of usage than other medical and pharmaceutical aims. From reviewing the literature, the 'querying' feature was rarely mentioned, thereby resulting in a low usage rate. For example, Birjali, Beni-Hssane and Erritali (2017) undertook a public health study that extracted Twitter data for semantic suicidal sentiments analysis. The authors did not explicitly mention using any querying techniques and the focus was on analysing data using a machine learning algorithm.

'Reusing ontology', 'reasoning', and 'restricted access sharing' are the third most used features in the literature. Around 50% of the reviewed papers in all the aims reused some ontologies, other than with 'public health', which was rated at the 40% usage level. A lack of public health-related ontologies in the reviewed public health focused papers was noticed. Public health is a broad aim that discusses different topics which were not necessarily represented before as ontologies. On the other hand, the available ontologies are mostly of different biomedical nature that are not necessarily be used in the public health projects such as the gene ontology (GO) and the infectious disease ontology (IDO). In addition, there are many available drugs-related ontologies such as the drug ontology (DrOn) which were mentioned by the medical and pharmaceutical papers. Another heavily used type of ontology is the clinical category, such as: i) the medical subject headings (MeSH), and ii) international classification of diseases (ICD). All of these types of ontologies were heavily used in all health aims, except for 'public health'.

Regarding the 'reasoning' feature, the medical aim used reasoning the most at more than 75%; while there were many using reasoning in both the 'health management' and 'pharmaceutical' health aims (50% - 75%). Public health was the aim that least mentioned 'reasoning', with a percent of 39%. This relatively low level can be related to the nature of the discussed topics with public health aims in oppose to the medical papers that demands some kind of decision-support and rules defining, such as in diagnosing systems (Pathak, Richard C Kiefer and Chute, 2012; Mohammadhassanzadeh *et al.*, 2017) or deciding on the most suitable treatment (Rung Ching Chen *et al.*, 2012). There were topics in health management that were also related to supporting decisions such as deciding on the best care plan for a patient (Alexandrou, Xenikoudakis and Mentzas, 2008) or the best way to communicate in hospital workflows (Nelson and Sen, 2014). Pharmaceutical topics, like finding adverse events between drugs (Natsiavas *et al.*, 2018), or discovering new drugs (McCusker *et al.*, 2014) also relied on finding new knowledge from a series of already known information. The nature of this kind of topic was very suitable for use in reasoning over defined rules.

Regarding 'the restricted data sharing' feature, more than half of the papers in each aim used this feature. The exception was the pharmaceutical articles, where the usage was around the 30% level. Pharmaceutical papers showed less usage than the others because the discussed topics that demanded some type of sharing information were related to the treatment of patients or finding the correct drug to treat a particular disease. Those types of topics were categorised under the medical aim in the health aims taxonomy. Therefore, the percentage of data sharing featured was only represented by topics addressing pharmaceutical questions from a biomedical point of view, rather than a medical perspective. For example, a general question like "what are the drugs that can cause an adverse event between them?", but not asking about an adverse event for a specific patient. The biomedical pharmaceutical type of questions usually do not depend on personal private information, thus there was no need for restricting any sharing of the data. In fact, between the four aims the pharmaceutical aim showed the highest percentage of publishing data in the LOD. All the aims had low citation rates in general; however, the pharmaceutical aim was the highest among them.

There was a little usage of the 'converting non-semantic data' feature evident in the literature. The medical and health management aims used it more than 'public health' and 'pharmaceutical'. This usage rate could be due to the fact that both aims (medical and health management) used some real patients' data, such as from electronic health records (EHR) saved in hospitals' data warehouses as relational databases. Therefore, to save time and effort automatic conversion tools were used to change this type of data into RDF. Examples of such cases include Pathak *et al.* (2012) in the medical

aim that used the Mayo clinic's data warehouse , or Hu et al. (2012) in the health management aim that mapped the electronic medical records (EMR) in Miyazaki University Hospital, Japan.

4.3.3.2 Examples

In this section, a list of examples for each of the addressed health questions demonstrates the different SW features that have been used. The SW features' taxonomy was used as a standard template to map the literature in all the examples.

Medical Examples

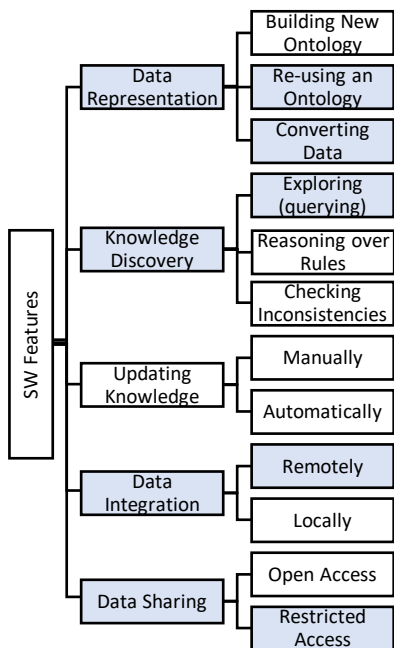


Figure 16: An Example for 'Diagnosis and decision support'

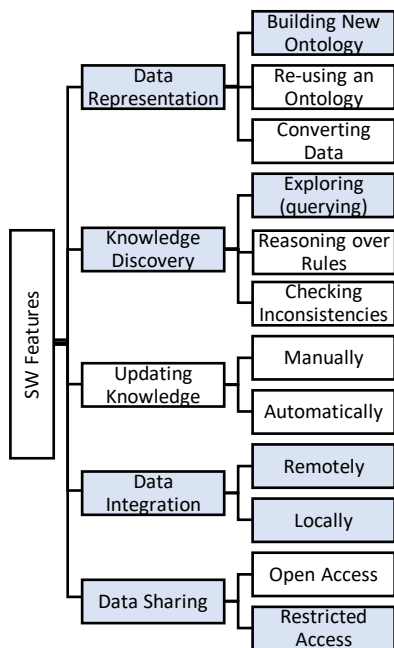


Figure 17: An Example for 'Monitoring patients via sensor devices'

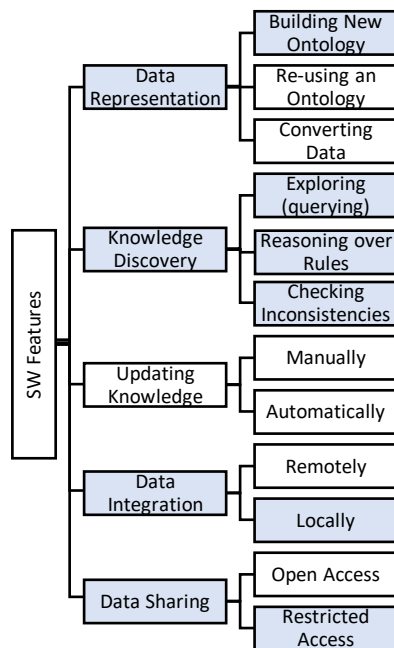


Figure 18: An Example for 'Treatment and drugs recommendation'

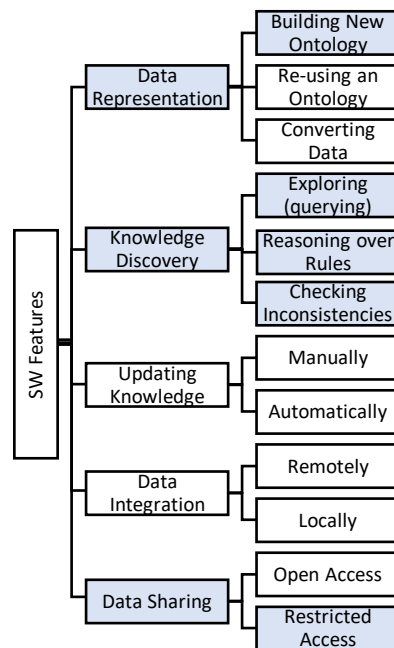


Figure 19: An Example for 'Examination using medical tools or devices'

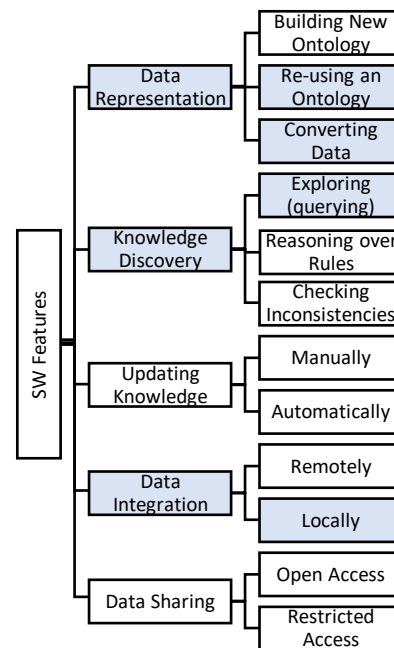


Figure 20: An Example for 'Finding causes of diseases'

Figure 16 highlights the used SW features in the work of Pathak *et al.* (2012). The authors in this paper aimed to identify genotype-phenotype associations to predict the possibility of developing type 2 diabetes and hypothyroidism in a cohort. The SW technologies were used to represent and query patients' data stored at the Mayo clinic's warehouse. The system mainly consisted of three parts: i) accessing data, ii) RDF mapping and iii) querying. The data was accessed from the Mayo clinic's biobank for the genomic data and patients' data from the EHR system. The retrieved data was converted into RDF data and then mapped into existing biomedical ontologies in the LOD; for example; a) translational medicine ontology (TMO) and ii) sequence ontology (SO). Finally, the system was provided with a SPARQL endpoint to enable the linked data to be queried both locally and remotely.

Figure 17 represents the work of Puustjarvi and Puustjarvi (2015). This paper discussed semantic interoperability in exchanging data between devices and services in an IoT and in cloud-based systems. The topic was about monitoring health data generated from sensors at home. The authors in this initiative developed a main ontology named 'smart home' that represented the exchanging process. Another ontology that was integrated with the home ontology was named the 'welfare' ontology that offered access to the user's personal welfare information. Moreover, the 'personal health record' ontology was linked to the 'welfare' ontology; a union which expanded the usability of both ontologies. The authors mentioned in their paper the advantage of using SPARQL for querying data as it enabled the user to search multiple datasets in one query. Also, it provided a mechanism to query data remotely by using the SERVICE keyword to query a SPARQL endpoint. This project mainly aimed to emphasise the use of 'linked data' principles in data integration. The integration took place on two levels: locally and remotely. The local integration was in linking different ontologies, such as 'home' and 'welfare', while the remote option was for integrating other datasets using federated querying.

Figure 18 shows the work of Rung Ching Chen *et al.* (2012) with a SW features taxonomy. The authors built a recommendation system that helped doctors to decide which anti-diabetic drug is most suitable for each individual patient. The system contained two ontologies: i) medicine and ii) patients ontologies. The 'medicine' ontology was an anti-diabetic drugs' ontology that represented: i) the drugs' attributes, ii) type of dispensing and iii) possible side effects, while the patients' ontology was built for testing purposes with symptoms data from 'fake' patients. Another important part of the system was the SWRL rules that defined the diabetes medication association's rules. Inference was performed into two levels: i) by using Pellet to check for inconsistencies or clashes in the anti-diabetic ontology and ii) via help from the Java Expert System Shell (JESS), through which

potential prescriptions for the patients was inferred from the rules and the patient's retrieved data. This project was limited in the level of integration and sharing processes.

Figure 19 represents the work of Maragoudakis, Maglogiannis and Lymberopoulos (2008) in developing medical tools. The authors aimed to develop an ontology for skin lesion images to be used with decision support systems. The ontology was developed using OWL to exploit the reasoning supporting features. The ontology's classes and properties were defined based on extracting features and descriptions from medical images. For example, the ABCD-rule in dermatology was used in the design meaning that: a) the asymmetry, b) border structure, c) variegated colour, and d) the differential structure features of the skin lesion were all considered. The ontology was beneficial by allowing dermatologists to retrieve and infer based upon existing database instances.

Figure 20 highlights the used SW features in the taxonomy of Gudivada *et al.* (2008). The authors aimed to identifying genes' mutations that cause diseases. They collected their data from multiple genomic and phenomic resources such as the gene ontology (GO), mammalian phenotype (MP) ontology among others. The non-RDF resources (text-based resources) were converted into RDF first, and then linked to the rest of the ontologies to form a BIO-RDF linked data space. Each resource was scored depending on its priority. SPARQL queries were used to retrieve disease-gene relationships, after finding the relative score of each result.

Public Health Examples

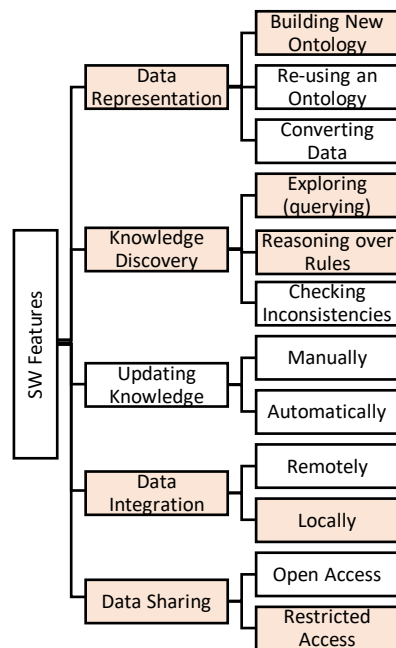


Figure 21: An Example for 'Promoting public awareness'

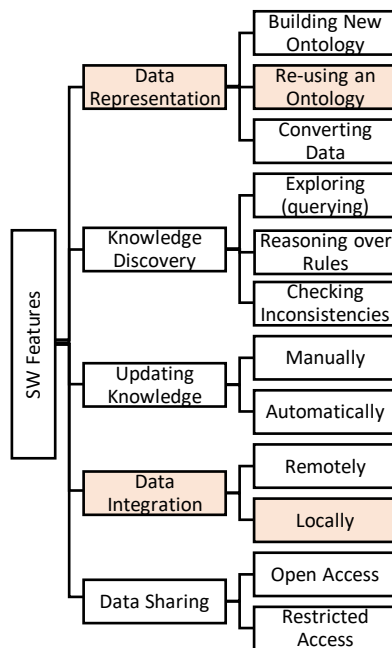


Figure 22: An Example for 'Epidemiology and environment surveillance'

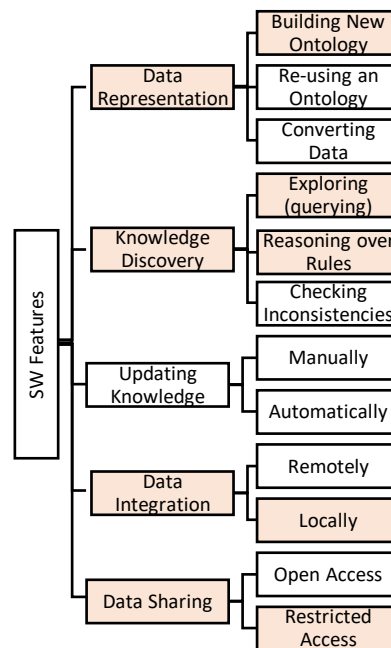


Figure 23: An Example for 'Supporting assistive technologies for special-needs people'

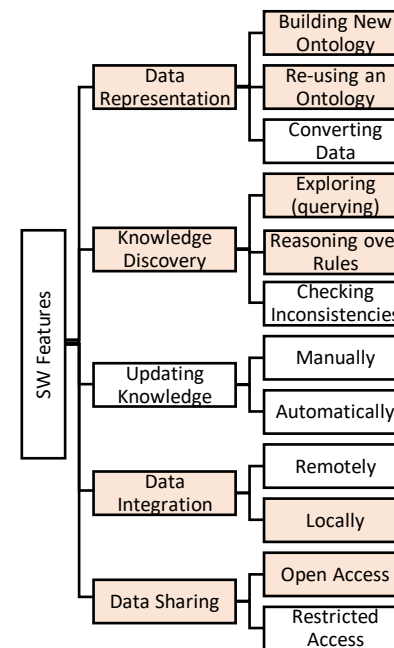


Figure 24: An Example for 'Understanding social behaviours'

Figure 21 represents the work of Wang *et al.* (2010) on promoting health related public awareness. A very popular topic in the public health domain is the subject of food/diet. This paper was an example of a health information system that answered queries regarding a user's health status, based on a diary of their food consumption. The aim was to have an improved healthier lifestyle resulting in disease prevention. The authors developed two ontologies: a) the first represented the food/nutrition domain and was named the 'food' ontology; b) the second was a personal profile ontology that represented some specific personal factors like gender, age, height, and weight. Moreover, the system was provided with expert designed rules that helped in the decision regarding the health status of the consumed food.

Figure 22 represents the used features in the work of Shaban-Nejad *et al.* (2012) on the topic of hospital surveillance: how advances in semantic technology could improve the analysis and detection of hospital-acquired infections (HAI). The researchers introduced the HAI ontology (HAIO), which represented a common understanding of the infection control domain and a means to achieve data interoperability in the area of hospital-acquired infections. HAIO was part of the HAIKU framework, which provided recommendations for infection control, risk detection and stratification. To build HAIO, many data resources were integrated, including: i) patients' records, ii) hospital guidelines and/ or iii) biomedical ontologies. SPARQL queries were used to retrieve data. The authors aimed for the HAI ontology to be re-used in many healthcare contexts.

Figure 23 represents the work of Baldassini *et al.* (2017) on assistive technology. The authors designed a system based on virtual reality and SW technologies. The system provided a domestic environment for elderly users where they could perform specifically tailored physical exercises based on their health conditions. The authors designed three ontologies to model: a) their health conditions, b) the living environment and the used sensor devices, and c) their physical measurements. The knowledge in the ontologies was extracted via a SPARQL endpoint. Moreover, reasoning over defined SWRL rules was performed to derive new information, such as the exercise workload needed for each user. The data collected by the system were saved as reports and shared with the intended caregivers or clinicians.

Figure 24 represents the work of Birjali, Beni-Hssane and Erritali (2017). The purpose of this paper was to analyse semantic suicidal sentiments expressed in social networks. A machine learning algorithm was proposed to compute semantic analysis for extracted Twitter data. For analysing the similarity of the words or terms, the authors used the WordNet lexical database. The extracted tweets were integrated locally and the terms within were mapped to an ontology.

Health Management Examples

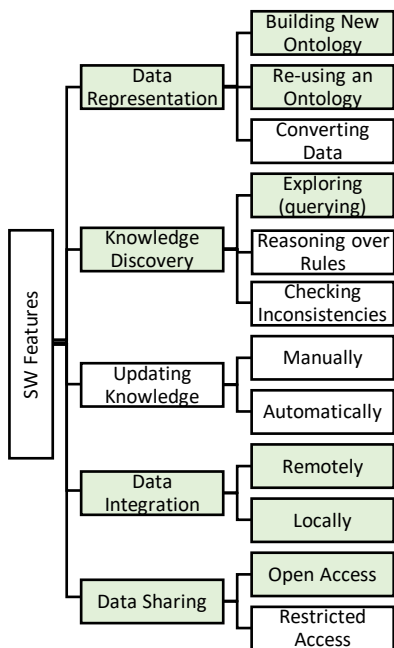


Figure 25: An Example for 'Supporting clinical trials and secondary use of EHRs'

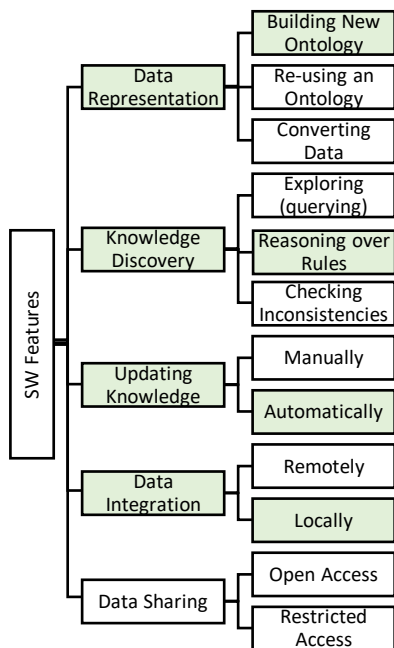


Figure 26: An Example for 'Clinical pathways and patients care plans'

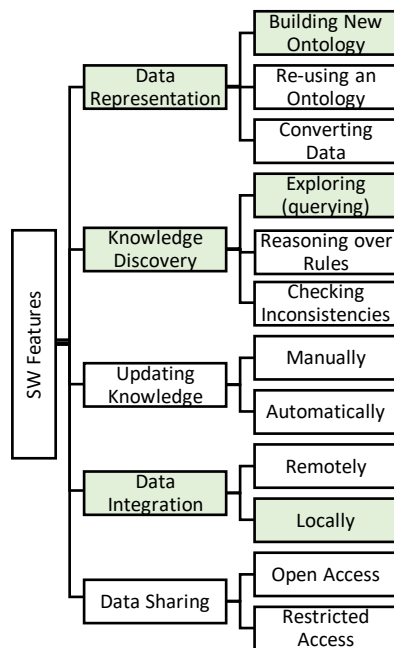


Figure 27: An Example for 'Clinical guidelines and policies'

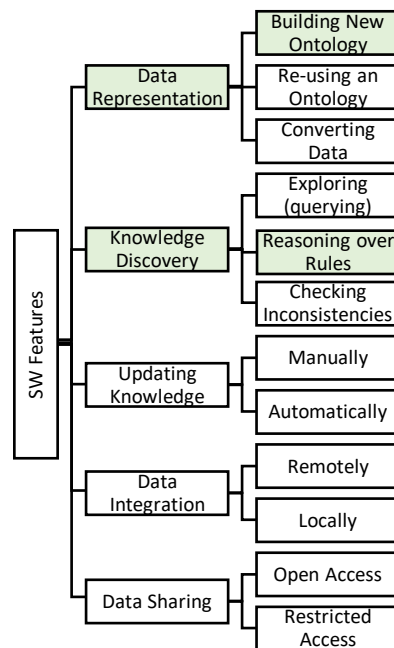


Figure 28: An Example for 'Training staff and supporting learning'

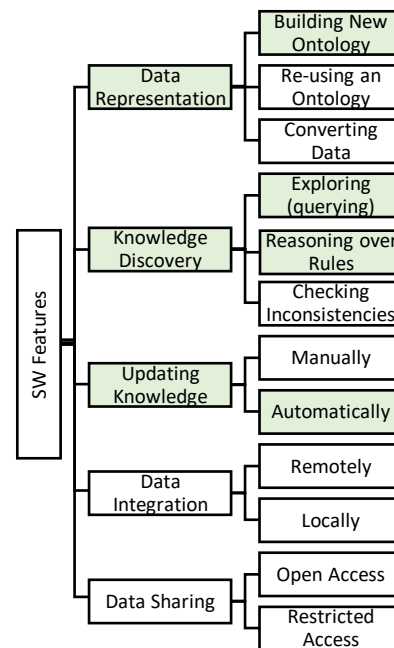


Figure 29: An Example for 'Workflow, communication and business processes'

Figure 25 highlights the used SW features on a taxonomy for Chondrogiannis *et al.* (2017). The authors demonstrated a system for automatically selecting patients for clinical trial recruitment by semantically representing the eligibility criteria. They developed an ontology namely the 'eligibility criteria' (EC) ontology, which was shared online. The terms in the ontology were linked to terms in international classification systems, such as the 10th version of the 'international classification of diseases' (ICD10), 'chemical entities of biological interest' (ChEBI) and LOINC. SPARQL was used for retrieving the criteria data after translating them from XML.

Figure 26 represents the work of Alexandrou *et al.* (2012) in modelling clinical pathways. The authors designed an ontology named SEMPETH (SEMantic PATHways) that conceptualised the domain of clinical pathways. SEMPETH linked three different domains under its umbrella: i) medical, ii) organisational and iii) financial. Interestingly, one of the motivations behind building SEMPETH was the ability to update knowledge in real-time. Moreover, semantic web rules (SWRL) were used in a case for discovering new knowledge about the human papillomavirus disease and its treatment pathway.

Figure 27 represents the work of Puustjarvi and Puustjarvi (2016) on a SW taxonomy. The authors aimed to disseminate relevant clinical guidelines to clinicians. They developed two ontologies: a) 'profile' ontology, which represented the clinicians' information and b) the 'guideline' ontology that represented the associated guidelines for each clinician's role. By linking and querying these ontologies, the relevant guidelines were connected with each clinician.

Figure 28 demonstrates the work of Bajenaru and Smeureanu (2015) on a SW taxonomy. The authors developed an e-learning system for the purpose of training healthcare managers. The ontology represented the student's profile and the learning materials. The educational content for each user was personalised, based on their profiles by inferring over defined rules.

Figure 29 represents an example of clinical workflow topic by Dang *et al.* (2008, 2009). The authors aimed to enable healthcare users and administrators by creating and managing medical workflows as well as personalising them. The team developed a healthcare ontology that represented hospital resources and business processes to be implemented within a workflow management system. The data in this project was retrieved and saved using relational database technology. Business rules were defined within the ontology to manage and update the available processes for physicians to handle.

Pharmaceutical Examples

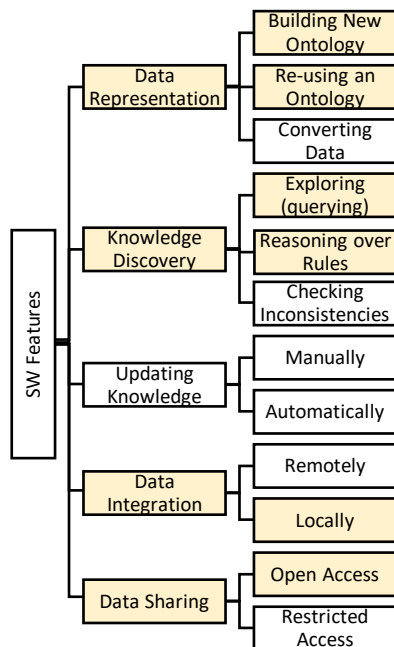


Figure 30: An Example for 'Finding adverse drug events'

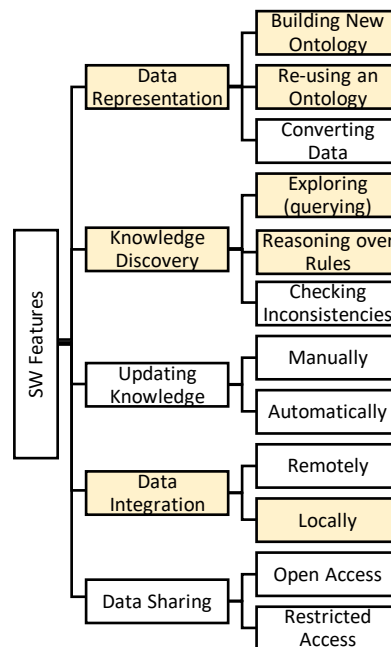


Figure 31: An Example for 'Drugs discovery'

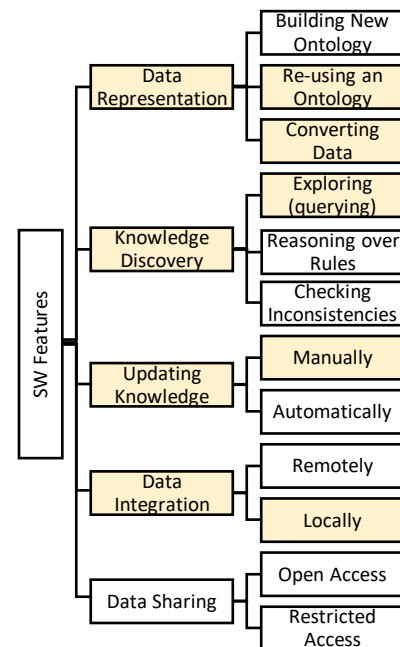


Figure 32: An Example for 'Testing drugs'

Figure 30 represents the work of Natsiavas *et al.* (2018) in finding / preventing adverse drug events. The authors built an ontology for the communication and representation of pharmacovigilance signals to help prevent the occurrence of any possible adverse drug events. This project was part of the pharmacovigilance (PV) field, specifically focusing on adverse events and signals detecting. The authors developed a new ontology named openPVsignal ontology that represented and shared signal information in the pharmacology field. The ontology's design was based on three types of PV signal data resources: i) the WHO's Pharmaceuticals Newsletter, ii) reports from the Netherlands pharmacovigilance centre's lab and iii) announcements contained in the FDA drug safety communication. The design of the ontology was based on the FAIR (findable, accessible, interoperable, and re-usable) data principles; thus it was publicly accessible. Moreover, the second aim in this project was to exploit the automatic reasoning features of the SW. Part of designing the ontology involved defining rules that represented the communicating of knowledge in PV. The reasoning was performed over the designed rules to find similarities in the data or interlinks between them.

Figure 31 highlights the work of McCusker *et al.* (2014) on a SW features taxonomy. The authors built a drug repurposing semantic framework to discover new links between existing drugs and diseases. They named their framework 'repurposing drugs with semantics' (ReDrugS). The authors built a data repository for the relationships between drugs, proteins and biological processes. The data was borrowed from existed biological, pharmacological, disease and gene ontologies such as 'drugbank' and 'gene ontology' (GO). An ontology representing a confidence measure was provided to be used as an indicator for the quality of the integrated information. The confidence measure was calculated based on the opinions of biochemistry experts. Reasoning was used via SPARQL queries to classify the asserted information, based on quality.

Figure 32 maps the work of Kohonen *et al.* (2013) on a SW features taxonomy. The authors developed a data warehouse named 'toxbank' for the purpose of representing and supporting test replacements for repeated dose systemic toxicity testing on animals. The 'toxbank' warehouse consisted of an integrated public toxicogenomics data such as the TGP (TG-GATEs) and the DrugMatrix datasets, and existing ontologies such as the gene ontology (GO). The authors developed a tool for converting the data into RDF. The needed data was accessed via a query engine. Any updated or new documents were added to the system via a provided GUI.

4.4 Discussion

By systematically reviewing the literature, two taxonomies were produced that mapped the health research literature exploiting the SW approach. The health aims taxonomy answered the first research question by addressing the types of health questions, while the second taxonomy answered the second research question by highlighting the SW features used in the reviewed literature. The third research question, concerning the affordances and challenges faced when using the SW approach, is answered in this discussion section.

The discussion is divided into three main sections, detailed below. The first section is about the health questions that were addressed in the literature, together with details of their main characteristics. The second section discusses using the SW features in the literature in general and is therefore spread across the four health aims. The last section analyses the noted affordances and challenges in the relevant literature.

4.4.1 Addressed Health Questions

The review revealed four main aims for the health questions in the literature that have been addressed, the aims being: i) medical, ii) public health, iii) health management and iv) pharmaceutical. Under each main aim there were a number of health questions (topics) that were addressed. In total, they were 17 health sub-aims or addressed questions. The biggest share of the literature, almost half, addressed medical questions. More than half of these medical questions were diagnostic in nature. Coming next in the percentage of usage in the literature is the topic of promoting public awareness within the contextual aim of public health. The next topics on the list of sub-aims with lower usage rates are: i) monitoring via devices, ii) treatment recommendations and iii) environment surveillance.

By analysing the aims behind using the SW to answer these types of questions, specific characteristics of the addressed topics were identified. The characteristics or the nature of the addressed questions were:

1. questions relating to decision making
2. questions depending on heterogeneous data integration
3. questions aiming for personalisation
4. questions relying on the availability of web information

The best example to illustrate the decision requiring nature in a question is in one that is diagnostic. During diagnosis, several decisions need to be reached based on factors including symptoms and health status, which can be represented as a logical rule. Shadbolt, Hall and Berners-Lee (2006) stated that one of the SW's capabilities was not just representing objects and any relationships between them, but also rules and logical statements. Thus, decision-based or decision-requiring questions were successfully addressed by using semantic rules and the power of reasoning using logic.

Many health questions depend on linking different topics to find an answer. This situation is very similar to connecting pieces of puzzle to achieve the final answer. The main promise of the SW and LD is to be able to connect data on the web (Bizer, Heath and Berners-Lee, 2009). Sagotsky *et al.* (2008) mentioned that interdisciplinary life sciences are powered by data integration on the web. The reviewed literature gave many examples of health-centric questions involving different types of data integration. For example, there was a type of question that involved integrating data generated from: i) hardware, such as patients' monitoring sensors (Puustjarvi and Puustjarvi, 2015), ii) medical examination devices (Maragoudakis, Maglogiannis and Lymberopoulos, 2008), or even iii) using assistive technology devices for special-needs people (Baldassini *et al.*, 2017). Other questions demanded integrating administrative data such as hospital guidelines and policies (Puustjarvi and Puustjarvi, 2016), or hospital process workflows (Dang *et al.*, 2008, 2009).

Another popular type of integration was linking patient records or EHRs with other forms of biomedical ontologies or clinical terminologies. Many of the medical-type questions used this type of integration: i) for diagnosing (Pathak *et al.*, 2012), ii) for prescribing treatments (Rung Ching Chen *et al.*, 2012), iii) for making care plans (Wang *et al.*, 2015), and / or iv) for selecting patients for clinical trials (Chondrogiannis *et al.*, 2017). Chondrogiannis *et al.* (2017) endorsed the importance of the existence of considerable amounts of controlled clinical terminologies and standards to be used for the purpose of representing and exchanging information for clinical studies. Zenuni *et al.* (2015) conducted a systematic review about the most used published healthcare ontologies and semantic data repositories. The review included examples like SNOMED CT and LOINC for ontologies and CardioSHARE and Bio2RDF for semantic repositories. The authors believed that the growth of health data in the web is rapid but still in non-semantic proprietary formats. It was also pointed out there is still a lack of user-friendly semantic data interfaces. These two important usage-limiting factors play a big role in encouraging health researchers to start using the SW.

The third characteristic relevant for addressed questions was 'personalisation', which involves aggregating data on a personal-level. Personalisation can be considered as one type of data

integration but for personal data only. Traditional linking methods in the health field involved the linking of personalised data by third parties centres in order to maintain data anonymity (Kotwal *et al.*, 2016). The traditional linked data approach was used in the health research community to give an overview both breadth and depth when addressing the tested sample (Falster, Jorm and Leyland, 2016). In the SW approach, questions involving personalisation were found in clinical settings such as: i) enhancing a treatment's quality by creating personalised treatment plans for patients (Wang *et al.*, 2013) and ii) in personal health information web services that provide targeted health information for every individual (Wang *et al.*, 2010).

The latter example also described the final characteristic of the addressed questions in the SW approach. Berners-Lee, Hendler and Lassila (2001) stated in their article about the SW: "The Semantic Web is not a separate Web but an extension of the current one." In the review, the SW was also used to consume health information broadly distributed on the web. Public health systems that aimed to promote health information, were built to access the available information on the web and produce electronic personal health consultants, as in the previous example (Wang *et al.*, 2010). These types of papers were similar to the literature review carried out by Eysenbach (2003). The review concentrated only on papers discussing health information consuming using the SW. The author concluded that the SW may open more opportunities for health information consumers by finding and aggregating information. However, this outcome might lead to overlaying on the web as a source of health information.

4.4.2 The Semantic Web's Features

The second part of this current study's review was concerned with finding the most used SW features in the literature. A SW features taxonomy was provided, which contained 12 features of the SW technology, categorised under five main aims. The three most used features in the literature were: i) representing data via building ontologies, ii) local data integration and iii) discovering knowledge via 'querying'. Coming next in the list were: iv) reasoning, v) restricted data sharing and vi) re-using ontologies.

Berners-Lee listed four rules for linking data in what is known as "Linked Data Principles" (Berners-Lee, 2006). The first principle focused on the concept of data representation by using URIs for naming. The second and third principles were concerned with defining data properly to enable data sharing by using look able URIs and standards for providing useful information. The last principle was all about including links to others work to enable data integration and re-use.

From reviewing the SW health literature, the concept of data representation was used widely via the developing and creating of required ontologies. Ontologies play a crucial role in the SW

technology as they form the core technology that the rest of the SW features relies on. Shaban-Nejad *et al.* (2016) said that the representation, exchanging and discovering of data are enabled by ontology use. Thus, there was no surprise when seeing the high usage rate for building ontologies across all health aims, as reported in the reviewed literature. It seems that most of the literature fulfilled the first principle of LD.

The second and third principles were concerned with how to use standards and proper data representation to enable data sharing. Part of the principles reflected the concept of data representation but was mostly focused on the purpose and process of data sharing. The review showed a very low level of sharing data publicly in all types of health research. Rare examples of this sharing include: a) the Apollo structured vocabulary (Apollo-SV) project (Hogan *et al.*, 2016), b) the SWAN (semantic web applications in neuromedicine) ontology (Ciccarese, Wu, Wong *et al.*, 2008) and c) the ToxBank data warehouse (Kohonen *et al.*, 2013). It was noted that most of the projects used restricted data sharing for security reasons. The results of the reviewed literature showed that the second and third LD principles were not fully followed, unlike the situation with the first principle.

The last LD principle encouraged SW users to include links and mappings to others' work. This principle could be reflected by the use of both 'data integration' and 're-using ontologies' features'. Slightly more than half of the systems re-used others ontologies in all of the topics except for public health research, where the use rate was even lower. The types of re-used ontologies were more biomedical and clinical in nature and did not fit with the public health topics. Moreover, the concept of a united global database or interlinked global datasets was not met. Many of the reviewed systems failed to define mappings to others resources or provide access to their data. 'Federated querying' was not used a lot in the reviewed papers, while the majority of the projects used local data integration in their systems. The fourth principle for defining links for others' work was fulfilled to a certain degree according to the results of the systematic review. There was a moderate reusing by accessing others' ontologies, but most of the data integration happened at a local level.

Two more SW features were clearly evident in the reviewed literature: i) 'discovering knowledge via querying' and ii) 'reasoning'. Querying was the third most used feature in the literature and reasoning was the fourth. Most of the papers used SPARQL for querying their ontologies; however, there were some systems that preferred to save their data in relational databases and accordingly used a suitable querying language. As mentioned earlier, most of the querying took place locally, while federated queries had a very low usage.

The reasoning feature had some variance in its use across the four health aims. The use of reasoning over rules depended on the topic being discussed, as mentioned earlier in the addressed health questions section. Medical topics employed reasoning the most, then health management topics and finally pharmaceutical topics. In the pharmaceutical field, detecting drug interactions were suitable to be inferred from rules representing drug metabolic pathways (Arikuma *et al.*, 2007). Machado *et al.* (2013) also reviewed 11 translational medicine systems; less than half of them only used the inference approach. However, all of them addressed medical or pharmacological questions. Thus, the use of inference and reasoning features was not sufficiently exploited across the different health topics. There is an opportunity to fill this gap, perhaps with more health research considered through inference-suitable questions like decision-based ones. Zenuni *et al.* (2015) noticed that many systems in their review used data mining techniques and machine learning to analyse data and discover new knowledge. While these tools were effective in many cases, they were not strictly designed for semantic data.

4.4.3 The Semantic Web's Affordances in Health Research

From reviewing the literature and the attempt to understand the reasons behind health researchers using the SW, several affordances were identified. The three most noticed properties for the SW were its ability: a) to represent a domain of knowledge by using ontologies, b) to integrate heterogeneous data and c) to discover knowledge via exploring and inferring.

Data representation was used for a variety of topics and domains. Topics like clinical pathways and treatment plans (Hu *et al.*, 2012), or clinical guidelines (Puustjarvi and Puustjarvi, 2016), or neuro medicine (Ciccarese, Wu, Kinoshita, *et al.*, 2008). It was not just that ontologies were able to represent different domains, some of the literature mentioned that representing data using triples was simple, too. Pathak *et al.* (2012) stated that the simplicity of modelling data using RDF was an important factor in using the SW. Moreover, the authors continued to say that the RDF model was more flexible in updating, adding or deleting data than the relational model.

Supporting SW technology is the availability of converting and mapping tools to RDF. As Marshall *et al.* (2012) noted, the ideal situation is to map relational data into RDF; however, in some cases there would be a need for conversion using any of the many tools available.

Following the main vision of the SW in linking data on the web, the majority of the reviewed papers used the SW for its ability to integrate heterogeneous data. Integration took different shapes in the reviewed projects. Some systems used integration for aggregating personal data (personalisation), where the researchers aimed to create personalised treatment plans for patients (Wang *et al.*, 2013). Other papers integrated data remotely using federated querying as

in Marshall *et al.*(2012), who created a federated query requesting all language renderings of a specific drug product. Three main data resources were involved in this query: RxNorm, DrugBank and DBPedia. Another type of integration is to achieve semantic interoperability by exchanging data between heterogeneous data sources locally. In Puustjärvi and Puustjärvi (2009), the authors aimed to allocate clinical resources between healthcare managers; such an aim demands cooperation between several heterogeneous information systems within a healthcare institution.

Finally, one of the distinguished affordances of using the SW is the ability to use it to discover new knowledge. Knowledge can be discovered according to the literature by either exploring linked data or inferring new information. Linked data can be explored by traversing links between connected elements. Querying data is commonly used to explore linked data, which was used by 77% of the reviewed papers. Exploring linked data is one of the main concepts in the SW vision, where Berners-Lee introduced the “Linked Data Principles” (Berners-Lee, 2006).

Although the literature varied in the usage rate of reasoning among health aims, it is still considered an important reason to use the SW technology. The concept of reasoning is related primary to the early SW vision by enabling machines to understand and process the data using AI (Berners-Lee, Hendler and Lassila, 2001). Reasoning was used in the literature for mainly two reasons: i) either to check inconsistencies and errors in ontologies or ii) to infer new information based on defined rules. For example, Horridge *et al.* (2014) used OWL reasoning as a quality assurance technique for checking for inconsistencies in the large medical ontology ICD-11. Part of the reason behind using the SW for inconsistency checking was the ability to trace the source items that caused the logical inconsistency, especially by using ‘reasoners’ like Pellet (Huang *et al.*, 2014). According to Wang *et al.*(2015), the semantic reasoning improved the intelligence of their developed independent clinical pathway system. Baldassini *et al.*(2017) stated that the semantic reasoning tools allowed ontologies to infer new information according to the knowledge model. In this project, they were able to classify a user’s cardio-respiratory fitness condition according to environmental and physiological data collected by monitoring devices.

4.4.4 Challenges and Recommendations

The health projects that used the SW approach for integrating data were expected to utilise the full capabilities of this technology. However, from reviewing the systems several challenges were noticed that hinder the SW’s usage in many cases. The noticed challenges were related to the used data whether it was in the data’s accessibility, security, heterogeneity, or quality.

Firstly, accessing sensitive data was a main obstacle in sharing and integrating data. Across the different health aims, private data such as patient or administrative information was difficult to

share publicly or integrate remotely in a secure manner. In line with this finding, Machado *et al.* (2013) stated that the solution for this issue is to enforce some security protocols from the SW community to track the owners, users and editors of the data. Moreover, following the example of the traditional linking approach in allowing independent data centres to handle data between data owners and developers, the issue of the availability of private data for research could be overcome. The role of the independent data centres is to prepare, clean and anonymise patients' data in order for that information to be suitable for research purposes. Such centres work like a link between private data owners and health researchers and developers.

The second identified challenge from the literature is securing private data. As mentioned earlier, some of the reviewed literature used personal private data obtained from patients' records in hospitals. In addition to this type of projects, there are projects that aim to build personal health web services, which include private information for each registered user. Securing such projects containing sensitive private information is a challenge to be faced when attempting to employ the SW in health research. In addition, Eysenbach (2003) said the SW would magnify the opportunities, as well as the challenges, associated with the web; in particular, regarding the privacy of the consumers' data. The semantic representation of personal data would ease and automate the integration process but with a high risk of exploitation by spammers. Supporting academia and security related studies by the SW community is recommended for the purpose of providing appropriate security access protocols and mechanisms.

Data quality is another general challenge facing any health research. In projects that use some data analytics techniques to draw conclusions or find patterns in the data, the quality of this data is crucial and will affect any conclusions reached. The data needs to be scientifically accurate and technically well-formed. In many health cases, the decision that is based on these results could be both critical and sensitive, as in the case of diagnostic systems or choosing the best treatment plan for a patient. Zenuni *et al.* (2015) agreed with how extremely important was the data quality in terms of affecting the results of an analysis.

For a better understanding of the challenges being presented and faced in such an interdisciplinary domain, the co-operation between computer scientists and health experts is essential. For example, improving reasoning technologies to check for inconsistencies could help to improve the quality of the integrated data. Consulting health experts regarding ontologies design and testing results is also crucial before one can trust the efficiency of a system. In this context, McArthur (2009) explained the importance of using the experience of health information professional in developing the SW in five key areas: i) ontology development, ii) knowledge

translation, iii) information retrieval, iv) scientific publishing, and v) resource classification and indexing.

Data heterogeneity is also another challenge that was evident in most of the reviewed literature. In the literature, various sources and forms of used data were mentioned. For example, Stavropoulos *et al.* (2016) built a system for monitoring elderly's health by analysing collected data from wearable sensors, while Singh *et al.* (2013) built a system that analyse medical images and extracts identifiable features from them. Many papers used hospital and patients' records as part of the integarted data as well. Health data on the web is growing in number, however, most of the available data is in a non-semantic format (Zenuni *et al.*, 2015). To step forward into achieving linked health data, more effort should be aimed into converting and publishing health data in a semantic format.

4.5 Conclusion

By systematically classifying the literature, two mapping taxonomies were produced: a) health aims and b) SW features. The purpose of these taxonomies was to help health and computer science researchers in understanding the specific characteristics of health questions addressed in the SW literature, as well as understanding the affordances and challenges of using the SW in health research.

This review has revealed 17 addressed health questions that were divided into four main aims: i) medical, ii) public health, iii) health management and iv) the pharmaceutical field. The diagnostic medical questions achieved the biggest share of the literature. By analysing the relationships between the addressed health question (health topic) and the used SW feature, four characteristics of the addressed questions were concluded. The characteristics were: i) decision-based questions, ii) personalisation questions, iii) questions based on integrating heterogeneous data and iv) questions rely on accessing web information.

The second contribution in this review was the SW features taxonomy. The taxonomy included 12 sub-features categorised under five main semantic web features, which were: i) data representation, ii) knowledge discovery, iii) data integration, iv) data sharing and v) updating knowledge. The most used SW features in ascending order were: a) building new ontology, b) integrating data locally and c) exploring data via querying. Some of the research trends and gaps were noticed based on the usage rate of the SW features across health aims. All health aims agreed on the popularity (usage rate) of building new ontologies and integrating data locally. The aims also agreed on the shortage of usage for: i) checking inconsistencies, ii) updating knowledge manually, iii) updating knowledge automatically, iv) remote data integration and v) open access

sharing. Moreover, there was an interesting positive correlation between medical questions and the use of reasoning over rules. Public health questions had a shortage in re-using ontologies, while pharmaceutical projects had a lack in the rate of sharing data, even privately.

In conclusion, the semantic web was used for its ability to represent a domain of knowledge, integrate heterogeneous data and, to discover knowledge via exploring linked data and inferring new information. However, the vision of integrating the data on the web in one interlinked database was not fulfilled. The notions of re-using published ontologies, providing mapping between resources and enabling public data sharing were neglected in many cases. Four challenges were identified in the process of applying the SW approach in health research: i) data accessibility, ii) data security, iii) data heterogeneity, and iv) data quality. Encouraging cooperation between SW developers and health experts, as well as enabling independent data centres to prepare and anonymise sensitive data could help in overcoming some of the issues that are being faced in this interdisciplinary domain. Supporting security related studies and encouraging more publishing health data in a semantic format will also help in supporting the use of the SW technologies in the health research.

4.6 Summary

This chapter systematically reviewed the literature of health research employing the SW. One of the review's aims is to identify the main types of addressed health questions and topics to answer the first research question. By answering this question, a better understanding for what types of questions the SW is good at dealing with can be reached.

The answer to this question is represented in the health aims taxonomy that includes 17 types of addressed health questions categorised under four main health aims (medical, public health, health management and pharmaceutical). The diagnostic medical questions had the biggest share of the literature. It was noticed that the identified questions have four basic characteristics: i) decision-based questions, ii) personalisation questions, iii) questions based on integrating heterogeneous data, and iv) questions rely on accessing web information.

The second contribution in this chapter is the SW features taxonomy that answers the second research question. The taxonomy includes five main semantic web features: i) data representation, ii) knowledge discovery, iii) data integration, iv) data sharing and v) updating knowledge; together with an additional 12 sub-features. The most used SW features in the literature were: a) building new ontology, b) local data integration, and c) discovering knowledge via querying.

In addition to finding the usage rate for each SW feature generally, a cross-reference analysis of the usage rates for the SW features across the health aims was provided. The aim of this analysis is to highlight any research trends, opportunities or gaps in the literature. One of the relationships noted was a positive correlation between medical questions and the use of reasoning over rules. Public health papers showed lower usage rates than the other aims in re-using ontologies and applying reasoning over rules; a result which implies that the nature of these topics does not suit the properties of these two features.

The systematic review was also used to analyse any mentioned affordances or challenges facing the use of the SW, or its application in, the huge field of health research. This part of the review is dedicated to answering the third research question. Some of the affordances identified in the literature were: a) the ability to represent knowledge, b) integrate heterogeneous data and c) discover new knowledge via exploring and reasoning. Some of the identified challenges that hinder using the SW in health research were issues regarding data accessibility, security, heterogeneity, and quality. Co-operation between SW developers and health experts, enabling independent data centres to handle sensitive data, supporting security academia and encouraging semantic health data publishing were some initiatives recommended to overcome the challenges.

Chapter 5 **Semantic Web Demonstrator for Health Research**

In the last chapter, the interdisciplinary literature relevant to using the semantic web (SW) in health research was systematically reviewed. The review revealed some of the most and some of the least addressed health questions in the literature. Medical and public health topics were the most discussed, while the topics of health management and pharmaceutical issues came third and fourth. In addition to the discussed topics, the most and least used SW features were also analysed. The review concluded that the SW was mainly used for representing knowledge, integrating heterogeneous data and discovering new knowledge.

In an attempt to examine some of the least discussed topics in the literature and to reflect upon the uses of the SW in those topics, this chapter demonstrates the fabric of this scenario by building a proof-of-concept model. A semantic web-based demonstrator has been built to represent NHS prescriptions in England. This chapter continues in answering the third research question. The aim of building the demonstrator is to illustrate how the SW tools were used to represent a health-related topic, as well as to analyse any affordances and challenges faced in the process. The first section in this chapter discusses the motivation and background information informing the chosen topic, while the second section describes the design of the demonstrator.

5.1 Prescriptions in England

As part of the transparency movement promised by the British government, anonymised dispensed prescription data is available to the public at two levels: presentation and chemical. The ability to derive chemical-level data from presentation-level data was stopped in 2012. From 2010, dispensed prescription data was available online openly and is updated monthly. The formal authority responsible for publishing the data is NHS Digital, known formerly as the Health and Social Care Information Centre (HSCIC).

The data covers any dispensed prescription issued in England by a general medical practitioner (GP) or a non-medical prescriber, such as a nurse or a pharmacist. Therefore, the number of prescriptions published every month can be quite massive, totalling around 18 million. It should be noted that the smaller number of private prescriptions are excluded from the monthly published datasets (Health and Social Care Information Centre (HSCIC), 2015).

5.1.1 Why Prescriptions?

The prescription dataset includes three types of information: i) information about the prescriber, ii) the medication and iii) the prescription's cost and dose. Thus, this dataset can be thought of as a link between pharmaceutical medication information and health management information (the prescribers' information). This special property makes dispensed prescription data a good candidate for representing heterogeneous data integration between two different health fields (pharmaceutical and health management). Interestingly enough, these two health fields were the least cited topics according to the results of the systematic review presented in the last chapter; an outcome which supports the idea of further exploring these two fields.

In addition, there were a limited number of papers that specifically tackled the topic of prescriptions in the systematic review. Puustjärvi and Puustjärvi (2006) illustrated 'querying' features in an electronic prescription system built using the SW. The authors discussed prescriptions as an electronic system to create a link between healthcare practitioners and registered pharmacies. The situation was different with Papakonstantinou *et al.* (2011), where the authors proposed a semantic wiki system containing training material for healthcare providers involving 'the ePrescribing process'. Another example was a clinical decision support system (CDSS) that was used for supporting choosing the most suitable antibiotic prescription for a patient (Calvillo-Arbizu *et al.*, 2014).

The aim of the paper from Puustjärvi and Puustjärvi (2006) was to suggest a solution for managing clinical data, while Papakonstantinou *et al.* (2011) suggested a different approach for training staff. Calvillo-Arbizu *et al.* (2014) had a purely medical aim in deciding the best treatment. It is noticed that these papers tackled the topic of prescriptions from: a) clinical management, b) human resources and c) medical perspectives. However, the prescription topic in this current work represents historical prescription data (dispensed prescriptions) at a national level to be used as statistical data for research purposes. To the best of this author's knowledge, there no papers which presented or investigated the NHS dispensed prescription data, or the British National Formulary (BNF) medication data, by using SW technology.

5.1.2 About the NHS

The prescription data is published under the supervision of the National Health Service (NHS). The NHS Digital, known formerly as Health & Social Care Information Centre (HSCIC), is the national provider of NHS information and IT systems. The NHS was launched in 1948 with a budget of £437 million that grew to £101 billion in 2015/2016 (NHS Choices, 2017a). The prescriptions' annual cost was approximately £9.3 billion in 2015 (NHS Digital, 2016).

From April 2013, the NHS structure in England was changed. In the period prior to April 2013 there were 10 Strategic Health Authorities (SHA) that managed 151 Primary Care Trusts (PCTs) where the practices reside. The structure was changed to having four regional teams administering 24 Area Teams (ATs) which administer 211 Clinical Commissioning Groups (CCGs) that commission the practices. Also, over 150 non-CCG cost centres exist such as trusts, councils and private company providers. The ATs are responsible directly for these non-CCG cost centres (HSCIC, 2015).

The NHS Digital announced in their website (NHS Digital, no date) that from the first of April 2020, a new data structure will be in place. The NHS intends to move from flat structure data model to structured XML files. The NHS apparently understands the importance of providing structured data for public use, either in research or for developing systems. Hopefully, in a more advanced step forward, the NHS will consider publishing their data in a semantic-friendly form, such as RDF, in order to simplify the integration process with other SW systems.

5.1.3 About the BNF

In the NHS prescriptions datasets, the medication information is represented by using the standards of the British National Formulary (BNF). The BNF is a UK pharmaceutical reference book that discusses correct dosage, indication, interactions and side effects of drugs. It is published jointly by the British Medical Association and the Royal Pharmaceutical Society. Two versions are available every year in March and September (British National Formulary Publications, 2017).

Each BNF book is divided into four sections:

- 1) Front Matter: including how to use the BNF, changes in this edition, guidance on prescribing.
- 2) The Main Chapters: including drugs monograph, uses, doses, safety issues, drug classes, treatment summaries and prices. The chapters are divided according to different aspects of medical care like infections, vacancies and appliances.
- 3) Appendices: including interactions, substances, advisory labels and wound care.
- 4) Back Matter: including dental and nurse prescribing, guidance for life support and emergency doses.

The Business Services Authority (BSA) publishes an online version of the BNF as a spreadsheet. The spreadsheet includes the taxonomical structure of known drugs detailing their names and codes. Each row in the spreadsheet represents a presentation of a drug, and each presentation has a unique code to identify it. These codes are built by a hierarchical

aggregation of detailed information about the drug. For example, one of the Paracetamol presentations is Panadol_Tab 500mg, which has the BNF code 0407010H0BFADAM. Each digit in this code represents a piece of information, as can be seen in table 5 (BSA,2017).

Example '0407010H0BFADAM'	Digit Position	Representing
04	Digits 1 & 2	The BNF chapter
07	Digits 3 & 4	The BNF section
01	Digits 5 & 6	The BNF paragraph
0	Digit 7	The BNF sub-paragraph
H0	Digits 8 & 9	The chemical substance
BF	Digits 10 & 11	The product
AD	Digits 12 & 13	The strength and formulation
AM	Digits 14 & 15	The equivalent (generic or branded product)

Table 5: The representations of the Panadol tab 500mg BNF presentation code

5.1.4 Examples of Prescription Projects

This section presents a list of projects which used prescription data as part of their work.

5.1.4.1 Mapping English GP prescribing data: a tool for monitoring health-service inequalities

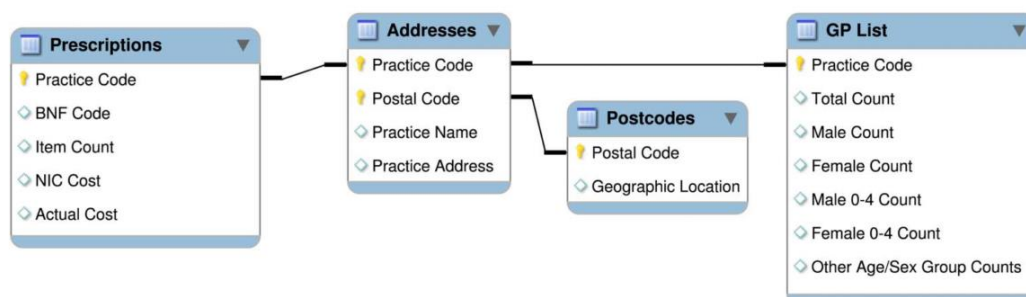


Figure 33: The linkage between the datasets in the 'Mapping English GP prescribing data: a tool for monitoring health-service inequalities' (Rowlingson et al., 2013)

(Rowlingson *et al.*, 2013) discussed the variations of prescribing performance on a national level. The authors demonstrated the variation of prescribing performance by using interoperable maps

relied on open data. The used datasets were downloaded from NHS information centres and the Ordnance Survey Website.

The main concept of their work depended on linking four datasets as shown in figure 33. The linking key between the datasets is the practice code which is a unique identifier for the available data. Spatial database (PostGIS, no date) was used to link datasets and produce maps representing their work. One of the use cases discussed was calculating the cost per person rate for the prescriptions of Metformin Hydrochloride and Methylphenidate across England. The results were represented as five coloured categories on the map of England.

5.1.4.2 OpenPrescribing

The OpenPrescribing project was provided by the Evidence Based Medicine (EBM) Data Lab at the University of Oxford (EBM DataLab - University of Oxford, 2017). OpenPrescribing.net offered customised prescription datasets for researchers, by relying on the NHS published open data. There were four provided services on the website. Two of the services were for finding multiple prescribing indicators for practices and Clinical Commissioning Groups (CCGs). The third service was for finding national trends for a specific drug. The most interesting part was in the last service, which allowed for a customised analysis by choosing the prescription's issuer and the prescribed medicine. This service can be very handy for researchers and policymakers. PostGIS spatial database (PostGIS, no date) was also used here for allowing geographical location querying.

5.1.4.3 iView

The NHS Digital aims to provide information and IT systems for health and care services. One of the services provided by the NHS Digital is the iView tool (NHS Digital, 2017b). iView is a web service designed for authorised health care users to: i) access information, ii) build reports, iii) generate charts, iv) export data and v) save reports for a variety of provided datasets. Due to the sensitivity of some of the datasets, accessibility was restricted to certain users. One of the datasets that had only one minor requirement, a mandatory registration step, was the CCG Prescribing dataset. The available information was similar to what was used in Rowlingson et al. (2013) and OpenPrescribing (EBM DataLab - University of Oxford, 2017). However, the difference was in the information presentation as the prescriptions can be viewed at CCG or Area Team levels only and not a practice level. This tool was useful in customising and filtering the viewed data to meet individual's needs as it offered a certain amount of flexibility.

5.1.4.4 Information Services Portal

Information Services Portal is a web service offered by the NHS Business Services Authority (NHS BSA) for NHS and public users (BSA, 2017). The portal offered many reports that can be downloaded, as well as different datasets. Sensitive data was restricted to NHS registered employees only. However, the prescriptions datasets had open access. The data can be filtered by a national, regional, area team or PCO level. Moreover, the viewed prescriptions can also be filtered by the BNF chapter. This system offered another method to view the prescriptions data but did not offer any data integration tools with other datasets.

5.1.4.5 Prescribing Analytics

Another project that used the open prescription dataset to analyse the prescriptions nationally was the Prescribing Analytics project (Prescribing Analytics, no date). Prescribing Analytics consisted of a group of computer scientists and NHS practitioners aiming to study possible prescription savings. They looked into the variety and range of prescriptions costs for branded against generic statin drugs and found that about 27 million pounds a month were wasted on branded versions in 2012. The results of the percentage of proprietary statin prescribing by CCG level was demonstrated on a map. The technology used for creating the maps was Leaflet open-source interactive maps (Agafonkin, 2017).

5.1.4.6 Medicine statistics: GP prescribing by constituency, 2015

In the 'Medicine statistics: GP prescribing by constituency, 2015' briefing paper Baker (2016) investigated the prescription rates via four approaches: i) type of medicine, ii) UK countries, iii) constituencies in England and iv) antibiotic resistance (Baker, 2016). The main source the work relied on was the open prescriptions datasets published by NHS Digital (HSCIC previously). The work aimed to measure the burden of illness and disease nationally by using the prescriptions rate. It was found that prescription rates demonstrate extraordinarily wide variations across England, and specifically in the examined constituency areas.

5.2 Demonstrator Architecture

This section describes the architecture of the prescription demonstrator. The aim of building this demonstrator is to illustrate, with a real case example, how the SW can be used in health research. This section mainly aims to answer the second research question in this thesis. The demonstrator represents the integration of heterogeneous pharmaceutical and health management data. The pharmaceutical data are represented in the British National Formulary (BNF) and the health management data are contained in both the dispensed prescriptions administration data and the NHS practices administration data. Finally, the system is tested with several health questions as queries.

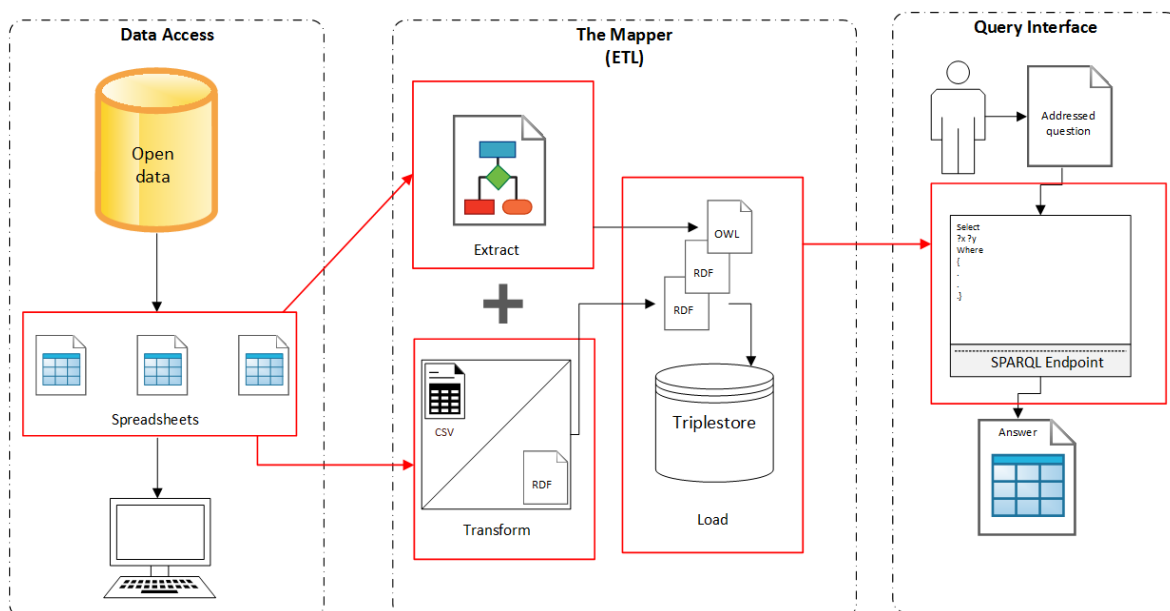


Figure 34: The prescription demonstrator architecture

The demonstrator contains three main components: a) Data Access b) Mapper and c) Querying Interface. Data representation and integration are illustrated in the first and second components, while the third component demonstrates data retrieval and discovery. Figure 34 shows the demonstrator architecture's three components; the next sections discuss the components in more detail.

5.2.1 Data Access Component

Data access is the first component in the prescriptions demonstrator. In this component, the steps of choosing, preparing and saving the right data are performed. The approach used to access the data is by firstly choosing the appropriate dataset from the available open data. The second step is to download it into a local machine and save it as a comma separated values (CSV) file. Prepare all the spreadsheets to be converted into RDF files in the second component.

5.2.1.1 Data Sources

All datasets chosen in this work were ‘open’, published under the open government licence (NHS Digital, 2017d). Seven datasets were downloaded from three different NHS departments: i) NHS Digital, ii) NHS BSA and iii) NHS Choices. NHS Digital is responsible for managing and analysing health data nationally (National Health Service, no date). NHS Business Services Authority (NHS BSA) supports other NHS organisations and the public by providing various electronic services (NHS BSA, 2017). NHS Choices is the official website NHS England with plenty of information and services offered to the public (NHS Choices, 2017b).








Dataset Name	Source	Licence	No. of Rows	Size
<i>Practice Detailed Prescribing Information</i>	NHS BSA	 Open Government Licence for public sector information	18,843,675	3.25 GB
<i>GP Branches in England</i>	NHS Choices	 Open Government Licence for public sector information	9,848	1.49 MB
<i>Number of Patients Registered with a GP Practice</i>	NHS Digital	 Open Government Licence for public sector information	7,493	344 KB
<i>GP Practices in England and Wales</i>	NHS Digital	 Open Government Licence for public sector information	13,386	1.54 MB
<i>GP Opening Times</i>	NHS Choices	 Open Government Licence for public sector information	139,696	4.65 MB
<i>GPs' Staff</i>	NHS Choices	 Open Government Licence for public sector information	83,992	3.01 MB
<i>BNF Code Information</i>	NHS BSA	 Open Government Licence for public sector information	76,813	16.8 MB

Table 6: The chosen open datasets for the prescriptions demonstrator

Table 6 lists some information about the seven chosen datasets for building the prescription demonstrator. These datasets are categorised into three types: i) prescription, ii) NHS practices and iii) BNF data. The three types are discussed in detail in the following sections.

5.2.1.2 Prescription data

Regarding the prescription data, there are several sources available online for prescription datasets. Within the NHS itself, there are different departments producing similar prescription datasets. For example, the *iView* tool provided by the NHS Digital (NHS Digital, 2017b) is used for viewing a customised version of the prescription data. Thus, the data can be viewed by a CCG or at Area Team level. Unfortunately, there is no filtering option which would allow viewing data at practice level.

Another source of the prescription data is the *GP practice prescribing data – Presentation level* that is published by the NHS Digital under the open government licence (NHS Digital, 2017d). This dataset contains information about the codes for the area team, CCG and prescriber practice along with the prescribed medicine, cost and quantity.

However, the *Practice Detailed Prescribing Information (PDPI)* dataset that can be accessed via the Information Services Portal (managed by the NHS Business Services Authority (BSA) under the Open Government License) contains similar information with some additions. The extra information includes the names and codes of the Regional Office, Area Team, CCG and practices, which gives it extra credit when compared to the NHS Digital dataset. The PDPI dataset has been chosen for this reason and also because the BSA provides other useful datasets that are consistent with this dataset. The chosen version of the dataset is from June 2017.

5.2.1.3 Practice information and NHS structure data

This data category relates to any organisational structural information in the NHS, or any supporting information about any of the NHS practices. The NHS's hierarchal structure information is taken from the same *PDPI* dataset published by the BSA. A more detailed level of the hierarchy information was added to the different practices' branches. This information is taken from the *GP branches in England (GP.csv)* dataset that is published by NHS Choices under the Open Government Licence (NHS Choices, no date b).

In addition to the hierarchal information, extra information about the GP practices is provided. For example, information about the total number of patients in each practice is added. The *Number of Patients Registered with a GP Practice* dataset is downloaded from the NHS Digital Website under the Open Government Licence under the name of *gp-reg-pat-prac-sing-age-all.csv* (NHS Digital, 2017c). Another dataset is the *GP Practices in England and Wales (epraccur.csv)* from NHS Digital under the Open Government Licence (NHS Digital, 2017a). This dataset includes information about the practices' addresses, postcodes and prescribing' settings.

On the practices' branches level, data about location, opening times and staff is added, too. From the previously mentioned dataset, *the GP (GP branches in England)*, information about the latitude and longitude measurements and postcodes for each branch are downloaded under the Open Government Licence (NHS Choices, no date b). The NHS Choices also published *the GP Opening Times* (NHS Choices, no date a) and *GPs' Staff* (NHS Choices, no date c) datasets under the Open Government Licence. *The GP Opening Times* dataset lists the opening and closing times for the weekdays per each branch, while *the GPs' Staff* dataset contains employees' information like title, given name, family name, job title and, for GPs only, the GMC number.

5.2.1.4 Medication data

In the prescription dataset (*PDPI*), the prescribed medicine is presented by the BNF coding system. Therefore, an online version of the BNF book is added to the prescription demonstrator. The BNF book version 72 (British Medical Association, Pharmaceutical Society of Great Britain and Joint Formulary Committee (Great Britain), 2016) is used. It was published in spreadsheet format in the Information Services Portal in the BSA website under the Open Government License (NHS Business Services Authority (BSA), 2017). In *the BNF Code Information* dataset the following are included: i) names and codes of chapters, ii) sections, iii) paragraphs, iv) sub-paragraphs, v) chemical substances, vi) products and vii) presentations of all registered drugs.

5.2.2 The Mapper

The Mapper is the second component in the prescription demonstrator. In this component, there are three main steps regarding the process for dealing with the data: Extract, Transform and Load (ETL). The ETL process is responsible for extracting data from various sources, then cleaning and customising the data to prepare it to be uploaded into a data warehouse (Skoutas and Simitsis, 2006).

The first step in the Mapper component is to extract the needed information from the previously mentioned datasets. Secondly, the chosen datasets in the Comma Separated Values (CSV) format need to be transformed into Resource Description Framework format (RDF). Finally, all converted data is loaded into a proper triple store. Designing the ontologies (the schema) is performed in the 'Extract' step, while the 'Transform' and 'Load' steps focus on the instances or the actual data.

5.2.2.1 Extract

The Extract step involves the approach followed to extract the knowledge from the chosen data for the purpose of designing ontologies or schema to incorporate into the prescriptions' demonstrator. The software used for designing the ontologies is Protégé version 5.2.0 that was developed at Stanford University (Musen, 2015). The ontologies were developed using the Web Ontology Language (OWL).

The approach for designing the ontologies is by studying and analysing the chosen data. The primary concepts are extracted; then those extracted concepts are represented as classes or data properties. Another type of information extracted from the datasets is the relationships between the concepts. These relationships are represented as data or object properties.

It is important to mention that the intention in designing the ontologies is not to have a comprehensive design for abstracting the NHS prescriptions in reality, but rather a simpler version of the real-world scenario for demonstration purposes only. In future, this work could be improved and expanded to reflect a precise version of the prescribing workflow in England.

In designing the ontologies, the focus was on the chosen datasets, which are categorised into three types: i) prescriptions' information, ii) NHS organisational structure and iii) medication data. From these three categories, three ontologies were designed: a) NHS ontology, b) BNF ontology and c) prescriptions ontology. The three ontologies are discussed in the following sections in detail.

5.2.2.1.1 Designing the Prescriptions Ontology

Mapping the prescriptions ontology from the *PDPI* dataset is straightforward as each row in the *PDPI* dataset represents an instance of the *Prescription* class. The *Prescription* class is the only defined class in this ontology with five linked data properties to it. *Quantity*, *ADQUsage*, *items*, *actualCost* and *NIC* are the prescription's properties that were represented as separated columns in the original dataset. Table 7 shows the mapping between the *Prescription* class in the prescriptions ontology and the extracted concept in the *PDPI* datasets.

Class Name	Column Name	Dataset Name
<i>Prescription</i>	Quantity	<i>PDPI</i>
	Items	<i>PDPI</i>
	NIC	<i>PDPI</i>
	Actual Cost	<i>PDPI</i>
	ADQ Usage	<i>PDPI</i>

Table 7: Extracting the prescriptions ontology information

The prescriptions ontology is considered as the bridge between the BNF ontology and the NHS ontology. Each instance of the *Prescription* class has two links. The first one is the object property *issuedBy* that points to a specific instance of the *Practice* class in the NHS ontology. The other property *contains* points to a specific instance of the *Presentation* class in the BNF ontology. Figure 35 shows the linkage between the prescriptions ontology and the NHS and BNF Ontologies.

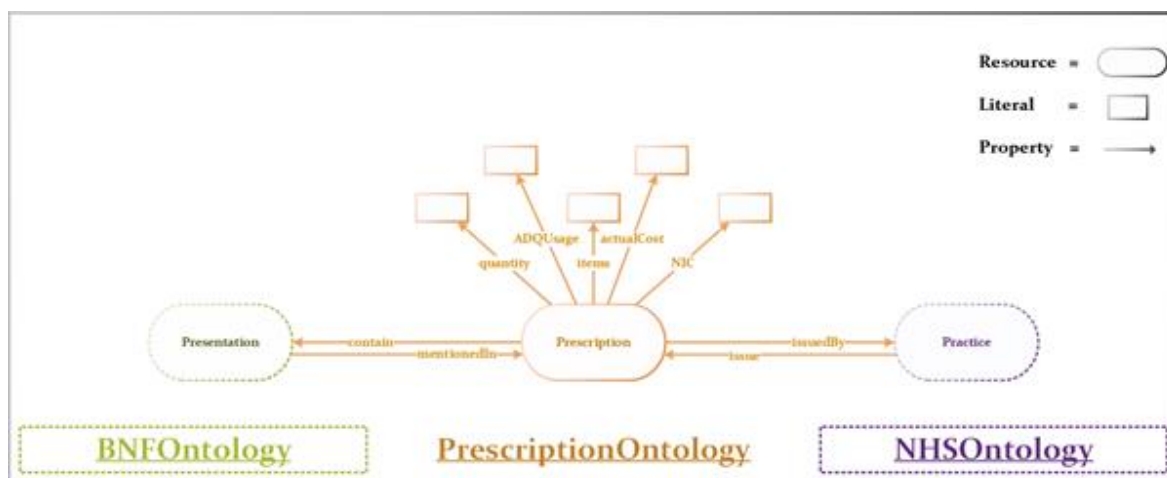


Figure 35: The prescriptions ontology

5.2.2.1.2 Designing the BNF Ontology

The mapping of the BNF ontology is extracted from the *BNF Code Information version 72* dataset. There are seven defined classes, two object properties and two data properties in the ontology. The main concepts extracted from the BNF dataset are: *i) Chapter, ii) Section, iii) Paragraph, iv) Sub-paragraph, v) Chemical Substance, vi) Product and vii) Presentation*. Each concept represents a class with two attached data properties namely *name* and *code*. Table 8 shows the mapping between a class in the BNF ontology and the extracted concept in the *BNF 72* dataset.

Class Name	Column Name	Dataset Name
<i>Chapter</i>	Chapter Name	BNF 72
	Chapter Code	BNF 72
<i>Section</i>	Section Name	BNF 72
	Section Code	BNF 72
<i>Paragraph</i>	Paragraph Name	BNF 72
	Paragraph Code	BNF 72
<i>Subparagraph</i>	Subparagraph Name	BNF 72
	Subparagraph Code	BNF 72
<i>ChemicalSubstance</i>	ChemicalSubstance Name	BNF 72
	ChemicalSubstance Code	BNF 72
<i>Product</i>	Product Name	BNF 72
	Product Code	BNF 72
<i>Presentation</i>	Presentation Name	BNF 72
	Presentation Code	BNF 72

Table 8: Extracting the BNF ontology information

Similar to the NHS ontology, the associations between the BNF classes were designed in a hierarchical manner as chapters include sections, which include paragraphs, which include subparagraphs, which include chemical substances, which include products and finally products include drug presentations. However, after testing the ontology with several queries including inferred data, there were logical inconsistencies in the results. Therefore, the type of relationships between the classes has been changed to *belongsTo* and *has* with all the classes remaining at the same hierarchal level, see figure 36.

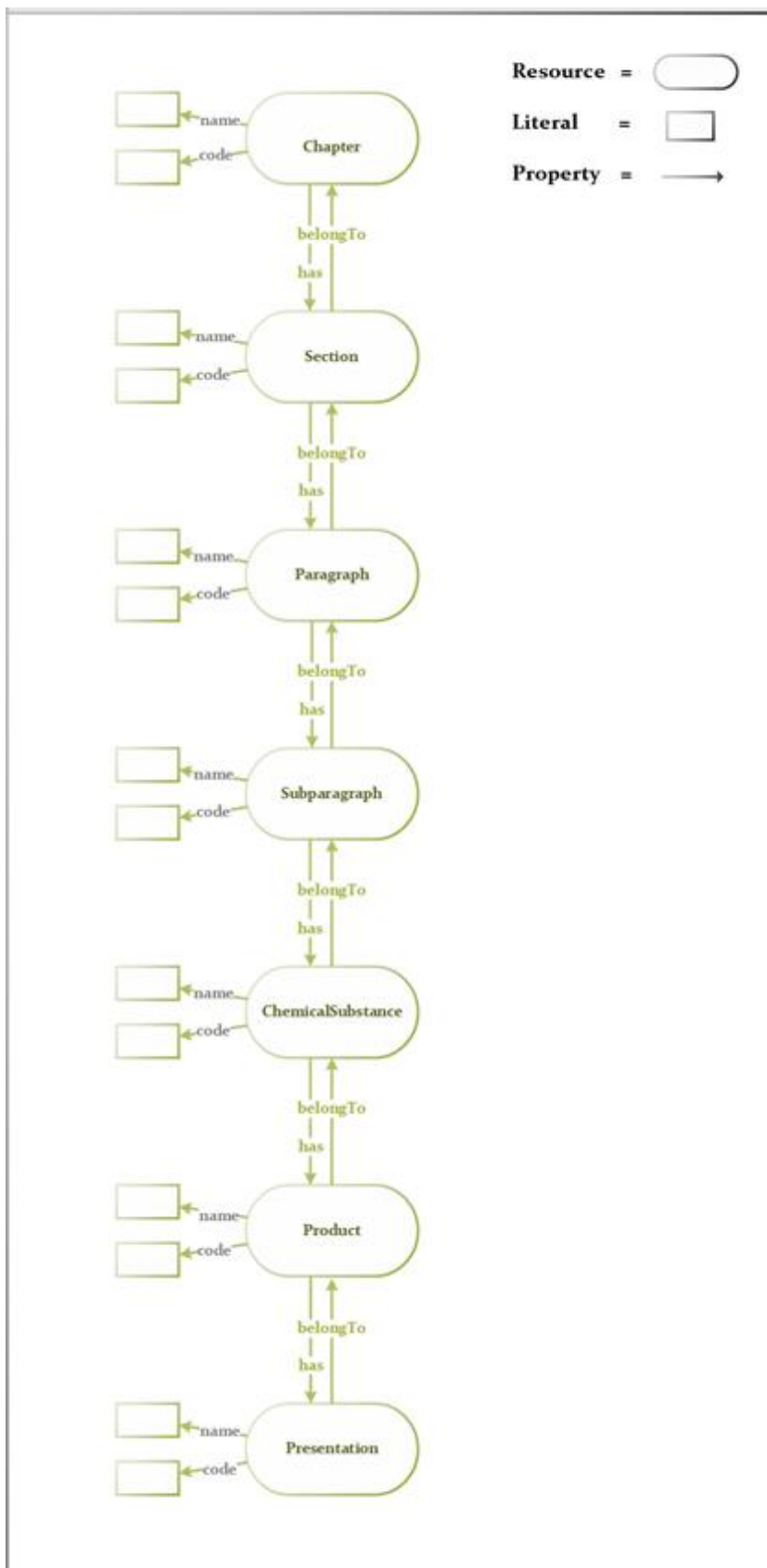


Figure 36: The BNF ontology

5.2.2.1.3 Designing the NHS Ontology

The NHS ontology represents the hierarchical structural information of the NHS units with more focus on the practices-related information. There are 11 classes in this ontology. Four of them are re-used from other published ontologies (FOAF and ORG), and the rest are defined classes mapped from the six chosen datasets. Table 9 shows the mapping between a class in the NHS ontology and the extracted concept in the chosen datasets.

Class Name	Column Name	Dataset Name
<i>RegionalOffice</i>	region office name	PDPI
	region office code	PDPI
<i>AreaTeam</i>	area team name	PDPI
	area team code	PDPI
<i>PCO</i> (<i>ClinicalCommissioningGroup</i> , <i>NonClinicalCommissioningGroup</i>)	PCO name	PDPI
	PCO code	PDPI
<i>Practice</i>	practice name	PDPI
	practice code	PDPI
	number of patients	Number of patients registered with a GP practice
	post code	GP practices in England and Wales
	prescribing settings	GP practices in England and Wales
<i>Branch</i>	branch code	GP branches in England
	branch name	GP branches in England
	post code	GP branches in England
	latitude	GP branches in England
	longitude	GP branches in England
<i>Employee</i>	title	GPs' staff
	first name	GPs' staff
	last name	GPs' staff
	job title	GPs' staff
	GMC number	GPs' staff
<i>WorkingTime</i>	day	<i>GP opening times</i>
	opening time	<i>GP opening times</i>
	closing time	<i>GP opening times</i>

Table 9: Extracting the NHS ontology information

The NHS ontology includes part of the NHS's organisational structure and some of the practices' information and properties. Regarding the NHS structure, it has been decided to rely on *the PDPI dataset* taxonomy because it is compatible with the dataset that will be used later in designing

the prescriptions ontology. *The PDPI dataset* shows the hierarchical relationships between: a) the regional offices, b) area teams, c) primary care organisations (PCO) and d) practices in order.

In the early design of the NHS ontology, these concepts were tied by containment relationships in a hierarchical design, which implies that: i) the practices are part of the PCOs, ii) the PCOs are part of the ATs, and iii) the ATs are part of the regional offices. Similar to the BNF ontology, the containment design led to inconsistent results, therefore, it has been replaced with the use of *belongTo* and *has* relationships between the concepts, see figure 37.

The NHS ontology also covers some of the practices' properties such as: i) addresses, ii) the number of registered patients, iii) staff information and iv) working hours. The practices' postcodes and prescribing settings types are extracted from *the GP Practices in England and Wales's dataset*, while the values of the data property *numberOfPatients* are taken from *the Number of Patients Registered with a GP Practice dataset*. The remaining concepts included in the NHS ontology are *Branch*, *Employee* and *WorkingTime*. They are extracted from the following datasets respectively: *GP branches in England*, *GPs' staff* and *GP opening times*.

Finally, some external vocabularies are borrowed and re-used with the defined classes in the NHS ontology. From the Organization ontology (ORG) (World Wide Web Consortium, 2014), *org:Organization*, *org:OrganizationalUnit*, *org:hasMember*, *org:memberOf* are integrated with our defined concepts. Other vocabularies are taken from the FOAF ontology (Brickley and Miller, 2010) like *foaf:Agent*, *foaf:Person*, *foaf:title*, *foaf:givenName*, *foaf:familyName*.

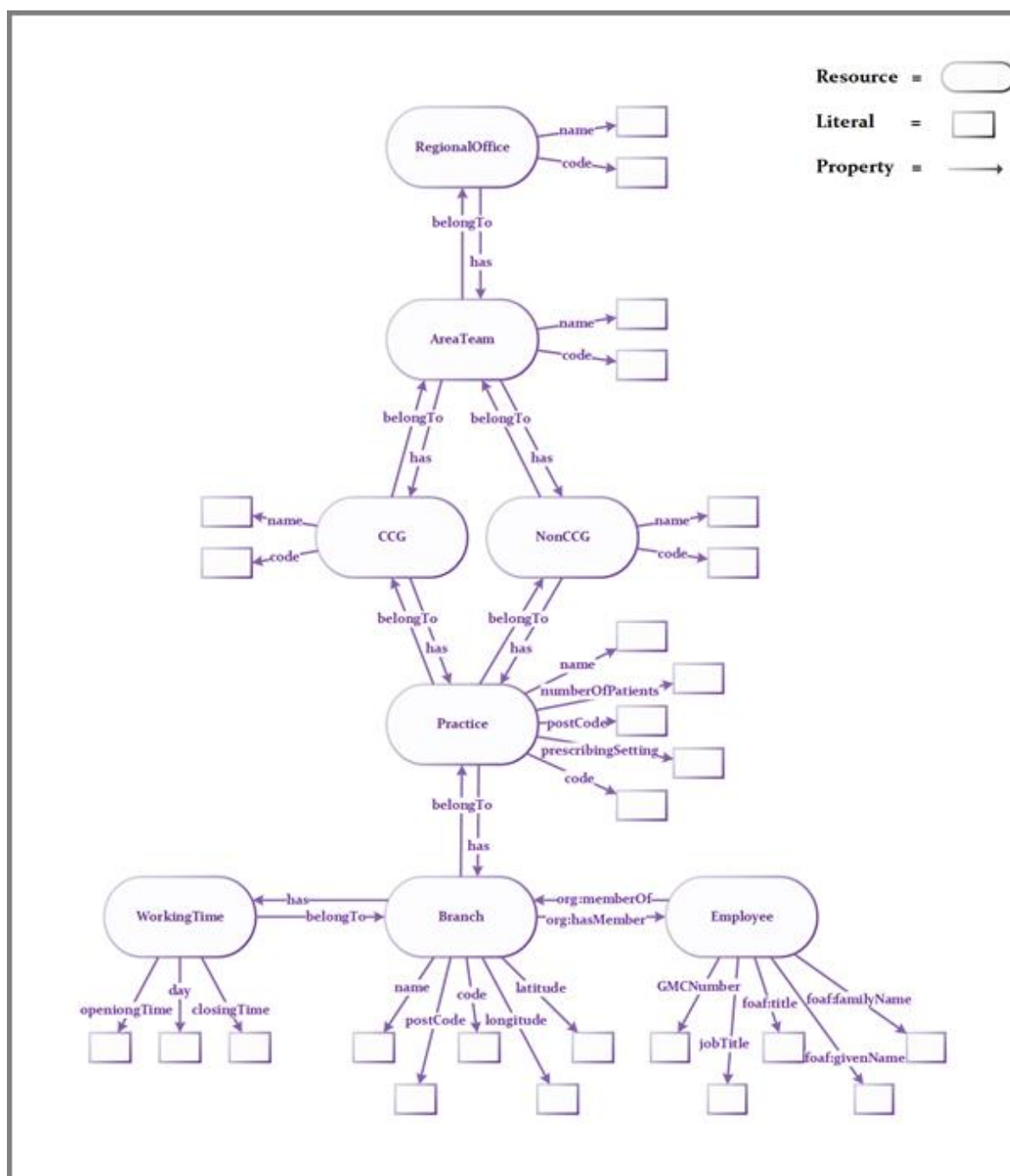


Figure 37: The NHS ontology

5.2.2.1.4 The Three Interlinked Ontologies

By integrating the three ontologies, the main design of the mapping ontologies in the prescriptions demonstrator is achieved. Figure 38 shows the result of the integration by having the NHS ontology in purple, BNF ontology in green and the prescriptions ontology in orange.

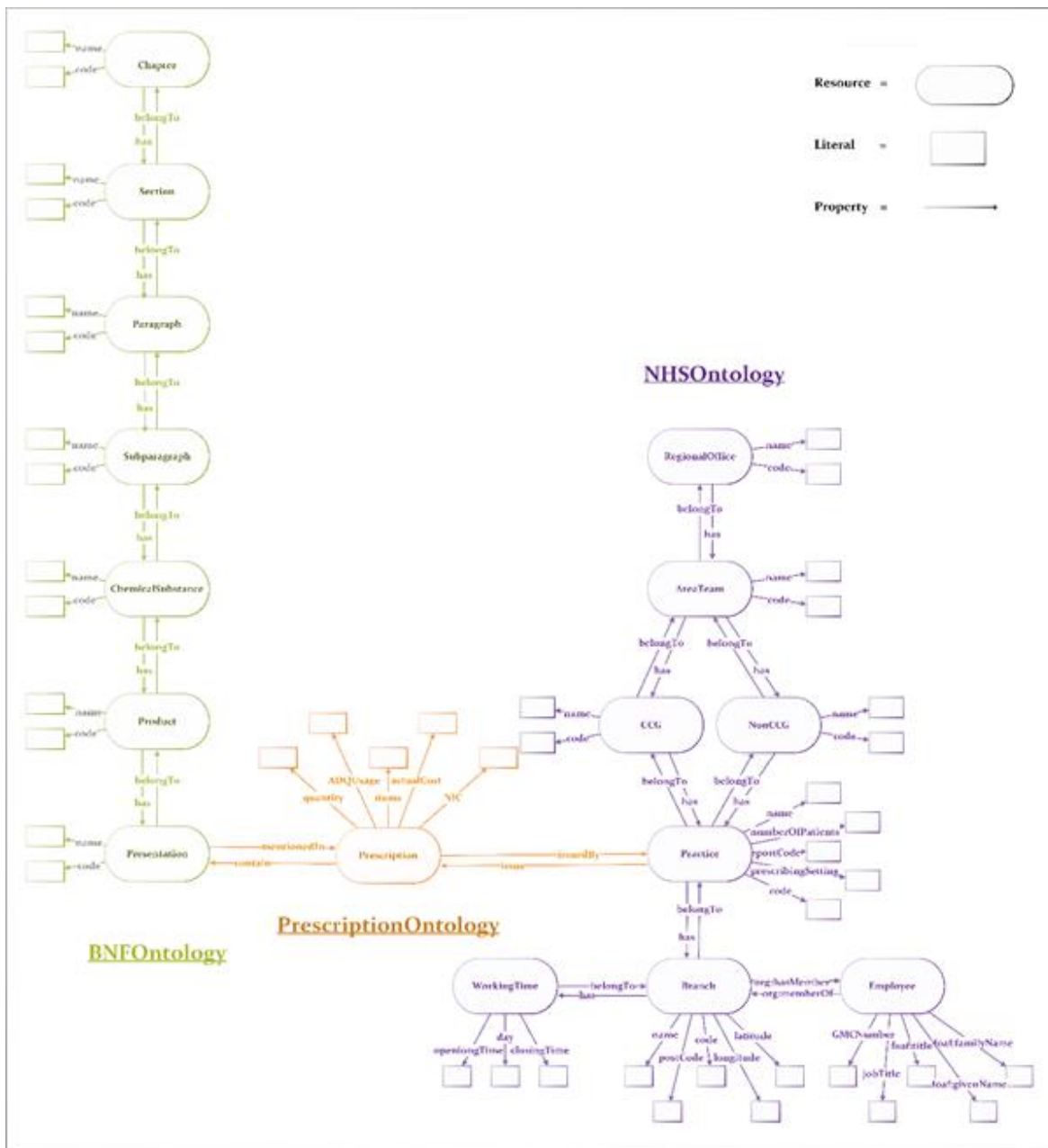


Figure 38: The integration of the NHS, BNF and prescriptions ontologies

5.2.2.2 Transform

After designing the three interlinked ontologies, it is time to prepare the actual data or instances to fit in the prescription demonstrator. In the transform step, the data in the seven identified spreadsheets is transformed into a more structured and friendly semantic form (RDF). The instances of the NHS ontology are transformed from the following datasets: *GP practices and surgeries*, *Number of Patients Registered with a GP Practice (gp-reg-pat-prac-sing-age-al)*, *GP Practices in England and Wales (epracur)*, *GP Opening Times and GPs' Staff*. The BNF ontology's instances are converted from the *BNF Code Information, the 72 edition*. The *Practice Detailed*

Prescribing Information (PDPI), June 2017 dataset is the biggest dataset among the others. Each row in the spreadsheet is mapped to an instance of the *Prescription* class in the prescriptions ontology.

A summary of the transforming step for the NHS, BNF and Prescriptions ontologies is found in figure 39.

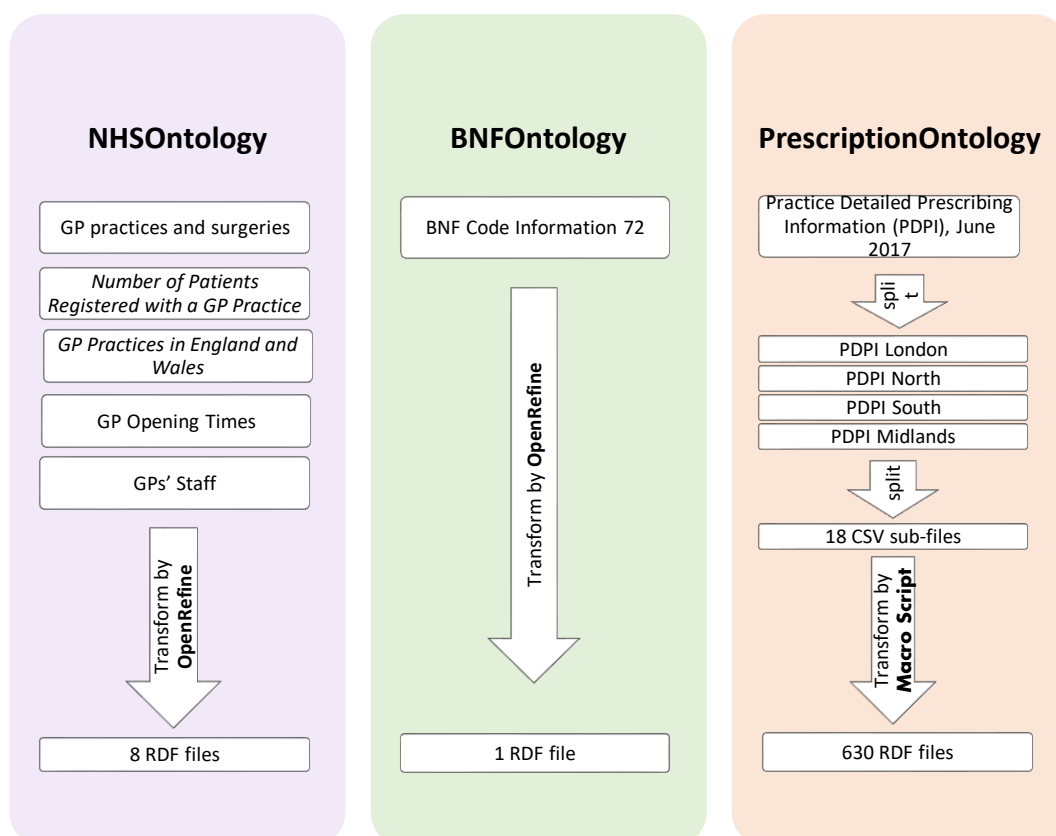


Figure 39: Transforming the CSV files into RDF files for NHS, BNF and prescriptions ontologies

For the NHS and BNF data, an open source application named OpenRefine (previously Google Refine) version 2.5 combined with RDF extension version 0.6 is used to convert the tabular CSV data into RDF. Open Refine is a tool provided by Google for cleaning and transforming data (Google, 2010). It can be linked with different extensions in order to perform specific type of transformation such as converting CSV data into RDF. The used RDF extension in this task provides graphical mapping for the targeted CSV data using RDF skeleton. The mapped data is then exported as RDF text file(s).

To extract the information from the available CSV files, several mapping rules were designed and applied to the files via OpenRefine. The following steps were followed for the mapping process:

- 1) Uploading the targeted CSV file into an OpenRefine project.
- 2) Uploading the OWL ontology file(s) which include the classes definitions into the project.
- 3) Choosing a suitable identifiable column of the CSV data to define URIs for the RDF instances of the OWL classes.
- 4) Declaring the type (the class) of the defined instances in step 3 using *rdf:type*.
- 5) Declaring any data properties linked to the defined instances in step 3 using the available data properties in the OWL file(s).
- 6) Declaring any object properties between defined instances using the available object properties in the OWL file(s).
- 7) Starting the converting process and exporting the results locally as an RDF file(s).

Region Code	Region Name	Area Team Code	Area Team Name
Y54	North of England	Q44	Cheshire, Warrington and Wirral Area Team
Y54	North of England	Q45	Durham, Darlington and Tees Area Team
...
Y55	Midlands and East of England	Q53	Arden, Herefordshire and Worcestershire Area Team
Y55	Midlands and East of England	Q54	Birmingham and The Black Country Area Team
...

Table 10: Part of the region office - area team CSV file

RDF Schema Alignment

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

Base URI: <http://www.semanticweb.org/ma8g13/ontologies/2017/6/untitled-ontology-19#> edit

RDF Skeleton RDF Preview

Available Prefixes: rdf owl xsd nhs rdfs foaf +add prefix manage prefixes

Region code URI

[x nhs:Regional_Office](#)

[add rdf.type](#)

a

[x >nhs.code->](#) **Region code cell** b

[x >nhs.name->](#) **region name cell**

[x >nhs.has->](#) **area team code URI**

[x nhs:Local_Office](#)

[add rdf.type](#)

a

[x >nhs.code->](#) **area team code cell**

[x >nhs.name->](#) **area team name cell** b

[add property](#)

[add property](#)

Figure 40: An example of a mapping rule using OpenRefine

Table 10 and figure 40 shows examples of mapping instances of the *RegionalOffice* and *AreaTeam* classes in the NHS ontology from a CSV file. Part of the *PDPI dataset* that shows the relationship between regional offices and area teams is represented in table 10. Two Regional offices and four linked area teams are shown in this example. The columns *Region Code* and *Area Team Code* are unique; thus, they were used as identifiers in the URIs instances definitions, see figure 40 (a). The *Region Code* instances are declared as type of the class *RegionalOffice*, while the *Area Team Code* instances are of type *AreaTeam*. Regarding the data properties, there are two used ones: *nhs:code* and *nhs:name*. For both of the *RegionalOffice* and *AreaTeam* defined instances, the region and area team code and name columns were mapped via the *nhs:code* and *nhs:name*, correspondingly, see figure 40 (b). Finally, the *nhs:has* object property was used to link the *RegionalOffice* and *AreaTeam* defined instances, see figure 40 (c).

After applying the defined mapping rule on the uploaded data in OpenRefine, an RDF file was exported locally with the converted data. Table 10 above shows two regional offices coded Y54 and Y55. The regional office with the code Y54 is linked to area teams with the codes Q44 and Q45. This piece of information with the names of the instances are shown in RDF in the following segment of the RDF file:

```
<rdf:Description rdf:about="http://www.semanticweb.org/ma8g13/NHSONtology#region_Y54">
  <rdf:type rdf:resource="http://www.semanticweb.org/ma8g13/NHSONtology#RegionalOffice"/>
  <nhs:code>Y54</nhs:code>
  <nhs:name>North of England</nhs:name>
</rdf:Description>

<rdf:Description rdf:about="http://www.semanticweb.org/ma8g13/NHSONtology#areaTeam_Q44">
  <rdf:type rdf:resource="http://www.semanticweb.org/ma8g13/NHSONtology#AreaTeam"/>
  <nhs:code>Q44</nhs:code>
  <nhs:name>Cheshire, Warrington and Wirral Area Team</nhs:name>
</rdf:Description>

<rdf:Description rdf:about="http://www.semanticweb.org/ma8g13/NHSONtology#region_Y54">
  <nhs:has rdf:resource="http://www.semanticweb.org/ma8g13/NHSONtology#areaTeam_Q44"/>
</rdf:Description>

<rdf:Description rdf:about="http://www.semanticweb.org/ma8g13/NHSONtology#areaTeam_Q45">
  <rdf:type rdf:resource="http://www.semanticweb.org/ma8g13/NHSONtology#AreaTeam"/>
  <nhs:code>Q45</nhs:code>
  <nhs:name>Durham, Darlington and Tees Area Team</nhs:name>
</rdf:Description>

<rdf:Description rdf:about="http://www.semanticweb.org/ma8g13/NHSONtology#region_Y54">
  <nhs:has rdf:resource="http://www.semanticweb.org/ma8g13/NHSONtology#areaTeam_Q45"/>
</rdf:Description>
```

The second part demonstrated in table 10 is regards the Y55 region office. This region office is linked to two area teams coded Q53 and Q54. This information is represented in RDF as the following:

```
<rdf:Description rdf:about="http://www.semanticweb.org/ma8g13/NHSONtology#region_Y55">
  <rdf:type rdf:resource="http://www.semanticweb.org/ma8g13/NHSONtology#RegionalOffice"/>
  <nhs:code>Y55</nhs:code>
  <nhs:name>Midlands and East of England</nhs:name>
</rdf:Description>

<rdf:Description rdf:about="http://www.semanticweb.org/ma8g13/NHSONtology#areaTeam_Q53">
  <rdf:type rdf:resource="http://www.semanticweb.org/ma8g13/NHSONtology#AreaTeam"/>
  <nhs:code>Q53</nhs:code>
  <nhs:name>Arden, Herefordshire and Worcestershire Area Team</nhs:name>
</rdf:Description>

<rdf:Description rdf:about="http://www.semanticweb.org/ma8g13/NHSONtology#region_Y55">
  <nhs:has rdf:resource="http://www.semanticweb.org/ma8g13/NHSONtology#areaTeam_Q53"/>
</rdf:Description>

<rdf:Description rdf:about="http://www.semanticweb.org/ma8g13/NHSONtology#areaTeam_Q54">
  <rdf:type rdf:resource="http://www.semanticweb.org/ma8g13/NHSONtology#AreaTeam"/>
  <nhs:code>Q54</nhs:code>
  <nhs:name>Birmingham and The Black Country Area Team</nhs:name>
</rdf:Description>

<rdf:Description rdf:about="http://www.semanticweb.org/ma8g13/NHSONtology#region_Y55">
  <nhs:has rdf:resource="http://www.semanticweb.org/ma8g13/NHSONtology#areaTeam_Q54"/>
</rdf:Description>
```

The rest of the NHS and BNF CSV files were converted in the same manner as discussed above in the examples. However, due to the size of the prescriptions dataset, it was not possible to convert this data in the same manner as the rest of the datasets. As a solution to this problem, the main dataset was twice split into smaller files. The first split divided the data according to the region that issued the prescriptions. The result was to have four relatively big CSV files, one for each region of England as shown in table 11.

The File Name	Size on Disk	No. of Rows	No. of CSV Sub-files	No. of converted RDF files
<i>PDPI for London</i>	421 MB	2,716,655	3	91
<i>PDPI for South of England</i>	785 MB	4,549,124	5	152
<i>PDPI for Midlands of England</i>	1.00 GB	5,656,377	6	189
<i>PDPI for North of England</i>	1.06 GB	5,921,519	6	198
TOTAL	≈3,22 GB	18,843,675	20	630

Table 11: Details of the splitting and converting steps for the prescriptions dataset

A second split was needed to be able to handle these files in Excel spreadsheets as they have a limitation in viewing of 1,048,576 rows at most. Therefore, the second split divided the four CSV files into 20 smaller ones.

The transforming approach for these 20 files was by writing a Visual Basic script that loops through the entire spreadsheet row by row and identifies the cells needing to be converted into triples. The script produces several text files that include the RDF statements. Each spreadsheet produces around 35 RDF files. Table 11 shows more details about the files that were produced; the script code can be found in Appendix E.

5.2.2.3 Load

As the size of the data is big, a sufficient triple store to handle all the data is needed. After consulting an expert in semantic web technologies about which triple store to use, GraphDB (Ontotext, 2017) was suggested. GraphDB is a free triple store that supports SPARQL queries and inference. In the UNIPROT study, they managed to upload 17 billion triples via GraphDB (World Wide Web Consortium, no date).

After downloading and installing the free edition of GraphDB on a local computer, a repository named *PrescriptionRepository* was created locally. Firstly, the NHS, BNF and Prescriptions ontologies (three OWL files) were uploaded into the *PrescriptionRepository*. Secondly, all produced RDF files from the previous step in total 639 RDF files were also uploaded. The total number of statements in the repository so far is 227,563,112 statements, see figure 41.



Figure 41: The PrescriptionRepository information in GraphDB

5.2.3 Querying Interface

The last component in the prescriptions demonstrator is the querying interface, where retrieving data and discovering features take place. GraphDB, the chosen triplestore, contains a SPARQL endpoint for querying the uploaded semantic data in the repository. This service is used to test the uploaded data and ontologies in the prescriptions demonstrator.

In this section, several testing queries are provided and implemented across the three ontologies to test different SW features. In each example: a) the question to be asked, b) the translated SPARQL query of this question, c) the resulted output and d) a diagram that shows the triples used in answering are illustrated.

5.2.3.1 Testing the NHS Ontology

This test case is about checking the validity of information retrieval from the NHS ontology. This ontology is by itself interesting as it was designed by integrating several spreadsheets. The question to be answered in this test case is: “Who are the doctors that work in London?”

```

PREFIX nhs: <http://www.semanticweb.org/ma8g13/NHSOntology#>
PREFIX org: <http://www.w3.org/ns/org#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT
  ?givenName ?familyName ?GMC
WHERE
{
  ?region a nhs:RegionalOffice.
  ?AT a nhs:AreaTeam.
  ?ccg a nhs:ClinicalCommissioningGroup.
  ?practice a nhs:Practice.
  ?branch a nhs:Branch.
  ?employee a nhs:Employee.

  ?region nhs:has ?AT.
  ?AT nhs:has ?ccg.
  ?ccg nhs:has ?practice.
  ?practice nhs:has ?branch.
  ?branch org:hasMember ?employee.

  ?region nhs:name "London".
  ?employee nhs:jobTitle "General Practitioner".

  ?employee foaf:givenName ?givenName.
  ?employee foaf:familyName ?familyName.
  ?employee nhs:GMCNumber ?GMC.
}

```

Figure 42: A SPARQL Query for Testing the NHS Ontology

To answer the test question, a SPARQL query is implemented as can be seen in figure 42. In this query, the first step is to identify the regional office that has the name ‘London’. Then, finding all the area teams that are connected with this specific regional office is the next step. Going down to the next level of the taxonomy of the NHS, the CCGs linked to the identified area teams are chosen. After that is the need to identify all the practices that are linked to the identified CCGs followed by, if possible, identifying any branches for these practices is next. Finally, the set of

employees who are working in one of the identified practices' branches and satisfying the condition of having a 'General Practitioner' title is retrieved.

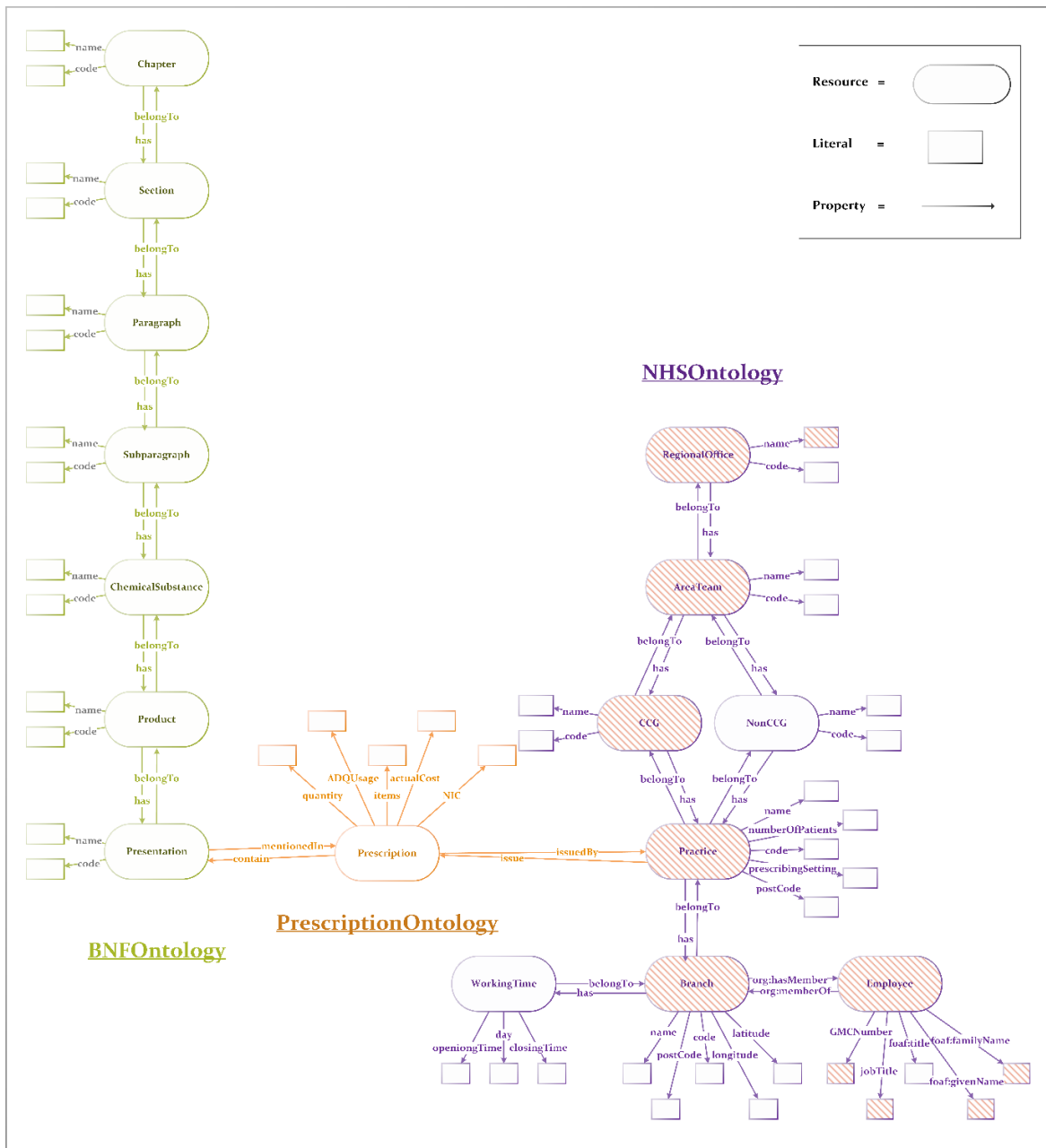


Figure 43: The elements used in testing the NHS ontology

Figure 43 shows the data elements (classes and properties) used to answer the test question. Six classes from the NHS ontology are involved in answering the query, which are: i) *RegionalOffice*, ii) *AreaTeam*, iii) *CCG*, iv) *Practice*, v) *Branch* and vi) *Employee*.

	givenName	familyName	GMC	↕
1	Adedayo	Adedeji	3509777	
2	Simisade	Adedeji	6148746	
3	Kamrun	Hossain	5176732	
4	Mohammed	Fateh	1431043	
5	Mohammed	Fateh	1431043	
6	Alex	Duodu	5199971	
7	Mina	Goyal	3253777	
8	Sneha	Amin	7084988	
9	Ndalai M.	Abaniwo	4058564	
10	Abdul	Qureshi	1487312	
11	Shansun N	Ahmad	1428452	
12	A	Adewole	6142239	
13	A	Sharma	6070168	
14	R	Hara	5191937	
15	S	Raza	6040246	
16	A	Adewole	6142239	
17	R	Hara	5191937	
18	S	Raza	6040246	
19	Chandima	Thalahitiya	7041169	
20	Humayan	Ahmad	3473485	

Figure 44: A snapshot of the results of the NHS ontology testing use case

The full results of the query are downloaded as a CSV file that has more than 5000 rows. Figure 44 shows only a relatively small part of the results.

5.2.3.2 Testing the BNF Ontology

This section is for testing the integrity of the BNF ontology. The BNF ontology includes a hierarchical representation of the supported BNF drugs. Therefore, the test question is chosen to demonstrate the taxonomy's representation of some BNF products. The question to be answered is: "What are the available vitamins' presentations?"

```

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

PREFIX bnf: <http://www.semanticweb.org/ma8g13/BNFOntology#>

SELECT
?presentation_name ?presentation_code
WHERE
{
    ?section a bnf:Section.
    ?section bnf:name "Vitamins".

    ?paragraph bnf:belongsTo ?section.
    ?subparagraph bnf:belongsTo ?paragraph.
    ?chemical_substance bnf:belongsTo ?subparagraph.
    ?product bnf:belongsTo ?chemical_substance.
    ?presentation bnf:belongsTo ?product.

    ?presentation bnf:name ?presentation_name.
    ?presentation bnf:code ?presentation_code.
}

```

Figure 45: The implemented query for the BNF ontology testing use case

Figure 45 shows the implemented query for answering the test question. The first step in answering the question is to find the vitamins section out of all the available BNF's sections. The next step is concerned with finding all the sub-categories of the vitamins section, starting from the section-level up to the presentation-level. The involved sub-levels in order are: *a) paragraph*, *b) subparagraph*, *c) chemical substance* and *d) product*. The final step in this query is to identify all

the presentations' names and codes for the products that: i) *belongTo* the chemical substances that ii) *belongTo* the sub-paragraphs that iii) *belongTo* the paragraphs that iv) *belongTo* the vitamins section.

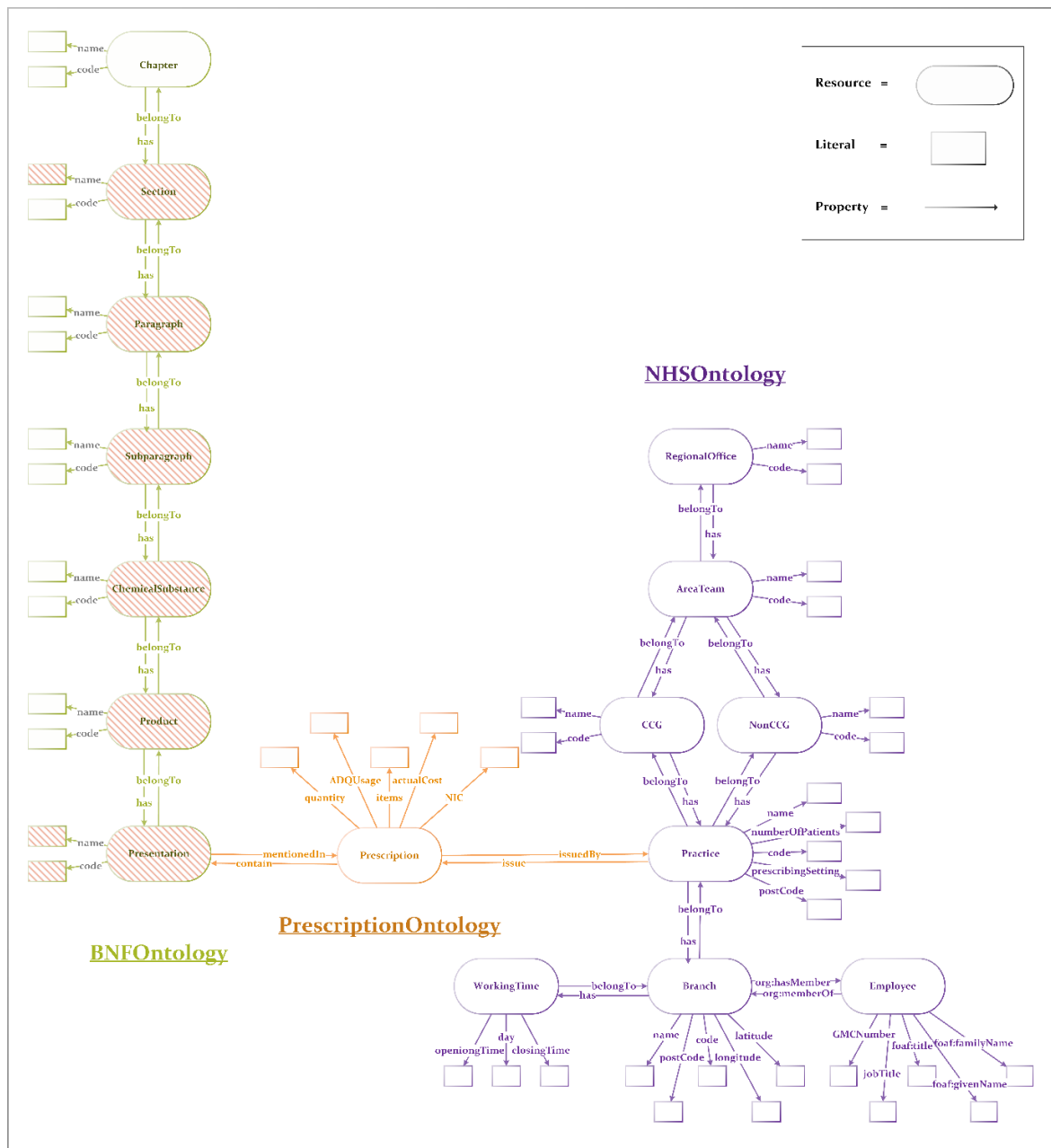


Figure 46: The data used in the BNF ontology testing query

Figure 46 shows the used data elements in the BNF ontology. The used classes in this test case are: i) *Section*, ii) *Paragraph*, iii) *Subparagraph*, iv) *ChemicalSubstance*, v) *Product* and vi) *Presentation*.

	presentation_name ⇅	presentation_code ⇅
1	Vit A_Inj 300,000u/ml 1ml Amp	0906011D0AAADAD
2	Vit A_Soln 25,000u	0906011D0AAAEAE
3	Vit A_Strong Cap 50,000u	0906011D0AAAFAF
4	Vit A_Tab 50,000u	0906011D0AAAGAG
5	Vit A & D_Cap	0906011D0AAAHAH
6	Vit A_Soln 100,000u/ml	0906011D0AAAIAI
7	Vit A_Inj 50,000u/ml 2ml Amp (Old)	0906011D0AAAIAJ
8	Vit A_Tab 10,000u	0906011D0AAAKAK
9	Vit A & D_Cap 2,500u/250u	0906011D0AAALAL
10	Vit A_Cap 4,500u	0906011D0AAAMAM
11	Vit A_Cap 5,000u	0906011D0AAANAN
12	Vit A_Cap 4,000u	0906011D0AAPAP
13	Vit A_Tab 5,000u	0906011D0AAAQAA
14	Vit A_Tab 50,000u (Import)	0906011D0AAARAR
15	Vit A_Oral Dps 5,000u/0.1ml	0906011D0AAATAT
16	Vit A_Liq Spec 25,000u/5ml	0906011D0AAAUAA
17	Vit A_Soln 150,000u/ml	0906011D0AAAVAV
18	Vit A_Inj 50,000u/ml 2ml VI	0906011D0AAAWAW
19	Retinol_Oral Dps 150,000u/ml	0906011D0AAAXAX
20	Vit A_Inj 50,000u/ml 2ml Amp	0906011D0AAAYAY
21	Ro-A-Vit_Tab 50,000u	0906011D0BBAAAG
22	Ro-A-Vit_Inj 300,000u/ml 1ml Amp	0906011D0BBABAD

Figure 47: Snapshot of the results of the BNF ontology testing use case

Figure 47 shows part of the results retrieved in order to answer the implemented query. The original downloaded CSV file has more than 3000 rows.

5.2.3.3 Testing the Prescriptions Ontology

The third and last test problem is for the prescriptions ontology. The prescriptions ontology has a simple design with only a single class namely the *Prescription* class. To test the prescriptions ontology, the following question was posed: “What are the NIC and number of items of all the prescriptions?”

```

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX pres: <http://www.semanticweb.org/ma8g13/prescriptionOntology#>

SELECT
  ?prescription ?NIC ?items
WHERE
  {
    ?prescription a pres:Prescription.
    ?prescription pres:NIC ?NIC.
    ?prescription pres:items ?items.
  }

```

Figure 48: The implemented query for the prescriptions ontology test case

Figure 48 shows the used SPARQL query used to retrieve all the prescription instances. This query is simple as it only identifies all the uploaded prescriptions' NIC and number of items.

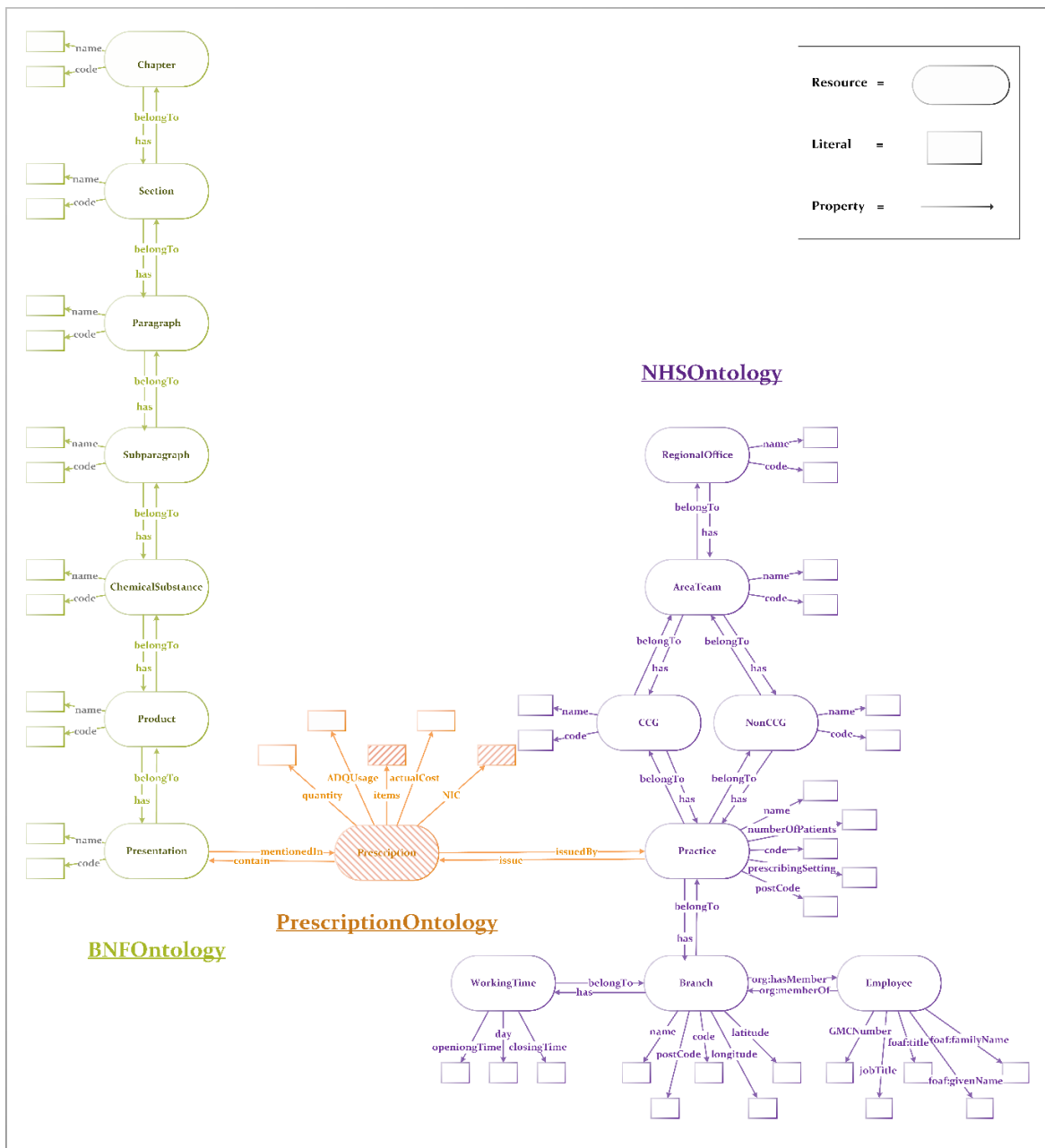


Figure 49: The used data in the prescriptions ontology test query

Figure 49 shows the involved class and data properties in the prescriptions ontology, which are the *Prescription* class and *NIC* and *items* data properties.

	prescription	NIC	items
1	pres:prescription_london_1	17.88	1
2	pres:prescription_london_10	10.7	5
3	pres:prescription_london_100	163.56	3
4	pres:prescription_london_1000	3.7	1
5	pres:prescription_london_10000	5.2	1
6	pres:prescription_london_10001	109.48	17
7	pres:prescription_london_10002	28.1	2
8	pres:prescription_london_10003	280.08	36
9	pres:prescription_london_10004	43.56	6
10	pres:prescription_london_10005	82.86	3
11	pres:prescription_london_10006	3.91	1
12	pres:prescription_london_10007	12.51	1
13	pres:prescription_london_10008	46.7	2
14	pres:prescription_london_10009	46.1	2
15	pres:prescription_london_1001	15.6	2
16	pres:prescription_london_10010	60.68	1
17	pres:prescription_london_10011	4.12	2
18	pres:prescription_london_10012	51.56	1
19	pres:prescription_london_10013	36.4	40
20	pres:prescription_london_10014	257.8	10
21	pres:prescription_london_10015	29.62	1

Figure 50: A snapshot of the results of the prescriptions ontology test case

Although this query is a simple one, the number of retrieved data instances is huge. Such a total demonstrates the capability of the SPARQL technology to handle big data. Figure 50 shows a part of the retrieved results as the number of rows in the CSV files exceeds 18 million.

5.2.3.4 Testing the Integration of the Three Ontologies

This test case is about investigating the integration process for several ontologies which can show the capability of the SW in handling such integration. The case design involves: a) the BNF ontology, b) the NHS ontology and c) the prescriptions ontology. The integration test question is: “What are the prescriptions that were issued by the ‘P82022’ practice and contain the drug coded ‘0601022B0AAABAB’?”

```

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX nhs: <http://www.semanticweb.org/ma8g13/NHSONtology#>
PREFIX pres:<http://www.semanticweb.org/ma8g13/prescriptionOntology#>
PREFIX bnf: <http://www.semanticweb.org/ma8g13/BNFONtology#>

SELECT
  ?prescription ?NIC ?items
WHERE
  {
    ?practice a nhs:Practice.
    ?prescription a pres:Prescription.
    ?bnf_presentation a bnf:Presentation.

    ?prescription pres:contain ?bnf_presentation.
    ?bnf_presentation bnf:code "0601022B0AAABAB".

    ?prescription pres:issuedBy ?practice.
    ?practice nhs:code "P82022".

    ?prescription pres:NIC ?NIC.
    ?prescription pres:items ?items.
  }

```

Figure 51: The implemented query for testing the integration of all three ontologies

Figure 51 shows the query used in answering the test question. This query involves integration of the prescriptions ontology with the BNF ontology and the NHS ontology. Firstly, in the BNF

ontology the BNF presentation which has the '0601022B0AAABAB' code is identified. Secondly, in the NHS ontology, the practice which has the 'P82022' code is identified. Finally, it is time for the integration step where all the prescriptions that are specifically linked to the identified presentation and practice are chosen to view their NIC and number of each items' properties.

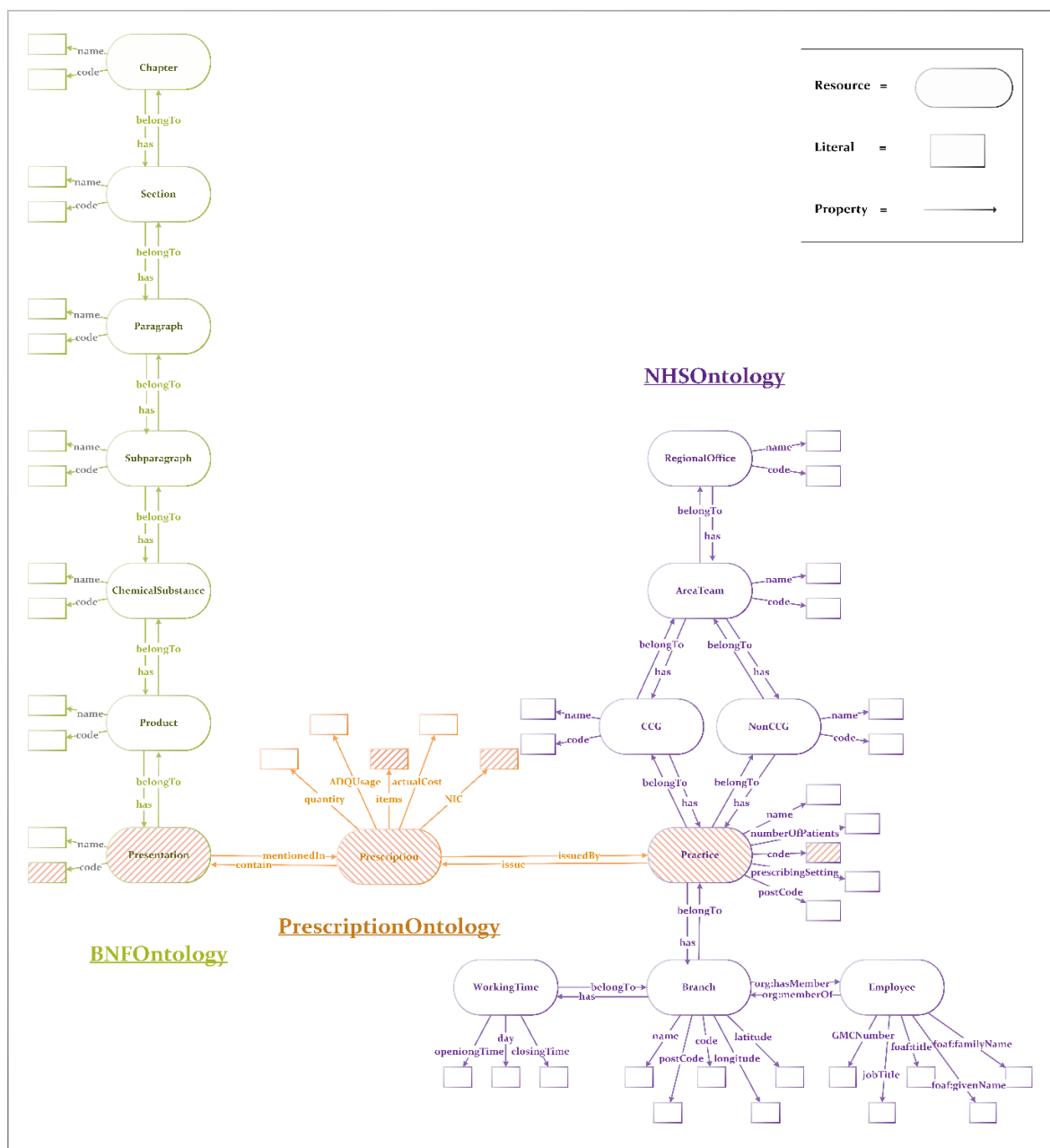


Figure 52: The data used to test the integration of all three ontologies

The classes involved in the three ontologies are: a) *Presentation* from the BNF ontology, b) *Prescription* from the prescriptions ontology and c) *Practice* from the NHS ontology, see figure 52.

	prescription ⚡	NIC ⚡	items ⚡
1	pres:prescription_north_2727002	58.32	18
2	pres:prescription_north_2727009	9.72	4
3	pres:prescription_north_2727016	3	15
4	pres:prescription_north_2727162	21.06	13
5	pres:prescription_north_2727165	1.64	4
6	pres:prescription_north_2727184	84.24	13
7	pres:prescription_north_2727197	21.06	26
8	pres:prescription_north_2727198	2.44	4
9	pres:prescription_north_2727199	43.74	9

Figure 53: A snapshot of the results of the integration of all three ontologies

The results of this specific integration are shown in figure 53. Nine specific prescriptions are identified out of millions of records.

5.2.3.5 Testing Inference

This test case demonstrates an example of the SW's inference feature. Discovering new relationships out of known ones is what is tested here. The test case used in this section illustrates inferring an object relationship from a known one based on OWL's inverse object property. The known facts defined in the BNF ontology are:

- 1) a *Product belongsTo* a *ChemicalSubstance*
- 2) the *has* object property is the inverse of the *belongsTo* property

The aim for this use case is to test if the reasoner is able to infer that a *ChemicalSubstance has a Product* based on the two known facts mentioned above. The testing question in this case is: "What are the BNF products that have the Metformin Hydrochloride chemical substance in their ingredients?"

```

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

PREFIX bnf: <http://www.semanticweb.org/ma8g13/BNFOntology#>

PREFIX pres: <http://www.semanticweb.org/ma8g13/prescriptionOntology#>
SELECT
  ?product ?name ?code
WHERE
{
  ?chemical_substance a bnf:ChemicalSubstance.
  ?chemical_substance bnf:name "Metformin Hydrochloride".

  ?chemical_substance bnf:has ?product.

  ?product bnf:name ?name.
  ?product bnf:code ?code.
}

```

Figure 54: The query that was implemented to test for inference

This query tests if the reasoner is able to add the inferred triple: *ChemicalSubstance has Product* to the already-known facts. Figure 54 shows the implemented query identifying the chemical substance named 'Metformin Hydrochloride' at the beginning. Then, all the associated products with the identified chemical substance via the inferred property are also identified.

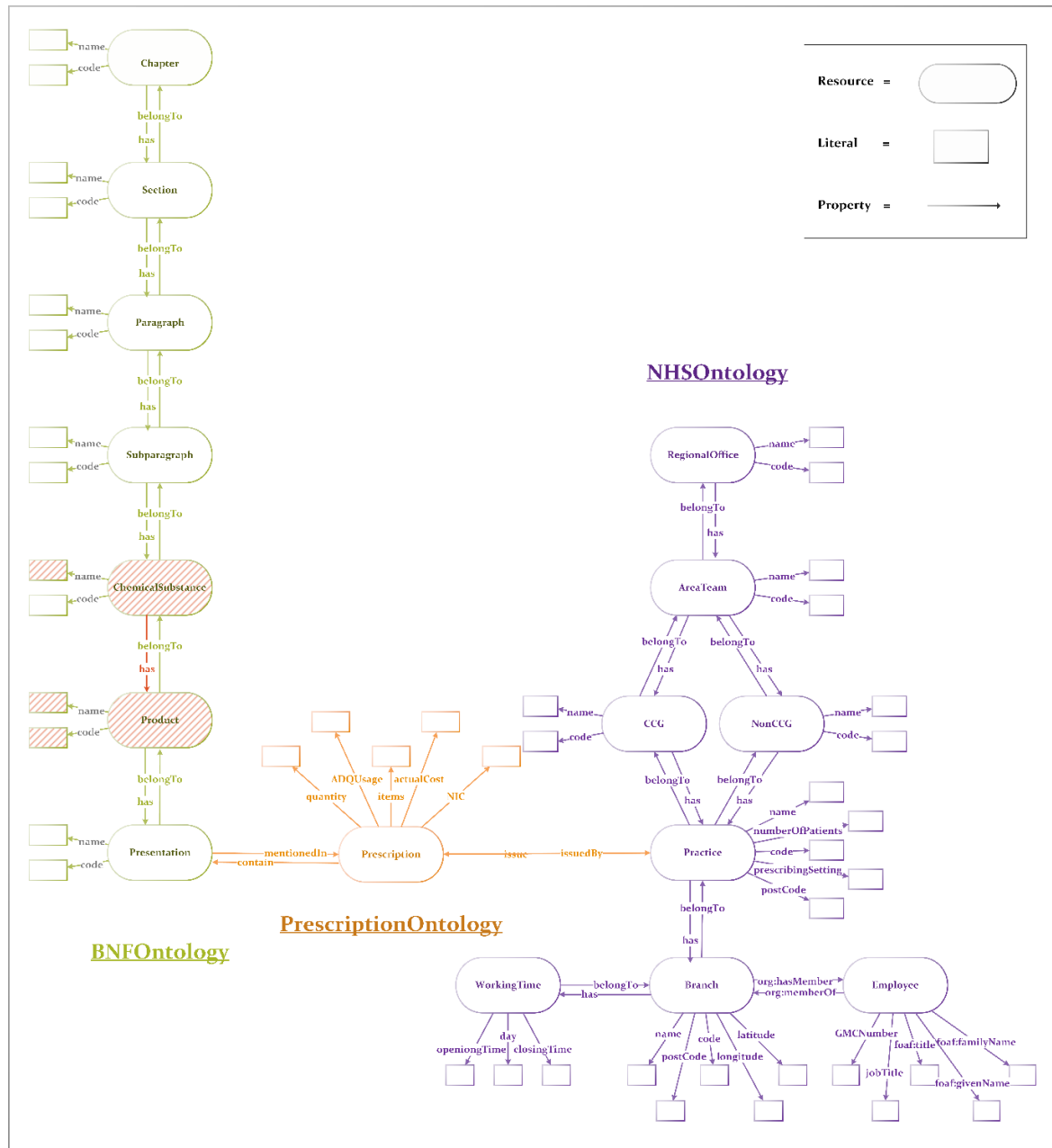


Figure 55: The data used in testing the inference query

This test case involves the *ChemicalSubstance* and *Product* classes with the *belongTo* and its inverse *has* object properties, see figure 55.

	product	name	code
1	bnf:BNF_product_0601022B0AA	Metformin HCl	0601022B0AA
2	bnf:BNF_product_0601022B0BB	Glucophage	0601022B0BB
3	bnf:BNF_product_0601022B0BC	Orabet	0601022B0BC
4	bnf:BNF_product_0601022B0BD	Glyformin	0601022B0BD
5	bnf:BNF_product_0601022B0BE	Ledermetin	0601022B0BE
6	bnf:BNF_product_0601022B0BF	Glucamet	0601022B0BF
7	bnf:BNF_product_0601022B0BG	Milform	0601022B0BG
8	bnf:BNF_product_0601022B0BH	Metsol	0601022B0BH
9	bnf:BNF_product_0601022B0BI	Bolamyn	0601022B0BI
10	bnf:BNF_product_0601022B0BJ	Metabet	0601022B0BJ
11	bnf:BNF_product_0601022B0BK	Glucient SR	0601022B0BK
12	bnf:BNF_product_0601022B0BL	Diagemet XL	0601022B0BL
13	bnf:BNF_product_0601022B0BM	Sukkarto SR	0601022B0BM

Figure 56: A snapshot of the results of the inference test case

A list of 13 Metformin Hydrochloride's products was produced, see Figure 56.

5.2.3.6 Testing Aggregation

The final use case is to test the aggregation feature in SPARQL. Aggregation is to apply expressions like COUNT, SUM, MIN, MAX over some selected groups of solutions using the GROUP BY keyword. The chosen testing question is: “Which practice has issued the biggest number of prescriptions?” The key word in this question is the word ‘biggest’. Biggest means there will be some sort of counting (summing) involved, as well as some sort of a comparison techniques.

```

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX nhs: <http://www.semanticweb.org/ma8g13/NHSONtology#>
PREFIX pres: <http://www.semanticweb.org/ma8g13/prescriptionOntology#>
PREFIX bnf: <http://www.semanticweb.org/ma8g13/BNFONtology#>
PREFIX org: <http://www.w3.org/ns/org#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT
  ?practice (COUNT(?prescription) as ?MAXnumberOfPrescriptions)
WHERE
{
  ?prescription pres:issuedBy ?practice.
}
GROUP BY ?practice
ORDER BY DESC(?MAXnumberOfPrescriptions)
LIMIT 1

```

Figure 57: The implemented query for testing aggregation

The query that answers the testing question is shown in Figure 57. The first step in this query is to find all the prescriptions that have been issued by practices. The answer to this part of the query is a single solution with millions of records in it. However, what is needed is to group these records into sub-solutions based on some criteria. This is performed in the GROUP BY part, where the identified prescriptions are grouped based on the issuing practice. After grouping the prescriptions based on the practices, the number of prescriptions in each group is counted using the COUNT keyword. The final step is to order the groups into descending order and shows the first group only as it will be the one with the maximum number of prescriptions.

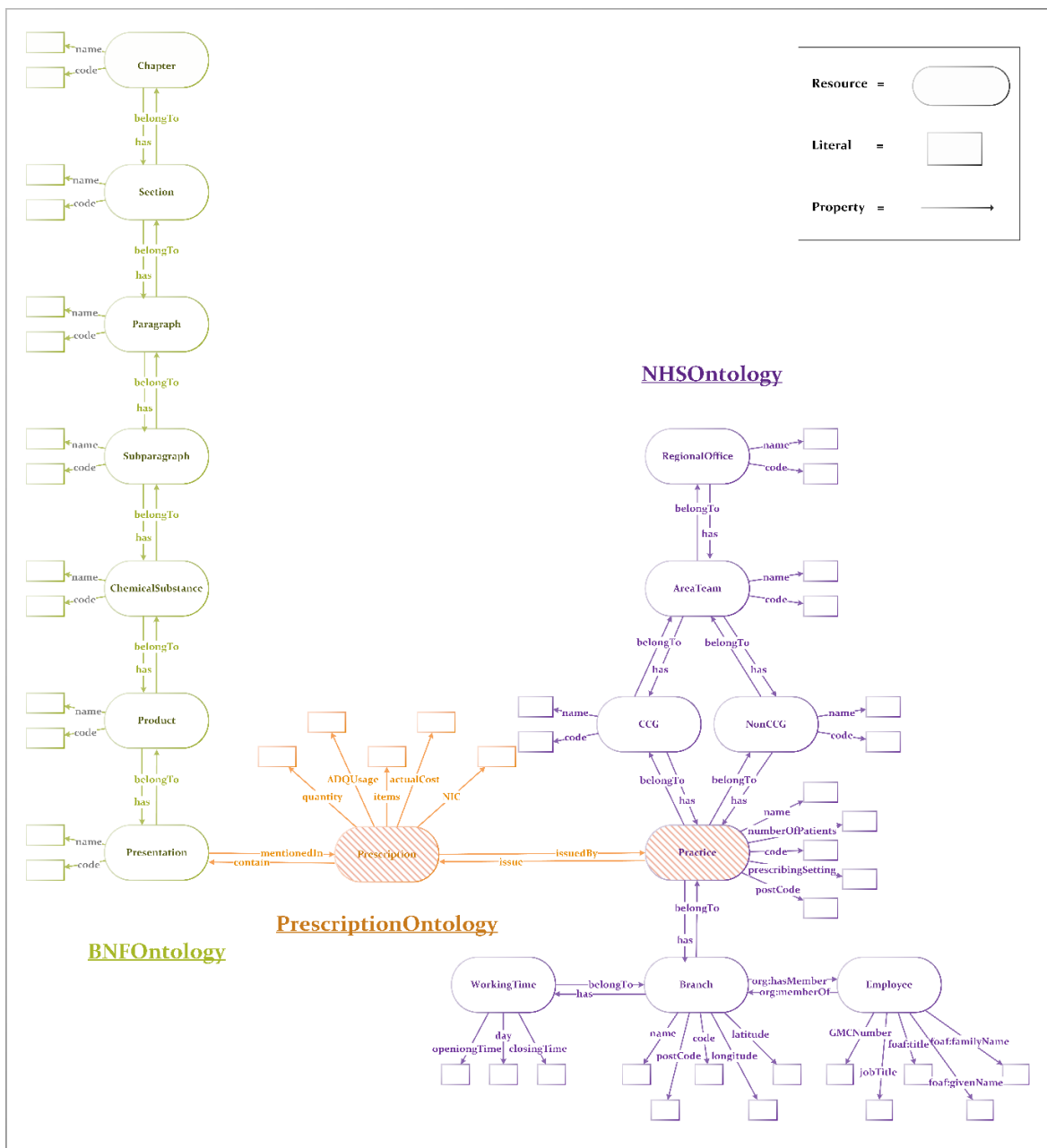


Figure 58: The data used in testing aggregation query

Figure 58 shows the involved data concepts in this query: i) *Prescription* and ii) *Practice*.

	practice	MAXnumberOfPrescriptions
1	nhs:practice_M85063	"10567"^^xsd:integer

Figure 59: A snapshot of the results of testing aggregation use case

The result of the asked question is a single record that shows the practice that issued the maximum number of prescriptions, see Figure 59.

5.3 Analysis

The aim for building this demonstrator was to test and illustrate different uses for the SW in a health-related topic. Specifically, this demonstrator focused more on testing representing, integrating and exploring data because these features were the ones most used in the literature. The SW features taxonomy presented in the systematic review chapter lists 12 sub-features categorised under five main features. In this analysis, the features used in the prescriptions demonstrator are discussed with the main affordances and challenges faced during the process. Figure 60 highlights the features used in the prescriptions demonstrator.

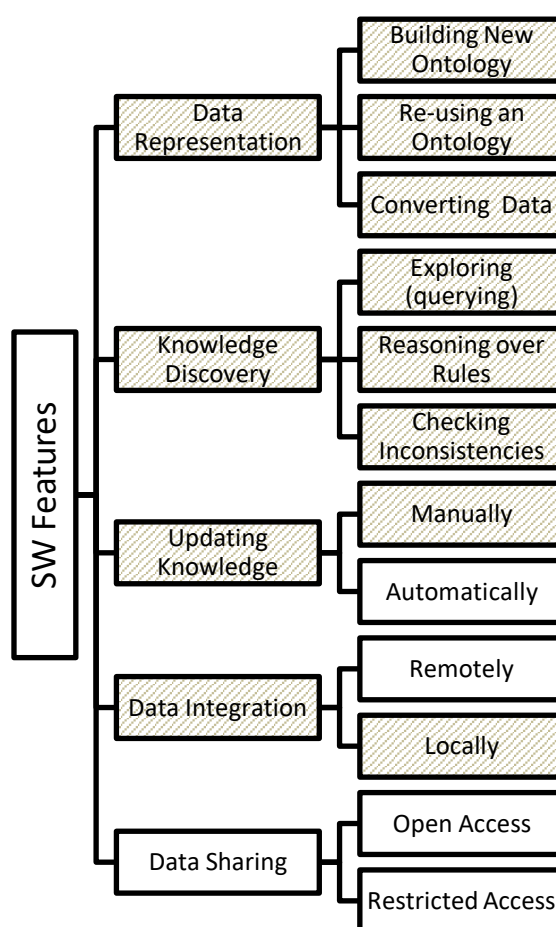


Figure 60: The SW features taxonomy for the prescriptions demonstrator

5.3.1 Data Representation

Regarding data representation, the prescriptions demonstrator represents the domain of dispensed prescriptions in England. Data representation was achieved by building three

ontologies representing three targeted, interlinked domains: i) dispensed prescriptions information, ii) NHS organisational structure and iii) BNF medication data. The built ontologies for representing these domains are the NHS, BNF and Prescriptions ontologies.

From a technical point of view, designing these demonstration ontologies was an affordable task for a person with some technical background, especially by using supporting tools like Protégé (Musen, 2015) to simplify the job. The availability of user-friendly tools and interfaces for browsing and displaying data would encourage further adopting for the SW approach in health research. Moreover, the process of designing ontologies can benefit from the availability of supporting tools for handling the syntax of the of the design. However, in real-case scenarios designing ontologies is a complicated task that involves decisions on the semantics of the design as well. Deciding on the most suitable design for an ontology could be a challenging job as the designing process demands some sort of human creativity and not only following uniform steps (Hu *et al.*, 2012). Computer scientists, as well as domains experts (in this case: health experts) should cooperate in planning for the design to achieve satisfactory results.

Besides building ontologies, the acts of re-using ontologies and converting instances were tested as well. There were some borrowed vocabularies from the Organization ontology (ORG) (World Wide Web Consortium, 2014) and the FOAF ontology (Brickley and Miller, 2010) used when building the NHS ontology. The re-using feature in this experiment was limited in usage but could be expanded by mapping more pharmaceutical terms with the defined terms in the BNF ontology. There are many drug-related ontologies available in the LOD that could be re-used in this scenario.

Regarding the converting feature, the data transformation was on the 'instances' level and two different approaches were employed. The first approach was automatic by using OpenRefine (Google, 2010) combined with designed mapping rules. The second approach used manually scripted code to traverse a spreadsheet row by row and transform data within it into structured data in RDF form. The existence of automatic efficient converting tools that can handle big amounts of data helps in facilitating the SW experience.

5.3.2 Data Integration

The integration feature in this topic played an important role. The prescriptions topic was chosen based on its suitability to represent heterogenous data integration. The prescriptions data works as a link between pharmaceutical medication information and health management information about the prescriber. The ability to link different types of data by using standards is considered one of the most significant advantages of the SW. It is the main concept in the SW vision,

implemented by using the Linked Data paradigm. An important condition in applying this linkage, that the success of the process relies on, is the ability to handle big data. The amount of data available nowadays is huge and so to achieve the basic idea of the SW vision in linking data concepts within documents, there is a crucial need for technologies to support this aim. In the prescriptions demonstrator for example, the triple store was able to handle the big amount of the data and SPARQL queries were successfully retrieved, but with delays in some cases. However, in one of the tools responsible for generating graphs for the uploaded data in the repository in GraphDB (Ontotext, 2017), the tool was not able to handle the big amounts of data presented to it. It is reasonable to conclude from the evidence gathered that the SW vision is still hindered by some technical challenges that hopefully can be resolved in the not-too-distant future.

5.3.3 Knowledge Discovery

The third component in the prescriptions demonstrator was a querying interface, a tool to search and explore the uploaded data in the triplestore. In this experiment, the querying feature in the SW was focused on more than reasoning. Searching for information in a database is one of the primary tasks in any data system. SPARQL managed to retrieve different answers for a range of test questions. For example, SPARQL was used successfully in aggregating information and performing some arithmetic calculations on the targeted data using the endpoint provided in GraphDB (Ontotext, 2017).

Reasoning and checking inconsistencies were tested and used narrowly. The Pellet reasoner was used in checking for inconsistencies in the ontologies design. It was mentioned that the designs of both BNF and NHS ontologies were updated based on the revealed inconsistencies. The SW has a range of flexibility in handling adding and deleting facts from the knowledge base. Defining schema in the SW happens by defining a group of triples just like the instances themselves, which makes updating the schema just as flexible as updating the instances. Updating data was performed easily via the UPDATE and DELETE properties of SPARQL.

Regarding reasoning over rules, in this experiment there was no definition for semantic rules in SWRL. However, inference was performed by testing OWL object properties' conditions, such as '*inverse of*' as explained earlier in the inference testing case. The system succeeded in adding new knowledge to the already known knowledge at run time.

The data sharing feature was not tested in this experiment because the prescription demonstrator was built as a simple proof-of-concept model that was not intended to be shared publicly. The design of the ontologies is very primitive and in its early stage. Perhaps, in the future, this model could be improved and the ontologies for BNF, NHS or prescriptions can be

shared in the LOD. This type of project needs a team of health experts from the three topics working together with ontology engineers. Having said that, during the time of the study the designed ontologies and the converted instances, were shared with the social sciences, social data and the semantic web (S3W) project that aim to identify the challenges and opportunities of semantic linked data for social science research. A SW-based demonstrator will be developed to address health inequalities questions by converting sections of the English Longitudinal Survey of Ageing (ELSA) and the Great British Class Survey (GBCS) into semantic linked data. Several ontologies will be linked to the demonstrator such as the developed BNF ontology in this thesis.

5.3.4 Limitations

The proof-of concept system that was implemented has some limitations. The main shortage in the prescriptions demonstrator is the simplicity of the ontological design. The design followed the Extract, Transform and Load (ETL) process, which extracted the main concepts from the chosen datasheets and implemented them as classes and properties. Only one version of each of the chosen datasets was chosen for modelling the system. Regarding the chosen health use cases, they were sample cases not necessarily representing real case scenarios. Indeed, to properly evaluate the efficiency of a technology there are many factors to be considered; for example: a) performance, b) query response time and c) the precision of the retrieved results. These factors were not focused on in this current work. The SW was evaluated on its ability to perform a certain task. Some of the SW features that were not tested in the prescriptions demonstrator were: i) data sharing and ii) reasoning over defined rules.

5.4 Summary

The aim of this chapter is to illustrate the SW features used in a proof-of-concept health research demonstrator in order to analyse any affordances or challenges that may face the user of the SW in addressing health questions. The third research question is continued to be answered from a practical point of view. The chosen health topic for the demonstrator is based on the results of the identified topics in the systematic review, as well as choosing a topic that is qualified to represent the different SW features, such as data integration. The issue of dispensed NHS prescriptions in England was chosen as it links pharmaceutical and health management domains.

The prescriptions demonstrator consists of three main parts: i) Data Access, ii) Mapper and iii) Querying Interface. The data used in developing the demonstrator was basically seven open datasets representing the BNF drugs-related data, NHS dispensed prescriptions and NHS organisation information. The procedure for mapping the seven data sets into semantic data was Extract, Transform and Load (ETL). Three interlinked OWL ontologies were designed: i) NHS ontology, ii) BNF ontology and iii) prescriptions ontology. Instances were converted to RDF manually and automatically and then uploaded into the GraphDB triplestore. Several testing SPARQL queries were successfully run across the ontologies and data.

The learnt affordances from this experiment were: i) the ability and flexibility in representing different health-related topics, ii) the ease of integrating data locally and iii) the possibility to explore linked data via querying. Certain technical issues were noticed in the process of implementing the SW demonstrator. Not all available tools were able to handle big data efficiently such as in querying, graph displaying or converting to RDF.

Chapter 6 Health Research: Two Use Cases

In the previous chapter, the process of developing the prescriptions demonstrator was discussed. The demonstrator represents the dispensed prescriptions in England, where various pharmaceutical or health management questions could be addressed via the system. In this chapter, two real case scenarios are addressed using the prescriptions demonstrator. The aim of this chapter is to demonstrate how health questions can be translated into SPARQL queries and answered using SW standards. In addition, analysing any accompanying affordances or challenges that emerge during the process is aimed as well. Therefore, the third research question continues to be answered in this chapter.

The purpose of choosing the cases that are used is to find interesting health-related questions from a health researcher's point of view and then submit them to the prescriptions demonstrator. The first case comes from the traditional health literature and addresses the issue of health inequalities in prescribing anti-diabetic medicines across England. The second case was suggested by health researchers in a focus group. It is about investigating the relationship between prescribing antidepressants and living in a coastal city/town in England. Both of these cases are translated into SPARQL queries and submitted to the prescriptions demonstrator. The following sections discuss in details the followed approach and results for both cases.

6.1 Case One: Prescribing Inequalities for Diabetic Medications

Studying and analysing variations in prescribing patterns across geographical locations is an interesting health topic featured in the literature. One of the papers that discussed this topic in the health literature is *'Mapping English GP prescribing data: a tool for monitoring health-service inequalities'* (Rowlingson *et al.*, 2013). In this paper, prescription data, together with other datasets, were integrated to answer health inequality questions and place the answers on interpretable maps.

6.1.1 Case Description

The chosen question in this case is to address any health inequalities in prescribing diabetes medication across England. To provide an answer the cost per person rate for metformin hydrochloride's prescriptions is calculated for each NHS practice in England as it was introduced in (Rowlingson *et al.*, 2013). The metformin hydrochloride is used for treating diabetes. There are different products to prescribe derived from the metformin hydrochloride formula. The cost per

person rate was calculated by summing the net ingredient cost (NIC) for all metformin hydrochloride prescriptions per practice. The next step is to divide the total costs by the total number of people registered in the same practice.

To be able to calculate the cost per person rate, different pieces of information in several datasets need to be brought together. Rowlingson *et al.* (2013) used several datasets from open sources, such as the NHS dispensed prescriptions data. In this current case, the prescriptions demonstrator is used to address the same problem as in the paper.

The next step is to translate the question: “What is the cost-per-person rate for metformin hydrochloride prescriptions in England?” or in another words “Is there a dissonance in prescribing diabetic medication in England?”

6.1.2 The Query

To answer the cost per rate question, data concepts from the three interlinked ontologies in the prescriptions demonstrator are employed. To find the the net ingredient cost for metformin hydrochloride prescriptions, integrated data from the prescriptions and BNF ontologies are retrieved, while the number of registered people in a practice that prescribed metformin hydrochloride is calculated from the prescriptions and NHS ontologies.

The cost per person rate for prescribing metformin hydrochloride in England is calaculated by writing a SPARQL query consisting of five steps:

- 1) Identifying the metformin hydrochloride’s products and presentations.
- 2) Identifying the linked prescriptions with the metformin hydrochloride’s presentations.
- 3) Calculating the total NIC .
- 4) Finding the total number of registered patients.
- 5) Calculating the cost-per-person rate.

Figure 61 shows the translated SPARQL query for the cost-per-person rate question, while figure 62 shows the five steps and the classes and ontologies used in the prescriptions demonstrator. Next, each step is discussed in detail with a focus on the ontological dimension.

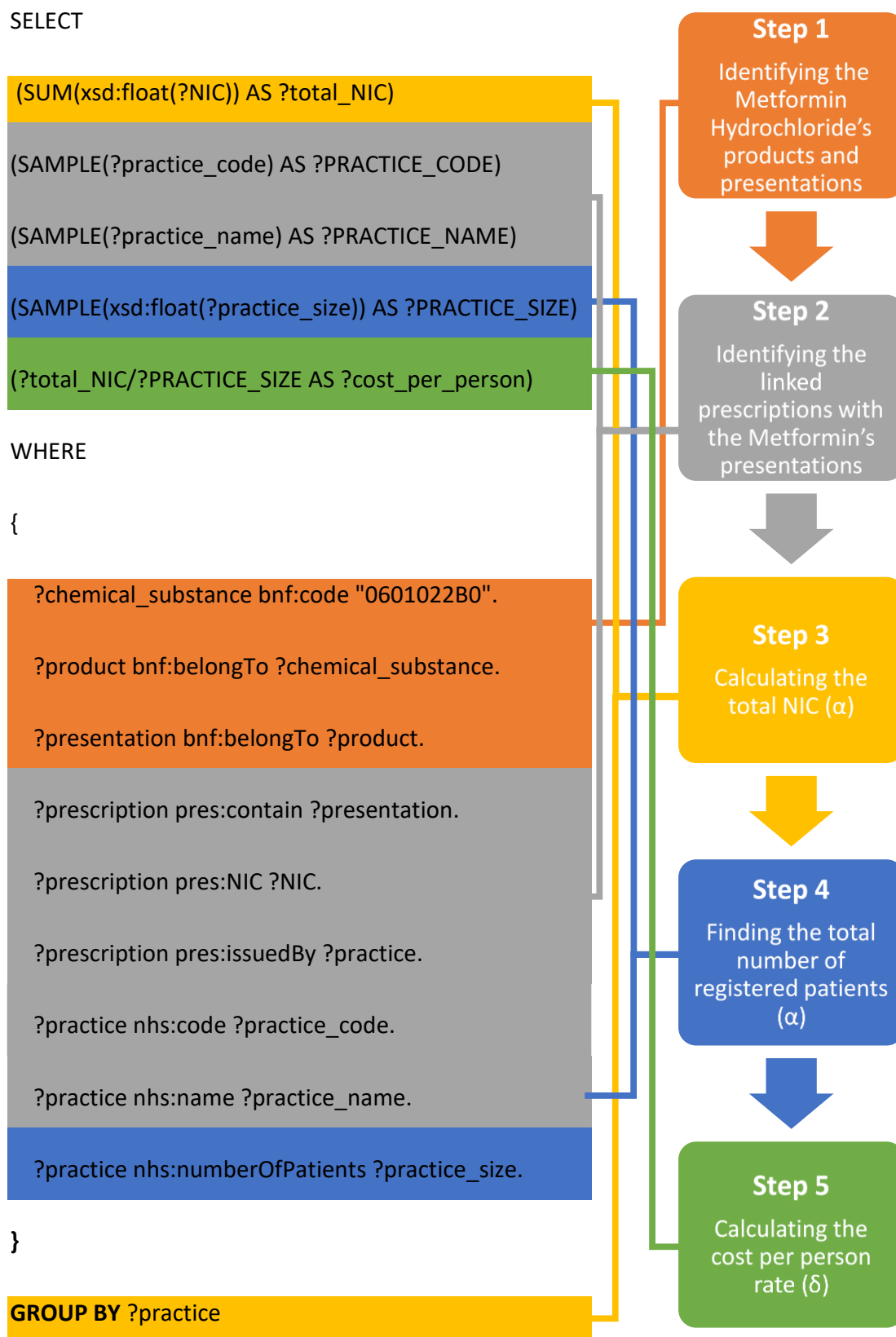


Figure 61: The SPARQL query for finding the cost-per-person rate

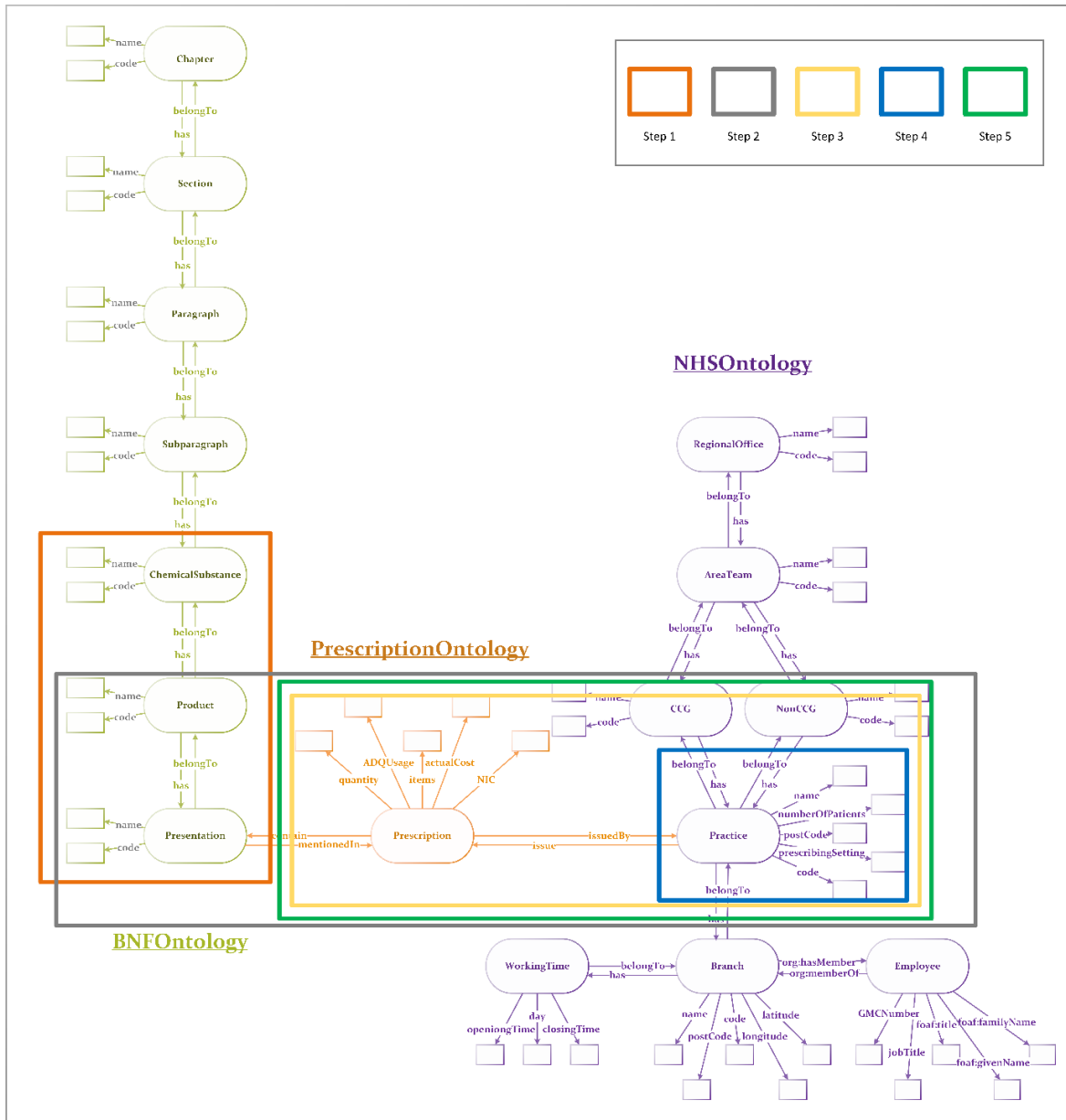


Figure 62: The data concepts used in the five steps of the cost-per-person query

Step 1: Identifying metformin hydrochloride products and presentations

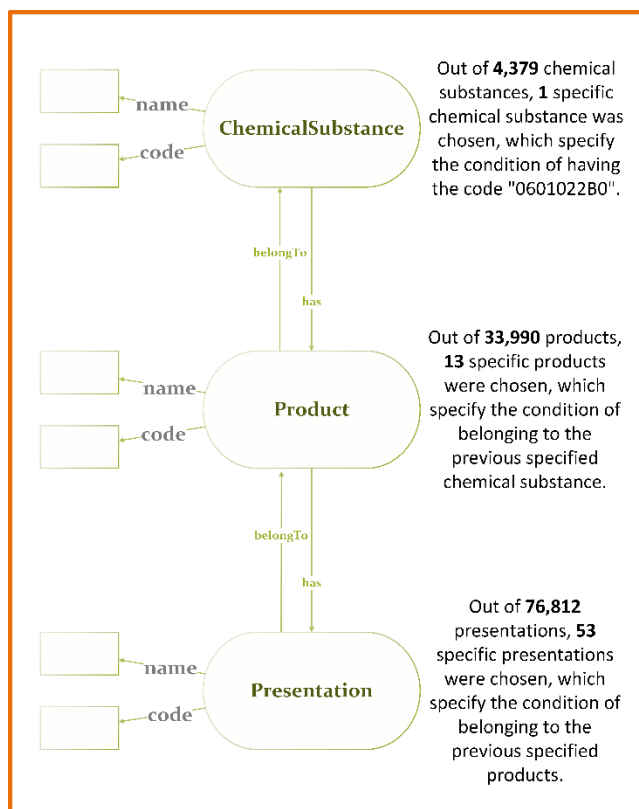


Figure 63: Step 1 of the cost-per-person query

Figure 63 shows the first step in the cost-per-person query representing a series of relationships between *ChemicalSubstance*, *Product* and *Presentation* classes in the BNF ontology.

The problem addressed in this case is to find the cost-per-person rate for metformin hydrochloride's prescriptions nationally. To solve this problem, first, the metformin chemical substance BNF code needs to be identified. It is '0601022B0' according to the BNF 72. In the BNF 72 dataset, there are 4,379 different listed chemical substances. The first part of the query is saying that from these 4,397 instances of the class *ChemicalSubstance*, one specific instance needs to be identified that uniquely assigned the BNF code '0601022B0'.

After identifying this specific instance of the *ChemicalSubstance* class, all the attached products that belong to this instance need to be found. In the BNF ontology there are 33,990 declared instances of the class *Product*. Out of these instances, there are only 13 different instances of the class *Product* attached to the specific identified chemical substance.

Similarly, the next step is to find all the presentations that belong to the 13 identified products. In the original BNF dataset, there are 76,812 different presentations of medical products. Only 53

specific presentations are identified that belong to the 13 identified products that belong to a specific chemical substance that has the code '0601022B0'. The role of those presentations is explained in the next step by showing the relationship between the BNF presentations and the prescriptions.

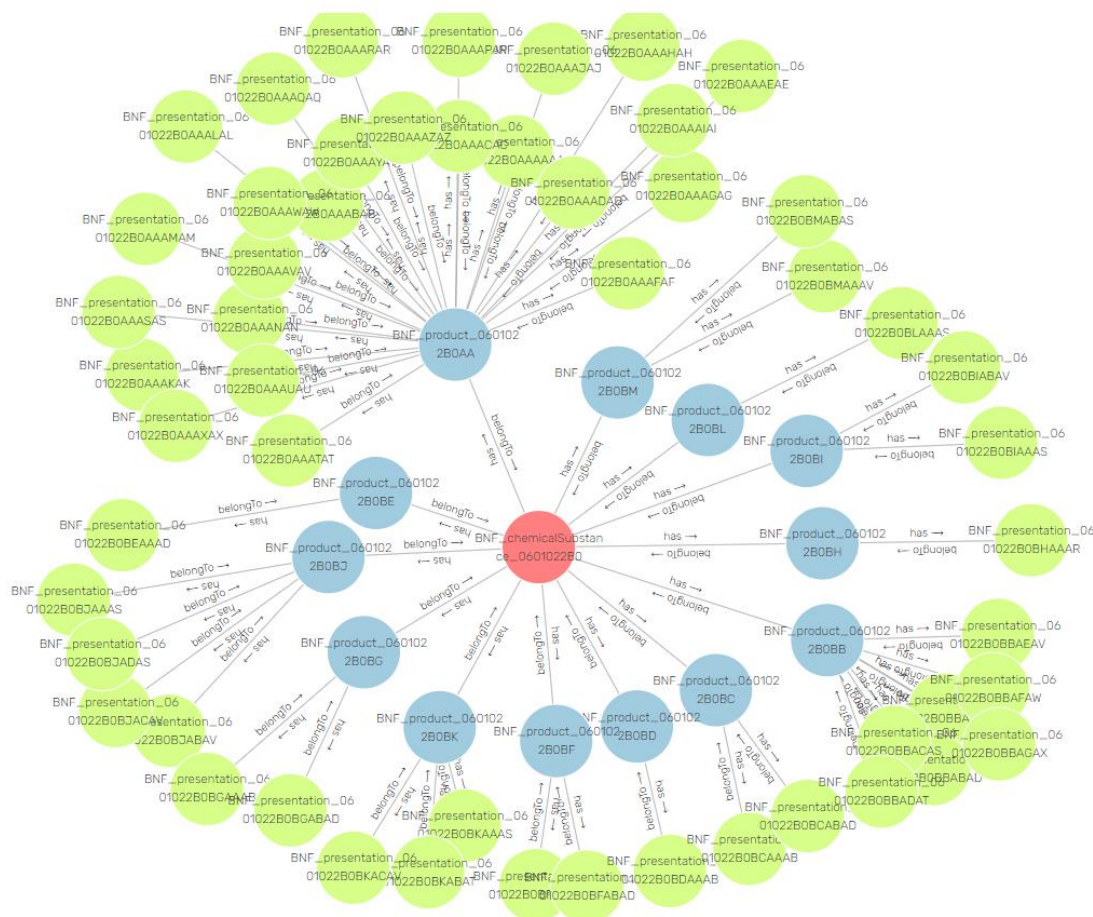


Figure 64: the links between the chemical substance coded '0601022b0' and its related products and presentations

For more clarification GraphDB (the triplestore used for the prescriptions demonstrator) has an automatic tool for producing graphs from queries. Figure 64 shows an overview of the relationships' series explained earlier between the chemical substances, products and presentations. The red coloured circle in the centre represents the only *ChemicalSubstance* instance that has the code '0601022B0', while the blue circles show the 13 identified products that are attached to this chemical substance by the *belongsTo* property. Finally, the green circles represent the 53 drugs' presentations that are used in the next step, in association with the *Prescription* class in the prescriptions ontology.

Step 2: Identifying the linked prescriptions with the metformin hydrochloride's presentations

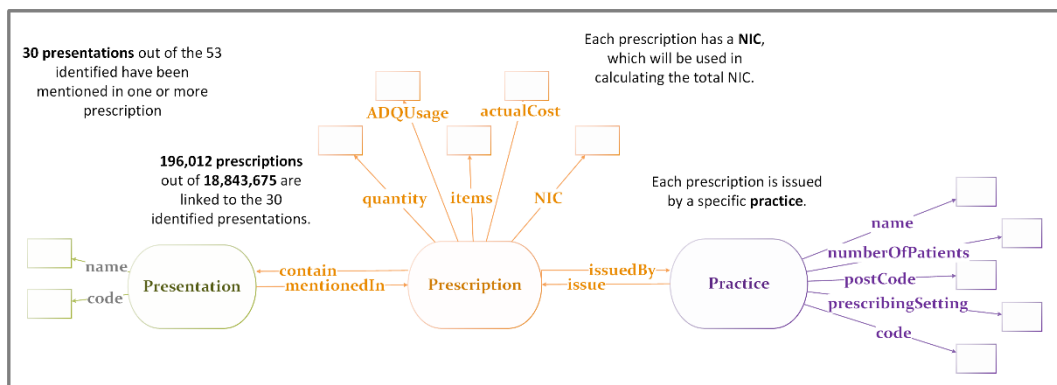


Figure 65: Step 2 of the cost-per-person query

From the previous step, 53 instances of the *Presentation* class are identified as those instances that belong to the metformin hydrochloride chemical substance. In this step, the prescriptions that are linked to these presentations are identified (see figure 65).

In the grey section of the main cost-per-person query (see figure 61) the system performs a search on more than 18 million prescriptions to identify a subset of 196,012 instances. There are more than 18 million uploaded instances of the class *Prescription* in the triplestore. In this step, only those instances that are linked to one of the 53 BNF presentations via the *contain* object property are identified. 196,012 instances of the *Prescription* class are identified as well as 30 instances of the *Presentation* class, which are linked specifically to these prescriptions.

Furthermore, in the design of the prescriptions ontology, each instance of the *Prescription* class is linked to several data properties. The data property that is important in this use case is the NIC, which is used in the next step to calculate the total NIC per each practice.

Finally, each one of the 196,012 instances of the *Prescription* class is linked to one instance of the *Practice* class via the object property *issuedBy*. 7,767 identified GPs' practices issued prescriptions for metformin hydrochloride. For each practice, two data properties are retrieved: *code* and *name* as these practices play a significant role in the next steps.

Step 3: Calculating the total NIC per practice

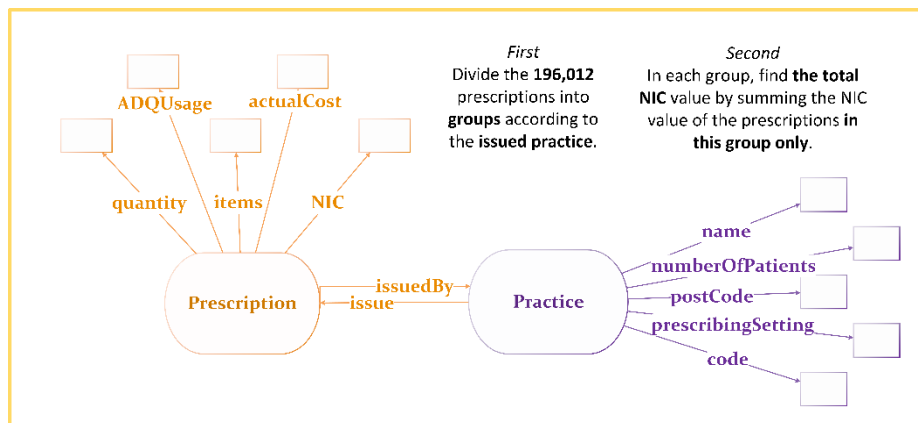


Figure 66: Step 3 of the cost-per-person query

After identifying the 196,012 prescriptions for metformin hydrochloride, the next step is to calculate the total NIC for all these prescriptions per practice. To perform this step, two sub-steps are needed, see figure 66.

First, those 196,012 instances of the *Prescription* class must be divided into groups according to the issuing practice linked to them. This means that each group contains all prescriptions for a specific practice. SPARQL provides the keyword *GROUP BY* to perform such a task.

The second step is to sum the NIC values attached to each prescription for each group. This means that if the number of the issuing practices in this query is 100, then 100 groups and accordingly 100 total NIC values must exist. To perform this step, the built-in summation function in SPARQL *SUM* is used.

Table 12 shows ten examples out of 7,767 groups for practices' prescriptions. These groups are divided according to the issuing practice. Each group is associated with only one practice, but with one or more prescriptions. The total NIC value in the last column is the result of summing the NIC values of the linked prescriptions in each group. For example, *THE DENSHAM SURGERY* prescribed 26 prescriptions for different presentations of the chemical substance metformin hydrochloride, and these drugs cost a total of £ 710.65.

Group ID	Practice Code	Practice Name	Number of linked prescriptions	Total NIC per practice
1	A81001	THE DENSHAM SURGERY	26	710.65
2	A81002	QUEENS PARK MEDICAL CENTRE	41	2799.01
3	A81004	WOODLANDS ROAD SURGERY	29	1479.3298
4	A81005	SPRINGWOOD SURGERY	24	902.99994
5	A81006	TENNANT STREET MEDICAL practice	30	4034.42
6	A81007	BANKHOUSE SURGERY	39	1280.7402
7	A81009	VILLAGE MEDICAL CENTRE	22	1079.1201
8	A81011	CHADWICK PRACTICE	27	2916
9	A81012	WESTBOURNE MEDICAL CENTRE	29	801.9599
10	A81013	BROTTON SURGERY	23	886.37006

Table 12: An example of the prescriptions' division into groups

Step 4: Finding the total number of registered patients

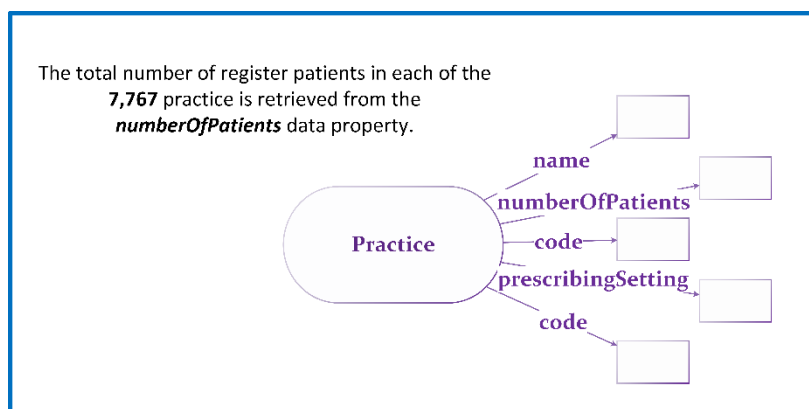


Figure 67: Step 4 of the cost-per-person query

This step concentrates on finding the total number of registered patients in each practice. To find the total number of registered patients, the data property *numberOfPatients* attached to the *Practice* class is used in the query see figure 67.

Step 5: Calculating the cost per person rate

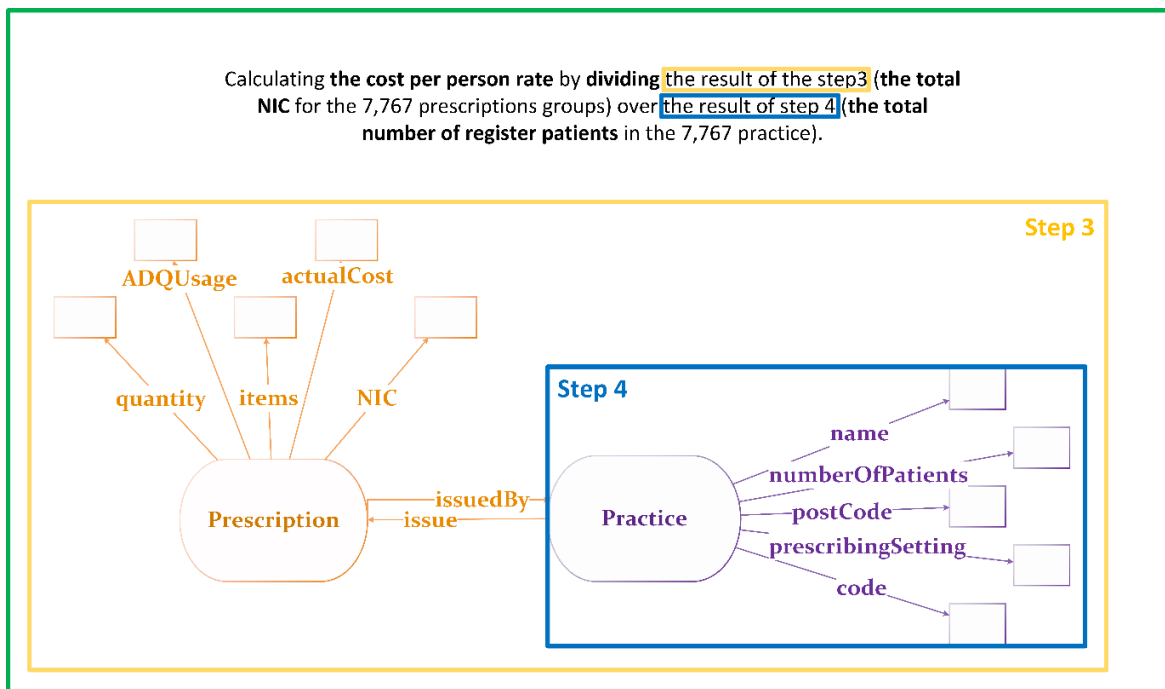


Figure 68: Step 5 of the cost-per-person query

In the final step, the cost-per-person rate is calculated by dividing the total NIC per practice computed in step 3, by the total number of registered patients in each practice found in step 4 (see figure 68). SPARQL provides the arithmetic operation division (/) as shown previously in the green highlighted section in the cost-per-person query (see figure 61).

6.1.3 Results

In Rowlingson *et al.* (2013) the results were presented as colour coded maps where the bright blue points presented the lowest cost-per-person rate, through dull blue, grey, dull red and bright red representing the highest rate. Figure 69 shows the cost-per-person rates on a London map (source: Rowlingson *et al.*, 2013).

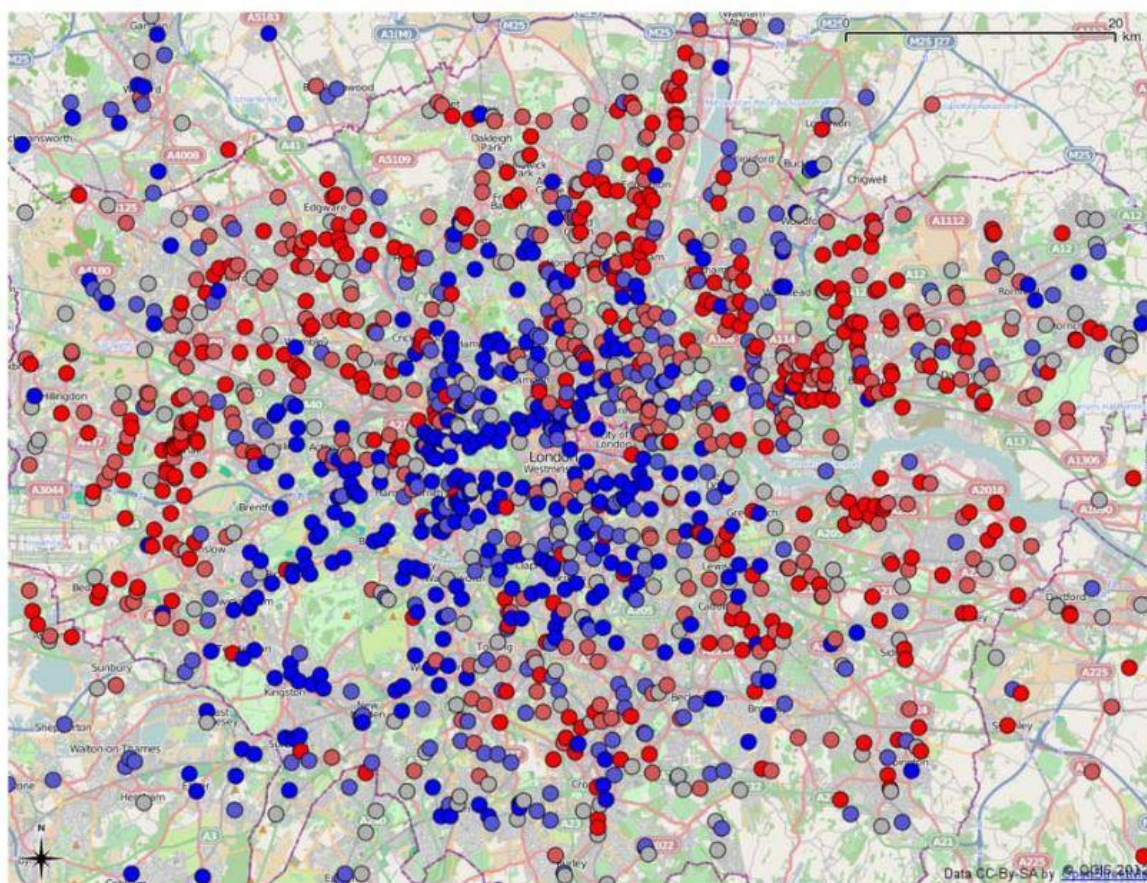


Figure 69: Cost-per-person for Metformin spending in London (Rowlingson *et al.*, 2013)

In an attempt to simulate some of the results reported by Rowlingson *et al.* (2013), the results from the cost-per-person query were downloaded as a CSV file. Then, the results were colour coded in the same manner from light blue to dark red according to the cost-per-person rate. In the Rowlingson paper, rates higher than 0.5 were excluded; thus, the same process is performed on this current example's results. Rates higher than 0.5 were excluded and were then represented as purple results in figure 70.

Figure 70 shows the percentage of each of the colour coded rates in the results. Most of the practices in England have a spending rate for metformin in the rate range from 0.1 to 0.2. Less than 10% of the practices have high spending rates. Table 13 shows a glance of the colour coded results in the CSV downloaded file.

As seen from the figures, the SW technologies were able to deliver similar results as in other traditional linked data approaches. despite the existence of some discrepancies in the exact results. Using different temporal datasets could be the reason behind these differences, which is expected in this scenario.

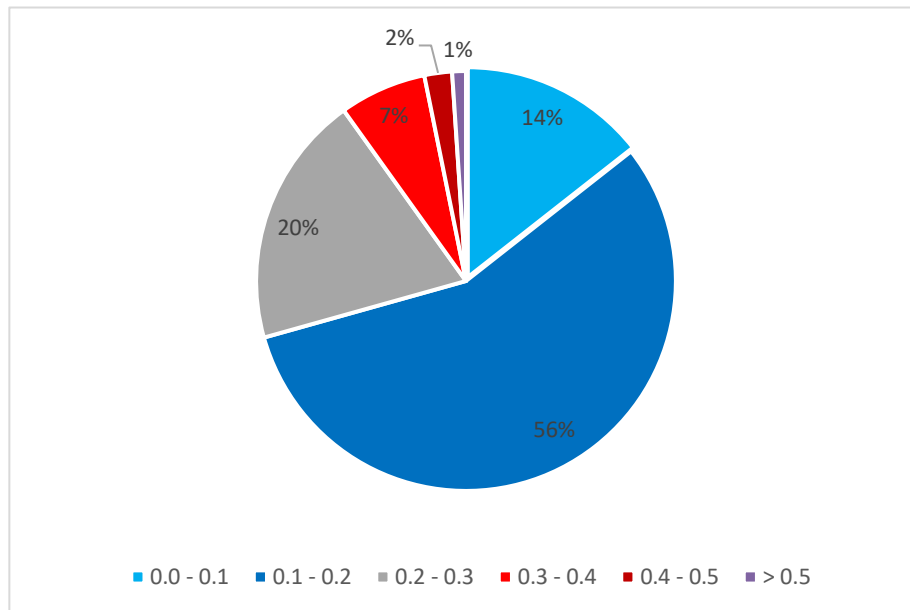


Figure 70: The percentages of the five colour coded cost-per-person rates

Practice Code	Practice Name	Practice Size	Total NIC	Cost per Person
A81033	OAKFIELD MEDICAL PRACTICE	3382	1619.5004	0.48
A81034	THORNABY & BARWICK MEDICAL GROUP	21744	2862.48	0.13
A81035	NEWLANDS MEDICAL CENTRE	10300	1284.81	0.12
A81036	NORTON MEDICAL CENTRE	17658	2478.4397	0.14
A81037	THE ERIMUS practice	6836	2675.6396	0.39
A81038	HIRSEL MEDICAL CENTRE	3332	860.2	0.26
A81039	EAGLESLIFFE MEDICAL practice	9859	1961.44	0.2
A81040	MARSH HOUSE MEDICAL practice	8641	1808.9998	0.21
A81041	HART MEDICAL practice	9128	3180.1191	0.35
A81042	SOUTH GRANGE MEDICAL CENTRE	10791	3123.8699	0.29
A81043	THE MANOR HOUSE SURGERY	8140	1077.2601	0.13
A81044	MCKENZIE HOUSE SURGERY	19272	2282.75	0.12
A81045	THE COATHAM ROAD SURGERY	7712	1306.3103	0.17
A81046	WOODLANDS FAMILY MEDICAL CENTRE	12845	2120.7502	0.17
A81047	MARSKER MEDICAL CENTRE	5056	252.82	0.05
A81048	ZETLAND MEDICAL PRACTICE	5172	2446.1602	0.47
A81049	KINGS MEDICAL CENTRE	5570	1190.4698	0.21

Table 13: A snapshot of the colour coded results for the cost-per-person query

6.2 Case Two: The Effect of Living in a Coastal City or Town upon the Prescribing of Antidepressants

Some public health questions tend to search for correlations between the people's health and different environmental conditions. One example in this area is the relationship between water surfaces and mental health; specifically, the incidence of depression. Dempsey *et al.* (2018) studied the possibility of linking exposing elderly to coastal blue space and depression, while Völker and Kistemann (2011) investigated the impact of blue space on well-being.

The topic of this case is very similar to the previous examples. It was suggested by a health expert in a designed focus group. The following section discusses the design of the focus group and how, based on the focus group's output, this topic was chosen.

6.2.1 Focus Group Design

The intention of this case's topic is to present an example from health researchers who were required to offer interesting suggestions for a health question that could be addressed by the prescriptions demonstrator. The followed methodology to produce some interesting health questions related to prescriptions is the outcome of conducting a focus group comprised of health researchers. The aimed output of the focus group is a list of suggested health questions that are interesting from the researchers' point of view.

After applying for an ethical approve from the Ethics Committee of the University of Southampton (Research Ethics Number ERGO/FEPS/49893: see Appendix F for more details) several health researchers were contacted. All participants were health researchers in the University of Southampton. They were contacted by email asking for their willingness to participate after providing them with some information about the study. After obtaining their individual approvals, a suitable time and place was organised.

The focus group consisted of three health researchers to give an opportunity for each participant to explain and discuss his/her opinion. To obtain an adequate range of suggestions, as well as to have a richer discussion, the participants were chosen with different health backgrounds, namely: a) nursing, b) audiology and c) physiotherapy.

Before starting the focus group, written approvals were collected from the participants, in line with the Ethics Committee's regulations. The author gave an overview of the study's aim and explained the main concepts in the prescriptions demonstrator. A paper-based model of the demonstrator, along with several open datasets and ontologies, were provided to the

participants. The participants were asked to think about any interesting questions that might be produced from these datasets but not limited to them. The participants then brainstormed several ideas and suggested different questions related to NHS prescriptions. The discussion took one hour and it was recorded and then transcribed. The transcribed discussion was then analysed and a list of 13 suggested questions were concluded. These questions are discussed in the next section.

6.2.2 The Focus Group's Results

The focus group's brainstorming resulted in a list of 13 suggested health-centric questions relating to the prescriptions demonstrator. The resulted questions are listed below in table 14.

Suggested Questions	Excluding Reasons
1. <i>Who prescribe better? The specialists or GPs? The experienced or junior doctors? The medical or non-medical prescribers?</i>	<ul style="list-style-type: none"> • Data restricted access • Depends on people's views
2. <i>Do the prescribers ask the right questions to the patients?</i>	<ul style="list-style-type: none"> • Data restricted access • Depends on people's views
3. <i>Is sophisticated way of prescribing by paying attention to the secondary characteristics of the medication been performed? And by what type of prescribers?</i>	<ul style="list-style-type: none"> • Data restricted access
4. <i>Do prescribers who are limited in prescribing perform better than unlimited ones?</i>	<ul style="list-style-type: none"> • Data restricted access
5. <i>How following different guidelines in different countries/regions affect the prescription of a specific disease? E.g. incurable skin diseases?</i>	<ul style="list-style-type: none"> • Data restricted access
6. <i>Is following guidelines or relying on doctor's experience is better?</i>	<ul style="list-style-type: none"> • Data restricted access • Depends on people's views
7. <i>How often the guidelines are been followed? How many patients were diagnosed with condition X and treated as in the guidelines?</i>	<ul style="list-style-type: none"> • Data restricted access
8. <i>What are the factors that can affect the prescribing decision? E.g. blood tests, patient preference?</i>	<ul style="list-style-type: none"> • Depends on people's views Broad aim
9. <i>How can guidelines resolve the issue of variation in treatment for the same condition?</i>	<ul style="list-style-type: none"> • Data restricted access • Broad aim
10. <i>How doctors make their treatment decision?</i>	<ul style="list-style-type: none"> • Depends on people's views
11. <i>Is there a pattern in prescribing branded and generic drugs?</i>	<ul style="list-style-type: none"> • Data restricted access
12. <i>What are the factors that affect the number of people suffering from a specific disease between different regions? E.g. weather & flu?</i>	<ul style="list-style-type: none"> • Data restricted access • Broad aim
13. <i>Is there a relationship between living in a coastal city or town and the prevalence of depression?</i>	Chosen question

Table 14: The list of suggested questions in the focus group with their exclusion reasons

The opinions and suggestions of each participant were valuable and helped the researcher to understand possible scenarios where the prescriptions demonstrator could be useful. Some examples of the suggested topics were around the effect of the prescriber's experience on the prescription decision. Others were on the procedure of prescribing and the relationship between the prescriber and the patient. Other questions were around questioning the feasibility of following clinical guidelines on all the cases and arguing the importance of the prescriber's experience in this equation. Some questions were about the sophisticated procedure that doctors follow in order to decide the best treatment for each case; particularly taking into consideration the side effects and the special characteristics that differentiate each drug product. Some questions were from a health inequality perspective in understanding patterns of prescribing branded and generic drugs. Another question was about the effect of weather or location on the prevalence of certain health conditions, such as depression.

The questions were analysed in order to find a suitable illustrative example for the purpose of this case study. Questions were excluded based on the practicality of addressing them using the SW approach. From table 14, there were three main reasons for excluding a question: 1) depending on primary data and people's views, 2) not having a narrow aim (too general), and 3) restrictions in data access.

There were five suggested questions that mainly depended on primary data such as surveys or experts' views. This type of questions was excluded because answering the question do not require using any of the SW tools. It relies on collecting data from original sources (human-source) by surveying people in the investigated matter. For example, to answer question 10, "*How doctors make their treatment decision?*", asking doctors by interviews for instance would be the most suitable method to answer this type of questions.

Another reason for excluding three questions was the generality and broadness of the question's aim. An example for this case would be question 8: "*What are the factors that can affect the prescribing decision?*". Listing the factors affecting prescribing decision is a very broad aim that cannot be represented as SPARQL query. Writing a query would require being specific and finite in describing the relationships between elements.

The last reason for excluding questions in this case study was the restrictions in accessing certain needed datasets. Ten questions involved integrating data with restricted access such as patients or prescribers' data. Such questions are interesting, but they are beyond the scope of this research. The main feature of the data used in this research is being open.

Analysing all the questions led to choosing question 13 in this illustrative case study, because it is suitable to be represented using the SW technologies as well as the needed data are available in open sources. Question 13 investigates the possible relationship between living in a coastal city or town and the depression rates of the inhabitants. The integration in this example is not limited to the three ontologies in the prescriptions demonstrator, but also involves integrating data remotely via federated queries. This case demonstrates another interesting SW feature that was not tested before in the thesis; one of the strong advantages of the SW in comparison to any other data linking techniques. Thus, the question of “Is there a relationship between living in a coastal city or town and the prevalence of depression?” was chosen to be addressed in this case. Another way of expressing the question is “What are the rates of prescribing antidepressants in coastal and non-coastal cities or towns in England?”

6.2.3 Case Description

The aim of this case is to check the effect of living in a coastal city or town on the depression rates evident in that city/town. The depression rate for a city/town is calculated by summing the number of antidepressants prescription per city/town over the city’s or town’s population. The geographical location for this case includes all practices in England.

To answer the question of the effect of living in a coastal city/town on the depression rate, data from the prescriptions demonstrator is needed, as well as some extra external data from the LOD. The local data used from the prescriptions is about NHS dispensed prescriptions for antidepressants by specific practices in certain cities/towns. Antidepressants drugs can be identified from the BNF ontology, specifically from the *Section* class that specifies the type of medicine.

The other part of the needed data is about defining coastal cities/towns in England. Wikidata provides massive crowdsourcing information about many things (The Wikimedia Foundation, 2019). One of the provided data in Wikidata is geographical data about countries, cities/towns, water bodies in a city/town and population information as well. All these types of data are needed to answer this use case. More details about the data that were used are in the next section.

6.2.4 The Query

The addressed question in this use case is “What are the rates of prescribing antidepressants in coastal and non-coastal cities/towns in England?” To translate this question into a SPARQL query, two main data sources are used: prescriptions demonstrator data and Wikidata. The purpose of

this query is to calculate the antidepressants prescription rate for a particular city/town. The rate is calculated by summing the number of antidepressants prescription per city/town over the city's or town's population.

The second important information task in this query is to identify coastal cities/towns from non-coastal ones. To understand if the geographical location of being by the coast affects the number of prescribing antidepressants, both coastal and non-coastal cities/towns' antidepressant rates need to be found, for comparison purposes. Thus, two similar queries are run for the two types of cities/towns, with 'location' being the only difference in the city/town definition.

Figure 71 shows the SPARQL query for calculating antidepressant rates. The query consists of eight main steps, as follows:

- 1) Identifying cities and towns located in England from Wikidata.
- 2) Identifying cities/towns' population from Wikidata.
- 3) Defining coastal cities/towns according to Wikidata.
- 4) Identifying the list of NHS practices in the selected coastal cities/towns.
- 5) Selecting all BNF presentations for antidepressants.
- 6) Calculating the antidepressant prescribing rates for each city/town.

Figure 72 shows the six steps, together with the classes and ontologies used in the prescriptions demonstrator and Wikidata. The following sections discuss each step in details.

```
SELECT DISTINCT
```

```
?practice_city
(SAMPLE(?population) AS ?POPULATION)
(COUNT (DISTINCT ?prescription) AS ?NumberOfPrescriptions)
(?NumberOfPrescriptions/?POPULATION*100 AS ?DepressionRate)
WHERE
{
SERVICE <https://query.wikidata.org/bigdata/namespace/wdq/sparql>
{
SELECT DISTINCT
?coastalCityLabel
?population
WHERE
{
?coastalCity wdt:P31/wdt:P279 wd:Q7930989;
wdt:P17 wd:Q145;
wdt:P131/wdt:P131/wdt:P131/wdt:P131 wd:Q21;
wdt:P1082 ?population.
wdt:P206 ?water;
VALUES ?waterType { wd:Q165 wd:Q39594 wd:Q47053 }
?water wdt:P31 ?waterType.
SERVICE wikibase:label
{ bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
}
?practice nhs:code ?practice_code;
nhs:name ?practice_name;
nhs:city ?practice_city.
FILTER (lcase(?practice_city) = lcase( str(?coastalCityLabel) ) )
?section a bnf:Section;
bnf:name "Antidepressant Drugs".
?presentation a bnf:Presentation;
bnf:belongTo/bnf:belongTo/bnf:belongTo/
bnf:belongTo/bnf:belongTo ?section.
bnf:code ?presentation_code;
?prescription a pres:Prescription;
pres:issuedBy ?practice.
pres:contain ?presentation.
}
GROUP BY ?practice_city
```



Figure 71: The 'depression rates in coastal cities/towns' query

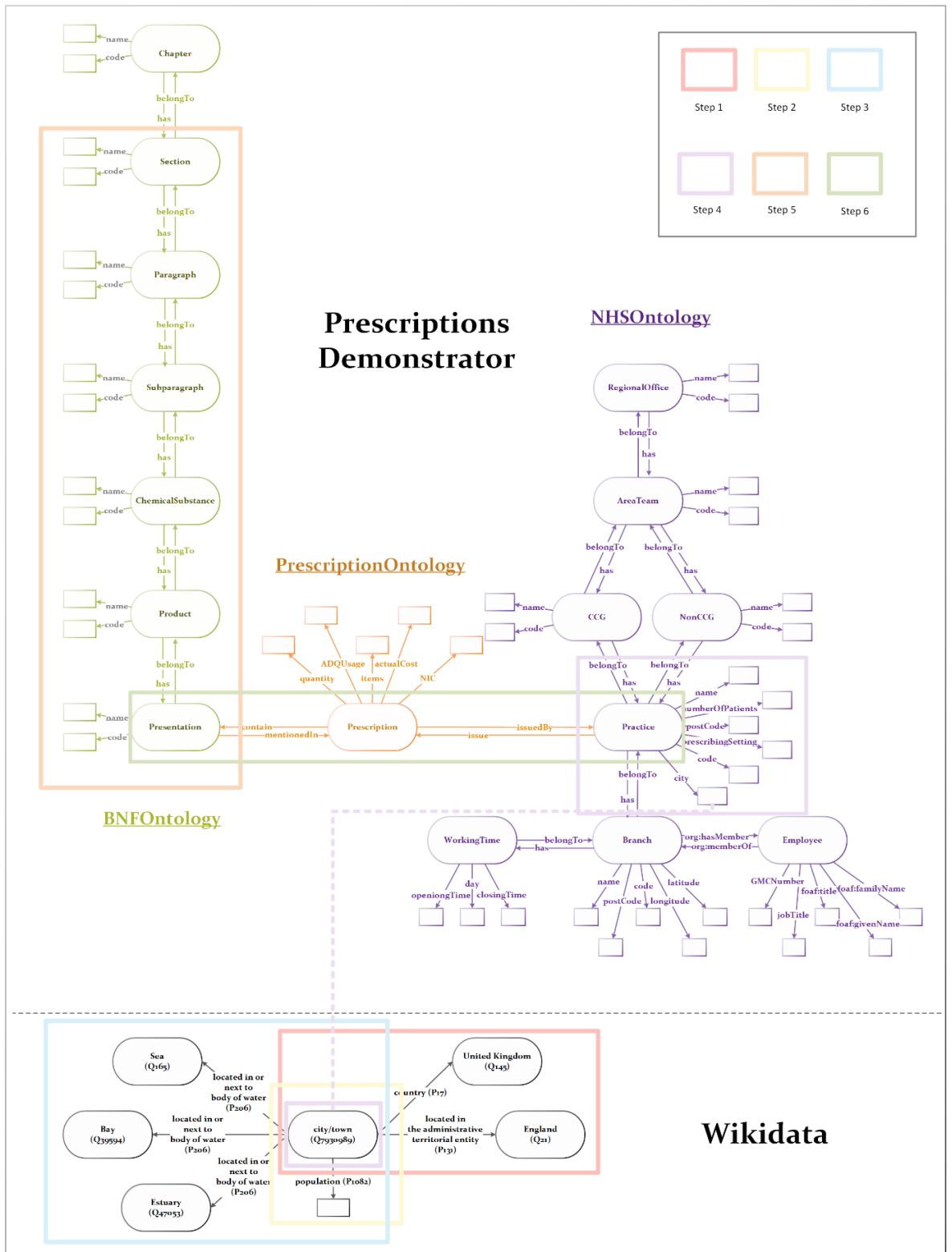


Figure 72: The data concepts used in the six steps of the ‘depression rates in coastal cities/towns’ query

Step 1: Identifying cities/towns and towns located in England from Wikidata

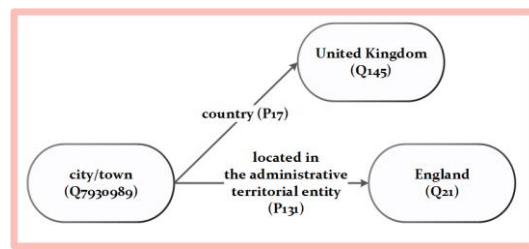


Figure 73: Step 1 of the 'depression rates in coastal cities/towns' query

To find the cities/towns or towns which are located in the United Kingdom, the instances of *city/town* class linked by the property *country (P17)* to the *United Kingdom (Q145)* instance are required. 1696 instances were found. 862 cities or towns are *located in the administrative territorial entity P131* named *England (Q21)*. This case focuses only on English territory within the UK because the prescriptions demonstrator data is only available across England. Figure 73 shows the first step in the query that identifies the list of towns or cities/towns in England.

Step 2: Identifying cities/towns' populations from Wikidata

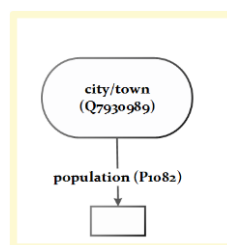


Figure 74: Step 2 of the 'depression rates in coastal cities/towns' query

This query aims to calculate the antidepressants prescribing rates in each selected city/town. Since cities/towns have different sizes and populations, the antidepressants prescribing rate is normalised by the population of each 3.

There is a defined data property in Wikidata named *population (P1082)* that is linked to some of the existing cities/towns in Wikidata. However, since Wikidata depends mainly on crowdsourcing for feeding the graph with data, there are many missing and outdated aspects that could mislead or hinder a research initiative. For example, in Wikidata only 177 English cities/towns out of 862 have the population information. Moreover, most of the defined population data is outdated,

based on the 2011 census. Figure 74 shows the part of the query of finding the population property for the identified cities/towns.

Step 3: Defining coastal cities/towns according to Wikidata

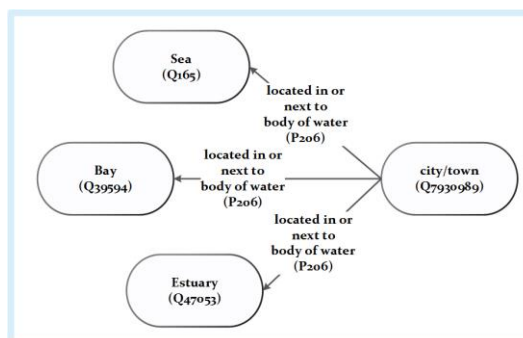


Figure 75: Step 3 of the ‘depression rates in coastal cities/towns’ query

One of the challenging steps in this query is deciding on what is the proper definition for a coastal city/town. One of the attempts to define a coastal city/town was by finding cities/towns that are located in a certain radius distance from the sea for example. However, this definition obviously does not work as some cities/towns would be within this radius distance, but not located on the sea; or they may only be small towns.

In Wikidata, there is an interesting property named *located in or next to body of water (P206)* that links geographical locations, like cities/towns with a sea or lake shore. By querying the available English towns or cities/towns that have the population information and located next to a body of water, 89 cities/towns were retrieved out of the 177 found in the previous step. By visually screening the results, only 5 cities/towns out of these 89 ones were located next to a salty water type. Most of the cities/towns in England are located next to rivers. Up to this point, there are two issues: a) how to define ‘salty’ water type and b) the very small number of retrieved results from this part of the query: 177 to 89 to 5.

Regarding defining salty water, Wikidata contains a class named *Marine Ecosystem (Q3304561)* that could be used for this definition. However, there are lots of missing data in the data, thus, not all the ‘salty’ water types are listed under the *Marine Ecosystem* class. The proper definition for a coastal city/town would be a city/town that is located next to a body of water and this body of water is a subclass of a marine ecosystem.

Returning to the visually filtered cities/towns located adjacent to a salty water type, the water type classes were either a *Bay* (Q39594) or *Estuary* (Q47053) or *Sea* (Q165). Therefore, the final decided definition for a coastal city/town in this specific case is “a *city or town* (Q7930989) that is *located in or next to body of water* (P206), and this body of water is either a *Bay* (Q39594) or *Estuary* (Q47053) or *Sea* (Q165)”. Unfortunately, only five instances match the definition of the English coastal city/town at the end. Figure 75 shows coastal cities/towns are identified by the three classes: *Bay* (Q39594) or *Estuary* (Q47053) or *Sea* (Q165).

Step 4: Identifying the list of NHS practices in the selected coastal cities/towns

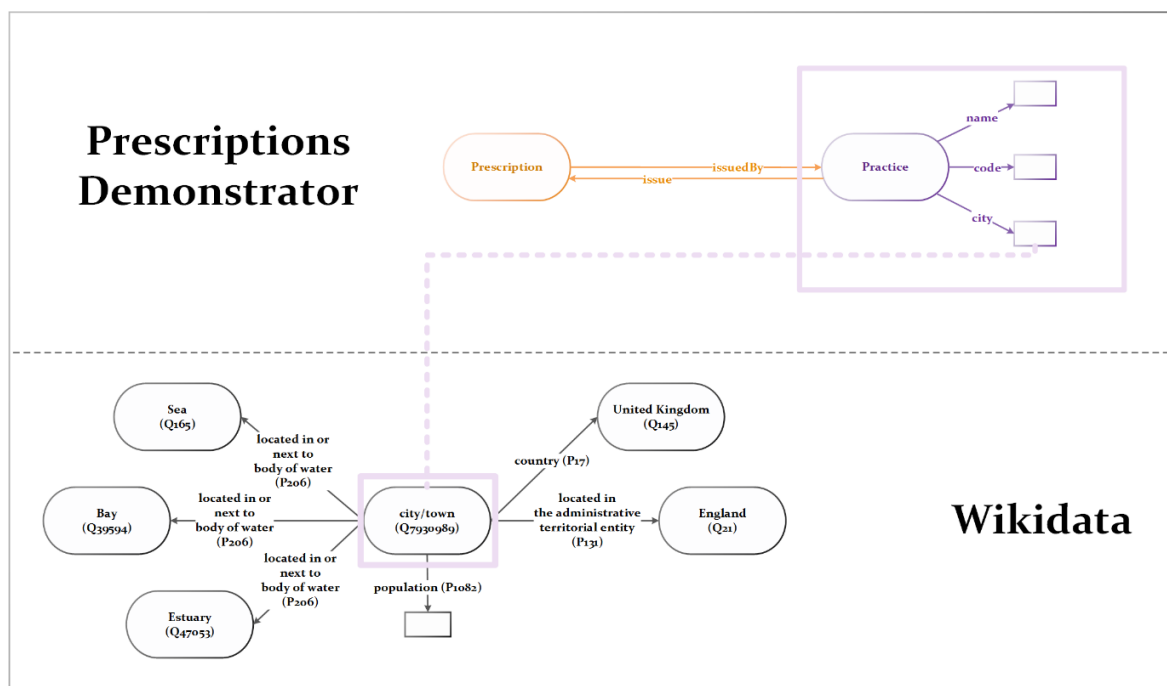


Figure 76: Step 4 of the ‘depression rates in coastal cities/towns’ query

In this step, the remote data integration is performed by querying the Wikidata SPARQL endpoint from the local prescriptions demonstrator query interface. The data retrieved from Wikidata is the list of English coastal cities/towns and their population property. The Wikidata and the prescriptions data are bound by the city/town name. In Wikidata, it is the coastal city/town’s name, while it is the NHS practice city/town in the NHS ontology. Figure 76 shows the linking between the Wikidata and the prescriptions demonstrator.

Step 5: Selecting all BNF presentations for antidepressants

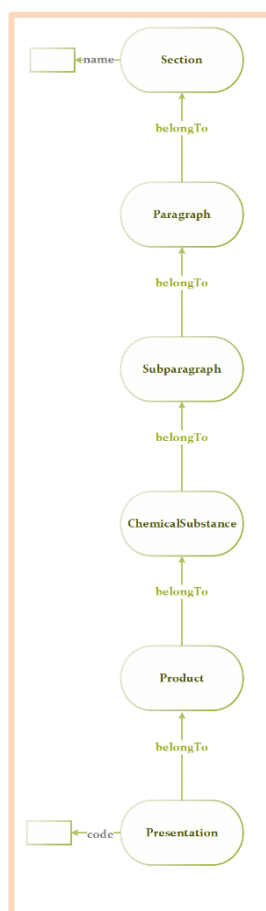


Figure 77: Step 5 of the ‘depression rates in coastal cities/towns’ query

Step 5 takes place in the BNF ontology. The aim of this step is to find all BNF presentations for antidepressants. To do so the BNF section named ‘Antidepressant’ is identified first. Then, all BNF presentations belonging to the antidepressant section are identified by traversing the links between BNF *Paragraph*, *Subparagraph*, *ChemicalSubstance*, *Product* and *Presentation* instances. The results of this step are 629 antidepressant presentations identified by their BNF presentation *codes*. Figure 77 shows how BNF presentations for antidepressants are identified in the BNF ontology.

Step 6: Calculating the antidepressants prescribing rate per each city/town



Figure 78: Step 6 of the 'depression rates in coastal cities/towns' query

The final step in this query is to find the antidepressants prescribing rates for each coastal city/town. The rate is defined as the total number of prescriptions in a city/town over this city/town's population. The city/town's population information is already retrieved from the federated query; however, the total number of prescriptions needs to be identified.

To count the prescriptions prescribed by an NHS practice in one of the coastal cities/towns list and issued for any of the identified antidepressants presentations, the prescriptions are grouped (aggregated) according to the issuing practices' city/town by using SPARQL's keyword *GROUP BY*. This means that there will be five groups of prescriptions for each coastal city/town. Secondly, the prescriptions are counted using the keyword *COUNT* in each group and the total is saved as *?NumberOfPrescriptions* variable. The final step is to use the arithmetical operation, division (/), to calculate the antidepressants prescribing rate by dividing the total number of prescription in a city/town over the population. Figure 78 shows the final step in the query by calculating the antidepressants prescribing rate per city/town.

6.2.5 Results

This query was run twice: i) firstly for finding the antidepressant prescribing rates for coastal cities/towns, and ii) secondly for finding the rates for a list of similar non-coastal cities/towns. Five non-coastal cities/towns were chosen to be tested to correspond to the five coastal cities/towns identified from Wikidata. The five non-coastal cities/towns are: 1) Truro, 2) Letchworth Garden City, 3) Derby, 4) Bolton and 5) Chiddingfold in comparison to the following five coastal cities/towns: 1) Wells-Next-The-Sea, 2) Sunderland, 3) Southampton, 4) Falmouth and 5) Morecambe. Table 15 show the results for both queries.

The queries were executed using the GraphDB SPARQL endpoint, where in some parts of the queries the Wikidata Query Service SPARQL endpoint was also invoked. The results are download as a CSV file.

coastal city/town	rate	Non-coastal city/town	rate
Falmouth	1.9	Bolton	1.2
Morecambe	1.1	Chiddingfold	3.9
Southampton	1.6	Derby	1.3
Sunderland	1.4	Letchworth Garden City	0.6
Wells-Next-The-Sea	3.6	Truro	3.5

Table 15: The results for the antidepressants prescribing rates for coastal and non-coastal cities/towns

	practice_city	NumberOfPrescriptions	POPULATION	DepressionRate
1	FALMOUTH	"419"^^xsd:integer	"21797"^^xsd:decimal	"1.922282882965545717300500"^^xsd:decimal
2	SOUTHAMPTON	"3946"^^xsd:integer	"253651"^^xsd:decimal	"1.555680837055639441594900"^^xsd:decimal
3	MORECAMBE	"398"^^xsd:integer	"34768"^^xsd:decimal	"1.144730786930510814542100"^^xsd:decimal
4	WELLS-NEXT-THE-SEA	"78"^^xsd:integer	"2165"^^xsd:decimal	"3.602771362586605080831400"^^xsd:decimal
5	SUNDERLAND	"3905"^^xsd:integer	"277417"^^xsd:decimal	"1.407628227541931460581000"^^xsd:decimal

Figure 79: The results for the coastal cities/towns query

Figure 79 shows the details for the coastal cities/towns results, while figure 80 shows the results for the non-coastal cities/towns.

	practice_city	NumberOfPrescriptions	POPULATION	DepressionRate
1	LETCHWORTH GARDEN CITY	"208""xsd:integer	"33249""xsd:decimal	"0.625582724292459923606700""xsd:decimal
2	CHIDDINGFOLD	"87""xsd:integer	"2211""xsd:decimal	"3.934871099050203527815500""xsd:decimal
3	BOLTON	"3495""xsd:integer	"285372""xsd:decimal	"1.224717211219040410411700""xsd:decimal
4	TRURO	"756""xsd:integer	"21555""xsd:decimal	"3.507306889352818371607500""xsd:decimal
5	DERBY	"3424""xsd:integer	"255394""xsd:decimal	"1.340673625848688692764900""xsd:decimal

Figure 80: The results for the non-coastal cities/towns

One of the main issues faced in applying this case was the limited information and missing data from Wikidata. Therefore, another ‘manual’ attempt to compare the coastal and non-coastal antidepressant prescribing rate was performed. Data was manually fed to the prescriptions demonstrator for 10 coastal and 10 non-coastal cities/towns in England. The data includes the names of coastal cities/towns with their populations and the names of non-coastal cities/towns with their populations. The data was saved as a text file (RDF) and uploaded into the prescriptions repository.

coastal city/town	Number Of Prescriptions	Population	Depression Rates	Round
<i>BOURNEMOUTH</i>	2199	183491	1.198423901	1.2
<i>BRIGHTON</i>	3514	290395	1.210075931	1.2
<i>HERNE BAY</i>	555	38563	1.439203381	1.4
<i>SOUTHPORT</i>	1312	90381	1.451632533	1.5
<i>WEYMOUTH</i>	968	58200	1.663230241	1.7
<i>FALMOUTH</i>	419	21797	1.922282883	1.9
<i>BLACKPOOL</i>	3160	139720	2.26166619	2.3
<i>WITHERNSEA</i>	159	6159	2.58158792	2.6
<i>VENTNOR</i>	196	5976	3.27978581	3.3
<i>SALTBURN-BY-THE-SEA</i>	412	5958	6.915072172	6.9

Table 16: The results of the manually added coastal cities/towns for the antidepressants query

Two queries were run across the new data to calculate the antidepressants prescribing rate for both coastal and non-coastal sources. Table 16 shows the results of coastal cities/towns, while table 17 shows the non-coastal cities/towns.

Non-coastal city/town	Number Of Prescriptions	Population	Depression Rates	Round
<i>BIRMINGHAM</i>	10393	3701107	0.280807877	0.3
<i>MANCHESTER</i>	10280	2568711	0.400200723	0.4
<i>CORBY</i>	483	62400	0.774038462	0.8
<i>LICHFIELD</i>	325	33816	0.961083511	1
<i>MILTON KEYNES</i>	2728	229941	1.186391292	1.2
<i>COVENTRY</i>	4943	366785	1.347655984	1.3
<i>READING</i>	3471	230046	1.508828669	1.5
<i>BANBURY</i>	1074	43867	2.448309663	2.4
<i>CAMBRIDGE</i>	3408	123900	2.750605327	2.8
<i>KINGSTON</i>	86	2626	3.274942879	3.3

Table 17: The results of the manually added non-coastal cities/towns for the antidepressants query

After obtaining the results of all the queries, the outcomes were shown to the health expert who initially suggested the question, in order to obtain feedback. The health expert thought that the results showed no correlation between living in a coastal city/town and their populations' depression rates. Moreover, the health expert suggested that the results were not sufficient to derive any conclusions from the outcomes presented. However, the expert did add that the ability to link different datasets easily is interesting; that while the process is not yet ready to be employed it seems to have useful potential.

6.3 Analysis

The aim of applying these two cases was to test the SW's features in addressing health questions and to analyse any affordances or challenges faced in the process. The two cases were chosen by two different methods and represented two different topics.

The first case addressed a health inequality question representing the distribution of diabetic medication prescribing across England. The question was inspired by a paper from Rowlingson *et al.* (2013); specifically the authors addressed question: "What is the cost-per-person rate for metformin mydrochloride prescriptions in England?". A SPARQL query consisting of five main steps was applied to the three interlinked ontologies in the prescriptions demonstrator. The query successfully addressed and answered the targeted question.

The question addressed in the second case was based on some suggestions offered from a health experts' focus group. The question was chosen because answering it relies on using open available data on the web, while the other questions were excluded due to the lack of accessing some of the suggested used data such as patients datasets. Exploiting LOD data is considered one of the strengths of the SW approach. Similarly, the lack of access to certain datasets is considered an obstacle facing researchers.

The question was designed to discover the effects (if any) of living in a coastal city/town on the amount of depression cases in that same city/town. The chosen question to be translated into a query was "What are the rates of prescribing antidepressants in coastal and non-coastal cities/towns in England?". As in the first use case, the question was translated into a SPARQL query consisting of six main steps. SPARQL was also able to address and answer the chosen question. In fact, the special thing in this query is that it successfully demonstrated federated querying.

The cases concentrated on translating the addressed question into a query. Thus, the main objective was to test the SW on its ability to explore, traverse and discover knowledge. SPARQL is the main SW standard used in this chapter to demonstrate knowledge discovery through '*querying*' the system. One of SPARQL's tested feature was the ability to filter and handle big data based on specific conditions. This feature was demonstrated several times in both cases. For example, in the first use case, the second step aimed to identify all the prescriptions in England that contained any of the previously identified BNF presentations. The power to handle millions of records by the SW technology was demonstrated in this case. The system was able to search and identify the prescribed drug's code in more than 18 million prescriptions records. Around 200,000

prescriptions were identified that contain one of the 53 previously chosen metformin presentations.

However, handling big number of records is not exclusive to SW's querying features; database technology can also handle enormous amounts of linking data. For instance in research by Rowlingson *et al.* (2013) that the first case presented in this current thesis was built on, the 2013 research team handled the same amount of records using relational databases, as well as the analysis section in OpenPrescribing project (EBM DataLab - University of Oxford, 2017).

One of the advantages of using querying systems is the ability to ask related questions in parallel with the main research question. This benefit is unlike fixed tools such as: a) iView from NHS Digital (NHS Digital, 2017b) or b) the Information Services Portal by the NHS BSA (NHS Business Services Authority (BSA), 2017); in both these tools only results filtered by the available parameters can be achieved. However, if the investigator wanted to know more about the subject, there is no available way to acquire that information when using fixed tools. In the two cases presented in this chapter, the process of asking extra questions through writing extra queries is found helpful in some situations. It has been found that this process helps in understanding the whole picture of the research problem and can give the researcher a deeper understanding of the problem that is being investigated.

An example of asking for extra information came in the first case where a query was raised about which metformin's presentations are used in the found prescriptions and which are not. From the first step in the cost per person query, 53 different presentations were identified. However, only 30 presentations out of the 53 were prescribed in England. This bit of extra information could be useful for different users, such as decision makers or pharmaceutical companies, because it would seem that there are 23 presentations of products in the market which have not been prescribed by any doctor.

Another helpful feature in SPARQL is the built-in arithmetic functions. They were used in calculating the cost-per-person rates in the first case and the antidepressants prescribing rates in the second case. Some examples of the used functions are: i) the division (/) operator, ii) the SUM function and iii) the COUNT function.

Data integration was also demonstrated in the two cases. Both queries contained multiple smaller queries within them for integrating data. Even in the second case, the integration level was remote. 'Federated query' was used via the SPARQL keyword SERVICE to remotely retrieve data from Wikidata. This feature is one of the SW's most powerful advantages, as what makes the SW distinctive from other data linking techniques is the availability and possibility to link data from

the cloud. Utilising data through reusing others published data saves time and effort and encourages data sharing in scientific environments.

Although in principle the remote data integration via federated querying pushes forward the development of a global SW, the often poor or inadequate quality of the available data is still a major challenge. For example, the desired data from Wikidata was either missing or outdated. The data property *located in or next to body of water (P206)* in Wikidata was not added to the definition of most of the cities/towns this researcher was able to access. Only 89 cities/towns out of 862 cities/towns had this information. Moreover, the *population property P1082* is also missing in the definition of some cities/towns. Only 177 out of 862 cities/towns had the population property. Another problem relating to the population property Wikidata is the outdated information. Most of the defined population data is from the 2011 census. Because Wikidata is a crowd sourcing knowledge base, it is expected that the quality of some of the provided data is inadequate; with significant implications for the validity of any research findings.

One more interesting feature provided by SPARQL is aggregating data (solutions) via grouping them. The used keyword for this task is GROUP BY. The default case in SPARQL is to have a single group for the set of solutions (results of the query). However, by using the GROUP BY keyword, the results can be aggregated into groups based on certain expressions (World Wide Web Consortium (W3C), 2013b).

Aggregation was used in both use cases. The grouping in the first case was by 'the practice'. There was 7,767 practices in England prescribed Metformin, thus, there is the same amount of groups in the query. In this case, the sum of prescriptions costs issued by the practice of a certain group is added to the aggregated information. Regarding the second case, the grouping is performed on the level of each practice's city/town. The prescriptions of antidepressant drugs are grouped and counted according to the issuer practice's city/town. So, the final result for each group is the number of antidepressant prescriptions in each city/town.

One of the major issues faced in the second case was the inadequate or unsatisfactory quality and the lack of availability of the required data. Due to many missing data definitions, the results of the query were not sufficient or adequate. Mainly, the missing data related to the definition of coastal cities/towns. The cities or towns that are *located in or next to body of water (P206)*, and this body of water is either a *Bay (Q39594)* or *Estuary (Q47053)* or *Sea (Q165)*. Starting from 862 instances of cities or towns located in England, and ending up with only five "coastal" cities/towns, suggests there is a large amount of missing data definitions in the Wikidata resource.

Besides the data missing from Wikidata, there was also some data missing from the NHS ontology representation. The datasets used in producing the NHS ontology are data published by different NHS organisations. Many data elements in these datasets were not registered correctly. For example, the practices' addresses datasheet includes a column for the practice's city/town. However, when the data within this column were mapped to the data property city/town in the Practice class in the NHS ontology, some of the produced data were 'dirty' and incorrect. Human errors in inputting information in the datasheet are very common and unfortunately produce incorrect mapped data as an outcome. This problem resulted in the linking of some of the practices' cities/towns with the names of the identified coastal cities/towns being missed. For example, if the spelling did not match exactly, the linkage will be ignored.

Finally, representing results as connected graphs was found helpful in some situations. Some triple stores and SPARQL endpoints provide graphical interfaces to view the retrieved data, as with the GraphDB tool. This tool was used to produce parts of the query's results as in figure 64 in the first use case. In some cases, representing information as a graph is simpler for the user and can help in reducing the difficulty of understanding complicated relationships between data elements. However, there are still some limitations in such tools, particularly regarding the limited number of viewed nodes and links.

6.4 Summary

This chapter described the process of addressing two health cases via the prescriptions demonstrator to test more of the SW's features. Part of the third research questions was answered in this chapter, in terms of analysing any affordances and challenges revealed during the process.

The first use case was inspired from the traditional literature and investigated prescribing inequalities for diabetic medications in England. The second case was suggested by health experts in a focus group; the question was designed to discuss the effect of living in coastal towns and cities/towns upon the prescribing rates of antidepressants. SPARQL queries were implemented on the prescriptions demonstrator to address both cases; queries from which results were obtained in case 1 and case 2.

The main SW feature tested in this chapter is the possibility of discovering knowledge via 'querying'. Both of the implemented SPARQL queries were able to explore and retrieve big numbers of linked records locally and also remotely. In the second case, an interesting feature of the SW was successfully tested, 'federated querying' from Wikidata. SPARQL also provides useful features such as the built-in arithmetic functions and the ability to group certain solutions to apply an expression to these groups. On the other hand, the challenges that hindered the process of analysing feasible results, especially in the last case, were: i) lack of adequate data and ii) poor or unsatisfactory data quality.

Chapter 7 Discussion

The first research question addressed in this thesis is “*What are the main health questions being addressed in health research employing semantic web technologies?*”. Section 7.1 discusses how this question was answered by reviewing the literature and how the health aims taxonomy was produced. Section 7.2 discusses the second research question, which is “*How are the semantic web features being used in health research?*”. This question is answered from the literature’s perspective by finding the SW’s main features used by health research. The final section in this chapter answers the third research question that asks: “*What are the affordances and challenges in employing the semantic web for health research?*”. The noted affordances and challenges evident when applying the SW in health research, from both the literature and practical perspectives, are discussed in this section.

7.1 Types of Addressed Health Questions

From systematically reviewing the literature of health research employing the SW, a health aims taxonomy was produced to answer the first research question in this thesis:

What are the main health questions being addressed in health research employing semantic web technologies?

In Chapter 4 Section 4.3.1, the health aims taxonomy were introduced. The taxonomy consists of 17 types of health questions that were addressed in the literature categorised under four broad health aims: a) medical, b) public health, c) health management and d) pharmaceutical. Some examples of the reviewed topics were in the form of questions relating to: i) clinical pathways and treatment plans (Hu et al., 2012), ii) clinical guidelines (Puustjarvi and Puustjarvi, 2016), and iii) Neuro medicine (Ciccarese, Wu, Kinoshita, et al., 2008).

The most repeated category of questions in the literature related to the medical aim specifically the diagnosis questions. From the results of the cross-tabulation analysis between health aims and used SW features located in the literature, the medical questions showed a positive correlation with using reasoning over rules. This relationship implies that finding answers to the medical aimed questions demands a ‘knowledge discovery’ method that can offer logic-based reasoning. Because the diagnosis questions take the biggest part of the medical category, they require the use of logic and reasoning as well. The diagnosis procedure relies on making decisions based on the patient’s condition; thus, the diagnosis questions have a decision-based

characteristic. The SW can support decision-based questions, particularly those focused on diagnostic issues, by offering the use of 'if-then' semantic rules and logic.

An example of using defined SW rules in decision-based questions is the work of Alexandrou, Xenikoudakis and Mentzas (2008). The authors mentioned an example of one of the defined rules in their decision-based system for identifying the best treatment for each patient. The rule says that "if the patient is admitted in the healthcare organisation and there is a diagnosis of neurological deficit, then the patient has to be evaluated for thrombolysis eligibility". This statement is translated into a SW rule, which will enable the inferring of new knowledge and recommendations in the system.

One of the characteristics of the health questions addressed in the literature is relevant to questions depending on heterogeneous data integration. Interdisciplinary life sciences are powered by data integration on the web (Sagotsky et al., 2008). Data integration techniques were found to be used in the literature with topics such as: a) integrating data generated from monitoring sensor hardware (Puustjarvi and Puustjarvi, 2015), b) medical examination devices (Maragoudakis, Maglogiannis and Lymberopoulos, 2008), and c) integrating different administrative data such as hospital guidelines and policies (Puustjarvi and Puustjarvi, 2016).

Health questions with personalisation nature were another identified type of questions addressed in the literature. Aggregating data on a personal-level is a well-known activity in traditional health research. The traditional health linked data approach is defined as the process that involves aggregating data from different resources about a certain person, whilst maintaining his/her privacy (Kotwal et al., 2016). The linking in this approach is performed by independent data centres. However, linking data via the SW approach is achieved by using the SW's standards for defining data concepts and then forming links between them. Linking via the SW's approach can be performed between any related data concepts not necessarily linking via a personal identifier. However, the availability of a similar method to the one employed by independent data centres (in the traditional approach) would probably encourage more researchers to use the SW approach in their studies, due to the availability of anonymised personal sensitive data ready to use for linking.

In the literature dealing with SW approach, personalisation questions took two different forms: i) either the linking is based on an electronic version of a patient's record or ii) a user account in a web service. An example of a question relating to a patient's personal record can be found in the work of Wang et al. (2013), where personalised treatment plans for patients are provided to enhance the treatments' quality. Wang et al., 2010 also provided an example for the other personalisation question by aggregating information for each user's account. The authors

developed a personal health information web service that provides targeted health information for every registered user.

The final characteristic of the questions relevant to the SW approach that were addressed related to accessing distributed information on the web. Public health systems that aimed to promote health information were built to access the available information on the web and produce electronic personal health consultants, such as found in Wang et al. (2010). These types of papers were similar to those accessed by Eysenbach (2003) who concentrated only on papers discussing the consumption of health information using the SW. The author concluded that the SW may open more opportunities for health information consumers in finding and aggregating information. However, this option might lead to overlaying on the web as a source of health information.

7.2 The SW's Uses in Health Research

Part of understanding the relationships between the SW and health research is to understand how the SW can be used in health research. The second research question concentrates on the uses of the SW in health research as the following:

How are the semantic web features being used in health research?

From systematically reviewing the literature, a taxonomy including 12 SW features to map the targeted literature was produced: 'the SW features taxonomy', see Chapter 4 section 4.3.2. The 12 identified features were identified under five main categories: a) data representation, b) data integration, c) knowledge discovery, d) updating knowledge and e) data sharing. The most used features in the literature were: i) representing data, ii) discovering knowledge and iii) integrating data.

7.2.1 Data Representation

Data representation in the SW can be achieved by conceptualising and modelling a certain domain. Usually, ontologies are used for this purpose; Gruber (1995) defined data representation as "an explicit specification of a conceptualization". Building ontologies includes tasks like defining classes and relationships between those classes (World Wide Web Consortium (W3C), 2008).

In the literature, building ontologies was the most used feature of the SW; registering 93% of total usage. SW standards like RDFS, RDF and OWL were used to model various topics and build ontologies in different domains. For example Dang et al.(2008) developed an ontology to

represent hospital resources and processes, while Hu et al. (2012) built an ontology to represent clinical pathways and treatment procedures.

In practice, access to and analysis of the domain of dispensed prescriptions in England was demonstrated in the prescriptions demonstrator. The demonstrator represents an integration between pharmaceutical and health management domains. The pharmaceutical data was in the form of the British National Formulary (BNF) and represented as the BNF ontology. The health management data represented as the prescriptions and NHS ontologies. All ontologies were built using OWL. The building process was an affordable task for a person with some technical background. However, designing ontologies that are not for demonstration purposes is a more complicated job that involves co-operation between computer scientists and domains experts.

Part of the data representation process is choosing ontologies for re-using, if needed. Around half of the reviewed literature involved re-used ontologies or part of them. In Hogan et al. (2016), around half of the used classes in their developed ontology are borrowed from other ontologies. The level of re-using can differ from one case to another. Sometimes, re-using the whole ontology is necessary, while at other times importing only certain parts of the ontology is enough. In designing the NHS ontology in the prescriptions demonstrator, some vocabularies from the Organization ontology (ORG) (World Wide Web Consortium, 2014) and the FOAF ontology (Brickley and Miller, 2010) were borrowed.

In the prescriptions demonstrator, the three developed ontologies were built from scratch with just few borrowed classes from other ontologies as there were no similar ontologies on the cloud that represent the same domains. All the available data was in the form of tabular data. The lack of existing ontologies needed for developing a system could be the reason behinds the high percentage rate of building new ontologies in the literature and relatively low percentage of re-using ontologies. The number of systems re-used a whole ontology in the literature were low, but more systems used parts of other ontologies or mapped their implemented classes with others in the cloud. By pointing to others work via mapping, the fourth LD principle is applied. There is still a big gap to be bridged here in adding more links to others work to fulfil the vision of the SW and LD. In this matter, the BNF ontology for example could be mapped to other drugs-related ontologies available on the cloud. Thus, it can be discovered by others and the concept of re-using and sharing data will be followed as one step forward into fulfilling the SW vision.

Other than representing the schema of a domain, modelling instances (the actual data) is another level of representation. Modelling instances in the SW approach follows the same triples concept as in defining the schema. One of the advantages of the triples' definition is the simplicity in applying it and in converting data to it. There are some available automatic tools for converting

tabular data into RDF. The availability and efficiency of such tools would encourage and support more adoption of the SW approach. Some examples of the mentioned converting tools in the literature are R2RML, spyder, virtuoso and Bio2RDF. Also in the literature, there were some papers that used manual techniques for converting the needed data into RDF form. Odgers and Dumontier (2015), for example, used PHP and Python scripts to transform relational EHRs into RDF.

In the prescriptions demonstrator case, two approaches for converting instances were followed. Firstly, OpenRefine (Google, 2010), an automatic tool to convert data to RDF, was used with most of the datasets. However, a manually scripted code was used for big data that OpenRefine failed to convert. The availability of automatic efficient converting tools is believed to encourage researchers to use the SW, because such a facility would ease the process of using the already available relational data. Marshall et al. (2012) said that mapping relational data into RDF is ideal with SW systems. Moreover, a co-operation between technical engineers and health experts is also recommended to overcome any challenges in representing the domain data.

7.2.2 Data Integration

The SW was used for integrating data in 84% of the reviewed literature. It was the second most-used feature; no surprise as the SW vision considers the web as a global interlinked data repository. However, the data integration meant here was done on a local level. In many cases this context is due to the personal and sensitive nature of the integrated health data. For example, Shaban-Nejad et al. (2016) integrated different sensitive hospital-related datasets locally to develop the integrated Hospital Acquired Infections Ontology (HAI).

From reviewing the literature, a small portion of projects which used SW tools for remote data integration were noted. The integration is performed via ‘federated queries’, where a SPARQL endpoint is invoked remotely in run time (World Wide Web Consortium (W3C), 2009). Pathak et al. (2012) used federated querying by invoking biomedical ontologies in the LOD such as Translational Medicine Ontology (TMO) and Sequence Ontology (SO). Using federated queries opens opportunities for exploring remote data resources on the fly. However, studying ‘federated querying’ is still an open research topic according to the World Wide Web Consortium, 2009) (W3C).

The SW vision introduced the idea of globally connected data; however, integrating data was used heavily in the literature but mostly at a local, rather than global, level. The last principle in the “Linked Data Principles” proposed by Berners-Lee (2006) encouraged SW users to include links and mapping to others’ work. According to the results of the systematic review, there was a

moderate reusing of others' ontologies, and most of the data integration happened at a local level. Thus, the fourth principle for defining links for others' work was not fully validated.

From a practical point of view, data integration locally and remotely were tested in the developed prescriptions demonstrator. Three OWL ontologies were built from seven open datasets (spreadsheets) namely: i) NHS ontology, ii) BNF ontology and iii) prescriptions ontology. These ontologies were integrated locally in one system by providing links between prescriptions and the prescribed BNF drugs, as well as between the prescription's issuing NHS practice and the prescription itself.

The linking process in the SW approach is performed via the triples concept. Meaning that by defining a relationship between two data objects, linking data is accomplished. This concept is simple to understand and use, which made integrating local data an achievable affordable task. However, some of the few obstacles that was faced in the process of linking the three ontologies were the availability of ready-to-use RDF data. As mentioned earlier, the available datasets were only in tabular form (spreadsheets). Thus, there was a need to build the ontologies and feed them with converted RDF instances before integrating together. In that process, some linking mismatching occurred due to the existence of misinput/dirty data in the original datasets.

In the second use case applied to the prescriptions demonstrator, a 'federated query' was run from the local SPARQL endpoint in the prescriptions demonstrator to retrieve remote data from Wikidata. Implementing the query was a relative success; however, the quality of the data was an obstacle to successfully retrieving sufficient amounts of information to analyse the use case.

Wikidata is a collaborative database edited by different people not necessarily experts in the data field. For this reason, the resulted linked data in wikidata is big and in various topics, but were edited by non-experts with no strict standards to be followed. In health projects that aim to deduce critical decisions, such collaborative knowledge bases are not recommended to be used as the integrated data cannot be trusted to be valid and complete. However, ontologies designed and developed by a collaboration between domain experts are better candidates for health research integration. For example, remotely integrating biomedical ontologies with trusted sources is a better approach for applying remote data integration using the SW technologies.

7.2.3 Knowledge Discovery

From systematically reviewing the literature, three approaches of discovering knowledge were identified: a) exploring linked data b) reasoning over rules and c) checking inconsistencies.

Exploring data by traversing links between connected data elements is performed via querying. Querying was mentioned 77% of the reviewed literature cases.

SPARQL queries were also tested in the prescriptions demonstrator and illustrated in the two applied health cases presented in chapter 6. In the two cases, health-related questions were translated into queries with multiple steps. Both queries successfully answered the addressed questions. SPARQL showed some interesting features in supporting the addressing of health questions as follows:

- a) exploring and filtering big amounts of data,
- b) performing data integration multiple times by using sub-queries,
- c) retrieving integrated data locally and remotely via 'federated querying',
- d) aggregating data and grouping solutions in a query,
- e) performing arithmetic functions on the retrieved results.

The other identified approaches from the literature relating to discovering knowledge by using the SW, depended on inferring new information from known facts. The W3C website states that inference is an automatic procedure to discover new relationships from known ones, and is based on the data and set of rules (World Wide Web Consortium (W3C), 2008). Inference was used in the literature for two main reasons: i) to find new information by depending on a set of defined rules (reasoning over rules) or ii) to check inconsistencies in ontologies. Reasoning over rules was used in the literature in 64% of the accessed sources, while checking for inconsistencies was featured in only 6% of the chosen literature sources.

From analysing the usage rates of the SW's features across health aims, there was a variance in using reasoning over the rules feature across the four health aims. Medical topics used the reasoning feature the most, then health management and pharmaceutical. Public health was the aim to use 'reasoning' the least. From this finding, it can be concluded that using reasoning over the rules in the literature was not exploited enough across the different health topics. There is an opportunity to fill these gaps with information obtained by more health research informed by inference-suitable questions, like those queries that are decision-based.

From the practice perspective, 'reasoning' and 'checking inconsistencies' were tested and used narrowly. The Pellet reasoner was used to check for inconsistencies in the three designed ontologies. The initial design for both of the NHS and BNF ontologies was by defining the classes in hierarchical manner such as a BNF chapter include a BNF section. However, this type of definition entails logically that all instances of a BNF sections must be also instances of the parent BNF chapter. This type of entailment is logically inconsistent and therefore the design of the ontologies was updated accordingly. Defining classes using OWL has the advantage of being based

on descriptive logic that can be used as mentioned above in finding inconsistencies and flaws in the ontology design.

Regarding reasoning over rules, there was no use for semantic rules in SWRL in this experiment. However, inference was performed just by testing the OWL object properties' conditions; like *inverse of* as explained earlier in the inference testing case in Chapter 5. The system succeeded in adding new knowledge to known knowledge at run time by analysing the known facts and defined rules and infer new fact base on the known inputs. This addition will help in discovering new information and possibly answering the addressed question.

7.2.4 Updating Knowledge

In the literature updating knowledge either manually or automatically showed low rates of usage. Updating knowledge was probably used by a bigger number of projects than was mentioned, but mentioning using it explicitly was neglected because updating knowledge is a secondary feature in most testing phases. However, the review highlighted using the option because it reflects the advantages of ease and flexibility in updating knowledge when using the SW.

In practice, the designs of both BNF and NHS ontologies were updated after discovering some inconsistencies related to the initial hierarchical design of both ontologies. The design was changed from defining the classes in hierarchical matter into adding relationships between the classes. Using SPARQL properties such as UPDATE and DELETE was simple and easy to use. The SW approach using SPARQL properties was flexible when it comes to adding and deleting facts from the knowledge base. The reason behinds this flexibility is because defining schema in the SW happens by defining a group of triples. Thus, the schema definition as well as the definitions of instances can be updated in the same manner.

7.2.5 Data Sharing

The literature review revealed two approaches to sharing data, doing so: a) publicly and b) privately. Only 3% of the reviewed papers shared their data publicly. The Apollo Structured Vocabulary (Apollo-SV) project (Hogan *et al.*, 2016), the Semantic Web Applications in Neuromedicine (SWAN) ontology (Cicarese, Wu, Wong, *et al.*, 2008) and the ToxBank Data Warehouse (Kohonen *et al.*, 2013) were examples of the few researchers who published their work publicly. On the other hand, around 50% of the reviewed projects used restricted data sharing for security reasons, especially when dealing with personal/patient records.

The second and third principles in the “Linked Data Principles” (Berners-Lee, 2006) were concerned with how to use standards and proper data representation to enable data sharing. The low usage rates revealed from the literature showed that the second and third LD principles were not fully followed.

In practice, data sharing was not used because the prescription demonstrator was built as a proof-of-concept model that was not intended to be shared publicly. The designs of the BNF, NHS and prescriptions ontologies are still in their early stages. Nevertheless, the designed ontologies, along with the converted instances, were shared with the social sciences, social data and the semantic web (S3W) project members. The S3W aims to understand the challenges and opportunities of semantic linked data for social science research. The project will analyse the conversion of sections of the English Longitudinal Survey of Ageing (ELSA) and the Great British Class Survey (GBCS) into semantic linked data for the purpose of a better understanding of health inequalities. The SW-based demonstrator to be built for addressing health inequalities questions will be enhanced with relevant semantic linked data, one of which is the developed BNF ontology in this thesis.

7.3 The Affordances and Challenges of Using the Semantic Web in Health Research

The main aim of this thesis is to understand the relationships between the SW and health research currently and potentially. The previous two sections discussed how the first and second research questions answered part of the research aim. This section discusses the learnt affordances and challenges from the literature and practice points of view in an attempt to answer the third research question:

What are the affordances and challenges in employing the semantic web for health research?

7.3.1 The Semantic Web’s Affordances for Health Research

Studying the SW from the literature and practice point of views highlighted many opportunities and affordances that can be offered to the health research community. The literature revealed three main affordance areas: a) data representation, b) data linking and c) knowledge discovery discussed in the following sections.

7.3.1.1 Representing various topics and questions

The SW showed a good ability in representing various health topics and addressing different types of health questions. The systematic review revealed 17 types of health questions that the SW was able to handle. Moreover, the prescriptions demonstrator represents a different domain that was not revealed by the literature, which is the domain of dispensed prescriptions in England. Three interlinked ontologies were built for purpose, which were the NHS, BNF and Prescriptions ontologies.

7.3.1.2 Simplicity and flexibility in representing data

Simplicity and flexibility are two of the advantages mentioned for using the SW's standards in representing data. Pathak *et al.* (2012) stated that RDF was simple to use for modelling data. The authors also stated that RDF model was more flexible when updating, adding or deleting data than was the relational model. In the prescriptions demonstrator, it was found that updating the ontologies design was flexible and simple by using SPARQL keywords UPDATE and DELETE. The flexibility in the SW approach is shown in adding or deleting triples in the same manner for instances or schema definitions.

7.3.1.3 Representing logical conditions

Besides the ability to represent a domain's schema, the SW showed an ability to represent conditions and rules in a domain. The SW is not just used for representing objects, and relationships between them, but it can offer more by representing rules and logical statements (Shadbolt, Hall and Berners-Lee, 2006). Using defined SW rules to model a domain was only moderately evident in the literature, being mostly used in the medical diagnostic papers.

7.3.1.4 Heterogenous data integration

In the literature, data integration was reportedly used to achieve data interoperability between heterogenous data sources. Puustjärvi and Puustjärvi (2009) aimed to allocate clinical resources between healthcare managers; a challenging task that demands cooperation between several heterogeneous information systems within a healthcare institution. Heterogenous integration was also achieved in the practical use cases. For example, in the second use case, Wikidata was integrated with the interlinked data from the three ontologies in the prescriptions demonstrator.

7.3.1.5 Local and remote data integration

Moreover, the SW was able to perform data linking locally and remotely. The majority of the reviewed projects integrated their data locally; however, the strength point of using the SW as a

linking tool is its ability to integrate data remotely. For example, Marshall *et al.* (2012) used 'federated querying' to request all language renderings of a specific drug product from three remote data sources: i) RxNorm, ii) DrugBank and iii) DBPedia. Both of local and remote integration were shown in the illustrative use cases. The first use case illustrated nicely how local data integration is achieved between the three ontologies: NHS, BNF and Prescriptions ontologies. The remote integration was also demonstrated by using federated query to Wikidata within the prescriptions demonstrator.

7.3.1.6 Incorporation between public and private datasets

One of the advantages mentioned in using the SW approach in linking data is the incorporation of the public and private datasets due to the integration between public data sources from the LOD cloud and private institute-specific patients' data (Pathak, Richard C. Kiefer and Chute, 2012). This advantage is essential in the health research, because it relies heavily on analysing personal-nature data such as patient data. Ten out of 13 questions suggested in the focus group depended mainly on integrating some private with public data, which gives an indication of the importance of providing efficient and safe approach to achieve the incorporation.

7.3.1.7 Flexibility in data linking

The flexibility in the linking process is another advantage in using the SW; a benefit achieved by using the SW standards for defining data concepts and any links between them. Linking in the SW approach can be performed between any related data concepts, not necessarily linking via personal identifiers. For example, one of the links in the second use case was between the coastal city's label in Wikidata in one side and practice's city property on the other side.

7.3.1.8 Exploring linked data

There were two identified approaches regarding how to discover knowledge using the SW: a) exploratory and b) inference. The SW has the ability to explore and traverse linked data locally and remotely via querying. For instance, BNF data was explored for all related BNF presentations of metformin hydrochloride in the first use case, while it was searched for all BNF presentations under the antidepressants section in the second use case.

7.3.1.9 Checking for logical inconsistencies

According to the literature, inference was used for two main reasons: i) to check inconsistencies and errors in ontologies or ii) to infer new information based on defined rules. For example, Horridge *et al.* (2014) used OWL reasoning as a quality assurance technique when checking for inconsistencies in the large medical ontology ICD-11. Part of the reason behind using the SW for

inconsistency checking was the SW's ability to trace the source items that caused the logical inconsistency, especially by using reasoners like Pellet (Huang *et al.*, 2014). In the prescriptions demonstrator, Pellet revealed some logical inconsistencies in the early stages of the ontologies design. The BNF and NHS ontologies were designed in a hierarchical manner that caused inconsistencies in the results. Then, the design was updated into defining classes at the same level with defined object properties between them.

7.3.1.10 Inferring new information

According to Wang *et al.* (2015), the semantic reasoning ability helped to improve the intelligence of the independent clinical pathway system the team had developed. Baldassini *et al.* (2017) stated that the semantic reasoning tools allowed ontologies to infer new information according to the knowledge model. In their project they were able to classify each user's cardio-respiratory fitness condition according to environmental and physiological data collected by monitoring devices, some of which were worn by the patients. Inference was slightly tested in the prescriptions demonstrator by testing inferring the inverse of certain properties at run time.

7.3.2 More Affordances

From a practical point of view, extra affordances were analysed in the process of experiencing the use of the SW in a health case scenario. All identified affordances from the literature were noted in the process of developing the prescriptions demonstrator and addressing health questions to it. The exception was when semantic rules were not needed in the chosen health cases. By using SPARQL as a querying language for addressing the chosen health questions, multiple abilities and features were highlighted for using this tool.

7.3.2.1 Ability to handle big data

The main noted affordance was the efficiency in handling big data. The prescriptions dataset included more than 18 million records; a total that can be challenging to other data management tools. The SW was able to represent, integrate, explore and retrieve this big data with relatively few technical issues such as querying time out or limitations in graphical displaying.

7.3.2.2 Ability to integrate multiple data at a time

SPARQL had the ability to perform multiple data integrations in a single query by using sub-queries. In both case studies, sub-queries were implemented to across different datasets to perform complex data integration and results sets. Beyond that, in the second use case the data integration level was more complex by implementing federated query remotely within the main

query locally. By integrating data remotely from the cloud, the SW showed an advantage over other traditional linking data techniques. This advantage helps in utilising published data by re-using it.

7.3.2.3 Aggregating data based on certain condition

SPARQL also showed an interesting useful feature in grouping and aggregating the retrieved solutions and applying expressions over the grouped solutions. In the first use case, aggregation was performed on the practice level, while it was done on the practice's city in the second case study. By grouping the solutions in both cases, certain arithmetic calculations were able to be performed on the produced groups.

7.3.2.4 Supporting arithmetic functions

Finally, SPARQL supports a range of useful built-in arithmetic functions that are helpful when involved with calculations. As mentioned earlier in the previous point, arithmetic calculations such as division and summing were performed on both use cases to calculate the cost-per-person rate as well as the antidepressants prescribing rate.

7.3.3 Challenges Facing Employing the Semantic Web in Health Research

Besides the found affordances of using the SW in health research, several challenges were identified as a result of reviewing the literature.

7.3.3.1 Data accessibility

One of the challenges is the difficulty of gaining access to patients' personal and or private data which is needed in order to conduct a research initiative. The projects or initiatives in the reviewed literature used personal private data that were owned by the same organisations which performed the research. The data can contain information about patients, doctors, health situation and financial information that is usually shared within the institution and viewed by access-authorised people only.

It would appear evident that 'unauthorised' researchers are seriously disadvantaged in such a situation. As a solution to this issue, traditional health research supports independent data centres that play a crucial role in providing researchers with anonymous personal linked data which is ready to use for research purposes. A similar solution could be implemented with the SW

approach as well, by using SW standards in order to produce semantically de-identified personal linked data.

The SW and LOD offer public datasets that contain information in different fields such as Wikidata or DBpedia. Such datasets or ontologies are easy to access by anyone and can offer useful information for the targeted research question, however, this is not enough for most cases in health research. In the conducted focus group, most of the suggested questions by the health experts relied mainly on accessing patients and doctors' data besides other open datasets. This case indicates that private restricted-access datasets such as patients' records are major data source in health research, but many ethical and security issues can accompany using these datasets especially when dealing with identified data. Thus, finding a solution for this 'no-access' challenge is a priority taking into consideration any accompanying ethical and security challenges.

7.3.3.2 Data security

Private data is usually protected by security protocols and policies. The SW approach is based on the concept of linking data, hence, security policies can hinder accessing the data. At the same time, private and personal data needs to be protected in an open linked systems. In the literature review, there were different uses of private data such as in systems built upon patients' records in hospitals or web services aggregating personal health data for registered users. Balancing between allowing accessing such private data and protecting it against data thefts is a challenge to be faced when attempting to employ the SW in health research. As mentioned by Eysenbach (2003), the SW offers opportunities as well as challenges in regards to consuming private data. For this reason, SW community should encourage more research in the area of security access protocols and mechanisms.

7.3.3.3 Data quality

Data quality is another challenge facing any general health researcher. The accuracy of the data extracted from medical systems is particularly crucial. If a fault decision was made due to the quality of the available data, the consequences could be disastrous. For example, there were many health projects in the reviewed literature aiming to build decision-based systems for diagnostic purposes or to facilitate choosing the best treatment plan for a patient. Any misleading results in such systems could endanger the patient's health by suggesting unsuitable diagnosis or treatments.

In practice, issues regarding data quality were also faced. In the second use case, there were many missing data in the definition of English coastal cities or towns within Wikidata, which led to an insufficient number of retrieved records. Also, most of the population-related data in Wikidata

was outdated, as it was based on the 2011 national census. Knowledge bases such as Wikidata which is edited collaboratively by not necessarily experts are not the most suitable data source for critical and sensitive health research. Trusting the quality of the data and the data source is a crucial factor in choosing the most suitable dataset to be used in a project.

Another issue was noticed regarding the data that was used in the NHS open datasets. In the practices addresses datasheet, there were inconsistencies in registering the data which led to incorrect converting from CSV to RDF; an error which, in turn, led to mismatched linking during the query's run time. Two noticeable human errors relate to the issue of errors in data entry: a) spelling mistakes or b) inconsistencies in spelling keywords; with both issues negatively affecting data quality.

To achieve a better quality for the published data, recommendations and guidelines for inputting, defining and sharing data should be formed and applied. Following standards in the data entry process will also help in overcoming the data quality issue. A co-operation between technicians and domain experts bridge the gap in understanding the domain data and the best way to represent it aiming for a better quality of the data.

7.3.3.4 Data heterogeneity

Data heterogeneity is also a challenge that was evident in most of the reviewed literature. Different types of data were used in the literature such as i) patients' monitoring sensors, ii) medical examination devices, iii) hospitals administrative data, and iv) patients' records. Zenuni *et al.*, (2015) stated that although the quantity of health data on the web is constantly growing, the available data mostly is in a non-semantic format. This statement proves the still existing gap of having the data on the web in a semantic form.

The practical trial comes in line with the literature findings. In practice, all used data in building the prescriptions demonstrator was in CSV format; no RDF data was available. Another noted form of data heterogeneity is in the diversity of available data sources even for the same datasets. For example, seven datasets were used from three different NHS departments, the latter being: i) NHS Digital, ii) NHS BSA and iii) NHS Choices. Some of these datasets were even replicated but with different naming systems.

To overcome the data heterogeneity challenges, data standards in registering, representing and publishing data need to be prescribed and then followed. Co-operation between computer scientists and health experts is essential in order to standardise health data on the web. The data needs to be scientifically accurate as well as technically well-represented.

7.3.4 More Challenges

From a practical point of view, more challenges were identified during the process of building the prescriptions demonstrator and addressing the two health uses cases. The identified challenges from the literature were noted also in practice with an exception of the securing private data challenge as there was no use to any private data in the prescriptions demonstrator trial.

7.3.4.1 Data availability

In the use cases, two issues were faced in regarding with the data availability. Firstly, there was a lack in available ontologies and RDF files for the BNF, NHS and prescriptions topics. Thus, there was a need to build ontologies for these topics with their RDF instances. Several available CSV spreadsheets were converted into RDF files to build up the prescriptions demonstrator. The lack of semantic data on the web hinder re-using ontologies in newer systems. More public sharing and publishing ontologies and RDF data is recommended for the success of the SW vision.

The second issue faced was the incompleteness of the RDF data. Although wikidata has more than 90 million data items (The Wikimedia Foundation, 2019), there was a severe lack in the related information to the English coastal cities. Collaborative open knowledge bases such as wikidata depend on individuals' efforts in uploading and updating the data. Thus, there is no guarantee for the availability nor the quality of the provided data. Moreover, the data needed in this case was more of a geographical/map type of data which made mapping it to RDF a nontrivial task. The more structural the data is, the easier the process of mapping to RDF will be. Although the task of determining a coastal city from a map is considered an easy task for humans, it is not that straightforward task for a machine. A programming algorithm is needed to define a coastal city for a machine. The borders of the city need to be determined and checked if they are shared with a marine body of water or not. Developing such programs and tools to handle nontrivial RDF converting will help in supporting publishing more linked data on the web. The SW is based on the principle of collaboration and data sharing; thus, the efficiency of the SW approach is compromised when there is a lack of published data on the cloud.

7.3.4.2 Complexity of building ontologies

In the prescriptions demonstrator, an initial prototype for the BNF, NHS and prescriptions ontologies was provided. Designing an ontology to represent a specific domain is a complex task that requires analysing the syntax and semantics of the design. The designing process requires human creativity in picturing a proper representation of a domain. Domain experts must be involved in the designing process as they are the most qualified people that understand the

requires of their field. It is recommended that computer scientist along with experts in the investigated domain cooperate to achieve the best possible results.

7.3.4.3 Lack of efficient user-friendly tools

The final identified challenge in employing the SW in health research is the necessity to have a certain amount of technical knowledge in using the SW approach. A knowledge in employing the main SW standards such as RDF, OWL and SPARQL is needed if one wishes to: a) build ontologies, b) convert instances and c) retrieve data via queries. For example, when an issue was faced in converting big amounts of data to RDF via the automatic tool, a manual script was coded to solve this technical issue. There is a lack of the number of available efficient user-friendly tools for accessing, representing, retrieving and viewing semantic health resources. For instance, the used tool in GraphDB for generating graphs for the uploaded linked data was not efficient in displaying big amount of data at a time. The availability of more easy -to-use and efficient supporting tools would encourage further adoption of the SW approach in health research.

Chapter 8 Conclusion

This chapter concludes the thesis by answering the three addressed questions at the beginning of the work and listing the main contributions, recommendations and future work in the field of employing SW technologies in health research.

The aim of this thesis was to understand the usages, affordances and challenges facing applying SW technologies in health research. Three research questions were addressed regarding: i) the types of health questions evident in the literature, ii) the uses of the SW in the literature, and iii) the identified affordances and challenges of applying the SW in health research. The author of this thesis believes that the three research questions were answered through different stages of this work.

In figure 81, a summary of the research questions and answers is illustrated. More details can be found in the next sections that discuss the contributions, recommendations and future work in this thesis.

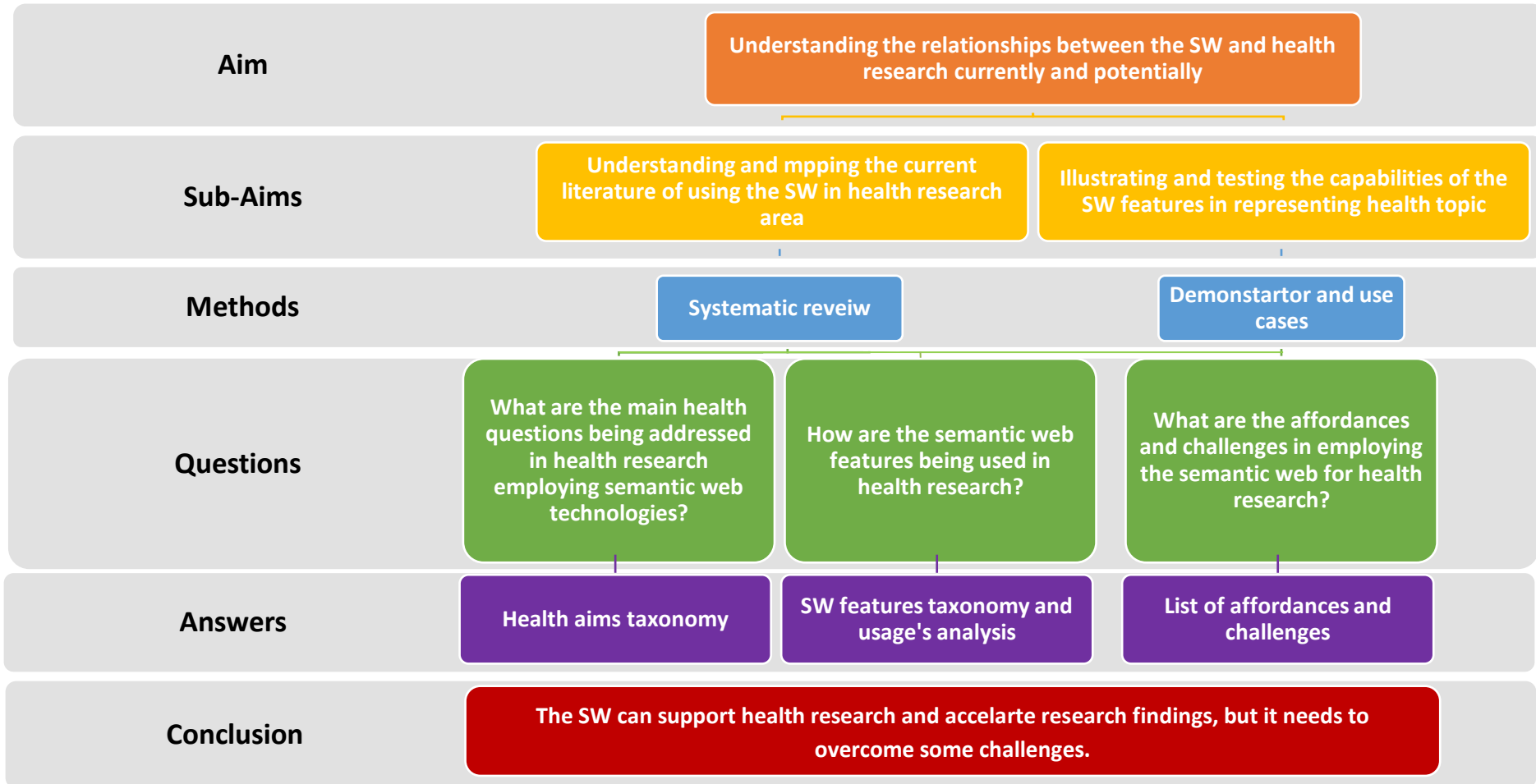


Figure 81: The thesis's summary figure

8.1 Contributions

The contributions of this work were achieved by answering the four research questions. Five contributions were achieved: 1) health aims taxonomy 2) SW features taxonomy 3) the prescriptions demonstrator, 4) a list of SW affordances and 5) a list of SW challenges.

8.1.1 Health Aims Taxonomy

The first research question asking about the main addressed health questions in the joint literature of the SW and health research. This question was answered in chapter 4 by systematically reviewing the literature. One of the main systematic review results was the health aims taxonomy that mapped the addressed health questions in the literature. The health aims taxonomy included 17 addressed health questions divided into four main aims: medical, public health, health management and pharmaceutical questions.

By analysing the relationships between the addressed health question and the used SW features, four characteristics of the addressed questions were deduced, which were 1) decision-based questions 2) personalisation questions 3) questions based on integrating heterogeneous data and 4) questions rely on accessing web information. This shows that the SW is able to represent various health topics and address different types of questions. Another found point in regards to the decision-based questions like diagnostic ones, there was a positive correlation between them and using reasoning over rules to discover new knowledge. This finding implies that that this technology has a potential in supporting health research especially in decision-based and diagnostic projects, which took a big share of the reviewed literature.

8.1.2 Semantic Web Features Taxonomy

The second research questions in this thesis concerned about how the SW features were used in health research. The answer of this question was in the systematic review in chapter 4. The systematic review revealed 12 SW features categorised under five main ones in the SW features taxonomy. The most used features according to the literature were building new ontology, integrating data locally and exploring data via querying.

By analysing the usage rate of the SW features across the health aims, research trends and gaps can be identified. For example, there was a trend in high usage rate of building ontologies and integrating data locally across all health aims. There was also a positive relationship between using reasoning over rules and medical aimed questions. Some of the noticed gaps were in the lack of re-using ontologies in the public health aim as well as a lack in sharing data in

pharmaceutical projects. Generally, there was a shortage in checking inconsistencies, updating knowledge manually and automatically, integrating data remotely and sharing data publicly by all health aims.

8.1.3 The Prescriptions Demonstrator

The aim of building the proof-of-concept prescriptions demonstrator was to show how the SW features can be used in representing and addressing health topic in order to analyse any affordances or challenges in the process. The demonstrator focused on testing representing, integrating and exploring prescriptions data. Three interlinked OWL ontologies with their accompanying RDF instances were developed namely the NHS, BNF and prescriptions ontologies. The main source for designing the ontologies was open datasheets available in the web. The used method for developing the system was the Extract, Transform and Load (ETL) process. GraphDB was the used triplestore that includes a built-in querying interface for retrieving and testing the system.

In addition to the prescriptions demonstrator, two health use cases were addressed using it. The uses cases meant to represent and address interesting health questions by translating them into SPARQL queries. The first use case was inspired from the literature and discussed the problem of prescribing inequalities for diabetic medications. The second use case was suggested by health expert and addressed the question of the effect of living in a coastal city in prescribing antidepressants. Both use cases were successfully applied to the prescriptions demonstrator, however, there was insufficient amount of retrieved data in the second one. The quality of the available data in the cloud was an obstacle in retrieving the needed results. There was many missing data as well as some outdated data.

8.1.4 List of the Affordances of Employing the Semantic Web in Health Research

The third research question concerned on addressing SW affordances and faced challenges in health research. A list of affordances was deduced from systematically reviewing the literature. In addition, more affordances were found in the process of developing the prescriptions demonstrator and addressing health questions to it.

The found affordances in the literature and practice were in the areas of: a) data representation, b) data linking and c) knowledge discovery. Some examples of the data representation affordances are: 1) the ability to represent various topics and questions, 2) simplicity and flexibility and 3) the ability to represent logical conditions. In regard to the affordances in data linking, the SW proved its ability in: 1) linking heterogenous data, 2) linking local and remote data,

3) linking public and private data and 4) flexibility in linking data. The third area where the SW proved its usefulness is in discovering knowledge. The SW was used to: 1) explore linked data, check for logical inconsistencies and finally 3) infer new information via using rules.

In developing the prescriptions demonstrator, a couple of affordances were noticed in addition to the identified ones from the literature. The SW show an ability to : 1) handle big data, 2) perform multiple integrations at a time, 3) aggregate data based on certain condition and 4) support arithmetic calculations.

8.1.5 List of the Faced Challenges of Employing the Semantic Web in Health Research

Continuing answering the third research question, a similar list to the affordances one was produced. A list of challenges facing employing the SW in health research was identified from a literature and practical perspectives. four challenges were identified from reviewing the literature: i) data accessibility, ii) data security, iii) data quality and iv) data heterogeneity. More challenges were identified in the process of building the prescriptions demonstrator, which were: 1) data availability, 2) complexity in developing ontologies and 3) lack of efficient user-friendly SW tools.

8.2 Recommendations

The SW community needs to benefit from the experts' experiences in dealing and managing health data to get the most of it in developing the SW. For example, health informatics researchers' expertise is invaluable in building ontologies, information systems and web services. Co-operation between computer scientists and domain experts is recommended to standardise health data on the web. The cooperation between different disciplinary is an essential step to the SW success.

Another example is in including some successful traditional solutions into the SW experiment. The traditional health linking approach used independent data centres for preparing and producing ready-to-use personal linked data for research purposes. enabling independent data centres to prepare and anonymise personal data should thereby overcome challenges related to data sensitivity and access. Such approach would encourage more health researchers to use the SW approach in their studies.

The SW community should support security-related research for developing more protocols in dealing with data sharing using SW technologies. Finally, to encourage further employing for the SW approach in health community, easy-to-use SW interfaces and tools should be provided to be used with health data resources.

8.3 Future Work

This section describes some proposed research directions in the field of employing SW technologies in health research.

8.3.1 Reasoning over rules in health topics addressing decision-based questions

Based on the findings of the systematic review, there is a correlation between studies addressing decision-based questions such as diagnostic questions and employing semantic rules to reason new information. However, there is a lack in using logic in the form of reasoning over rules in other types of studies. Knowing that the logic layer is one of the high-level layers in the SW layered stack, more research should be aimed into studying the possibilities of using semantic rules into different health topics.

8.3.2 Standardising the process of publishing linked data on the web

One of the main found challenges in the process of employing SW technologies in health research is the quality of the used data. Although there are some available standards for publishing data on the web, the review revealed that the standards are not followed in all the cases. Thus, more research should focus on how to improve the quality of the web data in any form.

8.3.3 Improving user-friendly tools for accessing linked data

In order to utilise from the available health linked data on the cloud, a proper technical knowledge is needed to handle exploring and retrieving knowledge from ontologies. To encourage non-technical researchers to use linked data at most, more user-friendly interfaces and querying tools should be produced.

8.3.4 Supporting security-related studies especially in the case of incorporation between private and public data

The vision of the SW is based on the concept of linking data on the web. While this vision is very ambitious, many challenges arise with it. The security of the integrated data is one of these challenges especially when dealing with sensitive health-related data. The data security community should focus on providing more solutions for the problem of integrating public and private datasets together.

8.3.5 Improving federated querying technology

The SW research community should focus more on improving SPARQL federated querying technology as it showed to be very handful in integrating data remotely. In the reviewed literature, few health studies used the federated querying feature, while it proved its usefulness in the practical example in the prescriptions demonstrator second use case. Research focus on the limitations, scope, challenges and opportunities for this feature in health research.

8.4 Ending Statement

It is the author's belief that the SW can support health research and accelerate research findings by representing various health topics and questions, integrating local and remote as well as private and public data sources and finally by discovering new knowledge via exploring linked data and inferring new information.

The SW has a potential in supporting health research, however, it is not yet ready to offer needed data for different types of health research. The data on the web is still suffering from various issues that considered to be obstacles in the face of conducting any relevant research. Some of the related challenges are concerned in the accessibility, security, quality, heterogeneity, availability and complexity of the used data.

More policies and protocols should be implemented by the SW community in managing linked data as well as supporting academia in finding further solutions in securing and sharing data on the web. A co-operation between computer scientists and domain experts is highly recommended to achieve a better understanding of the faced issues and produce scientifically accurate and technically well-formed data. Enabling independent data centres to produce ready-to-use linked semantic data as in the traditional approach would encourage more engagement from health researchers into adopting the SW approach.

Bibliography

Agafonkin, V. (2017) *Leaflet*. Available at: <http://leafletjs.com/> (Accessed: 16 November 2017).

Al-Nazer, A., Albukhitan, S. and Helmy, T. (2016) 'Cross-Domain Semantic Web Model for Understanding Multilingual Natural Language Queries: English/Arabic Health/Food Domain Use Case', in *Procedia Computer Science*. Elsevier Masson SAS, pp. 607–614. doi: 10.1016/j.procs.2016.04.138.

Alexandrou, D. A. *et al.* (2012) 'SEMPATH ontology: Modeling multidisciplinary treatment schemes utilizing semantics', *IEEE Transactions on Information Technology in Biomedicine*, 16(2), pp. 235–240. doi: 10.1109/TITB.2011.2161588.

Alexandrou, D., Xenikoudakis, F. and Mentzas, G. (2008) 'Adaptive clinical pathways with semantic Web rules', in *1st International Conference on Health Informatics, HEALTHINF 2008*. Funchal, Madeira, pp. 140–147.

Antoniou, G. and Harmelen, F. Van (2008) *A Semantic Web Primer*. 2nd edn. MIT Press.

Arikuma, T. *et al.* (2007) 'Ontology-Driven Hypothetic Assertion (OHA) for drug interaction prediction', in *Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007)*. IEEE, pp. 1–8. doi: 10.1109/IMSCCS.2007.4392573.

Artz, D. and Gil, Y. (2007) 'A survey of trust in computer science and the Semantic Web', *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2), pp. 58–71. doi: 10.1016/j.websem.2007.03.002.

Bajenaru, L. and Smeureanu, I. (2015) 'An ontology based approach for modeling e-learning in healthcare human resource management', *Economic Computation and Economic Cybernetics Studies and Research*. Academy of Economic Studies, 49(1), pp. 1–17. Available at: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84944587705&partnerID=40&md5=cb48d569d12dd02bd11f78468bb0b812>.

Baker, C. (2016) *Medicine statistics: GP prescribing by constituency, 2015*. Available at: <file:///S:/Downloads/CBP-7161.pdf>.

Baldassini, D. *et al.* (2017) 'Customization of domestic environment and physical training supported by virtual reality and semantic technologies: A use-case', in *2017 IEEE 3rd International Forum on Research and Technologies for Society and Industry (RTSI)*. IEEE, pp. 1–6.

Benson, T. (2012) *Principles of Health Interoperability HL7 and SNOMED (Health Informatics)*. 2nd

editio. Springer Science & Business Media. doi: 10.1007/978-1-84882-803-2.

Berners-Lee, T. (2006) 'Linked Data - Design Issues'. Available at: <http://www.w3.org/>.

Berners-Lee, T., Fielding, R. and Masinter, L. (2005) 'Uniform Resource Identifier (URI): Generic Syntax Status', *IETF RFP 3986 (standards track), Internet Eng. Task Force*.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) 'The Semantic Web', *Scientific American*, 284(5), pp. 34–43. doi: 10.1038/scientificamerican0501-34.

Beyan, O. *et al.* (2014) 'Towards Linked Vital Registration Data for Reconstituting Families and Creating Longitudinal Health Histories', *Knowledge Representation for Health Care KR4HC*, (April 2016). doi: 10.13140/2.1.4013.1206.

Birjali, M., Beni-Hssane, A. and Erritali, M. (2017) 'Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks', in *Procedia Computer Science*. Elsevier B.V., pp. 65–72. doi: 10.1016/j.procs.2017.08.290.

Bizer, C., Heath, T. and Berners-Lee, T. (2009) 'Linked Data - The Story So Far', *International Journal on Semantic Web and Information Systems (Ijswis)*, 5(3).

Bray, T. *et al.* (2008) *Extensible Markup Language (XML) 1.0 (Fifth Edition)*, W3C. Available at: <http://www.w3.org/TR/REC-xml/> (Accessed: 17 May 2017).

Brickley, D. and Guha, R. V. (2014) *RDF Schema 1.1*, World Wide Web Consortium (W3C). Available at: <http://www.w3.org/TR/rdf-schema/> (Accessed: 20 September 2016).

Brickley, D. and Miller, L. (2010) *FOAF Vocabulary Specification, Namespace Document*. Available at: <http://xmlns.com/foaf/spec/> (Accessed: 1 June 2017).

British Medical Association, Pharmaceutical Society of Great Britain and Joint Formulary Committee (Great Britain) (2016) *BNF. 72, September 2016 - March 2017*. 72nd edn. London: BMJ Publishing Group Ltd and RPS Publishing.

British National Formulary (BNF) Publications (2017) *About – BNF Publications*. Available at: <https://www.bnf.org/about/> (Accessed: 17 October 2017).

Buckley, E. S. *et al.* (2016) 'The utility of linked cancer registry and health administration data for describing system-wide outcomes and research: A BreastScreen example', *Journal of Evaluation in Clinical Practice*, 22, pp. 755–760. doi: 10.1111/jep.12536.

Bukhari, A. C. and Baker, C. J. O. (2013) 'The Canadian health census as Linked Open Data :

towards policy making in public health', in *9th International Conference on Data Integration in the Life Sciences*, pp. 7–10.

Calvillo-Arbizu, J. *et al.* (2014) 'Design of a clinical decision support system for assisting in empiric antibiotic treatments', in Lackovic, I. *et al.* (eds) *International Conference on Health Informatics, ICHI 2013*. Springer Verlag, pp. 304–307. doi: 10.1007/978-3-319-03005-0_77.

Cancer Council Australia (2018) 'Understanding Clinical Trials and Research', p. 68. Available at: <https://www.cancercouncil.com.au/wp-content/uploads/2020/04/UC-pub-Clinical-Trials-CAN750-web-low-res-July-2018.pdf>.

Chen, Rung Ching *et al.* (2012) 'A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection', *Expert Systems with Applications*. Elsevier Ltd, 39(4), pp. 3995–4006. doi: 10.1016/j.eswa.2011.09.061.

Chen, R C *et al.* (2012) 'A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection', *Expert Systems with Applications*, 39(4), pp. 3995–4006. doi: 10.1016/j.eswa.2011.09.061.

Cheung, K.-H. *et al.* (2009) 'Semantic Web for Health Care and Life Sciences: a review of the state of the art.', *Briefings in bioinformatics*, 10(2), pp. 111–113. doi: 10.1093/bib/bbp015.

Chondrogiannis, E. *et al.* (2017) 'A novel semantic representation for eligibility criteria in clinical trials', *Journal of Biomedical Informatics*. Academic Press Inc., 69, pp. 10–23. doi: 10.1016/j.jbi.2017.03.013.

Chondrogiannis, E. *et al.* (2018) 'Dynamic service detection for automated patient selection for study recruitment purposes', *Proceedings - IEEE 34th International Conference on Data Engineering Workshops, ICDEW 2018*. IEEE, pp. 72–77. doi: 10.1109/ICDEW.2018.00019.

Christen, P. and Churches, T. (2006) 'Secure health data linkage and geocoding: current approaches and research directions', *National e-Health Privacy and Security Symposium*. Available at: <http://cs.anu.edu.au/~Peter.Christen/publications/ehPass2006.pdf>.

Ciccarese, P., Wu, E., Wong, G., *et al.* (2008) 'The SWAN biomedical discourse ontology', *Journal of Biomedical Informatics*. doi: 10.1016/j.jbi.2008.04.010.

Ciccarese, P., Wu, E., Kinoshita, J., *et al.* (2008) 'The SWAN Scientific Discourse Ontology', *Journal of Biomedical Informatics*, 41(5), pp. 739–751. doi: 10.1016/j.jbi.2008.04.010.

Cipiere, S. *et al.* (2014) 'Global Initiative for Sentinel e-Health Network on Grid (GINSENG):

Medical Data Integration and Semantic Developments for Epidemiology', in *2014 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*. IEEE, pp. 755–763. doi: 10.1109/CCGrid.2014.45.

Clark, R. A. *et al.* (2016) 'Heart Failure Following Cancer Treatment Characteristics, Survival and Mortality of a Linked Health Data Analysis', *Internal Medicine Journal*.

Cornish, R. *et al.* (2015) 'Using linked educational attainment data to reduce bias due to missing outcome data in estimates of the association between the duration of breastfeeding and IQ at 15 years', *International Journal of Epidemiology*, 44(3), pp. 937–945. doi: 10.1093/ije/dyv035.

Cruz, A. La, Espinoza, M. and Vidal, M. (2015) 'RDF-ization of DICOM Medical Images towards Linked Health Data Cloud', in *VI Latin American Congress on Biomedical Engineering CLAIB 2014*. Paraná: Springer International Publishing.

DailyMed (no date) *DailyMed, U.S. National Library of Medicine*. Available at: <https://dailymed.nlm.nih.gov/dailymed/index.cfm> (Accessed: 14 December 2016).

Dalwadi, N., Nagar, B. and Makwana, A. (2012) 'SEMANTIC WEB AND COMPARATIVE ANALYSIS OF INFERENCE ENGINES', *Int. J. of Computer Science and Information Technologies.*, 3(3), pp. 3843–3847. Available at: <http://www.cs.citc.edu/~Neha>, (Accessed: 23 May 2019).

Dang, J. *et al.* (2008) 'An ontological knowledge framework for adaptive medical workflow', *Journal of Biomedical Informatics*, 41(5), pp. 829–836. doi: 10.1016/j.jbi.2008.05.012.

Dang, J. *et al.* (2009) 'Personalized medical workflow through semantic business process management', in *ICEIS 2009 - 11th International Conference on Enterprise Information Systems*. Milan, pp. 122–127. Available at: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-74549129476&partnerID=40&md5=87c79d4f913e406ecb08e78a93686c80>.

Dempsey, S. *et al.* (2018) 'Coastal blue space and depression in older adults', *Health & Place*. Pergamon, 54, pp. 110–117. doi: 10.1016/J.HEALTHPLACE.2018.09.002.

Denaxas, S. C. *et al.* (2012) 'Data resource profile: Cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER)', *International Journal of Epidemiology*. doi: 10.1093/ije/dys188.

DuCharme, B. (2011) *Learning SPARQL*. 1st edn. O'Reilly Media.

Dumontier, M. *et al.* (2010) 'The Translational Medicine Ontology : Driving personalized medicine by bridging the gap from bedside to bench', *Proceedings of the 13th ISMB'2010 SIG meeting 'Bio-*

ontologies', pp. 120–123.

EBM DataLab - University of Oxford (2017) *OpenPrescribing.net*. Available at: <https://openprescribing.net/> (Accessed: 17 October 2017).

Eilbeck, K. *et al.* (2005) 'The Sequence Ontology: a tool for the unification of genome annotations.', *Genome biology*, 6(5), p. R44. doi: 10.1186/gb-2005-6-5-r44.

Ellis, R. P. *et al.* (2013) 'Explaining Health Care Expenditure Variation: Large-Sample Evidence Using Linked Survey And Health Administrative Data', *Health economics*, 22, pp. 1093–1110. doi: 10.1002/hec.

Eysenbach, G. (2003) 'The Semantic Web and healthcare consumers: a new challenge and opportunity on the horizon?', *International Journal of Healthcare Technology and Management*, 5(3), pp. 194–212. doi: 10.1504/ijhtm.2003.004165.

Falster, K. *et al.* (2015) 'What factors contribute to positive early childhood health and development in Australian Aboriginal children? Protocol for a population-based cohort study using linked administrative data (The Seeding Success Study).', *BMJ open*, 5(5), p. e007898. doi: 10.1136/bmjopen-2015-007898.

Falster, M. O., Jorm, L. R. and Leyland, A. H. (2016) 'Visualising linked health data to explore health events around preventable hospitalisations in NSW Australia', *BMJ Open*, 6(9). doi: 10.1136/bmjopen-2016-012031.

Fareedi, A. A. and Hassan, S. (2014) 'The impact of social media networks on healthcare process knowledge management (using of semantic web platforms)', in *The 14th International Conference on Control, Automation and Systems (ICCAS 2014)*, pp. 1514–1519. doi: 10.1109/ICCAS.2014.6987802.

Farinelli, F., Barcellos De Almeida, M. and Linhares De Souza, Y. (2014) 'Linked Health Data: How linked data can help provide better health decisions', *Studies in Health Technology and Informatics*, 216, p. 1122. doi: 10.3233/978-1-61499-564-7-1122.

Fernández-Breis, Jesualdo Tomás *et al.* (2013) 'Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts.', *Journal of the American Medical Informatics Association : JAMIA*, 20, pp. 288–296. doi: 10.1136/amiajnl-2013-001923.

Fernández-Breis, J T *et al.* (2013) 'Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts', *Journal of the American Medical Informatics Association*, 20(E2), pp. e288–e296. doi: 10.1136/amiajnl-2013-001923.

Fotopoulou, E. *et al.* (2016) 'Linked Data Analytics in Interdisciplinary Studies: The Health Impact of Air Pollution in Urban Areas', *IEEE Transaction and Content Mining*, 4, pp. 149–164. doi: 10.1109/ACCESS.2015.2513439.

Gandon, F. and Schreiber, G. (2014) *RDF 1_1 XML Syntax*, *World Wide Web Consortium (W3C)*. Available at: <http://www.w3.org/TR/rdf-syntax-grammar/> (Accessed: 15 May 2017).

Goh, K. *et al.* (2007) 'The human disease network', *Proceeding of the National Academy of Sciences of the United States of America (PNAS)*, 104(21), pp. 8685–8690. doi: 10.1073/pnas.0701361104.

Google (2010) 'Google Refine'. Available at: <http://openrefine.org/>.

Gruber, T. R. (1995) 'Toward Principles for the Design of Ontologies', *International journal of human-computer studies*, 43(5), pp. 907–928.

Gudivada, R. C. *et al.* (2008) 'Identifying disease-causal genes using Semantic Web-based representation of integrated genomic and phenomic knowledge', *Journal of Biomedical Informatics*, 41(5), pp. 717–729. doi: 10.1016/j.jbi.2008.07.004.

Hanna, J. *et al.* (2013) 'Building a drug ontology based on RxNorm and other sources', *Journal of Biomedical Semantics*, 4(1). doi: 10.1186/2041-1480-4-44.

Harzing, A. W. and Alakangas, S. (2016) 'Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison', *Scientometrics*, 106(2), pp. 787–804. doi: 10.1007/s11192-015-1798-9.

Health & Social Care Information Centre (HSCIC) (2015) 'General Practice Prescribing Data (Presentation Level) Glossary of Terms', pp. 1–7. Available at: http://digital.nhs.uk/media/10686/Download-glossary-of-terms-for-GP-prescribing---presentation-level/pdf/PLP_Presentation_Level_Glossary_April_2015.pdf.

Heath, T. and Bizer, C. (2011) *Linked Data: Evolving the Web Into a Global Data Space*. Morgan & Claypool Publishers. Available at: <https://www.w3.org/standards/semanticweb/data>.

Helfin, J. (2004) *OWL Web Ontology Language Use Cases and Requirements*, *World Wide Web Consortium (W3C)*. Available at: <http://www.w3.org/TR/webont-req/> (Accessed: 13 May 2017).

Hilder, L. *et al.* (2016) 'Preparing linked population data for research: cohort study of prisoner perinatal health outcomes.', *BMC medical research methodology*. BMC Medical Research Methodology, 16(1). doi: 10.1186/s12874-016-0174-7.

- Hogan, W. R. *et al.* (2016) 'The Apollo Structured Vocabulary: an OWL2 ontology of phenomena in infectious disease epidemiology and population biology for use in epidemic simulation', *Journal of biomedical semantics*, 7(1), p. 50. doi: 10.1186/s13326-016-0092-y.
- HorrIDGE, M. *et al.* (2014) 'Reasoning based quality assurance of medical ontologies: a case study', *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2014, pp. 671–680. Available at: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84964315996&partnerID=40&md5=82dbba10ce33cfe8dfe984c969489a27>.
- Hu, Z. *et al.* (2012) 'Ontology-based clinical pathways with semantic rules', *Journal of Medical Systems*, 36(4), pp. 2203–2212. doi: 10.1007/s10916-011-9687-0
10.1002/14651858.CD006632.pub2; Cheal, J., Development and implementation of a clinical pathway programme in an acute care general hospital in Singapore (2000) *Int. J. Qual. Health Care*, 12, pp. 403-412. , 10.1093/intqhc/12.5.403; Loeb, M., Carusone, S., Goeree, R., Walter, S., Brazil, K., Krueger, P., Simor, A., Marrie, T., Effect of a clinical pathway to reduce hospitalizations in nursing home residents with Pneumonia (2006) *J. Am. Assoc.*, 295 (21), pp. 2503-2510. , 1.
- Huang, Z. *et al.* (2014) 'Online Treatment Compliance Checking for Clinical Pathways', *Journal of Medical Systems*. Springer New York LLC, 38(10). doi: 10.1007/s10916-014-0123-0.
- Hurley, K. F. *et al.* (2007) 'Ontology engineering to model clinical pathways: Towards the computerization and execution of clinical pathways', in *20th IEEE International Symposium on Computer-Based Medical Systems, CBMS'07*. Maribor, pp. 536–541. doi: 10.1109/CBMS.2007.79.
- Husain, M. J. *et al.* (2012) 'HERALD (Health Economics using Routine Anonymised Linked Data)', *BMC medical informatics and decision making*, 12(24). doi: 10.1186/1745-6215-12-S1-A44.
- Hussain, S. *et al.* (2012) 'EHR4CR: A semantic web based interoperability approach for reusing electronic healthcare records in protocol feasibility studies', in *CEUR Workshop Proceedings*.
- Jovanovik, M., Najdenov, B. and Trajanov, D. (2013) 'Linked Open Drug Data from the Health Insurance Fund of Macedonia', in *10th International Conference for Informatics and Information Technology*. Available at: <http://e-tnc.com/etnc/Portals/3/papers/loddhifm-ciit2013.pdf>.
- Juurlink, D. N. *et al.* (2009) 'A population-based study of the drug interaction between proton pump inhibitors and clopidogrel', *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, 180(7), pp. 713–8. doi: 10.1503/cmaj.082001.
- Kaddari, A., Malki, M. O. C. and Elmdeghri, S. B. (2016) 'A pattern-based workflow to an automatic planning and monitoring of medical activities' processes', *International Journal of Innovative*

- Computing, Information and Control*. ICIC International, 12(4), pp. 1209–1225. Available at: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84982890355&partnerID=40&md5=92723614f4ecfb01e1a80401905b9214>.
- Khan, W. A. *et al.* (2013) 'Process interoperability in healthcare systems with dynamic semantic web services', *Computing*, 95(9), pp. 837–862. doi: 10.1007/s00607-012-0239-3.
- Kitchenham, B. (2004) *Procedures for Performing Systematic Reviews*. Keele. Available at: <http://www.it.hiof.no/~haraldh/misc/2016-08-22-smat/Kitchenham-Systematic-Review-2004.pdf> (Accessed: 15 January 2020).
- Kohonen, P. *et al.* (2013) 'The ToxBank data warehouse: Supporting the replacement of in vivo repeated dose systemic toxicity testing', *Molecular Informatics*, 32(1), pp. 47–63. doi: 10.1002/minf.201200114.
- Kotwal, S. *et al.* (2016) 'A review of linked health data in Australian nephrology', *Nephrology*, 21(6), pp. 457–466. doi: 10.1111/nep.12721.
- Lim, M., Kim, M. and Lee, K. (2013) 'WPAN Based Semantic-Web Health Monitoring', *The Journal of The Institute of Internet, Broadcasting and Communication*, 13(6), pp. 167–172.
- Liu, Z. and Wang, J. (2016) 'A fine-grained context-aware access control model for health care and life science linked data', *Multimedia Tools and Applications*. doi: 10.1007/s11042-016-3269-6.
- Machado, C. M. *et al.* (2013) 'The semantic web in translational medicine: Current applications and future directions', *Briefings in Bioinformatics*, 16(1), pp. 89–103. doi: 10.1093/bib/bbt079.
- Maragoudakis, M., Maglogiannis, I. and Lymberopoulos, D. (2008) 'A medical, description logic based, ontology for skin lesion images', in *2008 8th IEEE International Conference on Bioinformatics and BioEngineering*. IEEE, pp. 1–6. doi: 10.1109/BIBE.2008.4696706.
- Marshall, M. S. *et al.* (2012) 'Emerging practices for mapping and linking life sciences data using RDF - A case series', *Journal of Web Semantics*, 14, pp. 2–13. doi: 10.1016/j.websem.2012.02.003.
- McArthur, A. (2009) 'Health Information Professionals and the Semantic Web: A Symbiotic Relationship?', *Journal of the Canadian Health Libraries Association*, 30(3), pp. 81–84. Available at: <http://search.proquest.com/docview/57692185?accountid=142596%5Cnhttp://pubs.nrc-cnrc.gc.ca/jchla/jchla.html>.
- McCusker, J. P. *et al.* (2014) 'A nanopublication framework for biological networks using cytoscape.js', in *CEUR Workshop Proceedings*. CEUR-WS, pp. 90–92.

- McDonald, C. *et al.* (1995) 'Logical Observation Identifiers Names and Codes (LOINC) Users' Guide'. Regenstrief Institute, Inc. and the Logical Observation Identifiers Names and Codes (LOINC) Committee. Available at: <http://loinc.org/downloads/files/LOINCManual.pdf>.
- McGuinness, D. L. and Harmelen, F. van (2004) *Owl web ontology language overview*, *World Wide Web Consortium (W3C)*. doi: 10.1145/1295289.1295290.
- Melamed, R. D., Khiabani, H. and Rabadan, R. (2014) 'Data-driven discovery of seasonally linked diseases from an Electronic Health Records system.', *BMC bioinformatics*, 15(6). doi: 10.1186/1471-2105-15-S6-S3.
- Menezes, P. M., Cook, T. W. and Cavalini, L. T. (2016) 'Convergence of health level seven version 2 messages to semantic web technologies for software-intensive systems in telemedicine trauma care', *Healthcare Informatics Research*, 22(1), pp. 22–29. doi: 10.4258/hir.2016.22.1.22.
- Miller, P. (2000) *Interoperability: What is it and Why Should I Care?*, *Ariadne*. Available at: <http://www.ariadne.ac.uk/issue24/interoperability/> (Accessed: 10 December 2016).
- Miyazaki, M. *et al.* (2015) 'My Health Dictionary : Study on Web Service using Program Information Data-hub as Linked Open Data', in *CEUR workshop Proceedings*.
- Mohammadhassanzadeh, H. *et al.* (2017) 'Semantics-based plausible reasoning to extend the knowledge coverage of medical knowledge bases for improved clinical decision support', *BioData Mining*, 10(1), p. 7. doi: 10.1186/s13040-017-0123-y.
- Moher, D. *et al.* (2009) 'Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement', *PLoS Medicine*. Public Library of Science, 6(7), p. e1000097. doi: 10.1371/journal.pmed.1000097.
- Musen, M. A. (2015) 'The Protégé Project: A Look Back and a Look Forward', *AI Matters*. *Association of Computing Machinery Specific Interest Group in Artificial Intelligence Association of Computing Machinery Specific Interest Group in Artificial Intelligence*, 1(4), pp. 4–12. doi: 10.1016/S2215-0366(16)30284-X.Epidemiology.
- National Cancer Institute (no date) *NCI Thesaurus*. Available at: <https://ncit.nci.nih.gov> (Accessed: 14 December 2016).
- National Health Service (NHS) (no date) *NHS Digital*. Available at: <https://www.digital.nhs.uk/home> (Accessed: 20 February 2017).
- Natsiavas, P. *et al.* (2018) 'OpenPVSignal: Advancing information search, sharing and reuse on

pharmacovigilance signals via fair principles and Semantic Web Technologies', *Frontiers in Pharmacology*, 9(JUN). doi: 10.3389/fphar.2018.00609.

Nelson, M. L. and Sen, R. (2014) 'Business rules management in healthcare: A lifecycle approach', *Decision Support Systems*, 57(1), pp. 387–394. doi: 10.1016/j.dss.2012.10.044.

NHS Business Services Authority (BSA) (2017) *Information Services Portal*. Available at: <https://apps.nhsbsa.nhs.uk/infosystems/welcome> (Accessed: 17 October 2017).

NHS Choices (2017a) *About the National Health Service (NHS) in England*. Available at: <http://www.nhs.uk/NHSEngland/thenhs/about/Pages/overview.aspx> (Accessed: 17 October 2017).

NHS Choices (2017b) *NHS Choices*. Available at: <https://www.nhs.uk/pages/home.aspx> (Accessed: 21 October 2017).

NHS Choices (no date a) *GP Opening Times - NHS Choices*. Available at: <https://data.gov.uk/dataset/gp-opening-times> (Accessed: 14 July 2017).

NHS Choices (no date b) *GP practices and surgeries - NHS Choices*. Available at: <https://data.gov.uk/dataset/gp-practices-and-surgeries> (Accessed: 14 July 2017).

NHS Choices (no date c) *GPs' staff - NHS Choices*. Available at: <https://data.gov.uk/dataset/gps-staff> (Accessed: 14 July 2017).

NHS Digital (2016) *Prescriptions Dispensed in the Community, Statistics for England - 2005-2015*. Available at: <http://digital.nhs.uk/catalogue/PUB20664> (Accessed: 17 October 2017).

NHS Digital (2017a) *GP and GP practice related data*. Available at: <https://digital.nhs.uk/organisation-data-service/data-downloads/gp-data> (Accessed: 17 January 2017).

NHS Digital (2017b) *iView - NHS Digital*. Available at: <https://iview.hscic.gov.uk/> (Accessed: 25 October 2017).

NHS Digital (2017c) *Patients Registered at a GP Practice - NHS Digital*. Available at: <https://digital.nhs.uk/Patients-registered-at-a-GP-practice-GP-data-hub> (Accessed: 14 July 2017).

NHS Digital (2017d) *Primary Care Prescribing - NHS Digital*. Available at: <https://digital.nhs.uk/article/6735/Primary-care-prescribing> (Accessed: 25 October 2017).

NHS Digital (no date) *Organisation Data Changes*. Available at:

https://hscic.kahootz.com/connect.ti/O_D_S/grouphome? (Accessed: 6 February 2020).

Nie, H. *et al.* (2013) 'From healthcare messaging standard to semantic web service description: Generating WSMO annotation from HL7 with mapping-based approach', in *Proceedings - IEEE 10th International Conference on Services Computing, SCC 2013*. IEEE, pp. 470–477. doi: 10.1109/SCC.2013.74.

Noy, N. F. and McGuinness, D. L. (2001) *Ontology Development 101: A Guide to Creating Your First Ontology*, Knowledge Systems Laboratory Stanford University. Available at: http://bmir.stanford.edu/file_asset/index.php/108/BMIR-2001-0880.pdf (Accessed: 14 December 2016).

Odgers, D. J. and Dumontier, M. (2015) 'Mining Electronic Health Records using Linked Data.', in *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science*, pp. 217–21. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4525267&tool=pmcentrez&rendertype=abstract>.

Ontotext (2017) 'GraphDB Free'. Ontotext. Available at: <https://ontotext.com/>.

Ostankov, A., Rohrbein, F. and Waltinger, U. (2014) 'LinkedHealthAnswers: Towards Linked Data-driven Question Answering for the Health Care Domain', in *Lrec 2014 - Ninth International Conference on Language Resources and Evaluation*, pp. 2613–2620.

Oxford University Press (no date) *Oxford Dictionaries*. Available at: <https://en.oxforddictionaries.com/> (Accessed: 10 November 2016).

Papakonstantinou, D. *et al.* (2011) 'A semantic wiki for user training in ePrescribing processes', in *4th ACM International Conference on Pervasive Technologies Related to Assistive Environments, PETRA 2011*. Heraklion, Crete. Available at: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84858032946&partnerID=40&md5=3534592e0ef39be4e3f4c5279264bcb9>.

Pathak, J. *et al.* (2012) 'Applying semantic web technologies for phenome-wide scan using an electronic health record linked Biobank.', *Journal of biomedical semantics*, 3(1), p. 10. doi: 10.1186/2041-1480-3-10.

Pathak, J., Kiefer, Richard C. and Chute, C. G. (2012) 'Applying linked data principles to represent patient's electronic health records at Mayo clinic: a case report', *Proceedings of the 2nd ACM SIGHIT symposium on International health informatics - IHI '12*, pp. 455–464. doi: 10.1145/2110363.2110415.

Pathak, J., Kiefer, Richard C and Chute, C. G. (2012) 'Using semantic web technologies for cohort identification from electronic health records for clinical research.', *AMIA Summits Transl Sci Proc*, pp. 10–19. Available at:

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3392057&tool=pmcentrez&rendertype=abstract>.

Pathak, Jyotishman, Kiefer, R. C. and Chute, C. G. (2013) 'Using linked data for mining drug-drug interactions in electronic health records', *Studies in Health Technology and Informatics*, 192(1–2), pp. 682–686. doi: 10.3233/978-1-61499-289-9-682.

Pathak, J, Kiefer, R. C. and Chute, C. G. (2013) 'Using linked data for mining drug-drug interactions in electronic health records', in *14th World Congress on Medical and Health Informatics, MEDINFO 2013*. 1st–2nd edn. Copenhagen, pp. 682–686. doi: 10.3233/978-1-61499-289-9-682.

Pathak, Jyotishman, Kiefer, R. and Chute, C. (2013) 'Mining Anti-coagulant Drug-Drug Interactions from Electronic Health Records Using Linked Data', *Data Integration in the Life Sciences*, pp. 128–140. Available at: http://link.springer.com/chapter/10.1007/978-3-642-39437-9_11.

PostGIS (no date) *PostGIS — Spatial and Geographic Objects for PostgreSQL*. Available at: <http://postgis.net> (Accessed: 17 October 2017).

Poulymenopoulou, M., Malamateniou, F. and Vassilacopoulos, G. (2015) 'A LOD-based service for extracting linked open emergency healthcare data', in *International Conference on Bioinformatics and Biomedical Engineering*. Springer International Publishing, pp. 85–91. Available at:

<https://www.scopus.com/inward/record.uri?eid=2-s2.0-84925263778&partnerID=40&md5=9b3df7169331814b9452ba5c3f04f581>.

Prescribing Analytics (no date) *Prescribing Analytics*. Available at: <http://prescribinganalytics.com/> (Accessed: 15 November 2017).

Puustjarvi, J. and Puustjarvi, L. (2015) 'The Role of Smart Data in Smart Home: Health Monitoring Case', in Papasratorn, B. and Chignell, M. (eds) *7th International Conference on Advances in Information Technology, 2015*. Elsevier B.V., pp. 143–151. doi: 10.1016/j.procs.2015.10.015.

Puustjarvi, J. and Puustjarvi, L. (2016) 'Selective dissemination of clinical guidelines in healthcare communities', in *IEEE International Conference on Industrial Engineering and Engineering Management*. IEEE, pp. 706–710. doi: 10.1109/IEEM.2015.7385739.

Puustjarvi, J and Puustjarvi, L. (2016) 'Selective dissemination of clinical guidelines in healthcare communities', in *IEEE International Conference on Industrial Engineering and Engineering*

Management, IEEM 2015. IEEE Computer Society, pp. 706–710. doi: 10.1109/IEEM.2015.7385739.

Puustjärvi, J. and Puustjärvi, L. (2006) 'The Challenges of Electronic Prescription Systems Based on Semantic Web Technologies', *Eceh*, pp. 251–262.

Puustjärvi, J. and Puustjärvi, L. (2009) 'Transactional allocation of clinical resources for health care processes', in *11th International Conference on Information Integration and Web-based Applications and Services, iiWAS2009*. Kuala Lumpur, pp. 524–528. doi: 10.1145/1806338.1806437.

QSR International Pty Ltd. (2019) 'NVivo qualitative data analysis software'. Available at: <https://www.qsrinternational.com/nvivo/home> (Accessed: 15 January 2020).

Rinciog, O. and Posea, V. (2015) 'Publishing Romanian public health data as Linked Open Data', in *The 5th IEEE International Conference on E-Health and Bioengineering (EHB)*. IEEE, pp. 1–4.

Robertson, J. *et al.* (2012) 'The health services burden of heart failure: an analysis using linked population health data-sets.', *BMC health services research*, 12(1), p. 103. doi: 10.1186/1472-6963-12-103.

Rowlingson, B. *et al.* (2013) 'Mapping English GP prescribing data: a tool for monitoring health-service inequalities', *BMJ open*, 3(e001363), pp. 1–10. doi: 10.1136/bmjopen-2012-001363.

Sagotsky, J. A. *et al.* (2008) 'Life Sciences and the web: a new era for collaboration', *Molecular Systems Biology*, 4(1), pp. 201–210. doi: 10.1038/msb.2008.39.

Shaban-Nejad, A. *et al.* (2012) 'HAIKU: A semantic framework for surveillance of healthcare-associated infections', in *3rd International Conference on Ambient Systems, Networks and Technologies, ANT 2012 and 9th International Conference on Mobile Web Information Systems, MobiWIS 2012*. Niagara Falls, ON: Elsevier B.V., pp. 1073–1079. doi: 10.1016/j.procs.2012.06.151.

Shaban-Nejad, A. *et al.* (2016) 'From Cues to Nudge: A Knowledge-Based Framework for Surveillance of Healthcare-Associated Infections', *Journal of Medical Systems*. Springer New York LLC, 40(1), pp. 1–12. doi: 10.1007/s10916-015-0364-6.

Shadbolt, N., Hall, W. and Berners-Lee, T. (2006) 'The Semantic Web Revisited', *IEEE Intelligent Systems*, 21(3), pp. 96–101. doi: 10.1109/MIS.2006.62.

Shah, T. *et al.* (2014) 'Enhancing Automated Decision Support across Medical and Oral Health Domains with Semantic Web Technologies', in *24th Australasian Conference on Information Systems*. Available at: <http://arxiv.org/abs/1403.7766>.

Sifaki-Pistolla, D. I. *et al.* (2017) 'Geospatial and Spatio-Temporal Analysis in Health Research: GIS in Health', in *Handbook of Research on Geographic Information Systems Applications and Advancements*. IGI Global, pp. 466–487. doi: 10.4018/978-1-5225-0937-0.

Singh, P. *et al.* (2013) 'Domain ontology based efficient image retrieval', in *7th International Conference on Intelligent Systems and Control, ISCO 2013*. IEEE, pp. 445–452. doi: 10.1109/ISCO.2013.6481196.

Skoutas, D. and Simitsis, A. (2006) 'Designing ETL processes using semantic web technologies', in *Proceedings of the 9th ACM international workshop on Data warehousing and OLAP - DOLAP '06*. New York, New York, USA: ACM Press, p. 67. doi: 10.1145/1183512.1183526.

Smith, B. *et al.* (2007) 'The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration', *NATURE BIOTECHNOLOGY VOLUME*, 25. doi: 10.1038/nbt1346.

Sonsilphong, S., Arch-Int, N. and Arch-Int, S. (2012) 'Rule-based semantic web services annotation for healthcare information integration', in *2012 8th International Conference on Computing and Networking Technology (INC, ICCIS and ICMIC), ICCNT 2012*, pp. 147–152. doi: 978-1-4673-1326-1.

Sreekanth, V. K. and Biswas, D. (2014) 'Information Asymmetry Minimization System for Potential Clients of Healthcare Insurance in Indian Context using Semantic Web', in *Proceeding of the 2014 IEEE Students' Technology Symposium*. IEEE, pp. 282–285.

Stavropoulos, T. G. *et al.* (2016) 'Multimodal Sensing and Intelligent Fusion for Remote Dementia Care and Support', in *Proceedings of the 2016 ACM Workshop on Multimedia for Personal Health and Health Care*. ACM, pp. 35–39. doi: 10.1145/2985766.2985776.

The Centre for Health Record Linkage (CHeReL) (no date) *The Centre for Health record linkage (CHeReL)*. Available at: <http://www.cherel.org.au/> (Accessed: 10 November 2016).

the Insight Centre for Data Analytics (2019) *The Linked Open Data Cloud*. Available at: <https://lod-cloud.net/> (Accessed: 8 January 2020).

The Wikimedia Foundation (2019) *Wikidata*. Available at: https://www.wikidata.org/wiki/Wikidata:Main_Page (Accessed: 20 October 2020).

Thomson Reuters (2016) 'EndNote'. Thomson Reuters. Available at: <https://endnote.com/> (Accessed: 21 January 2020).

Tilahun, B. *et al.* (2014) 'Design and development of a linked open data-based health information representation and visualization system: Potentials and preliminary evaluation', *Journal of*

Medical Internet Research, 16(10), p. e31. doi: 10.2196/medinform.3531.

U.S. Department of Health and Human Services (no date) *Overview - Adverse Drug Events - health.gov*. Available at: https://health.gov/hcq/ade.asp#_ftn1 (Accessed: 5 July 2019).

Ullah, F. *et al.* (2017) 'Semantic interoperability for big-data in heterogeneous IoT infrastructure for healthcare', *Sustainable Cities and Society*. Elsevier, 34(March), pp. 90–96. doi: 10.1016/j.scs.2017.06.010.

Velmurugan, A. and Ravi, T. (2016) 'Allergy information ontology for enlightening people', in *2016 International Conference on Computing Technologies and Intelligent Data Engineering, ICCTIDE 2016*. IEEE, pp. 1–7. doi: 10.1109/ICCTIDE.2016.7725329.

Völker, S. and Kistemann, T. (2011) 'The impact of blue space on human health and well-being - Salutogenetic health effects of inland surface waters: A review', *International Journal of Hygiene and Environmental Health*. doi: 10.1016/j.ijheh.2011.05.001.

Wang, H. Q. *et al.* (2013) 'Creating personalised clinical pathways by semantic interoperability with electronic health records', *Artificial Intelligence in Medicine*, 58(2), pp. 81–89. doi: 10.1016/j.artmed.2013.02.005.

Wang, H. Q. *et al.* (2015) 'Research and Development of Semantics-based Sharable Clinical Pathway Systems', *Journal of Medical Systems*. Springer New York LLC, 39(7). doi: 10.1007/s10916-015-0257-8.

Wang, M. H. *et al.* (2010) 'Ontology-based multi-agents for intelligent healthcare applications', *Journal of Ambient Intelligence and Humanized Computing*, 1(2), pp. 111–131. doi: 10.1007/s12652-010-0011-5.

Winkler, W. E. (2006) *Overview of record linkage and current research directions, Bureau of the Census*. doi: 10.1206/3728.2.

Wishart, D. S. *et al.* (2006) 'DrugBank: a comprehensive resource for in silico drug discovery and exploration', *Nucleic Acids Research*, 34(90001), pp. D668–D672. doi: 10.1093/nar/gkj067.

World Health Organization (no date) *World Health Organization Regional Office for Europe - Public health services*. Available at: <http://www.euro.who.int/en/health-topics/Health-systems/public-health-services> (Accessed: 1 May 2019).

World Health Organization (WHO) (no date) *Research*. Available at: https://www.who.int/health-topics/research/#tab=tab_1 (Accessed: 12 October 2020).

- World Wide Web Consortium (2013) *W3C Semantic Web Activity Homepage, W3C*. Available at: <https://www.w3.org/2001/sw/>.
- World Wide Web Consortium (2014) *The Organization Ontology*, <http://www.w3.org/TR/vocab-org/>. Available at: <http://www.w3.org/TR/vocab-org/> (Accessed: 1 June 2017).
- World Wide Web Consortium (W3C) (2004) *SWRL: A Semantic Web Rule Language Combining OWL and RuleML*. Available at: <https://www.w3.org/Submission/SWRL/> (Accessed: 8 January 2020).
- World Wide Web Consortium (W3C) (2008) *Inference, World Wide Web Consortium (W3C)*. Available at: <http://www.w3.org/standards/semanticweb/inference> (Accessed: 16 May 2017).
- World Wide Web Consortium (W3C) (2009) *Feature: Basic Federated Query, W3C*. Available at: https://www.w3.org/2009/sparql/wiki/Feature:BasicFederatedQuery#Feature:_Basic_Federated_Query (Accessed: 22 May 2019).
- World Wide Web Consortium (W3C) (2013a) *RIF Rule Interchange Format, World Wide Web Consortium (W3C)*. Available at: http://www.w3.org/standards/techs/rif#w3c_all (Accessed: 19 May 2017).
- World Wide Web Consortium (W3C) (2013b) *SPARQL 1.1 Query Language*. World Wide Web Consortium (W3C). Available at: <https://www.w3.org/TR/sparql11-query/> (Accessed: 20 February 2020).
- World Wide Web Consortium (W3C) (no date a) *Large TripleStores*. Available at: https://www.w3.org/wiki/LargeTripleStores#GraphDB.E2.84.A2_by_Ontotext_.2817B.29 (Accessed: 1 November 2017).
- World Wide Web Consortium (W3C) (no date b) *Query, World Wide Web Consortium (W3C)*. Available at: <http://www.w3.org/standards/semanticweb/query> (Accessed: 18 May 2017).
- Ye, Y. *et al.* (2008) 'A semantics-based clinical pathway workflow and variance management framework', in *2008 IEEE International Conference on Service Operations and Logistics, and Informatics, IEEE/SOLI 2008*. Beijing, pp. 758–763. doi: 10.1109/SOLI.2008.4686499.
- Ye, Y. *et al.* (2009) 'An ontology-based hierarchical semantic modeling approach to clinical pathway workflows', *Computers in Biology and Medicine*, 39(8), pp. 722–732. doi: 10.1016/j.compbiomed.2009.05.005.
- Zaman, N. and Li, J. (2014) 'Semantics-Enhanced Recommendation System for Social Healthcare',

in *2014 IEEE 28th International Conference on Advanced Information Networking and Applications*. IEEE, pp. 765–770. doi: 10.1109/AINA.2014.93.

Zaveri, A. and Ertaylan, G. (2017) 'Linked Data for Life Sciences', *Algorithms*, 10(4), p. 126. doi: 10.3390/a10040126.

Zenuni, X. *et al.* (2015) 'State of the Art of Semantic Web for Healthcare', in *Procedia - Social and Behavioral Sciences*. Elsevier B.V., pp. 1990–1998. doi: 10.1016/j.sbspro.2015.06.213.

Appendix A The Uses of the Semantic Web in Health

Data Matrices

A.1 General Information

<i>No.</i>	<i>paper</i>	<i>publication year</i>	<i>overview</i>	<i>domain</i>	<i>specific aims</i>
1	& A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data	2013	This paper introduce a new model to represent data in the EHRR by adding the semantics to The clinical element model (CEM) in order to perform reasoning and check consistency.	technical	supporting health-related decisions by adding a semantic layer to the CEM
2	Social health data integration using semantic Web	2012	This paper discusses a system where health related data is extracted and intrgated from different communitis to provide health knowledge for the goal of better health related decisions.	technical	providing health knowledge by extracting health-related data from different communities
3	*Applying semantic web technologies for phenome-wide scan using an electronic health record linked Biobank	2012	This paper discusses how to identify subjects with specific diseases (Type 2 Diabetes Mellitus (T2DM) or Hypothyroidism) and phenotypes by identifying genotype-phenotype associations through the use of the semantic web technology applied to the clinical data in the form of EHRR.	Medical	identifying subjects with specific disease by mapping phenotype - genotype associations
4	*Using Semantic Web Technologies for Cohort Identification from Electronic Health Records for Clinical Research	2012	This paper discusses the potential of using the semantic web for identifying subjects with specefic diseeses (Diabetes Mellitus) in cohort studies.	Medical	identifying subjects with specific disease by mapping phenotype - genotype associations
5	The role of taxonomies in social media and the semantic web for health education: A study of SNOMED CT terms in youtube health video tags	2013	This paper disccses the use of the semantic web in searching health education content in the youtube	technical	searching for health education content in social media by using video tags
6	WPAN Based Semantic-Web Health Monitoring	2013	This paper demonstrates a wireless smart application that uses the semantic technology for the purpose of diagnoses (pregnancy).	technical	managing and transferring patient's data by using a wireless application
7	Convergence of Health Level Seven Version 2 Messages to Semantic Web Technologies for Software-Intensive Systems in Telemedicine Trauma Care	2016	To present the technical background and the development of a procedure that enriches the semantics of Health Level Seven version 2 (HL7v2) messages for software-intensive systems in telemedicine trauma care."	technical	supporting information exchange and communication in telemedicine trauma care by adding semantics in the HL7v2 messages
8	Enhancing Automated Decision Support across Medical and Oral Health Domains with Semantic Web Technologies	2014	This paper aims to integrate patient's information from two domains: general and oral health to support decision.	Medical	finding inter-dependent conditions between two health domains by reasoning over cross-domain patient's information
9	A Comparison of Mobile Rule Engines for Reasoning on Semantic Web Based Health Data	2014	This paper compares the 4 benchmarks in mobile reasoning to examine their scalability. Atrial Fibrillation (AF) data set was their case study.	technical	supporting self health-related decisions using mobile platforms by examining the scalability of

					4 benchmarks reasoning engines
10	**Semantic Web Services for medical health planning	2012	This paper suggests a model for the problem of health planning. The model has been shown from the patient's perspective and the health provider's side.	technical	supporting health planning (finding the proper health provider for a certain patient) by developing a SWS presenting the health provider side and the patient side
11	**Multi-Agent Based Semantic Web Service Model and Ontological Description for Medical Health Planning	2012	This paper suggests a model for the problem of health planning. The model has been shown from the patient's perspective and the health provider's side.	technical	supporting health planning (finding the proper health provider for a certain patient) by developing a SWS presenting the health provider side and the patient side
12	%Potential of Linked Open Data in Health Information Representation on the Semantic Web	2012	This paper discusses the potential of the linked open data in representing health data.	technical	supporting health management data by discussing the potential of linked data in representing health data
13	Semantic Web and the Future of Health Care Data in Family Practice	2015	This paper discusses the different terminologies used by computer scientists regarding the semantic web in the health domain and suggest the use of the semantic web techniques along with the natural language processing techniques to improve the family practice.	technical	supporting data management in health care by introducing semantic web terminologies to family practitioners
14	Cross-Domain Semantic Web Model for Understanding Multilingual Natural Language Queries: English/Arabic Health/Food Domain Use Case	2016	This paper discusses an approach of using Arabic as a natural language for querying a health/food system immediately.	technical	supporting multilingual semantic web technologies by helping users to query the system in Arabic
15	Healthcare System Based on Semantic Web and XML Technologies	2013	This paper aims to manage the medical data during the diagnose through the use of semantic and XML technologies.	technical	managing diagnostic data by presenting an ontology-based framework that integrates biological data sources
16	Senhance: a Semantic Web framework for integrating social and hardware sensors in e-health	2015	This paper aims to integrate self-reported data from social networks with sensor's data in order to create a space for e-health applications.	technical	supporting health professionals by integrating self-reported data from social networks with observations from sensors
17	A semantic cache for enhancing Web services communities activities : Health care case Study	2012	This paper uses semantic annotations in community web services in a case study in the health domain.	technical	supporting collective memories in web services community by indexing a healthcare Web services with semantic annotations
18	A Semantic-Based Health Advising System Exploiting Web-Based Personal Health Record Services	2015	This paper propose a model for Personal Health Record(PHR) web service for the purpose of self health advising.	technical	supporting self-health advising systems by proposing a model for Personal Health Record (PHR)
19	Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts.	2013	This paper proposes a method to classify patients by the risk of developing colorectal cancer using their (EHRs).	Medical	classifying patients with the risk of developing a certain disease (identifying patient cohorts) by using phenotyping algorithm

20	Process interoperability in healthcare systems with dynamic semantic web services	2013	This paper proposes a method to add a semantic layer over existing health web services.	technical	supporting semantic interoperability in healthcare systems by adding a semantic layer over existing health web services
21	From healthcare messaging standard to semantic web service description: Generating WSMO annotation from HL7 with mapping-based approach	2013	This paper proposes a novel method to generate semantic annotations automatically.	technical	utilizing healthcare domain knowledge by generating semantic annotations automatically
22	Information Asymmetry Minimization System for Potential Clients of Healthcare Insurance in Indian Context using Semantic Web	2014	This paper introduces an information system to health insurance clients to find the proper provider with the difficulty of asymmetric information.	technical	minimizing information asymmetry in Indian healthcare insurance market by proposing an information system (SWS)
23	& A semantic-web oriented representation of clinical element model for secondary use of electronic healthcare data	2013	This paper introduces an effort to add a semantic layer to the clinical element model (CEM), which is an information model designed for representing clinical information in electronic health records (EHR) systems.	technical	supporting health-related decisions by adding a semantic layer to the CEM
24	The impact of social media networks on healthcare process knowledge management (using of semantic web platforms)	2014	This paper demonstrates the impact of social networks on the health care domain by developing ontologies.	technical	improving information flow and supporting decision making in healthcare by acquiring knowledge from social networks and integrate it to construct a domain ontology
25	EHR4CR: A semantic web based interoperability approach for reusing electronic healthcare records in protocol feasibility studies	2012	This paper proposes a model for improving the clinical trials through the use of electronic health records (EHRs).	technical	improving clinical trials by providing a semantic interoperability system for reusing EHRs
26	Rule-based semantic web services annotation for healthcare information integration	2012	This paper proposes a framework for supporting interoperability between incompatible electronic patients records (EPRs) in independent health care systems.	technical	proposing a framework for interoperating health care systems by generating semantic-rules with the help of annotations
27	*Applying linked data principles to represent patient's electronic health records at Mayo clinic: a case report	2012	This paper discusses how to convert the EHRs in Mayo clinic from relational database form to virtual RDF form.	Pharmacology	discovering potential drug-drug interactions by mining patient information
28	*Using linked data for mining drug-drug interactions in electronic health records	2013	This paper aims to find potential drug drug interactions (PDDI) in Mayo clinic's electronic health records (EHRs) for prescribed cardiovascular and gastroenterology drugs.	Pharmacology	discovering potential drug-drug interactions by mining patient information
29	*Mining Anti-coagulant Drug-Drug Interactions from Electronic Health Records Using Linked Data	2013	This paper aims to find potential drug drug interactions (PDDI) in Mayo clinic's electronic health records (EHRs) for prescribed cardiovascular and gastroenterology drugs. (same as above)	Pharmacology	discovering potential drug-drug interactions by mining patient information
30	Design and Development of a Linked Open Data-Based Health Information Representation and Visualization System: Potentials and Preliminary Evaluation	2014	This paper aims to evaluate the semantic web technology for representing, querying and visualizing the health data.	technical	evaluating linked data technologies for health care domain by developing an LOD-based health information system

31	The Canadian health census as Linked Open Data : towards policy making in public health	2013	This paper aims to re-publish the open Canadian health census data as linked open data.	technical	re-publishing the Canadian open health data by using the linked data technologies
32	Linked Open Drug Data from the Health Insurance Fund of Macedonia	2013	This paper focuses on transforming a dataset from the Health Insurance Fund (HIF) of the Republic of Macedonia website into an open linked data interlinked with the DrugBank domain.	technical	re-publishing the Macedonian Health Insurance Funds data by using the linked data technologies
33	Towards Linked Vital Registration Data for Reconstituting Families and Creating Longitudinal Health Histories	2014	The aim for this project is to create a knowledge base for historical information for individuals such as birth, marriage, death records. This will help the historians (researchers) to question the official reports for maternal and infants mortality rates the data in depth(longitudinal data).	Social	epidemiological planning by analysing longitudinal historical data
34	RDF-ization of DICOM Medical Images towards Linked Health Data Cloud	2015	This paper proposes a process for semantifying a Digital Imaging and Communications in Medicine (DICOM) medical image by extracting and serializing the metadata.	technical	Semintifying medical images by extracting and serializing metadata
35	Mining Electronic Health Records using Linked Data	2015	This paper aims to transform the data in the Stanford's STRIDE database to a semantic version to illustrate basic cohort selection, phenotypic profiling, and identification of disease genes.	Medical	selecting patient cohort by phenotyping and identifying genes
36	A fine-grained context-aware access control model for health care and life science linked data	2016	This paper introduce a new model for access control to the Health Care and Life Sciences (HCLS) linked data.	technical	controlling access for HCLS linked data by presenting an access control model
37	Linked Data Analytics in Interdisciplinary Studies: The Health Impact of Air Pollution in Urban Areas	2016	This paper introduce an approach for producing linked data analytics in urban environments.	technical	analysing linked data to serve research by using linked data technologies in an urban environment study
38	LinkedHealthAnswers: Towards Linked Data-driven Question Answering for the Health Care Domain	2014	This paper presents a natural language processing (NLP) system that utilize the querying process for the health data cloud.	technical	supporting quering the health data cloud by using natural language processing
39	Publishing Romanian public health data as Linked Open Data	2015	This paper discusses the process of publishing the Romanian health data as RDF.	technical	re-publishing the Romanian open health data by using the linked data technologies
40	My Health Dictionary : Study on Web Service using Program Information Data-hub as Linked Open Data	2015	This paper aims to develop an information hub for health related programmes on TV, where the user can explore programme related information and content information.	technical	exploring content information in TV health programmes by developing an information hub using linked open data
41	Personal healthcare record integration method based on linked data model	2014	This paper propose a system that integrates personal health records (historically and currentally) (EHRs) to help doctors in assessing the patient's condition.	technical	assessing patient's conditions by integrating historical and currental personal health records
42	A LOD-based service for extracting linked open emergency healthcare data	2015	This paper presents a method to create an emergency LOD by exporting emergency health data semantically and linking it with LOD.	technical	making emergency healthcare data available to authorized users by describing a LOD-based cloud service that automatically export emergency healthcare data

A.2 Domain Information

No.	Health domain	Health sub-domain(systems)	Health Issue	users
1	health management (decision support)	decision support system	general health	health professionals, decision makers
2	health information (public health)	social networking	general health	patients, care givers
3	clinical (diagnosis)	genome-wide association studies (GWAS), electronic health records (EHRs)	Type 2 Diabetes and Hypothyroidism, genetics	academic medical centers
4	clinical (diagnosis)	genome-wide association studies (GWAS)	Diabetes Mellitus, genetics	academic medical centers
5	health information (public health)	social networking	educating videos on surgery	health learners
6	clinical (diagnosis)	smartphone healthcare systems, Diagnostic Systems	pregnancy	patients, healthcare providers
7	health management (communication)	Healthcare information systems (HIS), telemedicine	trauma care	emergency services, the receiving hospital
8	clinical (diagnosis)	decision support system	general and oral health	healthcare practitioners
9	clinical (diagnosis)	decision support system	Atrial Fibrillation (AF)	healthcare practitioners
10	health management (planning)	Semantic Web Service (SWS)	general health	patients, doctors, hospitals
11	Health management (planning)	Semantic Web Service (SWS)	general health	patients, doctors, hospitals
12	health information (health data, representation)	decision support system	general health	health professionals, decision makers
13	health information (public health)	health information systems (HIS), electronic health records (EHRs), natural language process (NLP)	general practice (GP)	general practitioners (GPs)
14	health information (public health)	Semantic Web Service (SWS), natural language processing (NLP)	nutrition	the public
15	clinical (diagnosis)	diagnosing	Jaundice disease	personal, health providers
16	health information (public health)	social networking,	nutrition	health professionals
17	health information (public health)	Semantic Web Service (SWS) communities	general health	health professional
18	health information (public health)	Semantic Web Service (SWS), Personal Health Record (PHR), advising system	general health	the public
19	clinical (diagnosis)	electronic health records(EHRs), decision system	cancer	doctors, decision makers
20	health information (health data, representation)	Semantic Web Services (SWS)	laboratory	hospitals, clinicians, medical support staff, and patients.
21	health information (health data, annotation)	Semantic Web Services (SWS)	laboratory	developers
22	health information (health data, integration)	health information systems (HIS)	health insurance	clients, patients
23	health information (health data, representation)	electronic health records (EHRs)	general health	health practitioners
24	health information (public health)	health information systems (HIS), social media, knowledge management (KM)	general health	patient, health practitioners
25	clinical (research)	electronic health records (EHRs)	general	researchers

26	health information (health data, interoperability)	electronic patient records(EPRs), semantic web services (SWS)	general health	patient, health practitioners
27	clinical (diagnosis)	genome-wide association studies (GWAS), electronic health records (EHRs)	type 2 diabetes, genetics	academic medical centers
28	pharmacology (drugs interactions)	electronic health records (EHRs), genetics, potential drug-drug interactions(PDDIs), decision support system	cardiovascular and gastroenterology drugs	biomedical researchers
29	pharmacology (drugs interactions)	electronic health records (EHRs), genetics, PDDI, decision support system	anti-coagulant drugs	biomedical researchers
30	health information(health data, representation)	decision support system (DSS), health information system (HIS)	human immunodeficiency virus infection/acquired immunodeficiency syndrome (HIV/AIDS)	health managers
31	health information (health data, publishing)	decision support system (DSS), health information system (HIS)	The Canadian health census	health managers, policy makers
32	health information (health data, publishing)	publishing LOD	drugs	developers, public
33	clinical (research)	health history	maternal and infant mortality rates	historians, researchers
34	clinical (diagnosis)	imaging	general	clinicians, experts, patients
35	clinical (research)	electronic health records (EHRs)	genetics	researchers
36	health information (health data, accessibility)	access control system	general health	publishers
37	health information (public health)	publishing and analyzing LOD, decision support system	pollution	managers, policy makers
38	clinical (research)	natural language processing (NLP) system	general	researchers, clinicians
39	health information (health data, publishing)	publishing governmental data	general health	health managers, patients, researchers
40	health information (public health)	health TV programmes	general health	public
41	clinical (diagnosis)	electronic health records (EHRs)	general health	doctors
42	health information (health data, representation, mapping)	emergency medical services (EMS), personal health records (PHRs)	emergency	managers, policy makers, researchers

A.3 Aims / Goals

No.	<i>support health administration</i>	<i>support research</i>	<i>support diagnosis</i>	<i>support learning</i>	<i>support e-health systems</i>
1	1 (decision making by the healthcare provider)	0	0	0	1 (secondary use of EHRs)
2	1 (self-decision making)	0	0	1 (self-learning)	1 (extract and integrate social networks data)
3	0	1 (mapping phenotype - genotype associations)	1 (identify subjects with disease)	1 (support linked data)	1 (secondary use of EHRs)
4	0	1 (mapping phenotype - genotype associations)	1 (identify subjects with disease)	1 (support linked data)	1 (secondary use of EHRs)
5	0	0	0	1 (self-learning)	1 (tagging YouTube videos)
6	0	0	1 (checking the patient's situation wirelessly)	1 (self-monitoring)	1 (a smartphone application)

7	1 (communication within different healthcare services)	0	0	0	1 (HIS)
8	1 (decision making by the healthcare provider)	1 (finding associations between oral health and general health)	0	1 (support linked data)	0
9	1 (self-decision making)	0	0	0	1 (mobile semantic reasoners)
10	1 (self-decision making)	0	0	0	1 (SWS)
11	1 (self-decision making)	0	0	0	1 (SWS)
12	1 (decision making in general)	0	0	1 (support linked data)	0
13	1 (decision making in general)	0	0	1 (support linked data)	0
14	0	0	0	1 (support linked data)	0
15	0	0	1 (by presenting a disease ontology)	1 (support linked data)	0
16	0	0	0	1 (self-monitoring)	1 (integrating sensor's data with social media data)
17	0	0	0	0	1 (searching relevant resources by using annotations)
18	1 (self-decision making)	0	1 (self-diagnosis)	0	1 (self-advising system)
19	0	1 (mapping phenotypes)	1 (identify subjects with diseases)	1 (support linked data)	1 (secondary use of EHRs)
20	1 (decision making by the healthcare provider)	0	0	0	1 (adding semantic layer in e-health services)
21	0	0	0	1 (support linked data)	0
22	1 (self-decision making)	0	0	0	1 (comparing various health insurance services)
23	1 (decision making by the healthcare provider)	0	0	0	1 (secondary use of EHRs) (adding a semantic layer in the clinical model)
24	1 (communication within different healthcare services)	0	0	1 (self-learning)	0
25	0	1 (clinical trials)	0	0	1 (secondary use of EHRs)
26	0	0	1 (reviewing patient's history)	1 (support linked data)	1 (secondary use of EPRs)
27	0	1 (mapping phenotype - genotype associations)	1 (identify subjects with diseases)	1 (support linked data)	1 (secondary use of EHRs)
28	0	1 (finding potential drug drug interactions)	1 (finding a proper treatment)	0	1 (secondary use of EHRs)
29	0	1 (finding potential drug drug interactions)	1 (finding a proper treatment)	0	1 (secondary use of EHRs)
30	1 (decision making by the healthcare provider)	0	0	1 (support linked data)	0
31	1 (decision making in general)	1 (providing research data)	0	1 (support linked data)	1 (health services)
32	1 (decision making in general)	1 (providing research data)	0	1 (support linked data)	1 (health services)
33	1 (epidemiological planning)	1 (create longitudinal health data)	0	0	0
34	1 (communication within different healthcare services)	0	1 (pattern prediction)	1 (support linked data)	0

35	0	1 (cohort selection, phenotypic profiling)	1 (identification of disease genes)	1 (support linked data)	1 (secondary use of EHRs)
36	1 (planning accessibility control)	0	0	1 (support linked data)	0
37	1 (decision making by the healthcare provider)	0	0	0	0
38	1 (decision making in general)	0	0	1 (support linked data)	0
39	1 (decision making in general)	1 (providing research data)	0	1 (support linked data)	1 (health services)
40	0	0	0	1 (self-learning)	1 (TV programmes)
41	0	0	1 (reviewing patient's history)	0	1 (secondary use of EHRs)
42	1 (decision making by the healthcare provider)	1 (providing research data)	0	0	1 (emergency health services)

Appendix B The Traditional Linked Data Matrices

B.1 General Information

<i>No.</i>	<i>paper</i>	<i>resource</i>	<i>overview</i>	<i>specific aims</i>	<i>study type</i>	<i>study year</i>	<i>study period</i>	<i>sample size</i>
1	The utility of linked cancer registry and health administration data for describing system-wide outcomes and research: A BreastScreen example	(Buckley et al. 2016)	This paper discussed the results of linking cancer registry data with administrative data. The study resulted in an estimate of the invasive breast cancer (IBC) risk following a screen-detected ductal carcinoma in situ (DCIS).	predicting the risk of developing a disease by tracking patient's history	retrospective cohort study	1989-2010	21	9544
2	Heart Failure Following Cancer Treatment: Characteristics, Survival and Mortality of a Linked Health Data Analysis.	(Clark, R. A. et al. 2016)	This paper discusses the characteristics of the issue of heart failure in cancer's patients after exposing to chemotherapy.	predicting the risk of developing a disease by finding associations between two medical conditions (comorbidity)	retrospective study	1996-2009	13	15987
3	Data Resource Profile: Cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER)	(Denaxas et al. 2012)	This paper introduces the CALIBER project which provides the clinical research with ready-use data variables from the joint data repository consists of linked data extracted from electronic health records (EHR) from primary care , coded hospital records, social deprivation information and cause-specific mortality data. An example has been provided in the paper as well.	not related (introducing the idea of a new project)	a linked data project	2012-up to now (the project initiated)	4	901629
4	What factors contribute to positive early childhood health and development in Australian Aboriginal children? Protocol for a population-based cohort study using linked administrative data (The Seeding Success Study)	(Falster et al. 2015)	This paper studies the factor contributes to positive early childhood development in Aboriginal children in Australia.	finding success factors for a specific cohort by finding associations between various administrative information with development outcomes	retrospective cohort study	2009-2012	3	9000
5	Visualising linked health data to explore health events around preventable hospitalisations in NSW Australia	(Falster et al. 2016)	This paper investigates visualizing health services provided for participants, in order to analyse any patterns in health services use.	evaluating health outcomes (performance) by identifying and predicting health services	cohort study using static timelines	2006-2009	3	266950

				usage's patterns				
6	Explaining Health Care Expenditure Variation: Large-Sample Evidence Using Linked Survey And Health Administrative Data	(Ellis et al. 2013)	This paper attempts to predict the possible healthcare expenditure of each individual. They found that the costs are highly related to the person's age as well as to the lifestyle and unhealthy habits.	understanding health costs (expenditures) by finding associations between expenditures and patients characteristics (demographic, health conditions..)	cross-sectional survey	2006-2009	3	267188
7	HERALD (Health Economics using Routine Anonymised Linked Data)	(Husain et al. 2012)	In this paper they study the potential implications of linking 1)Patient's questionnaire 2)Routine datasets 3)Experimental data to get ONE joined data resource. This can allow them to map the patient's journey retrospectively and prospectively. Moreover, they support their work with an example for a patient with the Ankylosing Spondylitis (AS) condition.	identifying early characteristics of developing a disease by tracking patient's history	a linked data project	2009 (the example)	0	1
8	Preparing linked population data for research: cohort study of prisoner perinatal health outcomes.	(Hilder et al. 2016)	This paper aims to describe the process of linking data in the case of pregnant prisoners. They aim to study a representative population of mothers by the help of linked data.	not related (linking data for specific cohort)	retrospective cohort study (see paper "Pregnancy, prison and perinatal outcomes in New South Wales, Australia: a retrospective cohort study using linked health data")	2000-2006	6	404000
9	The health services burden of heart failure: an analysis using linked population health data-sets	(Roberts et al. 2012)	This paper aims to understand the burden of heart failure admissions on health services and to estimate the possible relationships with age and gender.	finding burden of a specific disease on a health service by finding associations between health burdens and patients characteristics (demographic, admission..)	cohort study	2000-2007	7	29161
10	A population-based study of the drug interaction between proton pump inhibitors and clopidogrel	(Juurlink et al. 2009)	This paper aims to study the drug interaction between proton pump inhibitors and clopidogrel by mining the patients sample exposed to both.	discovering potential drug-drug interactions by identifying patterns in patient information				
11	Using linked educational attainment data to reduce bias due to missing outcome data in estimates of the	(Cornish et al. 2015)	This paper aims to find the association between the duration of breastfeeding and IQ at the age of 15 by reasoning and analysing the linked data.	finding inter-dependent conditions between health domains				

	association between the duration of breastfeeding and IQ at 15 years							
--	--	--	--	--	--	--	--	--

B.2 Domain Information

No.	Health domain	Health Issue	users
1	clinical	breast cancer	?
2	clinical	blood or breast cancer, heart failure	?
3	clinical	Cardiovascular disease	health care and public health policy
4	social health (public health)	early development in children	?
5	health management (health services)	preventable hospitalization	?
6	health management (health economics)	health expenditure	policy makers
7	1) health information, 2) clinical	tracking patients' history	?
8	1) health information, 2) clinical	pregnancy in prison	researchers
9	1) health information, 2) clinical	heart failure	managers
10	1) health information, 2) clinical		
11	1) health information, 2) clinical		

B.3 Aims / Goals

No.	administration	diagnosis	pharmacology	learning
1	0	1 (finding disease's risks)	0	0
2	0	1 (finding disease's risks)	0	0
3	1 (health policies)	0	1	1 (supporting research by promoting for linked data)
4	1 (health services)(health policies)	0	0	0
5	1 (health services)	0	1	0
6	1 (health economics)	0	0	0
7	1 (health economics)	1(tracking patient's history)	1	1 (supporting research by promoting for linked data)
8	1 (health services)	1 (recognising patterns of diseases)	0	1 (supporting research by promoting for linked data)
9	1 (health services)(health economics)	0	0	1 (supporting research by promoting for linked data) ("The major strength of this study is that patient-level analyses allowed us to calculate readmission rates, median survival and importantly, quantify the strong age-related trends in incidence of disease, LOS and mortality.")
10			1	
11				

B.4 Methods / Tools

No.	data integration	data representation (modelling)	re-using resources	reasoning	querying (exploring)	data description (annotation)	data visualization
1	1	0	1	1	1	0	1
2	1	0	1	1	1	0	0
3	1	1	1	1	1	1	0
4	1	0	1	1	1	0	1
5	1	1	1	1	1	0	1
6	1	0	1	1	1	0	1
7	1	1	1	1	1	0	0
8	1	0	1	1	1	1	1
9	1	0	1	1	1	0	1
10	1		1				
11	1		1				

B.5 Data Resources

No.	no. data sources	data sources	data type	data format	Linking methods	Linking tool
1	2	administrative data, the South Australian breast cancer screening programme (BSSA)	administrative data, cancer registry	?	?	?
2	3	Hospital Admitted Patient Data Collection, Queensland Cancer Registry (QCR) facility and Unit Record, Birth, Deaths and Marriages	Administration data, Cancer Registry, Death Registry	?	probabilistic matching methods	Linkage Wiz software
3	5	1)the Clinical Practice Research Datalink (CPRD), 2)the Myocardial Ischaemia National Audit Project (MINAP), 3)Hospital Episodes Statistics (HES), 4)the Office for National Statistics (ONS) mortality, 5)social deprivation data. EACH DATA CATEGORY CONSISTS OF MANY DATA SOURCES!	1) Longitudinal primary care data Diagnoses and symptoms irrespective of hospitalization, drug prescriptions, vaccinations, blood test results, risk factors, 2)National registry of Acute Coronary Syndrome admissions Phenotype (ST Elevation Myocardial Infarction, Non-ST Elevation Myocardial Infarction, Unstable Angina), severity and treatment data, 3)National data warehouse of hospitalizations recorded for administrative purposes Inpatient, outpatient, emergency, critical care and maternity admissions Operations and surgical procedures, 4)National census of all deaths Primary and underlying cause of death, 5)Small area patient social deprivation data.	?	?(NHS number, date of birth, sex and post code)	a Trusted Third Party
4	3	various administrative data (birth outcomes, congenital conditions, hospital admissions, emergency department presentations, receipt of ambulatory mental healthcare services, use of general practitioner services, contact with child protection and out-of-home care services, receipt of income assistance and fact of death), Australian Early Development Census	administrative data, census data, birth registry	?	probabilistic matching methods, identifiers (name, date of birth, sex and address)	2 stages: 1)The Centre for Health Record Linkage (ChReL) 2)The Australian Institute of Health and Welfare (AIHW) Data Integration Services Centre,

		data, perinatal and birth registration data sets,				
5	6	1) the Department of Human Services' Medicare system (Australia's national universal health insurer), 2) the NSW Admitted Patient Data Collection (APDC), 3) all NSW public and private sector hospitals and day procedure centres, 4) the NSW Emergency Department Data Collection (EDDC), 5) the Medical Benefits Schedule (MBS), 6) the NSW Registry of Births, Deaths and Marriages (RBDM)	1) sociodemographic and health characteristics of participants questionnaire, 2) Hospitalisations, 3) census of all hospital separations (discharges, transfers and deaths), 4) Emergency Department Data, 5) Medicare-funded claims for GP and specialist medical practitioner consultations, 6) Death Records	?	probabilistic matching methods	CHeReL using ChoiceMaker software
6	2	1) the Medicare Australia enrolment database, 2) annual healthcare costs calculated from several years of hospital, medical and pharmaceutical records, 3) NSW Admitted Patient Data Collection, 4) NSW Emergency Department presentation data, 5) Medical Benefits Schedule (MBS) data, 6) Pharmaceutical Benefits System (PBS) data,	1) cross-sectional survey 2) panel dataset for annual costs 3,4,5,6) Administrative data	?	probabilistic matching methods (first name, surname, date of birth and address.)	the Centre for Health Record Linkage (CHeReL), the Sax Institute,
7	3	1) questionnaire about the number of visits for healthcare services 2) the Secure Anonymised Information Linkage (SAIL) databank 3) the Welsh population-based ankylosing spondylitis (PAS) cohort.	1) questionnaire 2) routine data (clinical data) 3) questionnaire + routine data	?	probabilistic matching methods (NHS number or a mixture of forename, surname, gender, postcode, and date of birth)	Matching Algorithm for Consistent Results in Anonymised Linkage (MACRAL), SAIL
8	5	1) The Offender Integrated Management System (OIMS) 2) The Perinatal Data Collection (PDC) 3) The Admitted Patient Data Collection (APDC) 4) The Pharmaceutical Drugs of Addiction System (PDAS) 5) The Register of Congenital Conditions (RoCC).	1) prisoner location and transfer history, classification, security, self-harm, demographics, and biometric identification. 2) patterns of pregnancy care, childbirth and newborn outcomes. 3) administrative data from public and private health services (patient demographics, procedures and diagnoses). 4) the therapeutic substance, the prescriber, and patient demographics. 5) notifications of structural and chromosomal conditions diagnosed during pregnancy and 12 months after birth (name and address details for the mother and the child).	?	probabilistic matching methods (person's identifier)	The NSW Centre for Health Record Linkage (CHeReL)
9	2	1) the NSW Admitted Patient Data Collection (APDC) 2) the NSW Registry of Births, Deaths and Marriages (RBDM)	1) private and public hospital separations (administrative data) 2) death registrations (administrative data)	? (available for public)	probabilistic matching methods (patients' names and other identifiers)	the Centre for Health Record Linkage using ChoiceMaker software

10					not mentioned	
11						

B.6 Challenges

No.	<i>mismatching linkage (conflicting data)</i>	<i>data ambiguity (data quality)</i>	<i>limited access to data</i>
1	0	1	0
2	1	0	1
3	1	1	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	1	0	1
9	1	1	0
10			
11			

B.7 Strengths

No.	<i>large sample size</i>	<i>cost and time effectiveness</i>	<i>higher resolution approach(multiple sources)(bigger picture)</i>	<i>assessing real-world services</i>	<i>identifying population-level patterns (trends, relationships)</i>
1	0	1	0	0	0
2	1	0	1	0	0
3	1	0	1	0	0
4	1(maximises statistical power, increasing the 'visibility', to explore geographic variation)	0	1	1	0
5	1	0	0	1	1
6	1	0	0	0	1
7	0	1	1	0	0
8	1	0	1	1	0
9	0	0	1	1	1
10					
11					

Appendix C The Semantic Linked Data Matrices

C.1 General Information

<i>No.</i>	<i>paper</i>	<i>year</i>	<i>overview</i>	<i>domain</i>	<i>specific aims</i>
1	*Applying semantic web technologies for phenome-wide scan using an electronic health record linked Biobank	2012	This paper discusses how to identify subjects with specific diseases (Type 2 Diabetes Mellitus (T2DM) or Hypothyroidism) and phenotypes by identifying genotype-phenotype associations through the use of the semantic web technology applied to the clinical data in the form of EHRR.	Medical	identifying subjects with specific disease by mapping phenotype - genotype associations
2	*Using Semantic Web Technologies for Cohort Identification from Electronic Health Records for Clinical Research	2012	This paper discusses the potential of using the semantic web for identifying subjects with specific diseases (Diabetes Mellitus) in cohort studies.	Medical	identifying subjects with specific disease by mapping phenotype - genotype associations
3	Enhancing Automated Decision Support across Medical and Oral Health Domains with Semantic Web Technologies	2014	This paper aims to integrate patient's information from two domains: general and oral health to support decision.	Medical	finding inter-dependent conditions between two health domains by reasoning over cross-domain patient's information
4	Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts.	2013	This paper proposes a method to classify patients by the risk of developing colorectal cancer using their (EHRs).	Medical	classifying patients with the risk of developing a certain disease (identifying patient cohorts) by using phenotyping algorithm
5	*Applying linked data principles to represent patient's electronic health records at Mayo clinic: a case report	2012	This paper discusses how to convert the EHRs in Mayo clinic from relational database form to virtual RDF form.	Pharmacology	discovering potential drug-drug interactions by mining patient information
6	*Using linked data for mining drug-drug interactions in electronic health records	2013	This paper aims to find potential drug drug interactions (PDDI) in Mayo clinic's electronic health records (EHRs) for prescribed cardiovascular and gastroenterology drugs.	Pharmacology	discovering potential drug-drug interactions by mining patient information
7	*Mining Anti-coagulant Drug-Drug Interactions from Electronic Health Records Using Linked Data	2013	This paper aims to find potential drug drug interactions (PDDI) in Mayo clinic's electronic health records (EHRs) for prescribed cardiovascular and gastroenterology drugs. (same as above)	Pharmacology	discovering potential drug-drug interactions by mining patient information
8	Towards Linked Vital Registration Data for Reconstituting Families and Creating Longitudinal Health Histories	2014	The aim for this project is to create a knowledge base for historical information for individuals such as birth, marriage, death records. This will help the historians (researchers) to question the official reports for maternal and infants mortality	Social	epidemiological planning by analysing longitudinal historical data

			rates the data in depth(longitudinal data).		
9	Mining Electronic Health Records using Linked Data	2015	This paper aims to transform the data in the Stanford's STRIDE database to a semantic version to illustrate basic cohort selection, phenotypic profiling, and identification of disease genes.	Medical	selecting patient cohort by phenotyping and identifying genes

C.2 Domain Information

No.	Health domain	Health sub-domain(systems)	Health Issue	users
1	clinical (diagnosis)	genome-wide association studies (GWAS), electronic health records (EHRs)	Type 2 Diabetes and Hypothyroidism, genetics	academic medical centers
2	clinical (diagnosis)	genome-wide association studies (GWAS)	Diabetes Mellitus, genetics	academic medical centers
3	clinical (diagnosis)	decision support system	general and oral health	healthcare practitioners
4	clinical (diagnosis)	electronic health records(EHRs), decision system	cancer	doctors, decision makers
5	clinical (diagnosis)	genome-wide association studies (GWAS), electronic health records (EHRs)	type 2 diabetes, genetics	academic medical centers
6	pharmacology (drugs interactions)	electronic health records (EHRs), genetics, potential drug-drug interactions(PDDIs), decision support system	cardiovascular and gastroenterology drugs	biomedical researchers
7	pharmacology (drugs interactions)	electronic health records (EHRs), genetics, PDDI, decision support system	anti-coagulant drugs	biomedical researchers
8	clinical (research)	health history	maternal and infant mortality rates	historians, researchers
9	clinical (research)	electronic health records (EHRs)	genetics	researchers

C.3 Aims / Goals

No.	support research	support diagnosis	support learning	support e-health systems
1	1 (mapping phenotype -genotype associations)	1 (identify subjects with disease)	1 (support linked data)	1 (secondary use of EHRs)
2	1 (mapping phenotype -genotype associations)	1 (identify subjects with disease)	1 (support linked data)	1 (secondary use of EHRs)
3	1 (finding associations between oral health and general health)	0	1 (support linked data)	0
4	1 (mapping phenotypes)	1 (identify subjects with disease)	1 (support linked data)	1 (secondary use of EHRs)
5	1 (mapping phenotype -genotype associations)	1 (identify subjects with disease)	1 (support linked data)	1 (secondary use of EHRs)
6	1 (finding potential drug drug interactions)	1 (finding a proper treatment)	0	1 (secondary use of EHRs)
7	1 (finding potential drug drug interactions)	1 (finding a proper treatment)	0	1 (secondary use of EHRs)
8	1 (create longitudinal health data)	0	0	0
9	1 (cohort selection, phenotypic profiling)	1 (identification of disease genes)	1 (support linked data)	1 (secondary use of EHRs)

C.4 Methods / Tools

<i>data integration</i>	<i>data representation (modelling)</i>	<i>re-using resources</i>	<i>reasoning</i>	<i>querying (exploring)</i>	<i>data description (annotation)</i>	<i>data visualization</i>
1	1	1	1	1	0	0
1	1	1	1	1	0	0
1	1	1	1	1	0	0
1	1	1	1	1	0	0
1	1	1	1	1	0	0
1	1	1	1	1	0	0
1	1	1	1	1	0	0
1	1	1	0	1	1	0
1	1	1	1	1	1	0

C.5 Data Resources

<i>No.</i>	<i>no. data sources</i>	<i>data type (form)</i>	<i>data sources</i>	<i>sample size</i>	<i>data format</i>	<i>transformation technique</i>	<i>Access to data sources</i>
1	3	1) an integrated resource for various patient's data , 2) a large cohort of Mayo Clinic patients with clinical data (linked via their EHRs) and genotype data 3)biomedical terminologies and ontologies 4)extending the biomedical terminologies and ontologies	1)the Mayo Clinic Life Sciences System (MCLSS), which contains patient demographics, diagnoses, hospital, laboratory, flowsheet, clinical notes, and pathology data obtained from multiple clinical and hospital source systems within Mayo Clinic 2)Mayo Genome Consortia (MayoGC) 3)Translational Medicine Ontology (TMO), Sequence Ontology (SO) 4) the Ontology for Biomedical Investigations, Prostate Cancer Ontology, NCI Thesaurus, SNOMED CT	6307	1) relational database, CSV, Excel, TAB 2)relational database 3) RDF	R2RML, spyder	0,1
2	2	1) an integrated resource for various patient's data 2)biomedical terminologies and ontologies	1) the Mayo Clinic Life Sciences System (MCLSS), 2)mainly Translational Medicine Ontology (TMO)is used along with many of the Linked Open Drug Data (LODD)(DrugBank, LinkedCT, DailyMed, DBPedia, Diseaseome, RDF-TCM, RxNorm, SIDER, STITCH, ChEMBL, WHO Global Health Observatory, Medicare)	0	CSV, TAB, Microsoft Excel files, RDF	virtuoso	0,1
3	3	1) collected scientific literature, 2) experts views, 3)terminology and thesauri	1)various 2)a general practitioner and two dental surgeons 3)Systematised Nomenclature of Medicine (SNOMED-CT)	0	OWL, ?	0	0,1
4	6	1) medical encyclopedias, clinical guidelines 2)Terminological resources, 3) Archetype repositories 4) Source EHR schemas and data 5) Repositories of ontologies 6) Phenotyping archetype	1) openEHR Clinical Knowledge Manager 2)SNOMED CT 3) OBSERVATION.lab_test-histopathology, openEHR-EHR-EVALUATION.colorectal_screening.v1 4)? 5) The colorectal-domain ontology 6) ?	0	?, XML	?	?

5	4	1) patients records (EHRs) 2)terminology 3)mapping ontologies (from DB to RDF) 4)external biomedical datasets (LODD)	1) The Mayo Clinic Life Sciences System (MCLSS) (Medical Revenue Information Systems (MRIS), HealthQuest, Master Patient Identification Information (MPII)/Registration, Mayo Integrated Clinical Systems (MICS), Clinical Notes, Pathology Reports) 2) SNOMED CT , NCI thesaurus 3) Translational Medicine Ontology (TMO), the Ontology for Biomedical Investigations, Prostate Cancer Ontology 4)LODD (e.g. DrugBank, LinkedCT, DailyMed, DBPedia, Diseasesome, RDF-TCM, RxNorm, SIDER, STITCH, ChEMBL, WHO Global Health Observatory, Medicare)	millions	RDF,CSV, TAB or Excel files, XML	Virtuoso Universal Server	0,1
6	2	1) patient clinical and demographic data (EHRs) 2)a public drug data repository.	1) The Mayo Clinic Life Sciences System (MCLSS) 2) DrugBank	6758 patients	are available as RDF , but not origannaly RDF (relational database, text..)	Virtuoso Universal Server foMCLSS, Bio2RDF for DrugBank	0,1
7	4	1) patient clinical and demographic data (EHRs) 2) patient clinical information and genomics 3)a public drug data repository 4) terminology	1) The Mayo Clinic Life Sciences System (MCLSS) 2) Mayo's Electronic Medical Record and Genomics (eMERGE) 3) DrugBank 4) SNOMED CT	6758 patients	are available as RDF , but not origannaly RDF (relational database, text..)	Virtuoso Universal Server foMCLSS, Bio2RDF for DrugBank	0,1
8	6	1) birth, death and marriage records (BDM) 2)LOD 3)parish records for certain countries 4) legacy datasets from other funded research projects 5) census data (1901-1911) 6) information on Irish place names and street level information for Dublin.	1) the General Register Office (GRO) records in Dublin (1864-1913) 2) ? (model) 3) the Irish Genealogy platform 4) the Online Historical Population Reports 5) the National Archives of Ireland 6) Logainm	0	?, RDF	manually	1
9	3	1) clinical records (EHRs) 2)biomedical Linked Datasets 3)standard health care dictionaries	1) central repository for EHR data from the Lucile Packard Children's Hospital and Stanford Hospital and Clinics (STRIDE) 2)Online Mendelian Inheritance in Man (OMIM), SIDER 3) SNOMED-CT, ICD9, RxNORM, LOINC	0	MySQL, RDF	Virtuoso 7.1.0	0,1

C.6 Technologies

No.	ontologies	metadata	URI	RDFS	RDF	OWL-S	OWL	SWRL	SPARQL
1	1	0	0	1	1	0	1	0	1
2	1	0	1	1	1	0	1	0	1
3	1	0	0	0	1	0	1	1	1
4	1	0	0	0	0	0	1	0	1
5	1	0	0	1	1	0	1	0	1
6	1	0	1	1	1	0	1	0	1
7	1	0	1	1	1	0	1	0	1
8	1	1	1	1	1	0	0	0	1
9	1	1	1	0	1	0	1	0	1

C.7 Weaknesses

No.	query performance	scalability of reasoning engines	lack of data coverage
1	1	0	1 (TMO didn't cover all the needed medical terminologies, so they had to extend it)
2	1	0	1
3	0	0	0
4	0	1	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
9	1	0	0

C.8 Strengths

adequate Sample size (more accessible data) (utilising data on the web)	flexible representation (simple data model)	formal specification for domain knowledge	accuracy	accelerate scientific findings	incorporation between public and private datasets
1	1	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	1	1	0	0	0
0	0	0	1	1	1
0	1	0	1	1	1
0	0	0	0	1	1
1	0	0	0	0	0
1	1	1	0	1	1 (incorporation between clinical (EHRs) and biomedical data (published ontologies in LOD))

Appendix D The Thematic Matrix of the Systematic Review for the Semantic Web's Uses in Health Research

<i>no.</i>	<i>reference</i>	<i>aim</i>	<i>topic</i>	<i>detailed topic</i>
1	Abas, 2011 #10043	medical	diagnosis	clinical decision support system (CDSS) / acute postoperative pain management
2	Addy, 2015 #9021	public	awareness	food and exercise recommendation
3	Ahire, 2015 #10419	public	awareness	patient education (Breast Cancer)
4	Ahmed Benyahia, 2013 #10354	medical	diagnosis	telemonitoring / decision support system./ auscultation sounds
5	Ahmed Benyahia, 2013 #10638	medical	diagnosis	telemonitoring / decision support system./ auscultation sounds
6	Ajigboye, 2016 #10067	medical	monitoring	wearable sensors / transducers electronic data sheets / Health Information Systems (HIS) - monitoring
7	Al Manir, 2018 #9162	public	epidemiology	surveillance of malaria analytics - epidemiology
8	Al-Abad, 2009 #10423	public	awareness	food, nutrition and health
9	Albukhitan, 2013 #10406	public	awareness	food, nutrition and health
10	Alexandrou, 2008 #10297	clinical	care plans	clinical pathways
11	Alexandrou, 2012 #8970	clinical	care plans	clinical pathways
12	Al-Hamadani, 2014 #10325	medical	diagnosis	6 diagnosing coronary artery diseases
13	Ali, 2009 #10028	medical	examination	8 ambient intelligent medical devices / Pheochromocytoma and/or Neuroblastoma tumors
14	Ali, 2018 #8886	public	awareness	recommendation systems
15	Al-Nazer, 2012 #10555	public	awareness	food, nutrition and health
16	Al-Nazer, 2013 #10405	public	awareness	food, nutrition and health
17	Al-Nazer, 2014 #10404	public	awareness	patient education (Vaccine Information)
18	Al-Nazer, 2015 #10063	public	awareness	food, nutrition and health
19	Al-Nazer, 2016 #10397	public	awareness	food, nutrition and health
20	Alobaidi, 2018 #9040	medical	diagnosis	diagnosis and treatment - Knowledge extraction for clinical documentation
21	Alsomali, 2016 #9160	pharm	adverse events	Adverse drug events (ADEs)
22	Alsulmi, 2017 #10453	medical	diagnosis	Clinical Decision Support Systems (CDS)
23	Amith, 2015 #9120	public	assistive	physical training for elderly
24	Amith, 2016 #10193	public	awareness	Vaccines / human papillomavirus (HPV) vaccine / Improving patients knowledge
25	Andreasik, 2011 #10376	clinical	learning	10 controlling the correctness of medical procedures
26	Arch-Int, 2011 #10042	medical	diagnosis	to capture the complete clinical history of a patient EPR
27	Ardestani, 2012 #10551	medical	treatment	11 medical informatics / Bariatric Surgery / obesity - treatment
28	Argüello Casteleiro, 2009 #9221	clinical	guidelines	medical guidelines - EHRs
29	Argüello, 2008 #10084	clinical	guidelines	medical guidelines - EHRs
30	Argüello, 2009 #10012	clinical	guidelines	medical guidelines - EHRs
31	Argüello, 2009 #10133	clinical	guidelines	medical guidelines - EHRs
32	Argüello, 2011 #10621	medical	diagnosis	12 psychiatry / content based analysis - diagnosis

33	Arikuma, 2007 #10464	pharm	adverse events	Drugs interaction (DDIs)
34	Assele Kama, 2012 #10642	medical	diagnosis	13 to share and query clinical heterogeneous data - diagnosis - detecting bacteria
35	Bajenaru, 2015 #8919	clinical	learning	training - education
36	Baldassini, 2017 #10623	public	assistive	independent living support for elderly/disabled
37	Bamidis, 2013 #10014	medical	monitoring	Exergames as health monitoring tools
38	Bamparopoulos, 2016 #9118	public	assistive	Travelling plan for elderly
39	Batouche, 2012 #10549	public	assistive	Independent lifestyle for elderly
40	Becnel, 2017 #9197	clinical	EHR	
41	Benmimoune, 2016 #10355	medical	diagnosis	16 Clinical decision support systems (CDSSs) / assist clinicians
42	Benton, 2008 #10539	public	social behaviour	Health counselling for health behaviour change
43	Benyahia, 2012 #10510	medical	monitoring	17 telemonitoring and alerts detection / chronic disease
44	Berges, 2010 #10596	medical	diagnosis	to interpret on the fly data about medical observations - present in EHRs
45	Besana, 2009 #9193	clinical	guidelines	medical guidelines
46	Bhat, 2007 #10420	pharm	drug discovery	Drugs discovery
47	Bhat, 2016 #9011	medical	diagnosis	19 prediction of Flu / Data Extraction - diagnosis
48	Bhattacharjee, 2015 #10211	medical	treatment	Vaccines / To track vaccines using barcodes packaging
49	Bibault, 2018 #9272	medical	examination	21 radiation oncology - medical tools
50	Bickmore, 2011 #9092	public	assistive	Seniors health self-tracking
51	Billis, 2015 #10521	public	social behaviour	Suicide sentiment prediction
52	Birjali, 2017 #10389	public	awareness	Reviewing Food ontologies
53	Bonte, 2016 #10189	clinical	workflow	workflow
54	Borges, 2017 #10184	medical	causes	22 Organisms - medical studies - biological data
55	Boulos, 2015 #8942	public	epidemiology	Epidemiology
56	Bouslimi, 2017 #9310	medical	examination	23 medical image / a radiological collaborative social network
57	Boyce, 2010 #10348	pharm	adverse events	Drugs interaction/ adverse events
58	Bratsas, 2010 #10347	clinical	learning	Health education and health care
59	Briand, 2018 #9227	public	epidemiology	malaria - epidemiology
60	Calbimonte, 2017 #9064	medical	monitoring	25 wearable sensor data / diabetes monitoring
61	Calderone, 2017 #10160	medical	causes	26 disease similarities - literature - research - study biological phenomenon
62	Calvillo, 2013 #9033	public	awareness	Safety in food consumption
63	Calvillo, 2014 #10346	clinical	workflow	health alarms / unawareness and fatigue of health professionals / workflow
64	Calvillo-Arbizu, 2014 #10345	medical	treatment	clinical decision in prescribing
65	Cappellari, 2017 #10583	medical	treatment	28 discovering novel treatments
66	Cardillo, 2015 #10206	medical	diagnosis	the coding of health conditions in the patient summary
67	Carro, 2003 #10617	medical	examination	medical images
68	Casteleiro, 2008 #9222	clinical	guidelines	30 diabetic retinopathy - guidelines
69	Casteleiro, 2017 #10182	clinical	guidelines	31 diabetic retinopathy - guidelines
70	Ceccarelli, 2008 #10511	medical	diagnosis	33 Clinical Decision Support Systems (CDSSs) / oncology
71	Ceccarelli, 2009 #10565	medical	diagnosis	32 Clinical Decision Support Systems (CDSSs) / oncology

72	Çelebi, 2013 #10228	medical	diagnosis	34 SNP-disease / genetic variation / cause human diseases / genotype and phenotype - diagnosis
73	Celebi, 2017 #10169	pharm	adverse events	Drugs interaction (DDIs)
74	Çelik Ertuğrul, 2016 #9174	public	awareness	Awareness of diabetic foot
75	Çelik, 2014 #10477	public	awareness	Safety in food consumption
76	Celik, 2014 #10478	medical	monitoring	35 paediatric consultation and monitoring
77	Cerizza, 2006 #10287	medical	treatment	patients summaries - treatment
78	Ceusters, 2006 #9333	medical	diagnosis	36 schizophrenic patients / treatment / diagnostic
79	Chammas, 2013 #10589	public	awareness	Awareness of anti-hypertension drugs
80	Chapman, 2010 #9201	public	epidemiology	Standardized surveillance syndromes / Respiratory, gastrointestinal, constitutional, and influenza-like illness (ILI)
81	Chen, 2006 #10472	clinical	guidelines	38 radiation protection guidelines
82	Chen, 2012 #8933	medical	treatment	treatment
83	Chen, 2012 #8961	public	awareness	Road traffic injuries / road safety
84	Chen, 2016 #9172	clinical	learning	Diabetes health education
85	Chondrogiannis, 2012 #10120	clinical	EHR	automated patient selection for clinical trials / clinical research with healthcare /
86	Chondrogiannis, 2017 #9072	clinical	EHR	clinical research - selecting patients for clinical trials
87	Chondrogiannis, 2018 #10476	clinical	EHR	automated patient selection for clinical trials
88	Ciccarese, 2008 #9101	medical	treatment	42 neurodegenerative disorders / Developing cures / Alzheimer's - treatment
89	Cipière, 2014 #10415	public	epidemiology	large-scale epidemiology analysis / epidemiological studies / Medical data integration
90	Colacino, 2017 #10103	clinical	guidelines	Policy making for public health
91	Corrigan, 2012 #10243	medical	diagnosis	44 diagnostic decision support tools
92	Couch, 2011 #10489	public	epidemiology	the infestation of <i>Aedes aegypti</i> (dengue vector) / epidemiology
93	Courtot, 2013 #10235	medical	diagnosis	45 public health / adverse events following immunization / Diagnostic criteria and clinical guidelines
94	Cureí, 2013 #10227	pharm	adverse events	pharmacovigilance / text-mining clinical notes from EHRs
95	Dang, 2008 #9103	clinical	workflow	workflow
96	Dang, 2009 #10321	clinical	workflow	workflow
97	Daniulaityte, 2015 #8816	pharm	drug discovery	Drugs uses/ drug abuse research
98	Dao, 2013 #9087	clinical	learning	learning - Musculoskeletal System
99	Das, 2006 #8830	clinical	care plans	participation of the patient in the treatment plan
100	Dasmahapatra, 2006 #9024	medical	diagnosis	47 breast cancer screening / the diagnostics of imaging modalities
101	De Farias, 2017 #10157	medical	causes	48 orthology / evolutionary genetics - research
102	De Maio, 2011 #10334	medical	diagnosis	49 automatic disease diagnosis / medical decision making / dermatological diseases
103	De Mendonça, 2014 #10110	public	epidemiology	the infestation of <i>Aedes aegypti</i> (dengue vector) / epidemiology
104	De Mendonça, 2015 #9002	public	awareness	Health learning
105	De Potter, 2012 #8891	medical	diagnosis	50 medical decision making / patient information aggregation
106	Del Carmen Legaz-García, 2012 #10645	public	assistive	Monitoring elderly conditions remotely
107	Dessi, 2017 #10170	medical	diagnosis	medical reports - diagnosis
108	Deus, 2012 #10122	pharm	drug testing	the reporting of experimental context and results of gene expression studies - research
109	Deus, 2012 #9089	pharm	drug testing	the reporting of experimental context and results of gene expression studies - research
110	Dhombres, 2011 #10251	medical	treatment	53 rare diseases and orphan drugs - treatment
111	Dieng-Kuntz, 2004 #10377	medical	diagnosis	54 Virtual Staff / a cooperative diagnosis

112	Dingli, 2008 #10267	medical	monitoring	55 patient-centric health care services / direct staff effectively / monitors the patients
113	Dingli, 2014 #9016	public	assistive	Elderly nutrition
114	Dong, 2015 #10400	clinical	care plans	patient's circle of care
115	Doore, 2010 #10349	medical	monitoring	56 personal exposure history / Health monitoring and disease surveillance
116	Douali, 2012 #10520	medical	diagnosis	58 Genomics / personalized medicine / clinical decision support system (CDSS)
117	Douali, 2012 #10646	medical	diagnosis	57 Clinical Diagnosis / Clinical Decision Support Systems / Urinary Tract Infection
118	Douali, 2013 #10331	pharm	adverse events	Adverse drug events (ADEs)
119	Doucette, 2012 #10353	medical	diagnosis	59 medical decision support system
120	Dridi, 2018 #10508	medical	monitoring	60 Public personalized healthcare monitoring
121	Dridi, 2018 #10527	medical	monitoring	61 Public personalized healthcare monitoring
122	Duclos, 2007 #8828	pharm	drug discovery	antibacterial spectra / the antibiotic susceptibility - discovery
123	Dupplaw, 2009 #8806	medical	examination	63 Medical Imaging
124	Eccher, 2013 #8899	medical	diagnosis	64 cancer therapies / oncologic Electronic Patient Record / decision support system
125	Echeverría, 2015 #10398	clinical	workflow	65 provisioning of health services / respiratory chronic disease - business process
126	Eholié, 2016 #10192	public	social behaviour	breast cancer - public opinions - behaviours
127	Elkader, 2018 #9007	medical	diagnosis	67 chronic kidney disease diagnosis
128	El-Sappagh, 2018 #8867	medical	diagnosis	clinical decision support systems - General Medical Science
129	El-Sappagh, 2018 #9109	medical	diagnosis	68 diabetes mellitus treatment / clinical decision support system (CDSS)
130	El-Subaihi, 2013 #10678	medical	treatment	69 Child cancer - treatment plan
131	Ertugrul, 2017 #10497	medical	monitoring	70 Tracking System / Acute Respiratory Tract Infection / monitoring
132	Eshghishargh, 2018 #10101	medical	diagnosis	71 answering questions in neuroinformatics - research
133	Espín, 2016 #8932	public	epidemiology	Epidemiology / review
134	Falkman, 2007 #10131	medical	diagnosis	evidence-based medicine (EBM) - oral medicine
135	Falkman, 2008 #9166	medical	diagnosis	evidence-based medicine (EBM) - oral medicine
136	Faro, 2010 #10606	medical	examination	72 Public telemedicine system / healthy lifestyles and self-care - medical tools
137	Fenza, 2011 #10502	medical	monitoring	73 Pub Sensors / prophylactic and follow-up monitoring of patients
138	Fernández-Breis, 2013 #9199	medical	diagnosis	74 the identification of patient cohorts / colorectal cancer screening - diagnosis
139	Fisher, 2016 #9115	medical	diagnosis	75 dermatologic disease - diagnosis
140	Florczyk, 2010 #8920	pharm	drug testing	anti-microbial use
141	Forbes, 2014 #8883	clinical	workflow	communication with minority patients - medical consultation
142	Galán-Mena, 2016 #10070	medical	diagnosis	76 diagnosis / Autism Spectrum Disorders
143	Gangwar, 2012 #10046	clinical	care plans	medical health planning
144	Gemmeke, 2014 #10216	medical	examination	medical images
145	Gerhold, 2010 #10603	clinical	guidelines	guidelines and checklists - medical malpractice
146	Gordon, 2013 #8818	public	awareness	hypersensitivity disorder allergy / decision support system / awareness
147	Goynugur, 2018 #10455	clinical	guidelines	guidelines and policies
148	Grosjean, 2012 #10245	clinical	learning	education
149	Gudivada, 2008 #9098	medical	causes	77 Identifying disease-causal genes / genomic and phenomic knowledge / genomic studies
150	Hadzic, 2004 #10327	medical	causes	78 human disease research study

151	Hadzic, 2008 #10480	medical	diagnosis	79 mental health - diagnosis
152	Haider, 2015 #10441	medical	examination	80 medical laboratory data - medical tools
153	Halim, 2018 #8960	public	awareness	Nutrients / phytochemical
154	Hamiz, 2018 #9195	public	awareness	Public health awareness / emergency response
155	Handayani, 2010 #10411	public	awareness	Female-related diseases - awareness
156	Harris, 2007 #8895	public	awareness	Patient instructions
157	Hayuhardhika, 2013 #10409	medical	diagnosis	83 diagnosing Diabetes Mellitus
158	Heimonen, 2012 #10250	public	awareness	food, nutrition and health
159	Helmy, 2015 #9062	public	epidemiology	infectious disease epidemiology
160	Helmy, 2016 #9012	public	awareness	food, nutrition and health
161	Hogan, 2016 #9114	public	assistive	Elderly Well-being monitoring
162	Horst, 2011 #10035	public	epidemiology	environmental data in cancer-related risk studies
163	Hřebíček, 2012 #10338	public	awareness	Personal Health Recommender
164	Hu, 2012 #9182	clinical	care plans	clinical pathways
165	Hu, 2016 #10560	clinical	learning	Health learning
166	Huang, 2014 #9177	clinical	care plans	clinical pathways
167	Huang, 2016 #9060	clinical	care plans	Nutrition of chronic disease patients
168	Huang, 2017 #10452	medical	diagnosis	84 chronic obstructive pulmonary disease and lung cancer - diagnosis
169	Hulse, 2013 #8819	public	awareness	Public personalized education materials / pre-diabetes and metabolic syndrome- awareness
170	Hurley, 2007 #10495	clinical	care plans	clinical pathways
171	Hussain, 2012 #10246	clinical	EHR	clinical research - selecting patients for clinical trials - EHRs
172	Hussain, 2012 #10479	medical	diagnosis	86 urinary tract infections - diagnosis
173	Iftikhar, 2011 #10038	medical	diagnosis	87 Informatics Patient Registration Scenario - discover diseases - diagnosis
174	Iram, 2011 #10466	public	epidemiology	predict the spread of disease / epidemiology
175	Iskandar, 2015 #10059	medical	examination	89 Cardiac MRI in heart attack patients / Health Information System - medical tools
176	Islam, 2012 #10552	public	awareness	diabetes patients - awareness
177	Islam, 2013 #10671	public	awareness	diabetes patients - awareness
178	Ivaşcu, 2018 #10099	public	awareness	Health and exercise advising
179	Izumi, 2006 #10532	public	awareness	Awareness of healthcare
180	Jabbar, 2017 #9336	medical	monitoring	92 monitoring devices / health status
181	Jain, 2016 #10533	medical	diagnosis	93 human family tree - diagnosis
182	Jentzsch, 2009 #10261	clinical	care plans	tailored therapeutics / Enabling tailored therapeutics/ connect drug and clinical trials related data sources
183	Ji, 2013 #10052	public	awareness	Awareness of healthcare
184	Ji, 2014 #9031	public	awareness	Awareness of national health trends
185	Ji, 2017 #9151	clinical	learning	Diabetes education
186	Jiang, 2011 #8821	pharm	adverse events	Adverse Drug Events (ADEs) / phenotypes related to ADEs
187	Jiang, 2016 #9066	clinical	EHR	94 clinical study / health care and clinical research
188	Jiménez, 2017 #10175	clinical	learning	96 Medical Training - education
189	Jin, 2007 #10496	medical	examination	medical image
190	Jing, 2007 #10662	medical	causes	98 biological and clinical information / Cystic Fibrosis exemplar - research
191	Jing, 2007 #9234	medical	causes	97 biological and clinical information / Cystic Fibrosis exemplar - research
192	Jing, 2014 #9178	medical	diagnosis	decision - molecular genetic knowledge - EHR
193	Judkins, 2018 #9107	pharm	adverse events	Drugs interaction (DDIs) / pharmacokinetic-based natural product-drug interactions (PK-NPDIs)
194	Jupp, 2011 #9128	medical	diagnosis	99 kidney and urinary pathway / Chronic renal disease / diagnosis
195	Kaddari, 2016 #9028	clinical	workflow	workflow
196	Kamada, 2017 #10174	medical	treatment	100 Genomic Variants - treatment
197	Kamel Boulos, 2006 #8894	public	social behaviour	vaccine hesitancy for public health

198	Kang, 2017 #9322	public	awareness	Personal health advising
199	Karim, 2017 #10151	medical	diagnosis	102 Scientific Experiments / Dielectric Measurements of Biological Tissues - research
200	Kataria, 2010 #10591	medical	diagnosis	diagnosis - comprehensive picture of a patient's health status
201	Kergosien, 2017 #10530	medical	examination	103 histopathological image /
202	Khoozani, 2010 #10301	medical	diagnosis	104 Human stress / mental health - diagnosis - treatment
203	Kim, 2007 #10276	medical	diagnosis	105 heart disease detection / detection system - diagnosis
204	Kim, 2010 #10032	public	assistive	Health and environment monitoring for elderly
205	Kim, 2014 #10221	public	awareness	Overweight and obesity / exercise / marathon
206	Kim, 2015 #10498	public	assistive	Elderly monitoring at home
207	Kishino, 2014 #10112	clinical	learning	Health education
208	Kohonen, 2013 #9245	pharm	drug testing	testing on animals
209	Kolias, 2015 #10595	medical	diagnosis	106 carotid atherosclerosis / clinical diagnosis / Health Information Systems (HIS)
210	Konstantinidis, 2013 #9238	public	awareness	Promoting health for chronic conditions
211	Kontotasiou, 2011 #10487	clinical	learning	107 medical interventions / healthcare practice and education
212	Koutkias, 2014 #10225	pharm	adverse events	drug safety research / pharmacovigilance
213	Kozák, 2013 #10232	medical	treatment	Summaries of product characteristics (SPCs) /for better navigation in summaries of product characteristics
214	Krishnamurthy, 2016 #10561	medical	diagnosis	109 mental health and behavioural disorders / identification / Addiction - diagnosis
215	Krummenacher, 2007 #10494	medical	treatment	patients summaries - treatment
216	Krummenacher, 2009 #8893	medical	treatment	patients summaries - treatment
217	Kudama, 2017 #10179	medical	diagnosis	110 human phenotypic abnormalities - research
218	Kwon, 2014 #10281	medical	monitoring	111 sleep management / sensor- monitoring
219	Lasorsa, 2016 #10065	public	epidemiology	epidemiological indicators
220	Lavigne, 2013 #10236	public	awareness	Safety in construction sites
221	Le, 2014 #8850	public	assistive	Self-care ability
222	Lee, 2009 #8823	medical	diagnosis	diagnosis - cancer
223	Lee, 2013 #9181	public	awareness	Personalized healthcare recommender
224	Legaz-García, 2015 #9198	clinical	EHR	114 secondary use of clinical data / colorectal cancer - research
225	Legaz-García, 2016 #9219	clinical	EHR	113 secondary use of clinical data / colorectal cancer - research
226	Leroux, 2012 #10247	medical	causes	115 longitudinal clinical trial data / neurodegenerative diseases / Alzheimer's disease
227	Li, 2014 #10500	public	epidemiology	Epidemic management
228	Liaw, 2016 #9153	clinical	workflow	clinical research - workflow
229	Lima, 2017 #10388	medical	diagnosis	116 General guidelines / decision-making
230	Lin, 2015 #9157	medical	diagnosis	diagnosis - tools
231	Lin, 2017 #10457	medical	diagnosis	117 Emotions - diagnosis
232	Lô, 2017 #10167	medical	treatment	118 African Traditional Medicine / conventional medicine - treatment
233	Lopes, 2011 #10367	medical	causes	119 human genome - research
234	López-García, 2010 #10597	medical	monitoring	120 home tele-monitoring systems / chronic diseases
235	Lossio-Ventura, 2017 #10456	medical	treatment	121 obesity and cancer - treatment
236	Lossio-Ventura, 2018 #8868	medical	treatment	121 obesity and cancer - treatment
237	Mabotuwana, 2008 #10492	medical	treatment	123 hypertensive patients / nonadherence to prescribed medication - treatment
238	Maghsoud-Lou, 2017 #9171	clinical	workflow	workflow
239	Malas, 2017 #10159	pharm	drug discovery	Drug Repurposing

240	Malhotra, 2015 #9275	pharm	drug discovery	scientific literature / identify drug usage and comorbidities
241	Maragoudakis, 2008 #10091	medical	diagnosis	125 skin lesion images / medical science research / decision support systems
242	Marcos, 2013 #10226	pharm	adverse events	Adverse drug events (ADEs) / vaccine adverse events (VAEs)
243	Marquet, 2007 #8829	medical	treatment	126 Grading glioma tumours / treatment and prognosis
244	Marshall, 2012 #9212	pharm	drug discovery	drug uses - discoveries
245	Mayer, 2006 #10664	public	epidemiology	Public health surveillance
246	McCusker, 2012 #10249	public	social behaviour	enabling community health information / smoking prevalence and tobacco policies
247	McCusker, 2013 #10588	public	social behaviour	population statistics for tobacco
248	McCusker, 2014 #10214	pharm	drug discovery	Drug Repurposing
249	McGuinness, 2012 #10590	public	epidemiology	Air quality monitoring
250	Meditskos, 2015 #10199	medical	monitoring	130 clinical autonomy assessment / monitoring / dementia
251	Meditskos, 2016 #10105	public	awareness	question answering system / elderly - awareness
252	Meilender, 2012 #10641	medical	diagnosis	131 decision guidelines in oncology
253	Meneu, 2010 #10614	medical	monitoring	133 PUBLIC sensor / personalized health systems / chronic disease / lifestyle - monitoring
254	Meng, 2015 #10417	public	epidemiology	Anticipating health hazards
255	Mezghani, 2016 #10610	medical	diagnosis	134 patient treatment management / clinical decision support systems / personalized treatments / hyperglycemia in type 2 diabetes
256	Miori, 2012 #10470	public	awareness	Public health awareness
257	Mirhaji, 2004 #10531	public	awareness	the thyroid gland and obesity / food recommender
258	Mirza, 2008 #10329	medical	diagnosis	135 sensor / awareness / medical diagnostic imaging
259	Mohammadhassanzadeh, 2016 #10082	medical	diagnosis	137 clinical decision support / answering medical queries / disease diagnostic queries
260	Mohammadhassanzadeh, 2017 #10173	medical	diagnosis	136 clinical decision support / answering medical queries / disease diagnostic queries
261	Mohammadhassanzadeh, 2017 #8854	medical	diagnosis	clinical decision support
262	Moncrieff, 2013 #10296	medical	causes	138 spatial health research / factors that influence the disease
263	Monteiro, 2016 #10311	medical	examination	139 radiology reports / medical imaging reports
264	Moraes, 2012 #10553	clinical	learning	training professional knowledge
265	Murthy, 2011 #10468	public	social behaviour	Cancer / health groups / discussions - understanding public behaviour
266	Muthuraman, 2014 #8955	public	awareness	Health information recommender
267	Nachabe, 2018 #10078	medical	monitoring	142 Monitoring / diabetes patient
268	Nachawati, 2014 #10219	medical	monitoring	Health monitoring
269	Najeeb, 2016 #9059	clinical	EHR	Malaria patient records / LOD / EHR - clinical research
270	Natsiavas, 2017 #10181	pharm	adverse events	pharmacovigilance
271	Natsiavas, 2018 #8936	pharm	adverse events	pharmacovigilance
272	Nelson, 2014 #8912	clinical	workflow	healthcare processes - business rules management (BRM)
273	Nguyen, 2011 #10256	medical	causes	new hypotheses / research / the pathological processes underlying diseases / Chagas disease
274	Nithya, 2016 #9286	medical	diagnosis	146 health record / dental and general health / Health Information Systems (HIS) / decision support
275	Nyulas, 2012 #10241	medical	treatment	147 traditional medicine - treatment
276	Obermaisser, 2014 #10475	medical	diagnosis	148 active diagnosis

277	Osman, 2013 #10515	medical	diagnosis	149 decision support / dementia care
278	Pagkalos, 2014 #10343	medical	monitoring	Monitoring personal health
279	Panahiazar, 2015 #10442	medical	examination	150 GENERAL personalized medicine - medical tools
280	Pandiyani, 2011 #10040	public	epidemiology	infectious diseases reporting - epidemiology
281	Papakonstantinou, 2011 #10125	clinical	learning	training - education
282	Pappachan, 2015 #10278	medical	diagnosis	152 Public community health-care in underserved areas - diagnosis
283	Paraiso-Medina, 2013 #10306	medical	treatment	153 breast cancer trials / genomic information / translational clinical trials
284	Passi, 2015 #9022	medical	diagnosis	154 decision support system / colorectal cancer / Guidelines
285	Passornpakorn, 2016 #8974	public	awareness	Health self-care
286	Patel, 2007 #8826	clinical	EHR	156 clinical trials / EHR
287	Pathak, 2012 #10351	medical	diagnosis	159 genotype-phenotype associations / Type 2 Diabetes / EHR / hypotheses generation - diagnosis
288	Pathak, 2012 #8820	medical	diagnosis	157 genotype-phenotype associations / Type 2 Diabetes and Hypothyroidism / EHR / hypotheses generation - diagnosis
289	Pathak, 2012 #9126	medical	diagnosis	158 genotype-phenotype associations / Type 2 Diabetes / EHR / hypotheses generation - diagnosis
290	Pathak, 2013 #10439	pharm	adverse events	Drugs interaction (DDIs)
291	Pathak, 2013 #10635	pharm	adverse events	Drugs interaction (DDIs)
292	Paul Rupa, 2016 #8834	medical	diagnosis	160 Genetic Testing / predict the genetic conditions - diagnosis
293	Pellison, 2017 #10391	public	epidemiology	Tuberculosis / epidemiological surveillance
294	Piccinni, 2017 #10622	pharm	adverse events	pharmacovigilance / Adverse drug events (ADEs)
295	Pierce, 2012 #8907	clinical	EHR	162 clinical research / research data
296	Piñero, 2015 #8911	medical	diagnosis	163 human diseases and genes - diagnosis
297	Podgorelec, 2007 #10493	medical	diagnosis	165 Medical diagnostic process / the mitral valve prolapse syndrome
298	Podgorelec, 2009 #8892	medical	diagnosis	164 Medical diagnostic process / the mitral valve prolapse syndrome
299	Prior, 2009 #9243	medical	diagnosis	166 medical knowledge / military medicine / medical decision making / new procedures and therapies
300	Puustjärvi, 2009 #10356	clinical	workflow	workflow
301	Puustjärvi, 2010 #10300	medical	diagnosis	health decision
302	Puustjärvi, 2010 #10358	clinical	learning	learning
303	Puustjärvi, 2010 #10676	medical	monitoring	Health monitoring
304	Puustjärvi, 2011 #10124	clinical	learning	learning
305	Puustjärvi, 2011 #10302	medical	diagnosis	health decision
306	Puustjärvi, 2011 #8805	medical	monitoring	monitoring - telemedicine
307	Puustjärvi, 2012 #10118	medical	monitoring	monitoring - chronic care
308	Puustjärvi, 2013 #10305	medical	examination	telemedicine - medical devices
309	Puustjärvi, 2015 #10399	public	awareness	childhood obesity
310	Puustjärvi, 2016 #10074	public	awareness	Personal health information assistant
311	Puustjärvi, 2016 #10335	public	awareness	Health self-care
312	Puustjärvi, 2016 #10336	clinical	guidelines	guidelines and policies
313	Quinn, 2017 #8930	public	awareness	Personalized patient education

314	Quinn, 2018 #8995	public	awareness	Personalized patient education / diabetic patients
315	Rahimi, 2014 #9032	medical	diagnosis	167 Type 2 Diabetes Mellitus / EHR / diagnose
316	Rajapakse, 2008 #9102	public	epidemiology	168 dengue serotypes in scientific abstracts / research
317	Ramesh, 2015 #10447	clinical	EHR	169 neurology / secondary analysis / clinical research / epilepsy
318	Ramírez, 2011 #10037	public	awareness	Mental health/ information search - awareness
319	Ramzan, 2014 #10630	medical	diagnosis	171 Clinical decision support systems / human expertise / mental-health problems
320	Reda, 2018 #10100	medical	diagnosis	decision making - personal fitness data from wearable devices
321	Rhayem, 2018 #10526	medical	diagnosis	174 medical connected objects / detected vital signs / decision-making
322	Riazanov, 2012 #10123	public	epidemiology	surveillance - hospital infections
323	Riazanov, 2013 #9124	public	epidemiology	surveillance - hospital infections
324	Riga, 2014 #10111	public	epidemiology	Air quality
325	Robles-Bykbaev, 2016 #10071	clinical	guidelines	175 therapy plans / children / communication disorders / intervention guidelines
326	Rodrigues, 2016 #10509	medical	diagnosis	177 urinalysis / renal condition / tutoring and decision-support systems
327	Rodrigues, 2018 #10507	medical	diagnosis	176 urinalysis / renal condition / tutoring and decision-support systems
328	Rodríguez, 2009 #10536	medical	diagnosis	178 diagnosis / recommend the medications / diagnosis decision support systems
329	Rodríguez-González, 2011 #10039	medical	diagnosis	179 diagnosis / recommend the medications / diagnosis decision support systems
330	Rodríguez-González, 2011 #9014	medical	diagnosis	180 diagnosis / recommend the medications / diagnosis decision support systems
331	Rodríguez-González, 2012 #8906	medical	diagnosis	181 diagnosis / recommend the medications / diagnosis decision support systems
332	Rodríguez-Molina, 2013 #9304	medical	monitoring	monitoring physical parameters on a person
333	Roldán-García, 2016 #9112	clinical	EHR	clinical research - classifying diseases to determine the case fatality and morbidity rates
334	Rubin, 2008 #10095	medical	examination	182 Medical imaging
335	Ruttenberg, 2007 #8866	medical	causes	183 translational research / neuroscience researchers
336	Sabra, 2018 #10077	medical	diagnosis	184 clinical narratives / personalized diagnosis and treatment plan / clinical decision support system / EHR / venous thromboembolism
337	Sætre, 2016 #10069	clinical	EHR	185 acute patient histories / cohort identification / discover new hypotheses - research
338	Sahoo, 2014 #10443	medical	causes	186 clinical research / patient care / electrophysiological signals / epilepsy
339	Santana, 2014 #10559	medical	diagnosis	188 diagnosis / vascular system /clinical decision-making
340	Sasaki, 2013 #10438	public	assistive	Safety of walking routes
341	Schober, 2014 #10220	pharm	drug testing	Antibiotics resistance / enabling antibiotics resistance surveillance
342	Schweitzer, 2015 #10629	clinical	workflow	diabetes routine consultation / EHR / clinical workflows
343	Sethuraman, 2017 #10075	medical	diagnosis	190 Manag ?medical analysis / decide the most excellent connected specialist for a patient - diagnosis
344	Shaban-Nejad, 2012 #10407	public	epidemiology	surveillance - hospital infections
345	Shaban-Nejad, 2016 #9175	public	epidemiology	surveillance - hospital infections
346	Shaban-Nejad, 2017 #8839	public	epidemiology	Population health data / epidemiology
347	Shabo, 2005 #9058	medical	diagnosis	diagnosis - medical family history
348	Shah, 2011 #10431	medical	diagnosis	191 Clinical Decision Support System / medical and oral health domains
349	Shah, 2013 #10051	medical	diagnosis	194 Clinical Decision Support System / medical and oral health domains
350	Shah, 2013 #10566	medical	diagnosis	193 Clinical Decision Support System / medical and oral health domains
351	Shah, 2015 #8880	medical	diagnosis	192 Clinical Decision Support System / medical and oral health domains

352	Sherimon, 2013 #9026	medical	diagnosis	198 Clinical Decision Support System / Diabetes, hypertension/ clinical guidelines
353	Sherimon, 2016 #10523	medical	diagnosis	197 Clinical Decision Support System / Diabetes, hypertension/ clinical guidelines
354	Sherimon, 2016 #8842	medical	diagnosis	196 Clinical Decision Support System / Diabetes, hypertension / clinical guidelines
355	Sheydin, 2013 #9001	medical	treatment	199 medical research, clinical research
356	Shields, 2018 #9257	clinical	workflow	front-line nurses processes
357	Shojanoori, 2013 #9318	medical	diagnosis	201 PUBLIC personalized remote patient monitoring / Decision making / care home
358	Shojanoori, 2014 #10587	medical	diagnosis	202 PUBLIC Care homes / Assistive self-care / decision making - monitoring
359	Shu-Hui, 2008 #10542	medical	diagnosis	203 Lung Cancer Patients / medical tracking / diagnosing
360	Silachan, 2011 #10429	medical	diagnosis	diagnosis
361	Silalahi, 2015 #10316	medical	treatment	medicinal plant
362	Sim, 2014 #9083	clinical	EHR	Clinical Research / clinical studies
363	Sinaci, 2013 #9085	clinical	EHR	clinical research - secondary use of EHRs
364	Singh, 2013 #10090	medical	examination	medical images
365	Singh, 2013 #9337	medical	examination	206 Image retrieval /EHR
366	Sivamani, 2016 #8991	public	awareness	Balanced diet for livestock / nutrition of livestock
367	Slăvescu, 2014 #10344	medical	diagnosis	207 assisting medical decisions
368	Sojic, 2016 #9117	public	awareness	Personalized health / obesity
369	Solovieva, 2018 #9108	medical	diagnosis	209 genetic diseases / disorders of glycan metabolism - diagnosis
370	Sommaruga, 2011 #10126	public	assistive	independent living support for elderly
371	Sophia, 2018 #10143	public	assistive	Exergaming for elderly people
372	Sordo, 2013 #10632	medical	diagnosis	decision support
373	Soualmia, 2005 #9036	public	awareness	finding information - awarness
374	Spyropoulos, 2010 #9240	medical	monitoring	210 cardiorespiratory diseases / home monitoring
375	Srinivasan, 2006 #8832	public	epidemiology	Public health surveillance
376	Stavropoulos, 2015 #10522	medical	monitoring	211 PUBLIC health monitoring / sensors / dementia care
377	Stavropoulos, 2016 #10384	medical	monitoring	212 PUBLIC health monitoring / sensors / dementia care
378	Streibel, 2016 #10187	public	epidemiology	Epidemiology / infectious disease control
379	Sun, 2015 #9079	clinical	EHR	clinical research - EHRs
380	Supriyanto, 2011 #10620	public	epidemiology	tropical diseases information / epidemiology
381	Surján, 2006 #8901	public	social behaviour	Public health indicators
382	Sutar, 2018 #10563	pharm	adverse events	side effects of drugs
383	Sutcliffe, 2012 #10644	public	epidemiology	Simulation modeling of population health / epidemiology
384	Szalontai, 2014 #9249	pharm	drug testing	214 probiotic bacteria / counting method - research - testing approach
385	Szwed, 2013 #10049	clinical	guidelines	CLINICAL / medical guideline for asthma control assessment / personal health
386	Tablado, 2004 #10378	public	assistive	Health Assisting elderly
387	Tahmasebian, 2016 #8807	medical	diagnosis	215 INFO electronic discharge summary / decision making
388	Tamposis, 2015 #10558	medical	diagnosis	216 INFO? Medical imaging workflow / prognosis and diagnosis procedures
389	Tao, 2011 #10364	clinical	EHR	clinical research - secondary use of EHRs
390	Tao, 2012 #10350	clinical	EHR	clinical research - secondary use of EHRs
391	Tao, 2012 #10434	clinical	EHR	clinical research - secondary use of EHRs

392	Tao, 2013 #9200	clinical	EHR	clinical research - secondary use of EHRs
393	Tao, 2014 #8855	pharm	adverse events	vaccine adverse events
394	Tello, 2015 #10340	medical	examination	217 INFO medical images
395	Thermolia, 2013 #10015	medical	monitoring	219 patient monitoring system / bipolar disorder / TeleCare
396	Thermolia, 2015 #10450	medical	monitoring	220 patient monitoring system / bipolar disorder / TeleCare
397	Thermolia, 2015 #10593	medical	monitoring	218 patient monitoring system / bipolar disorder / TeleCare
398	Trinugroho, 2012 #10599	public	assistive	Well-being of inhabitants / elderly independent living
399	Trpkovska, 2014 #10440	medical	diagnosis	222 PUBLIC predictive health / predicting children's general diseases - diagnosis
400	Trpkovska, 2014 #10577	medical	diagnosis	223 PUBLIC predictive health / predicting children's general diseases - diagnosis
401	Tsai, 2007 #10295	clinical	guidelines	guidelines and policies
402	Ullah, 2017 #9316	medical	treatment	medicine recommendation - treatment
403	Usher, 2013 #9017	medical	monitoring	224 chronic disease / sensor- monitoring
404	Vadivu, 2012 #9134	medical	treatment	medicinal plants - treatment
405	Van Woensel, 2014 #10444	medical	diagnosis	225 PUBLIC Behavioural User / Assistive / decision support
406	Van Woensel, 2017 #10168	medical	diagnosis	227 PUBLIC Behavioural User / Assistive / decision support
407	Van Woensel, 2017 #10185	medical	diagnosis	226 PUBLIC Behavioural User / Assistive / decision support
408	Vandervalk, 2013 #9165	pharm	adverse events	Drugs interaction (DDIs)
409	Vassilev, 2013 #10085	public	assistive	Assisting disability
410	Vega-Gorgojo, 2016 #10191	pharm	drug testing	pharmacogenomic testing
411	Velentzas, 2008 #10537	public	assistive	Monitoring elderly at home
412	Velmurugan, 2016 #10072	public	awareness	Allergy information / improving people's education - awarness
413	Vergari, 2011 #9047	medical	monitoring	229 Telemedicine / personalisation / monitoring
414	Visser, 2011 #8861	pharm	drug discovery	bioassays / high-throughput screening (HTS) / to identify small molecule chemical probes and drugs
415	Wang, 2009 #10544	medical	examination	232 Child Psychiatry Neuroimaging /
416	Wang, 2009 #10674	clinical	workflow	business process
417	Wang, 2010 #9063	public	awareness	Healthy diet
418	Wang, 2012 #10548	clinical	EHR	231 Clinical trial
419	Wang, 2013 #8844	clinical	care plans	clinical pathways
420	Wang, 2013 #9334	medical	causes	233 Cancer research
421	Wang, 2014 #9082	clinical	care plans	clinical pathways
422	Wang, 2015 #9176	clinical	care plans	clinical pathways
423	Waqiialla, 2016 #10393	medical	treatment	234 Cardiac Rehabilitation - treatment
424	Webster, 2011 #9090	medical	causes	235 translational research / disease genes / riluzole and alcohol abuse
425	Weng, 2010 #10467	medical	monitoring	236 Alert monitoring / Psychological health
426	Wheeler, 2017 #10180	clinical	guidelines	237 Hypertension / clinical guidelines / behaviour change
427	Widmer, 2013 #10282	clinical	care plans	Patient guidance
428	Wiesner, 2011 #10488	public	awareness	Personalized health recommender
429	Willighagen, 2009 #10263	pharm	drug discovery	drug discovery
430	Wrighton, 2009 #10505	medical	diagnosis	238 mental-health / decision support system
431	Yao, 2009 #10017	public	awareness	Health awareness
432	Ye, 2008 #10519	clinical	care plans	clinical pathways
433	Ye, 2009 #8900	clinical	care plans	clinical pathways
434	Yilmaz, 2013 #9179	medical	diagnosis	decision support
435	Yu, 2013 #10229	medical	diagnosis	240 identify patients with diabetes / EHR / clinical trials
436	Yu, 2014 #10087	medical	monitoring	241 monitoring / sensors / collecting data / repository / data analytics / research

437	Yu, 2018 #10459	medical	monitoring	Health monitoring
438	Zaman, 2014 #10501	public	awareness	Personal Health information recommender
439	Zaveri, 2011 #10432	medical	treatment	242 Research-Disease Disparity
440	Zhang, 2015 #9119}	pharm	adverse events	vaccine adverse events
441	Zhang, 2016 #8889	medical	diagnosis	244 Clinical Decision Support System / patient assessments or recommendations / type 2 diabetes mellitus
442	Zhang, 2016 #9173	medical	diagnosis	243 Clinical Decision Support System / patient assessments or recommendations / type 2 diabetes mellitus
443	Zheng, 2010 #10655	medical	examination	medical tools - images - physiological signals electrocardiogram
444	Zhou, 2017 #8808	medical	diagnosis	245 fault diagnosis
445	Zhu, 2012 #9125	medical	treatment	Structured Product Labeling (SPL) / Profiling structured product labeling / exchanging / mapping
446	Zhu, 2012 #9183	pharm	drug testing	Pharmacogenomics research
447	Zhu, 2014 #10342	medical	monitoring	246 PUBLIC telehealth / hypertension / self-monitoring

Appendix E Converting CSV to RDF Script Code

Sub printAllTheRows()

```

Dim s As Long
Dim e As Long
Dim cnt As Integer
cnt = 1
'e = 1048576
e = Rows.Count
For s = 1 To e Step 30000
Call printSection(s, cnt)
cnt = cnt + 1
Next
End Sub

```

Sub printSection(sr As Long, fileCNT As Integer)

```

Dim FileNo As Integer
FileNo = FreeFile
Dim Filename As String
Filename = "C:\Users\ma8g13\Desktop\ontologies\SOUTH1\SOUTH1_section" & fileCNT & ".rdf"
Open Filename For Output As #FileNo
Dim msg As String
msg = "<?xml version=""1.0"" encoding=""UTF-8""?>" & vbCrLf
msg = msg & "<rdf:RDF" & vbCrLf
msg = msg & " xmlns:pres=""http://www.semanticweb.org/ma8g13/prescriptionOntology#" & vbCrLf
msg = msg & " xmlns:rdf=""http://www.w3.org/1999/02/22-rdf-syntax-ns#" & vbCrLf
msg = msg & " xmlns:owl=""http://www.w3.org/2002/07/owl#" & vbCrLf
msg = msg & " xmlns:xsd=""http://www.w3.org/2001/XMLSchema#" & vbCrLf
msg = msg & " xmlns:rdfs=""http://www.w3.org/2000/01/rdf-schema#" & vbCrLf
msg = msg & " xmlns:foaf=""http://xmlns.com/foaf/0.1/" > " & vbCrLf & vbCrLf & vbCrLf & vbCrLf & vbCrLf
Print #FileNo, msg

Dim practiceCode As String
Dim prescriptionCode As String
Dim medicineCode As String
'1048575
Dim r As Long, lr As Long
lr = sr + 30000 - 1
For r = sr To lr
If (r > Rows.Count) Then
Exit For
End If
'declaring a practice
practiceCode = Cells(r, 8)
msg = "<rdf:Description
rdf:about=""http://www.semanticweb.org/ma8g13/ontologies/2017/8/prescriptionOntology#practice_" &
practiceCode & """" > " & vbCrLf
msg = msg & " <rdf:type
rdf:resource=""http://www.semanticweb.org/ma8g13/ontologies/2017/8/prescriptionOntology#Practice""/
> " & vbCrLf

msg = msg & "</rdf:Description> " & vbCrLf & vbCrLf & vbCrLf

```

```

    'declaring a prescription
    'prescriptionCode = r
    prescriptionCode = Cells(r, 17)
    msg = msg & "<rdf:Description
rdf:about=""http://www.semanticweb.org/ma8g13/ontologies/2017/8/prescriptionOntology#prescription_
south_" & prescriptionCode & """"> " & vbCrLf
    msg = msg & "    <rdf:type
rdf:resource=""http://www.semanticweb.org/ma8g13/ontologies/2017/8/prescriptionOntology#Prescriptio
n""/>" & vbCrLf

    msg = msg & "</rdf:Description> " & vbCrLf & vbCrLf & vbCrLf
    'declaring a medicine
    medicineCode = Cells(r, 9)
    msg = msg & "<rdf:Description
rdf:about=""http://www.semanticweb.org/ma8g13/ontologies/2017/8/prescriptionOntology#BNF_present
ation_" & medicineCode & """"> " & vbCrLf
    msg = msg & "    <rdf:type
rdf:resource=""http://www.semanticweb.org/ma8g13/ontologies/2017/8/prescriptionOntology#Medicine""
/>" & vbCrLf
    msg = msg & "</rdf:Description> " & vbCrLf & vbCrLf & vbCrLf
    'practice -> has -> prescription & the rest data properties
    msg = msg & "<rdf:Description
rdf:about=""http://www.semanticweb.org/ma8g13/ontologies/2017/8/prescriptionOntology#prescription_
south_" & prescriptionCode & """"> " & vbCrLf
    msg = msg & "    <pres:has
rdf:resource=""http://www.semanticweb.org/ma8g13/ontologies/2017/8/prescriptionOntology#BNF_prese
ntation_" & medicineCode & """"/>" & vbCrLf & vbCrLf

    msg = msg & "    <pres:items>" & Cells(r, 11) & "</pres:items>" & vbCrLf
    msg = msg & "    <pres:quantity>" & Cells(r, 12) & "</pres:quantity>" & vbCrLf
    msg = msg & "    <pres:ADQ_usage>" & Cells(r, 13) & "</pres:ADQ_usage>" & vbCrLf
    msg = msg & "    <pres:NIC>" & Cells(r, 14) & "</pres:NIC>" & vbCrLf
    msg = msg & "    <pres:actualCost>" & Cells(r, 15) & "</pres:actualCost>" & vbCrLf & vbCrLf
    msg = msg & "</rdf:Description> " & vbCrLf & vbCrLf & vbCrLf

    'prescription -> has -> medicine
    msg = msg & "<rdf:Description
rdf:about=""http://www.semanticweb.org/ma8g13/ontologies/2017/8/prescriptionOntology#practice_" &
practiceCode & """"> " & vbCrLf

    msg = msg & "    <pres:has
rdf:resource=""http://www.semanticweb.org/ma8g13/ontologies/2017/8/prescriptionOntology#prescriptio
n_south_" & prescriptionCode & """"/> " & vbCrLf

    msg = msg & "</rdf:Description> " & vbCrLf & vbCrLf & vbCrLf & vbCrLf & vbCrLf
    'print the triples for this row
    Print #FileNo, msg
    Next r
'at the end of the file
msg = "</rdf:RDF>" & vbCrLf & vbCrLf
Print #FileNo, msg
Close #FileNo
End Sub

```


Appendix F Focus Group Inputs and Outputs

F.1 Information Sheet

Participant Information Sheet

Study Title: An Investigation of the Semantic Web Feasibility in Health Data Integration

Researcher: Mona Almofarreh

ERGO number: ERGO/FEPS/49893

You are being invited to take part in the above research study. To help you decide whether you would like to take part or not, it is important that you understand why the research is being done and what it will involve. Please read the information below carefully and ask questions if anything is not clear or you would like more information before you decide to take part in this research. You may like to discuss it with others but it is up to you to decide whether or not to take part. If you are happy to participate you will be asked to sign a consent form.

What is the research about?

This research is being carried out by Mona Almofarreh (a computer science PhD Student at the University of Southampton). The research aims to investigate the feasibility of using the Semantic Web as a tool for integrating health data. Linking heterogeneous datasets can be a challenging task due to the different available forms and types of data. The Semantic Web provides standards and tools that can support heterogeneous data linking. This research will examine the possibility of using the Semantic Web as a tool for linking prescription-related data by using open data.

Why have I been asked to participate?

You are invited to participate in this study to help the researcher with suggestions of interesting research questions for the provided data from a health expert perspective. These questions will be attempted to be answered where possible using the Semantic Web approach.

What will happen to me if I take part?

If you decide to take part of this study, a convenient time meeting will be arranged with other health researchers (3-4) for about an hour. A brief introductory explanation of the prescription model will be demonstrated. After that, you will be provided with samples of some prescription-related datasets and will be asked to brainstorm with the other participants some interesting research questions that can be asked across these datasets.

Are there any benefits in my taking part?

By taking part in this study, you will get to know more about the opportunities of using the Semantic Web in linking health data.

Are there any risks involved?

No.

What data will be collected?

The following is a complete list of all the data that will be collected through the study:

- 1) Participant name, which will only be used for identification purposes by the researcher. It will be stored on a password protected computer then deleted after finishing the research.
- 2) Participant e-mail, which will be used for requesting their participation. It will be stored on a password protected computer until the end of the research then deleted after finishing the research.
- 3) The participant's expertise area, which will be used to understand the relevance to the study. It will be stored on a password protected computer until the end of the research then deleted after finishing the research.
- 4) The Institute where the participants belong to, which will be used to understand the relevance to the study. It will be stored on a password protected computer until the end of the research then deleted after finishing the research.
- 5) Consent forms, which will be stored as paper-based version in a locked filing cabinet with limited access to the investigator only. At the end of the research, all consent forms will be destroyed.
- 6) Audio recording of the participant, which will be stored in the recorder. As soon as the recording is transcribed it will be deleted.
- 7) Recorder transcript, which will be stored on a password protected computer until the end of the research then deleted after finishing the research.

Will my participation be confidential?

Your participation and the information we collect about you during the course of the research will be kept strictly confidential.

Only members of the research team and responsible members of the University of Southampton may be given access to data about you for monitoring purposes and/or to carry out an audit of the study to ensure that the research is complying with applicable regulations. Individuals from regulatory authorities (people who check that we are carrying out the study correctly) may require access to your data. All of these people have a duty to keep your information, as a research participant, strictly confidential.

The information will be collected during this study will be stored on a password protected computer in a secured lab. Any collected data during this study will only be used for the purpose of this study. All collected data will be deleted at the end of this research. Individual responses will not be identified. All responses will be compiled together and analysed as a group.

Do I have to take part?

No, it is entirely up to you to decide whether or not to take part. If you decide you want to take part, you will need to sign a consent form to show you have agreed to take part.

What happens if I change my mind?

You have the right to change your mind and withdraw at any time without giving a reason and without your participant rights being affected. If you withdraw from the study, we will keep the information about you that we have already obtained for the purposes of achieving the objectives of the study only.

What will happen to the results of the research?

Your personal details will remain strictly confidential. Research findings made available in any reports or publications will not include information that can directly identify you without your specific consent.

The results of the study will be used in constructing proper queries (research questions) to be tested across our model. Our intention is both to provide feedback to the participants and to examine our findings from the perspective of those most closely involved in the process. The results of this study will be used for the purpose of achieving an academic degree and also it might be used in any potential academic publications. However, all collected data will be destroyed at the end of this research.

Where can I get more information?

For further information, please do not hesitate to contact us.

Mona AlmoFarreh: ma8g13@soton.ac.uk

Dr. Mark Weal: mjw@ecs.soton.ac.uk

What happens if there is a problem?

If you have a concern about any aspect of this study, you should speak to the researchers who will do their best to answer your questions.

If you remain unhappy or have a complaint about any aspect of this study, please contact the University of Southampton Research Integrity and Governance Manager (023 8059 5058, rqoinfo@soton.ac.uk).

The research Team contact information:

Mona AlmoFarreh: ma8g13@soton.ac.uk

Dr. Mark Weal: mjw@ecs.soton.ac.uk

Data Protection Privacy Notice

The University of Southampton conducts research to the highest standards of research integrity. As a publicly-funded organisation, the University has to ensure that it is in the public interest when we use personally-identifiable information about people who have agreed to take part in research. This means that when you agree to take part in a research study, we will use information about you in the ways needed, and for the purposes specified, to conduct and complete the research project. Under data protection law, 'Personal data' means any information that relates to and is capable of identifying a living individual. The University's data protection policy governing the use of personal data by the University can be found on its website (<https://www.southampton.ac.uk/legalservices/what-we-do/data-protection-and-foi.page>).

This Participant Information Sheet tells you what data will be collected for this project and whether this includes any personal data. Please ask the research team if you have any questions or are unclear what data is being collected about you.

Our privacy notice for research participants provides more information on how the University of Southampton collects and uses your personal data when you take part in one of our research projects and can be found at <http://www.southampton.ac.uk/assets/sharepoint/intranet/Is/Public/Research%20and%20Integrity%20Privacy%20Notice/Privacy%20Notice%20for%20Research%20Participants.pdf>

Any personal data we collect in this study will be used only for the purposes of carrying out our research and will be handled according to the University's policies in line with data protection law. If any personal data is used from which you can be identified directly, it will not be disclosed to anyone else without your consent unless the University of Southampton is required by law to disclose it.

Data protection law requires us to have a valid legal reason ('lawful basis') to process and use your Personal data. The lawful basis for processing personal information in this research study is for the performance of a task carried out in the public interest. Personal data collected for research will not be used for any other purpose.

For the purposes of data protection law, the University of Southampton is the 'Data Controller' for this study, which means that we are responsible for looking after your information and using it properly. The University of Southampton will keep identifiable information about you for 1 year after the study has finished after which time any link between you and your information will be removed.

To safeguard your rights, we will use the minimum personal data necessary to achieve our research study objectives. Your data protection rights - such as to access, change, or transfer such information - may be limited, however, in order for the research output to be reliable and accurate. The University will not do anything with your personal data that you would not reasonably expect.

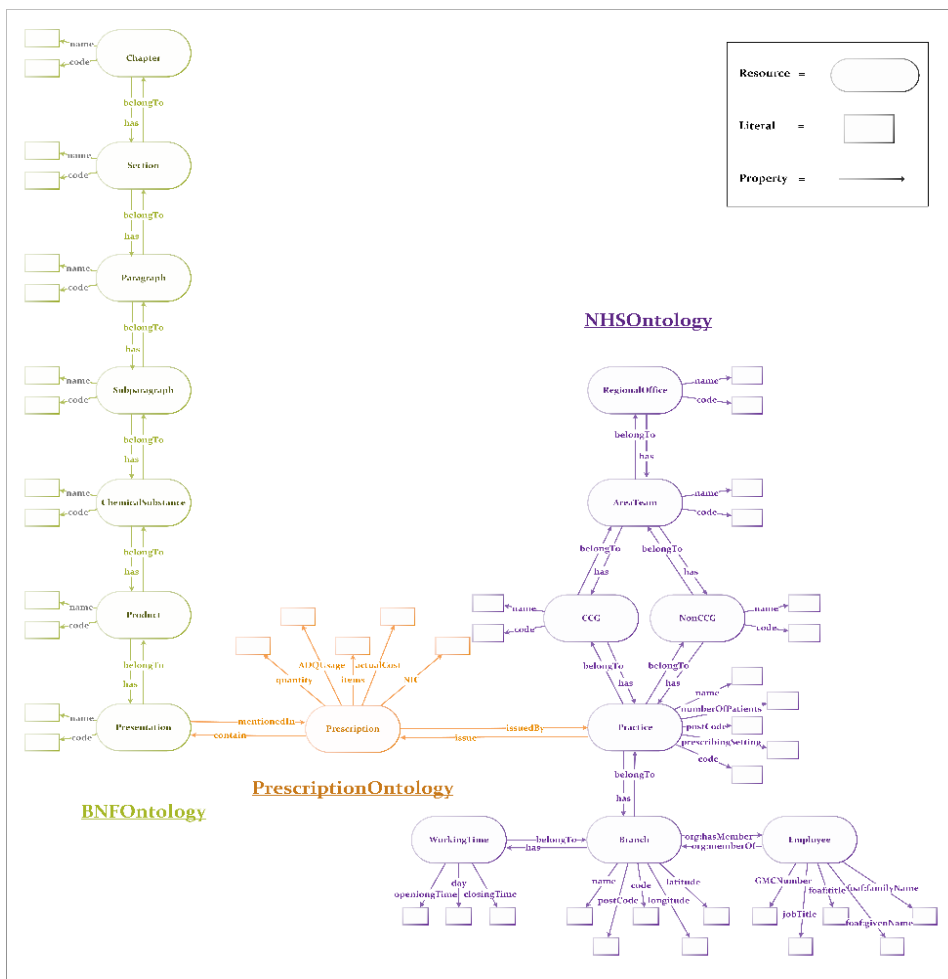
If you have any questions about how your personal data is used, or wish to exercise any of your rights, please consult the University's data protection webpage (<https://www.southampton.ac.uk/legalservices/what-we-do/data-protection-and-foi.page>) where you can make a request using our online form. If you need further assistance, please contact the University's Data Protection Officer (data.protection@soton.ac.uk).

Thank you for taking the time to read this information sheet and for considering taking part in this research

F.2 Consent Form

CONSENT FORM	
Study title: An Investigation of the Semantic Web Feasibility in Health Data Integration	
Researcher name: Mona Almofarreh	
ERGO number: ERGO/FEPS/49893	
<i>Please initial the box(es) if you agree with the statement(s):</i>	
I have read and understood the information sheet (2019-06-17 /v1) and have had the opportunity to ask questions about the study.	
I agree to take part in this research project and agree for my data to be used for the purpose of this study.	
I understand my participation is voluntary and I may withdraw at any time for any reason without my participation rights being affected.	
I understand that taking part in the study involves audio recording which will be transcribed and then destroyed for the purposes set out in the participation information sheet.	
I understand that if I withdraw from the study that it may not be possible to remove the data once my personal information is no longer linked to the data.	
I understand that I may be quoted directly in reports of the research but that I will not be directly identified (e.g. that my name will not be used).	
<p>Name of participant (print name).....</p> <p>Signature of participant.....</p> <p>Date.....</p> <p>Name of researcher (print name).....</p> <p>Signature of researcher</p> <p>Date.....</p>	

F.3 The Prescriptions Demonstrator



The BNF ontology:

The British National Formulary (BNF) is a United Kingdom (UK) pharmaceutical reference book that discusses correct dosage, indication, interactions and side effects of drugs. It is published twice a year, however we are speaking here about the online version of the BNF that mainly lists the names and codes of all the taxonomical structure of each drug. Each row in the spreadsheet represent a presentation of a drug (name and unique code). This BNF code contains some aggregated information about the drug specifying the chapter, section, paragraph, subparagraph, chemical substance and product of this BNF presentation.

The Prescriptions ontology:

NHS Digital published monthly anonymised prescriptions dataset. The dataset contains all prescriptions issued by a physician in England excluding the private sector. This data sets is very large that could reach 18 million prescriptions monthly. The type of information available in the dataset is about the prescriber (practice code), the medication (the BNF code) and prescription's information such as quantity and cost.

The NHS ontology:

This part of data represents the hierarchy of the NHS structure starting from the practices' branches, Clinical commissioning Groups CCG, Area Teams and Regional offices. It also has some additional information for the practices such as the number of patients registered.

F.4 Suggested Datasets**The Prescription Dataset:**

NHS Digital publishes monthly anonymised prescriptions dataset. The dataset contains all prescriptions issued in England excluding the private sector. This data set is very large that could reach over 18 million prescriptions monthly. The type of information available in the dataset is about the prescriber (practice code), the medication (the BNF code) and prescription's information such as quantity and cost.

Example:

Practice code	BNF code	BNF name	No. of itmes	NIC	Actual cost
Y04937	0103050P0AAAAAA	Omeprazole_Cap E/C 20mg	1	3.64	3.38
N81002	0102000N0AAABAB	Hyoscine Butylbrom_Tab 10mg	25	69.37	66.14

The BNF Dataset:

The British National Formulary (BNF) is a pharmaceutical reference book that discusses correct dosage, indication, interactions and side effects of drugs published in the United Kingdom (UK) twice a year. However, we are speaking here about the online open dataset of the BNF that mainly lists the names and codes of all the taxonomical structure of each drug. Each row in the spreadsheet represent a presentation of a drug (name and unique code). This BNF code contains some aggregated information about the drug specifying the chapter, section, paragraph, subparagraph, chemical substance and product of this BNF presentation. These book's sections are divided based on different categorising aims such as aspects of medical care, therapeutic use, treatment summaries, and the monographs of the drugs.

Example:

Chapter	Code	Section	Code	Paragraph	Code	Subparagraph	Code
Gastro-Intestinal System	1	Dyspep&Gastro-Oesophageal Reflux Disease	101	Antacids and Simeticone	10101	Antacids and Simeticone	101010
Gastro-Intestinal System	1	Dyspep&Gastro-Oesophageal Reflux Disease	101	Antacids and Simeticone	10101	Antacids and Simeticone	101010
Chemical Substance	Code	Product	Code	Presentation	Code		
Alexitol Sodium	0101010A0	Alexitol Sod	0101010A0A	Alexitol Sod_Tab 360mg	0101010A0AAAAAA		
Alexitol Sodium	0101010A0	Actal	0101010A0B	Actal_Tab 360mg	0101010A0BBAAAA		

The NHS Hierarchal Structure:

This part of data represents the hierarchy of the NHS structure starting from the practices' branches up to the regional office that includes them. Each data element in the dataset is represented by a name and a unique code.

Region Office code	name	Area team code	name	CCG code	name	Practice code	name	Branch code	name
Y54		Q44		01R		N81008		
...			A81004		3031	
...			A81004		3032	

Area Team Info:

There are around 54 ATs that are responsible for funding the CCGs. The following is an example of address information of the ATs.

a. Address:

Code	Name	address	City	County	Postcode	Latitude	Longitude
Q44

CCG Info:

Before 2013, the GP practices used to belong to Primary Care Trusts (PCTs), however, after April in that year all the PCTs have been replaced with Clinical Commissioning Groups (CCGs) and other non-CCGs centres such as trusts, councils and private companies. In the following dataset address information for the CCG are listed. The second example shows the coding for the CCGs from the Office for National Statistics (ONS).

a. Address:

Code	Name	address	City	County	Postcode	Latitude	Longitude
00M

b. CCG ONS code:

ONS code	CCG code	CCG name
E38000162	00M	NHS South Tees CCG
E38000163	00N	NHS South Tyneside CCG

Practice Info:

Each GP practice in England is provided with address information as well as information about the registered patients in it (gender and age).

a. Address:

Code	Name	address	City	County	Postcode	Latitude	Longitude
F82625

b. Practice size:

It also has some additional information for the practices such as the number of patients registered.

Practice Code	Male 0-4	Female 0-4	Male 5-14	Female 5-14	Male 15-24	Female 15-24	...	Total Number of Registered Patients
F82625	276	223	593	556	415	373		4150
F82018	511	448	914	912	826	842		19886
Y02583	673	667	852	836	472	561		13733

c. Prescribing Setting:

Moreover, each practice is categorised into one of the 25 different prescribing settings. The list of available prescribing settings are:

WIC Practice	Optometry Service	Secure Children's Home
OOH Practice	Urgent & Emergency Care	Immigration Removal Centre
WIC + OOH Practice	Hospice	Court
GP Practice	Care Home / Nursing Home	Police Custody
Public Health Service	Border Force	Sexual Assault Referral Centre (SARC)
Community Health Service	Young Offender Institution	Other – Justice Estate
Hospital Service	Secure Training Centre	Prison

And this is an example of categorising the practices.

Practice code	name
xxxxx	GP Practice
xxxxx	Public Health Service
xxxxx	Community Health Service
xxxxx	Hospital Service
...

Branch Info:

There are some practices that have more than one GP branch belong to them. Each branch has address, opening times and staff information. The following are some examples.

a. Address:

Code	Name	address	City	County	Postcode	Latitude	Longitude
2915

b. Opening Time:

code	Week Day	Times	opening time	closing time
2915	Monday	08:00-19:30	08:00	19:30
2915	Tuesday	08:00-19:30	08:00	19:30
2915	Wednesday	08:00-18:30	08:00	18:30
2915	Thursday	08:00-18:30	08:00	18:30
2915	Friday	08:00-18:30	08:00	18:30

c. Staff:

code	Title	GivenName	FamilyName	JobTitle	GMCNumber
2915	Dr	Anita	Syed	Administrator	
2915	Ms	Sabrina	Sellars	Administrator	
2915	Dr.	Babar	Farooq	General Practitioner	6083109
2915	Dr	Kish	Iqbal	General Practitioner	6074061
2915	Mr	Sameer	Butt	Practice Manager	
2915	Ms	Jamie	Gorman	Reception Staff	

HM Land Registry datasets:

This dataset lists the average prices for properties registered in HM Land Registry since 1968. The region codes here are the same as used in the ONS. Regions can be any geographical name such as Country, Regional, County/Unitary/District Authority or London Borough.

Example:

Date	Region Name	Area Code	Average Price	Monthly Change	Annual Change
01/03/2019	England	E92000001	243127.6439	-0.503148356	1.123000047
01/03/2019	South East	E12000008	318490.8938	-0.572313069	-0.432971692
01/03/2019	Hampshire	E10000014	311649.0272	-0.169487463	-0.79106943
01/03/2019	Southampton	E06000045	208691.8854	-0.593435708	-0.138978199

The ONS Address Directory (ONSAD):

The ONS Address Directory (ONSAD) relates the Unique Property Reference Number (UPRN) for each GB address to a range of current statutory administrative, electoral, health and other area geographies. It also links UPRNs to 2011 Census Output Areas (OA) and Super Output Areas (SOA), and in doing so helps support the production of area based statistics from address-level data. The UPRN is the unique identifier for every spatial address in Great Britain.

Example:

Unique Property Reference Number	County	Local Authority District	(Electoral) ward	Former Strategic Health Authority	country	region
1E+10	E9999999	E06000045	E05002470	E18000009	E92000001	E12000008
1E+10	E9999999	E06000046	E05008484	E18000009	E92000001	E12000008
Westminster parliamentary constituency	European Electoral Region	Travel to Work Area	LAU2 area local administrative units	National park	2011 Census Output Area (OA)	2011 Census Lower Layer Super Output Area (LSOA)
E14000955	E15000008	E30000267	E05002470	E99999999	E00087246	E01017281
E14000762	E15000008	E30000070	E05008484	E99999999	E00087369	E01017299
Middle Layer Super Output Area (MSOA)	Parish/ community	2011 Census Workplace Zone	Clinical Commissioning Group	Built-up Area	Built-up Area Sub-division	2011 Census rural-urban classification
E02003580	E43000036	E33041064	E38000167	E34999999	E35999999	C1
E02003597	E04001308	E33041154	E38000087	E34999999	E35999999	E1
2011 Census Output Area classification	Local Enterprise Partnership	Local Enterprise Partnership	Police Force Area	Index of Multiple Deprivation		
7B1	E37000029		E23000030	708		
1B3	E37000029	E23000030	8991	E34999999		

The Drug Ontology (DrOn):

An ontology of drugs. DrOn contains content developed by the National Library of Medicine in RxNorm. In creating DrOn, we have used RxNorm content only with SAB = RXNORM.

Classes	451,192
Individuals	19

The Drug Ontology

Last uploaded: February 15, 2019

Summary | **Classes** | Properties | Notes | Mappings | Widgets

Jump to:

metformin

Details	Visualization	Notes (0)	Class Mappings (75)	
Preferred Name	metformin			
	C4H11N5 InChI=1S/C4H11N5/c1-9(2)4(7)8-3(5)6/h1-2H3,(H5,5,6,7,8)			
Synonyms	1,1-Dimethylbiguanide N,N-dimethylimidodicarbonimidic diamide Metformin InChIKey=XZWYZXLIPIXDOLR-UHFFFAOYSA-N CN(C)C(-)N(C)N=N			
Definitions	A guanidine that has formula C4H11N5.			
ID	http://purl.obolibrary.org/obo/CHEBI_6801			
database_cross_reference	KEGG COMPOUND:657-24-9 ChemIDplus:657-24-9 KEGG COMPOUND:C07151 Wikipedia:Metformin			
has_exact_synonym	N,N-dimethylimidodicarbonimidic diamide Metformin			
has_related_synonym	C4H11N5 InChI=1S/C4H11N5/c1-9(2)4(7)8-3(5)6/h1-2H3,(H5,5,6,7,8) 1,1-Dimethylbiguanide InChIKey=XZWYZXLIPIXDOLR-UHFFFAOYSA-N			

Human Disease Ontology:

Creating a comprehensive hierarchical controlled vocabulary for human disease representation.

Classes	12,694
Individuals	0

Human Disease Ontology

Last updated: March 2, 2018

Summary Classes Properties Notes Mappings Widgets

Jump to:

- disease
 - disease by infectious agent
 - bacterial infectious disease
 - commensal bacterial infectious disease
 - actinobacillosis**
 - actinomycosis
 - chlamydia
 - gas gangrene
 - inclusion conjunctivitis
 - Lemierre's syndrome
 - lymphogranuloma venereum
 - pertussis
 - Ritter's disease
 - toxic shock syndrome
 - trachoma
 - opportunistic bacterial infectious disease
 - primary bacterial infectious disease
 - fungal infectious disease
 - parasitic infectious disease
 - viral infectious disease
- disease of anatomical entity
- disease of cellular proliferation
- disease of mental health
- disease of metabolism
- genetic disease
- physical disorder
- syndrome

Details	Visualization	Notes (0)	Class Mappings (24)
Preferred Name	actinobacillosis		
Synonyms	Actinobacillosis, NOS		
Definitions	A commensal bacterial infectious disease that results in infection, has_material_basis_in Actinobacillus ureae, which is a commensal of the human respiratory tract. The pathogen causes meningitis, endocarditis, bacteremia, atrophic rhinitis, bronchitis, pneumonia, conjunctivitis, peritonitis, and otitis media.		
ID	http://purl.obolibrary.org/obo/DOID_4974		
database_cross_reference	UMLS_CUI:C0001247 MESH:D000187 SNOMEDCT_US_2016_03_01:16140007		
definition	A commensal bacterial infectious disease that results in infection, has_material_basis_in Actinobacillus ureae, which is a commensal of the human respiratory tract. The pathogen causes meningitis, endocarditis, bacteremia, atrophic rhinitis, bronchitis, pneumonia, conjunctivitis, peritonitis, and otitis media.		
has_exact_synonym	Actinobacillosis, NOS		
has_obo_namespace	disease_ontology		
id	DOID:4974		
in_subset	http://purl.obolibrary.org/obo/doi#zoonotic_infectious_disease http://purl.obolibrary.org/obo/doi#gram-negative_bacterial_infectious_disease		
label	actinobacillosis		

F.5 The Resulted Questions

1. *Who prescribe better? The specialists or GPs? The experienced or junior doctors? The medical or non-medical prescribers?*
2. *Do the prescribers ask the right questions to the patients?*
3. *Is sophisticated way of prescribing by paying attention to the secondary characteristics of the medication been performed? And by what type of prescribers?*
4. *Do prescribers who are limited in prescribing perform better than unlimited ones?*
5. *How following different guidelines in different countries/regions affect the prescription of a specific disease? E.g. incurable skin diseases?*
6. *Is following guidelines or relying on doctor's experience is better?*
7. *How often is the guidelines are been followed? How many patients were diagnosed with condition X and treated as in the guidelines?*
8. *What are the factors that can affect the prescribing decision? E.g. blood tests, patient preference?*
9. *How can guidelines resolve the issue of variation in treatment for the same condition?*
10. *How doctors make their treatment decision?*
11. *What are the factors that affect the number of people suffering from a specific disease between different regions? E.g. weather & flu?*
12. *How weather or geographical location can affect the number of people prescribed anti-depression drugs?*
13. *Is there a pattern in prescribing branded and generic drugs?*