Semantic Framework for Regulatory Compliance Support

By

Krishna Sapkota

A thesis submitted in partial fulfilment of the requirements of
Oxford Brookes University
for the degree of
Doctor of Philosophy

Oxford Brookes University
Department of Computing and Communication Technologies
Faculty of Technology, Design and Environment
Wheatley Campus
Oxford, OX33 1HX, UK

June 2013

Abstract

Regulatory Compliance Management (RCM) is a management process, which an organization implements to conform to regulatory guidelines. Some processes that contribute towards automating RCM are: (i) extraction of meaningful entities from the regulatory text and (ii) mapping regulatory guidelines with organisational processes. These processes help in updating the RCM with changes in regulatory guidelines. The update process is still manual since there are comparatively less research in this direction. The Semantic Web technologies are potential candidates in order to make the update process automatic. There are stand-alone frameworks that use Semantic Web technologies such as Information Extraction, Ontology Population, Similarities and Ontology Mapping. However, integration of these innovative approaches in the semantic compliance management has not been explored yet. Considering these two processes as crucial constituents, the aim of this thesis is to automate the processes of RCM. It proposes a framework called, RegCMantic.

The proposed framework is designed and developed in two main phases. The first part of the framework extracts the regulatory entities from regulatory guidelines. The extraction of meaningful entities from the regulatory guidelines helps in relating the regulatory guidelines with organisational processes. The proposed framework identifies the document-components and extracts the entities from the document-components. The framework extracts important regulatory entities using four components: (i) parser, (ii) definition terms, (iii) ontological concepts and (iv) rules. The parsers break down a sentence into useful segments. The extraction is carried out by using the definition terms, ontological concepts and the rules in the segments. The entities extracted are the core-entities such as subject, action and obligation, and the aux-entities such as time, place, purpose, procedure and condition.

The second part of the framework relates the regulatory guidelines with organisational processes. The proposed framework uses a mapping algorithm, which considers three types of

entities in the regulatory-domain and two types of entities in the process-domains. In the regulatory-domain, the considered entities are regulation-topic, core-entities and aux-entities. Whereas, in the process-domain, the considered entities are subject and action. Using these entities, it computes aggregation of three types of similarity scores: topic-score, core-score and aux-score. The aggregate similarity score determines whether a regulatory guideline is related to an organisational process.

The RegCMantic framework is validated through the development of a prototype system. The prototype system implements a case study, which involves regulatory guidelines governing the Pharmaceutical industries in the UK. The evaluation of the results from the case-study has shown improved accuracy in extraction of the regulatory entities and relating regulatory guidelines with organisational processes. This research has contributed in extracting meaningful entities from regulatory guidelines, which are provided in unstructured text and mapping the regulatory guidelines with organisational processes semantically.

Acknowledgement

My sincere gratitude goes to my first supervisor, Dr Arantza Aldea for her inspiration and motivation, without which, I could not have started this PhD. I am very much indebted and thankful to her for her continuous help, support, supervision, inspiration and motivation during this thesis. Likewise, I would like to thank my second supervisor, Dr Muhammad Younas for his continuous help, support, supervision and motivation. Despite their busy schedule, Arantza and Younas are giving me their valuable feedback regularly (mostly weekly), which kept me motivated all the time during this thesis.

I am very grateful to my director of study, Prof David A. Duce for finding funding for extended period of this thesis and conference attendance, and his valuable advice and feedback on this thesis. Similarly, I would like to thank my advisor, Dr Rene Banares-Alcantara at University of Oxford for his valuable advice and feedback on this thesis and publications related to this thesis. In particular, I am thankful to him for making me understand the case study in the Pharmaceutical industry, providing all the resources needed to carry out the experiment on the case study and contributing to some funding for conference attendance.

I really appreciate the help of Dr. Nigel Crook for arranging funding for conference attendance and extended period of my thesis. Similarly, I am thankful to all the staff in Department of Computing for helping me on my learning, teaching, funding and admin related tasks. I am also thankful to Dr Julian Hunt, Dr Aidid Tahir and Berkan Sesen at University of Oxford for providing the information needed for the case study. I would also like to thank Basel Yousef, Bazil Solomon, Anzar Ahamed and Shakil Ghori for their motivation, support and friendship.

At last but not least, my thank goes to my wife Bindu Devkota Sapkota for her love and support, to my children Kribin Sapkota and Eva Sapkota for being inspiration to me, to my parents (Umakanta Sapkota and Lilawati Sapkota), father-in-law (Bhimsen Devkota), mother-in-law(Chumkala Devkota), sisters (Rita Bhandari and Divya Sharma), brother-in-laws (Kanchan Devkota, Guru Prasad Bhandari) , sisters-in-laws (Rachana Devkota Parajuli and Srijana Devkota), nephews (Gurav Bhandari & Gyalab Bhandari) and nieces (Aarya Sharma, Sherya Sharma and Swechhya Bhandari) for inspiring and encouraging me.

Table of Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

A-Box	Assertion Box
BF	Baseline Framework
BPN	Business Process Management
CLO	Core Legal Ontology
CMM	Conditional Markov Model
CRF	Conditional Random Field
CSPL	Common Pattern Specification Language
CSS	Cascading Style Sheet
DL	Description Logic
DOLCE	Descriptive Ontology for Linguistic and Cognitive Engineering
DOM	Document Object Model
DSA	Document Structure Analysis
EF	Extended Framework
EMEA	European Medicines Agency
FBO	Frame Based Ontology
FDA	Food and Drug Administration
GATE	General Architecture for Text Engineering
GLP	Good Laboratory Practice
GMP	Good Manufacturing Practice
HMM	Hidden Markov Model
HTML	Hypertext Markup Language
IC	Information Content
IE	Information Extraction
IR	Information Retrieval
JAPE	Java Annotation Pattern Engine
KM	Knowledge Management
LCS	Lowest Common Subsumer
LKIF	Legal Knowledge Interchange Format
LLD	Language of Legal Discourse
LR	Language Resources
MEMM	Maximum Entropy Markov Model
MHRA	Medicines and Healthcare products Regulatory Agency
NLP	Natural Language Processing
OM	Ontology Mapping
OWL	Web Ontology Language
PDF	Portable Document Format

POS	Part of Speech
PR	Processing Resources
RCM	Regulatory Compliance Management
RDF	Resources Description Framework
SA	Semantic Annotation
SOX	Sarbanes-Oxley Act
SQL	Structured Query Language
SWRL	Semantic Web Rule Language
T-Box	Terminological Box
UMLS	Unified Medical Language System
URN	User Requirements Notation
W3C	Worldwide Web Consortium
XML	Extensible Markup Language

1 Introduction

1.1 Research Overview and Motivation

Regulatory Compliance Management (RCM) is a management process, which an organization implements to conform to the relevant regulations. RCM is applied to all sizes and sectors of businesses and organisations from small to corporate level (Haider 2002, Watts 2006). Managing the regulatory compliance manually is a laborious and extensive task and necessitates expertise in the field. It costs a huge amount of capital investment to the organisations and still remains as an error prone process (Haider 2006). Legislations impose stringent compliance requirements, and organisations have to make heavy investments in order to meet the requirements. For instance, in 2002, the enforcement of the Sarbanes-Oxley Act (SOX) on organisations in the USA cost around $1.4 trillion (Zhang 2007).

RCM comprises several processes such as extracting regulatory entities from regulatory guidelines, modelling regulatory guidelines, mapping the regulatory guidelines with organisational processes, compliance enforcement, compliance monitoring and compliance audit (El Kharbili 2012). An improvement in each process plays a significant role in the improvement of the overall compliance management. Various commercial tools such as Microsoft Security Compliance Manager[1], LexisNexis Compliance 360[2], IBM SCORE[3] and Xactium Compliance Manager[4] are developed in order to help in compliance enforcement, monitoring and auditing. In addition, frameworks for enforcing, monitoring and auditing (Agrawal et al 2006, Liu et al 2007, Namiri and Stojanovic 2007, Uszok et al 2004) have played a useful role in improving the RCM processes.

[1] http://technet.microsoft.com/en-us/library/cc677002.aspx
[2] http://www.lexisnexis.com/en-us/products/compliance-360.page
[3] http://www-03.ibm.com/software/products/us/en/score
[4] http://www.xactium.com/compliance-management-software/

However, the current (commercial) tools and frameworks fall short of appropriately addressing the issues of extraction, modelling and mapping in the RCM (Breaux et al 2008, Gao et al 2011, Ghanavati et al 2007, Kiyavitskaya et al 2009, Mu et al 2009, Sapkota et al 2012). This is because of the complexity of the text, unavailability of standard modelling techniques and inadequate mapping algorithms (Breaux et al 2008, Ghanavati et al 2007, Kiyavitskaya et al 2008).

Among the RCM processes, this research focuses on two processes: (i) extraction of meaningful entities from the regulatory text and (ii) mapping regulatory guidelines with organisational processes. These two processes help in updating a system when new regulatory guidelines are introduced or some changes in the existing guidelines are made. When a new regulatory guideline is introduced, a compliance manager needs to find out the organisational processes affected by the guidelines. In other words, she has to relate the regulatory guidelines with the organisational processes. The relating process is manual that means the compliance manager needs to compare each organisational process with each regulatory guideline, which is a time consuming, laborious work and prone to errors. Improving the automation and accuracy of these processes helps in the accuracy and the automation of RCM. Therefore, this thesis aims to contribute towards the automation on these two processes in the RCM: (i) extraction of the regulatory entities and (ii) mapping regulatory the guidelines with the processes.

Extracting regulatory entities from regulatory guidelines poses several challenges such as (i) processing complex document structure, (ii) determination of external and internal references (iii) anaphora and cataphora resolution and (iv) processing sentence continuations. Extraction frameworks have to work against these challenges in order to identify entities correctly and completely. Current frameworks extract entities in the regulatory guidelines mostly by using indicator terms (Gao et al 2011) and rules (Mu et al 2009). In this research, the combination of four components: (i) parser, (ii) definition terms, (iii) ontological concepts and (iv) rules is proposed. There is research that uses the combination of some of these components, and are

found useful in regulatory entity extraction. However, the combination of these four components - in order to extract regulatory entities - has not been explored yet. The author believes that this combination increases the correctness and completeness of the regulatory entity extraction.

The mapping between a regulatory guideline and an organisational process comes with challenges such as (i) ambiguity and complexity of the regulatory text, (ii) implicit information in the description of organizational processes and (iii) absence of a standard framework to work with regulation and process ontologies in order to facilitate the processes such as mapping. There are similarity algorithms, which are potential methodologies for the mapping such as the sentence to sentence mapping (Agirre et al 2012, Barzilay and Elhadad 2003, McCarthy et al 2012, Mohler et al 2011), concept to concept mapping (Chen et al 2010, Ge and Qiu 2008, Hawalah and Fasli 2011) and word to word mapping (Pedersen et al 2004). These similarity algorithms do not consider the specific nature of regulation and process ontologies for the mapping process. An Ontology is a semantic knowledge representation format, and utilizing ontological structure in the similarity computation between two entities makes the similarity computation process meaningful and accurate. This research has utilised the ontological structure of the regulation and process ontologies in order to map the regulatory guidelines and organisational processes.

The rest of the chapter is organised as follows. Section 1.2 highlights the aims and objectives of this research. The overview of the framework is provided in Section 1.3. The contributions and originality of this thesis are pointed out in Section 1.4. The scope of the research is provided in Section 1.5. Section 1.6 presents the overview of the structure of this thesis. Finally, Section 1.7 lists the publications made during this research.

1.2 Aims and Objectives

The aim of this research is to propose a generic semantic framework in order to support RCM system. In order to achieve this aim, the following objectives are set:

- To explore the advancements and issues associated with: (i) the extraction of regulatory entities from regulatory guidelines and (ii) the mapping between the regulatory guidelines and organisational processes.
- To investigate and develop methodologies in order to extract regulatory entities from the text in regulatory guidelines. Extraction of the entities helps in relating the regulatory guidelines with organisational processes with more accuracy and completeness.
- To investigate and develop methodologies in order to map regulatory guidelines with organisational processes. The mapping process helps to update RCM when a new regulatory guideline is introduced, or there are changes in existing guidelines.
- To implement and evaluate the extraction and mapping methodologies in a case study which provides a sufficient level of complexity and can be validated within the scope, resources and time of the research.

1.3 Summary of Proposed Approach

In this thesis, the proposed approach is called RegCMantic, where RegCM refers to "Regulatory Compliance Management" and CMantic refers to "Semantic". The processes of the RegCMantic framework are divided into two main phases; (i) extraction phase and (ii) mapping phase. In the first phase, the regulatory entities are extracted from regulatory guidelines. In the second phase, the regulatory guidelines are mapped with organisational processes.

In the first part of the framework, the extraction of regulatory entities is carried out with a combination of four components: (i) parser, (ii) definition terms, (iii) ontological concepts and

(iv) rules. Prior to the extraction, the framework identifies the structure of the regulatory document and important document components such as regulatory paragraphs, topics and titles. Once the extraction is completed, the extracted regulatory entities are populated in regulation ontology. In the second part of the framework, the regulatory guidelines, which are represented in a regulation ontology, and organisational processes, which are represented in a process ontology are mapped. Mapping is a process to determine whether a regulatory guideline is related to an organisational process.

The validation of the framework is carried out by implementing it in a case study related to the Pharmaceutical industry in the UK. Within this domain, an Aspirin production process and the EU regulatory guidelines to govern the process, Eudralex (Eudralex 2013) are selected since the regulatory guidelines comprise a fair amount of complexity for the framework to be tested. Likewise, the Aspirin production process has a comparatively clear and simple structure and are modelled into ontological concepts in a process ontology (Sesen et al 2010).

The extraction part of the framework is evaluated using precision, recall and f-measure. In order to use these techniques, the result of the framework is compared with manual annotations. Since there is no annotation benchmark to compare with system-generated annotations, the annotations are compared with annotations created by the user. The comparison has created three types of annotations: correct annotations (true positive), incorrect annotations (false positive) and missing annotations (false negative). Likewise, the mapping part of the framework has also been evaluated using the same techniques. The mappings generated by the framework are compared with the manual mappings, and three types of mappings are identified: correct mappings (true positive), incorrect mappings (false positive) and missing mappings (false negative). The performances of the both parts of the RegCMantic framework are compared with the other related frameworks.

1.4 Contribution and Originality

The contribution of this research lies in the proposed RegCMantic framework. This framework is put forward as a result of reviewing the existing frameworks, and particularly the future-work suggestion presented in the papers by Sesen *et al*. (2010) and Kharbili *et al*. (2010). The suggestion is to improve the RCM by providing the automation in extracting semantic-regulation and relating the organisational processes with the applicable regulations. There are stand-alone frameworks for Information Extraction (IE) (Castillo et al 2003, Sarawagi and Agichtein 2006), ontology population (Müller et al 2004), similarities (Richardson et al 1994, Slimani et al 2006, Yang and Powers 2007) and ontology mapping (Doan et al 2003, Kalfoglou and Schorlemmer 2003, Noy 2004). However, integration of these innovative approaches in the RCM has not been explored yet. Considering these two processes as crucial constituents in the RCM, this thesis claims the following contributions and the originalities. The originalities are the innovations provided by this research. The contributions are the combination of the innovations and the adaptation of the existing approaches in this thesis.

1) **Algorithm to Identify Document Components and Predicting Document Structure:** A document contains various document components, which constitutes the structure of the document. Some examples of the components are title, paragraph, headers and footers. In order to extract meaningful regulatory entities from the regulatory text, it is essential to identify the document-components that contain regulatory guidelines. This thesis has created some algorithms to identify these components and the document structure.
2) **Algorithm to Identify the Regulatory Guidelines:** From the document structure, it identifies the regulatory guidelines in the document.
3) **Algorithm to Identify Meaningful Entities in the Regulatory Guidelines:** With in the regulatory guidelines, this framework identifies the important regulatory entities such as the subject, object, action and obligation. Identification of the regulatory

entities helps in relating the regulatory guidelines with organisational processes automatically.

4) **Tools for Constructing Regulatory Ontology and Representing the Regulatory Entities and Regulatory Guidelines in the Ontology:** An ontology to represent the regulatory guidelines and regulatory entities is essential for further processing the information in a semantic way. This research has constructed a regulatory ontology by extending an existing upper level legal ontology.
5) **Computing Similarity between the Entities of Regulatory Guidelines and Organisational Processes:** In order to compute similarity between a regulatory guideline and an organisational process, it is essential to identify the similarity between their entities. For example, determining the similarity between the subjects and actions of a regulatory guideline and an organisational process helps in determining the similarity between the guideline and the process. This research computes the similarity between the entities in regulatory guidelines and organisational processes.
6) **Computing Similarity between Regulatory Statements and Organisational Processes:** A regulatory guideline contains one or more regulatory statements. Before relating the regulatory guideline to organisational processes, it is essential to relate its statement with the processes. This framework computes the relatedness of a statement with processes.
7) **Computing Similarity between Regulatory Guidelines Organisational Processes:** Finally, this research determines the relatedness between a regulatory guideline and an organisational process.

Among the above contributions, the author claims the originality in the followings:

1) **Identifying the Regulatory Guidelines and Entities in a Regulation Document:** The proposed framework identifies the regulatory guidelines from various document

structures. Furthermore, within the regulatory guidelines, it identifies the regulatory entities.

2) **Relating Regulatory Guidelines to Organisational Processes:** In order to facilitate the compliance manager with automation in the update process in RCM, the proposed framework relates the regulatory guidelines with organisational processes with the help of regulatory entities and process entities.

1.5 Scope

RegCMantic is a general framework, and likely to be applied to any domain where organisational processes are represented in an ontology. Some of the boundaries under which this research is carried out are listed below.

1) The RCM framework assumes that the organisational processes are represented in an ontology in order to relate the organisational processes with regulatory guidelines. In the case study of this framework, a process ontology called OntoReg is used. The OntoReg is developed by a team at University of Oxford (Sesen et al 2010) during their continuous research into Pharmaceutical processes.
2) The extraction of regulatory guidelines from the regulation text requires manual intervention. The extraction process generates suggestions, and the users need to select or modify the suggestions. In other words, the ultimate decision should always be allowed to the users since the extraction does not produce 100% accurate results.

3) In the mapping process, which is the process of finding the relevant regulation for an organisational process, the notion of ontology mapping cannot be applied as such, since regulatory guidelines and organisational processes are not similar concepts. Ontology mapping can be applied to determine the similarity between two similar concepts in two similar ontologies. Since the regulatory ontologies and process ontologies are completely different ontologies, and regulatory guidelines and organisational processes are different concepts, the state of art ontology mapping algorithms cannot be applied in the RegCMantic as such.
4) The mapping process needs expert's intervention to select the mapping. Similar to the extraction process, the system only generates suggestions, and the user should either select or modify the suggestion.

1.6 Organisation of the Thesis

The remaining part of the thesis is organised into various chapters, which is given below.

Chapter 2 describes the technologies used in this thesis and reviews the related frameworks. In particular, it describes the natural language processing, Semantic Web technologies and similarity measures. Introducing these technologies, this chapter then reviews various approaches to the compliance management and justifies the selection of the proposed framework.

Chapter 3 describes the proposed RegCMatic framework for RCM. It describes the contribution, scope and limitation of the framework. This chapter also explains how the proposed framework can be implemented. In particular, it clarifies various algorithms, implementation requirements, guidance and warnings.

Chapter 4 describes a case study implementing the proposed framework. It examines the application of the RegCMantic framework in a real life scenario, the Pharmaceutical industry

as its case study. In particular, application of the Eudralex regulation to the Pharmaceutical processes is observed.

Chapter 5 discusses and analyses the findings of the case study. Each phase of the framework is evaluated with expertise in the domain. Finally, the Chapter 6 concludes the finding of this research and highlights the potential directions of the future research.

1.7 Publications

While working on this thesis, some papers are published, and they are listed below.

Sapkota K, Aldea A, Younas M and Duce DA (2013) RP-Match : A Framework for Automatic Mapping of Regulations with Organizational Processes. *The 10th IEEE International Conference on e-Business Engineering (ICEBE 2013)*. Coventry, UK: IEEE Computer Society Press, (Accepted).

Sapkota K, Aldea A, Younas M, Duce DA and Banares-Alcantara R (2012) Extracting Meaningful Entities from Regulatory Text. *Proceedings of the Fifth International Workshop on Requirements Engineering and Law (RELAW '12)*. Chicago: IEEE Computer Society Press, 29–32.

Sapkota K, Aldea A and Banares-Alcantara R (2012) Semantic Knowledge Mapping: An Extension of Compendium with Semantic Knowledge Representation. *International Journal of Artificial Intelligence & Applications (IJAIA)* 3(5): 1–12.

Sapkota K, Aldea A, Younas M, Duce DA and Banares-Alcantara R (2011a) Towards Semantic Methodologies for Automatic Regulatory Compliance Support. *Proceedings of the 4th workshop on Workshop for Ph.D. Students in Information & Knowledge Management (PIKM '11)*. Glasgow: ACM Press, 83–86.

Sapkota K, Aldea A, Younas M, Duce DA and Banares-Alcantara R (2011b) Semantic-ART: a framework for semantic annotation of regulatory text. *Proceedings of the Fourth Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR '11)*. Glasgow: ACM Press, 23–24.

2 Literature Review

2.1 Introduction

The purpose of this chapter is to review and critically analyse current Regulatory Compliance Management (RCM) systems, in particular how the existing systems manage the changes due to new regulatory guidelines. Furthermore, it reviews the current RCM systems and some methodologies, techniques and tools connected to them such as Semantic Web technologies, IE and similarity measures. When some regulatory guidelines are introduced, or some changes in the existing regulatory guidelines are made, a RCM has to relate the regulatory guidelines to the existing organisational processes. This chapter investigates how the semantics embedded in the regulatory guidelines are identified, extracted and represented in order to relate them with the organisational processes.

In order to identify the semantics of the regulatory guidelines, it is needed to identify the structure of the regulatory documents and the meaningful entities in the regulatory guidelines. Identification of the structure of a document is called Document Structure Analysis (DSA), and that of meaningful and relevant entities is called Information Extraction (IE). Determining the similarity between the regulatory entities and process entities helps in mapping the regulatory guidelines with organisational processes. Application of the Semantic Web technologies can make the extraction and mapping processes semantic. Therefore, this chapter describes the approaches to RCM, Semantic Web technologies, DSA, IE and similarity measures.

The rest of the Chapter is organised as follows. Section 2.2 reviews the RCM approaches and the issues. The main goal of this thesis is to help in RCM. This section looks into the strength and limitations of the current frameworks. Semantic Web technologies and their uses in the RCM are described in Section 2.3. The approaches to the DSA are explored in Section 2.4. In

Section 2.5, the processes of regulation extraction and semantic annotations are analysed, and the selection of a method is justified. Various techniques of semantic similarity are reviewed in Section 2.6. This section reviews the current similarity measures and justifies why some similarity measures are selected for this thesis.

2.2 Regulatory Compliance Management

This section reviews various RCM approaches. In particular, it analyses the strength and limitations of the existing approaches and describes how the proposed framework can fulfil some of the shortcomings of the existing approaches.

RCM is a process that aims to ensure that requirements are satisfied with organisational processes (Yip et al. 2007). RCM ensures public health, safety and security and maintains the quality of the products. RCM is applied to all sizes and sectors of businesses and organisations from small to corporate level (Haider 2002, Watts 2006). RCM is very important to businesses and organisations since, on one hand, failure to comply with the regulatory guidelines results into heavy penalties (Breaux et al 2006); and on the other hand, managing regulatory compliance also costs substantially. For an instance, according to a survey (Zhang 2007), enforcement of the Sarbanes-Oxley Act (SOX) on organisations in the USA in 2002 cost around $1.4 trillion. Managing the regulatory compliance manually (Conley 2000, Haider 2006, Muhammed 2007) is a laborious and extensive task and necessitates expertise in the field which costs a huge amount of capital investment to the organisations and still remains as an error prone process (Haider 2006).

Improving RCM requires improvement on its underlying processes. RCM comprises several processes such as (i) extracting regulatory entities from regulatory guidelines, (ii) modelling regulatory guidelines, (iii) mapping the regulatory guidelines with organisational processes, (iv) compliance enforcement, (v) compliance monitoring and (vi) compliance audit (El Kharbili 2012). Therefore, the improvement in each process plays a significant role in

improvement of the overall compliance management. For instance, providing automation in some of these processes increases the automation of the overall RCM process.

There are several contributions, which provide automation in RCM processes. For instance, the work presented in Kiyavitskaya et al. (2007) extracted rights and obligations from regulations by applying a framework called Cerno, which involved textual semantic annotation. Knowledge Acquisition in Automated Specification (KAoS) services were applied by Uszok et al. (2004) to RCM for Semantic Web services. Although these frameworks extract regulatory entities from regulatory guidelines, mapping the regulatory guidelines with organisational processes has not been addressed.

Equally, the database technologies appeared as useful in the process when Agrawal et al. (2006) applied these technologies to check compliance in RCM. User Requirements Notation (URN) based framework was explored by Ghanavati et al. (2007) to track legal compliance in healthcare. Liu et al. (2007) applied statistic methods for compliance checking in Business Process Managing (BPM). One of the interesting implementations of the logical approach was described in Namiri & Stojanovic (2007) by implementing internal controls in business processes with an introduction of a semantic layer without changing the original business process. Likewise, logical exploration of Logrippo (2008) and Governatori et al. (2009) on RCM was found encouraging. Although these frameworks contributed towards checking the compliance and noncompliance processes, extracting and relating regulatory guidelines with organisational processes has not been addressed.

Similarly, Semantic Web technologies are found useful in RCM. One of the major contributions towards it is (El Kharbili et al 2008), where the authors investigated creating a policy based framework for semantic business process and compliance management. A semantic based RCM for the Pharmaceutical industry is described in Contreras & Banares-Alcantara (2009), where the authors developed a pharmaceutical regulatory knowledge base to support the validation process of new chemical products. The same work is further extended

by Sesen et al. (2010) using ontology and rule based compliance checking in pharmaceutical processes. These frameworks are focusing on managing compliance knowledge manually and checking compliance. They have not addressed the issues of extracting regulatory guidelines and mapping them with organisational processes automatically.

Although, there is a considerable amount of improvement in some parts of the compliance management, some processes are still in need of improvement. For example, there are some commercial tools such as Microsoft Security Compliance Manager[5], LexisNexis Compliance 360 [6], IBM SCORE [7] and Xactium Compliance Manager [8] which help in compliance enforcement, monitoring and auditing. However, there are less improvements in the extraction, modelling and mapping processes related to RCM (Breaux et al 2008, Gao et al 2011, Governatori et al 2009, Kiyavitskaya et al 2009, Mu et al 2009, Sapkota et al 2012). This is because of the complexity of the text, unavailability of standard modelling techniques and inadequate mapping algorithms.

Updating RCM with changes in regulatory guidelines automatically needs automation in the extraction, modelling and mapping processes. In order to automate these processes, RCM has to exploit the advancement in technologies, which can streamline the processes such as knowledge management, IE and similarity measures. Standardising the modelling of regulatory knowledge requires expertise in the regulatory domain to agree in some formats and techniques, which requires another area of exploration and the time and effort needed for the exploration exceeds the scope of this thesis. Therefore, this research focuses on automating two processes: extracting regulatory guidelines and mapping the guidelines with organisational processes.

[5] http://technet.microsoft.com/en-us/library/cc677002.aspx
[6] http://www.lexisnexis.com/en-us/products/compliance-360.page
[7] http://www-03.ibm.com/software/products/us/en/score
[8] http://www.xactium.com/compliance-management-software/

This section has described some of the RCM approaches, their strength and limitations. The next section discuses about the advantage of representing regulatory knowledge in a semantic format and analyses the various approaches in knowledge representation.

2.3 Knowledge Management and Ontologies

This section reviews the technologies that can be used to represent and manage the regulatory information. In particular, it describes knowledge management (KM) in terms of its role in a semantic RCM. It introduces a semantic knowledge representation format, ontology and describes how some of the legal ontologies represent the knowledge in regulatory guidelines.

KM is a process of organising the knowledge, which involves three important tasks: (i) identifying the required and the available knowledge, (ii) planning the processes on the basis of the available knowledge, and (iii) urging the appropriate actions based on the planning (Castillo et al 2003). The implicit knowledge can be made explicit by using the Semantic Web technologies. When the knowledge carries the maximum semantics, it can be utilised to its maximum potential. Ontologies can be used as Semantic Web technologies.

2.3.1 Ontology for Knowledge Representation

Representing knowledge in ontologies optimises the usefulness of the knowledge (Gruber 1995). Gruber (1993) defines ontology as "an explicit specification of a conceptualization". Another definition (Borst 1997) states that it is "a formal specification of a shared conceptualisation". Based on these definitions, there arose another definition (Guarino 1998), which combined the earlier ones and presented as "an explicit and formal specification of conceptualisation", which is the most referenced definition (George 2010). The conceptualisation refers to the vocabulary and the intentional meanings of the domain; explicit

refers to making the knowledge in the ontology explicit, and formal refers to making the model of the domain machine interpretable (Horrocks 2003).

Structurally, an ontology is a taxonomy of concepts and description of their relationships and attributes which captures the intended meaning of a domain. There had been several attempts to standardize the format of ontology since its popularity grew in late 20th century such as Simple HTML Ontology Extensions (SHOE), Ontology Inference Layer (OIL) and DAPRA Agent Mark-up Language (DAML). Finally, the Worldwide Web Consortium (W3C) recommended the Web Ontology Language (OWL) as a standard format for writing ontologies. OWL is based on the description logic and comes in three flavour such as OWL-Lite, OWL DL and OWL Full. Recently, a new version, OWL 2.0 is recommended by W3C. Implementing OWL ontology to represent the regulatory guidelines can be the best option considering the following reasons.

(1) The popular legal ontologies such as LRI-Core (Breuker and Hoekstra 2004) and LKIF-Core (Hoekstra et al 2007) are represented in OWL, which can be extended to represent the knowledge in the regulatory guidelines.
(2) The organisational processes are represented in OWL ontology; representing regulatory guidelines in the same format will make the processing easier.
(3) OWL is the default ontology in the popular ontology editor, Protégé.

Ontologies are categorized differently by various authors considering different criteria. According to Guarino (1998), it can be classified into three types based on the level of their generality. The first type is the top-level ontology, which is the most generic one and captures the domain-independent knowledge. It is also called the upper-ontology or foundation-ontology. The examples of upper ontologies are Cyc[9] and WordNet[10]. The second type is the domain or task ontology, which captures the knowledge of a generic domain or task. The third type is called the application ontology, which captures the knowledge of a domain. The use of

[9] http://www.cyc.com/about-cyc
[10] http://wordnet.princeton.edu/

the appropriate level of generality can save the expertise time while creating an ontology for an application. Similarly, reusing the existing ontology eases a knowledge engineering process.

Representing regulatory guidelines in an ontology makes the guidelines machine understandable and helps in further processes such as relating the guidelines with organisational processes. The next section describes how ontologies are used to represent the knowledge in the regulatory guidelines.

2.3.2 Ontology for Representing Regulatory Guidelines

While creating an ontology for an application, it is recommended that one should make use of the existing ontologies. Therefore, this research has considered reusing or extending an existing legal ontology. In this section, some state of art legal ontologies are analysed and justified why one is more suitable than the others are, for this research.

Language of legal discourse (LLD) (McCarty 1989) implements rules and formulas in order to capture "deep conceptual models" of a particular legal domain in terms of practical as well as theoretical application. Similarly, Gangemi et al. (2003) created core legal ontology (CLO), which is based on a foundational ontology, descriptive ontology for linguistic and cognitive engineering (DOLCE) (Gangemi et al 2002) with the inclusion of description and situation. Breuker & Hoekstra (2004) developed LRI-Core, a core ontology that covers the main concepts that are common to all legal domains.

There are several amendments to the top level and core level concepts in the legal ontologies. The legal knowledge interchange format core ontology (LKIF-Core) (Hoekstra et al 2007) was built on the LRI-Core (Breuker and Hoekstra 2004) ontology and developed considering all the legal ontologies and is regarded as a standard for the core legal ontologies. It is a legal core ontology and translates legal knowledge bases written in different representation formats and languages. It aims to play two roles: (i) interpreting legal knowledge bases in different languages and (ii) formalising the legal knowledge. Legal knowledge acquisition can be made

easy by defining the concepts such as `norm`, `judge`, `liability`, `document` and `claim`. The ontology is described under three layers: (i) top level, (ii) intentional level and (iii) legal level.

2.3.3 Relevancy of the Knowledge Representation Technologies

In relation to this thesis, the ontologies can be used to represent the knowledge of regulatory guidelines and organisational processes. The regulatory guidelines and organisational processes can be modelled into ontologies in order to make them machine understandable by creating and interpreting the relationships of their concepts. Most of the legal ontologies have attempted to standardize the top-level concepts by adhering to the concepts in the popular top-level ontologies, and they seemed to have adopted three tier ontologies (top, core and domain). Among the top concepts found in these ontologies are those representing `space`, `time`, `abstract`, `physical` and `mental-entities`. The top level ontologies only comprise the pure generic concepts, whereas the core ontology contains a rich set of low-level, reusable terms in the legal domain such as `norm`, which is a core concept placed under the top concept `mental-entity`. The domain ontology extends the above two ontologies (i.e. top and core ontologies), and designed to work in a specific legal domain such as regulations for the Pharmaceutical industry.

The LKIF-Core ontology is found to be more suitable than other ontologies because (1) it is the latest development in the legal KR community and (2) It has useful concepts, which can be extended to represent the knowledge in regulatory guidelines.

This section has described how knowledge representation in a semantic format helps in RCM and justified the selection of OWL ontology format and LKIF-Core ontology. The next section discusses about the various techniques, which can be adapted to identify the structure of a regulatory document.

2.4 Document Structure Analysis (DSA)

This section reviews the techniques to identify the structure of the regulatory guidelines. Identifying regulatory entities in regulatory guidelines helps RCM to relate these guidelines to organisational processes. If the structure of the regulatory document is identified, the regulatory entities can be extracted more accurately.

A document is composed of various document-components such as sections, paragraphs, titles and page numbers. Identifying these components within a document is referred to as DSA.

Processing regulatory guidelines needs to identify the structure of the documents which are published in different document-formats such as PDF, HTML, XML, text and doc. Mostly these documents are in electronic format (digital documents). Understanding the structure of a digital document is a challenge underpinned by different formats and not properly defined structures. Other challenges on DSA as identified in Nojoumian & Lethbridge (2007) are given below.

- Although statutes are laid out in a particular way, in general practice, it is not rigorously followed.
- Cross-referencing and signposting are the good practices in presenting coherent information in documents. There are two types of cross references; explicit and implicit. In the explicit cross-references, the target information is represented by clear indications such as hyperlink or section numbers. However, in implicit cross-references, there is a lack of clear representation of the target and poses challenges such as misleading and inconsistent information.
- There is a considerable amount of noise on the documents provided online. Most of the documents on the web are PDF and HTML, and these documents are created by some kinds of document generators, which contain unnecessary information.

2.4.1 DSA Approaches

Various techniques are proposed for tackling the challenges in DSA, and the approaches are classified according to the type of algorithm used. In general, DSA can be divided into two phases (Song Mao et al 2003): (1) document physical structure analysis and (2) document logical structure analysis.

2.4.1.1 Document Physical Structure Analysis

The physical DSA identifies various physical entities or regions in a document such as text-blocks, lines, words, figures, tables and backgrounds (Namboodiri and Jain 2007). The approaches to document physical layout analysis are categorized into top-down, bottom-up or hybrid approaches (Song Mao et al 2003). In the top-down approaches, the document-components discovery starts from the whole document, and the smaller components are discovered iteratively. The process of iteration continues until a breaking criterion is met (Kise et al 1998, O'Gorman 1993). In the bottom-up approaches, the process starts from the smaller units in a document such as letters, lines and ultimately the whole document is identified (Baird et al 1990, Nagy et al 1992). There are some approaches which utilise the both paradigms and are called mix approaches (Pavlidis and Zhou 1992).

2.4.1.2 Document Logical Structure Analysis

The logical or functional DSA determines the logical components of a document such as titles, authors, affiliations, authors, keywords, introduction and conclusion (Luong et al 2010, Song Mao et al 2003). In document logical structure, document-segments are arranged in a hierarchy. A document-segment can be defined as a unit of a document or a document-component in a specific position with specific features (Agichtein and Ganti 2004). Similar to the steps in the physical structure analysis, logical structure analysis comprises the two crucial steps; segmentation and classification (M-w Lin et al 2006). In the segmentation stage, various types of document-segments are identified such as paragraph, title, graphics and images. In particular, the text-segments are identified with the help of font-features, position and

indicators (Anjewierden 2001). The font-feature is the morphological observation of a text such as font-size, font-weight, font-style and font-colour. The position is the location of the text with respect to the vertical and horizontal axis of the document as well its adjacent segments. The indicators specify the nature of a text-segment such as list numbers or segment prefixes or labels. These attributes of a segment help to determine the classification of the segment. The classification is the process of arranging the segments in a hierarchy and needs to determine containment of each segment.

The document logical analysis approaches can be categorised into three basic classes: rule based, grammar based and other approaches (Stoffel et al 2010). In the rule based approaches a set of rules is defined to identify and label the document-components (Ishitani 1999, Kim et al 2001, C. C. Lin et al 1997). Likewise, various approaches (Conway 1993, Krishnamoorthy et al 1993, Tateisi and Itoh 1994) proposed to specify grammar rules in different ways. These grammars also identify the document-components and assign labels to them. In addition, there are various other approaches such as machine learning (Esposito et al 2008, Paaß and Konya 2012) and probabilistic approaches (Klink et al 1999).

2.4.2 Relevancy of DSA Approaches

Since the current regulatory guidelines are presented in electronic formats, this research is related to the approaches of the logical document analyses. However, it is important to know where the document logical analysis process falls under the overall document analysis process. The earlier approaches of converting the printed documents to digital documents have shown holistic approaches from physical to logical analysis. The steps in the logical analysis are directly related to the current research since it aims to identify the document-component where the regulatory guidelines are embedded.

The two basic steps of the logical structure analysis: segmentation and classification (Pavlidis and Zhou 1992) are selected for this research. Likewise, the combination of bottom-up and top-down approach has also been found appropriate to this research as well as the

identification of segments by defining rules (Ishitani 1999, Kim et al 2001, C. C. Lin et al 1997). The above approaches proposed methods to process the segmentation step with the help of font-features, positions and indicators. However, it is observed that the research is still in need of exploring the special nature of regulatory documents. This research aims to analyse the special nature of the regulatory documents such as excessive use of model verbs, passive voice and segment-indicators.

This section has described the approaches, which can be used to identify the structure a regulatory document, which can help in the extraction of regulatory entities. The next section discusses various extraction techniques and frameworks, which can be used to extract regulatory entities from the regulatory guidelines.

2.5 Information Extraction (IE)

This section reviews the techniques that deal with the identification, annotation and extraction of information in regulatory guidelines. One of the processes, which can improve RCM, is the extraction of regulatory entities from regulatory guidelines. Once the document-structure is identified, the regulatory entities in the regulatory document should be extracted to make the regulatory guidelines meaningful. The meaningful regulatory guidelines helps to update RCM with changes or updates in the regulatory guidelines.

IE is a branch of Natural Language Processing (NLP), which is used to extract specific information from the provided natural language documents. IE is used to extract information from the unstructured and semi structured text and convert them into a structured format. The unstructured text is the text available in most of the traditional, non-Web format such as text, word and PDF documents, and they are harder to process. The text in the web is organised into some distinct structures such as tables, lists, paragraphs and headings. These texts are called semi-structured text and are easier to process. The structured texts are stored in a well organised and computer interpretable formats such as database tables, XML documents and

ontologies. Since the information on the web is increasing exponentially, it is desirable that the information is in a machine interpretable format, and IE helps to convert the static information on the web to the computer understandable format.

IE is often confused with the similar process called Information Retrieval (IR). IR is concerned about looking for the relevant documents based on the specific criteria and then finding specific information within the resulting documents (Baeza-Yates and Ribeiro-Neto 1999). IE is about searching for the specific text within the documents and returning the exact answers to the user's query instead of returning the list of related documents (Sarawagi 2007). IR is more mature field than IE. Although they look similar, IR was evolved by the influence of information theory, probabilistic theory and statistics, whereas IE was developed by the influence of rule-based systems in computational linguistics and natural language processing (Lee 2005).

Semantic Annotation (SA) is a process of tagging ontological concepts in the provided text, which will be used to populate the instances of the concepts in the ontology. The SA is not only used to extract the concepts and instances, but also the relationship among the concepts. It can be considered as a subfield of the IE or an Ontology Based Information Extraction (OBIE), albeit, some schools of thought consider it as a separate field.

IE is applied in various area such as personal information management, scientific applications, web oriented applications and commercial applications. Its application in commercial software led its popularity and promoted more research in this direction. The most popular applications in such a direction are news tracking, customer care, data cleaning and classified advertisements (Sarawagi 2007).

The approaches to IE are classified into different categories by different surveyors. Some are considering the basis of automation for the classification, whereas the others are considering whether an approach is a rule based or a statistic based. Based on the automation, the

approaches are categorised into the manual Knowledge Engineering approach and the Automatic Training approach. Each of these approaches is described below.

2.5.1 Knowledge Engineering Approach

In this approach, the experts in knowledge engineering and the domain in question are expected to create the extraction rules manually (Sarawagi 2007). This is an iterative process, and needs continued modification until the acceptable level of the expected result is acquired.

There are various representation formats for the rules; however, the basic idea and structure are similar in all of them. Some of the systems and tools have gained popularity in creating and representing in a specific format such as Common Pattern Specification Language (CSPL) (D Appelt et al 1993). CSPL is implemented in Java Annotation Pattern engine (JAPE) (Cunningham et al 2002), regular expressions are used in WHISK (Soderland 1999), SQL expressions are used in Avatar (Jayram et al 2006) and Datalog expressions are used in DBLife (Shen et al 2007).

Similar to logic programming, the general structure of the rule consists of two parts: (1) contextual patterns as premise and (2) actions as consequences (Sarawagi 2007). The patterns are defined by rules, which are more or less similar to the regular expression or by simply defined by a list of terms or entities. The actions part of the rule performs various useful actions that are related to entity annotation such as modifying the annotation, inserting start and end nodes in the entity and grouping the entities in a specified order. When multiple entities are grouped together to form a single entity, it requires the creation of the boundaries. The boundaries are the start and the end nodes (points) and all the tokens inside the boundaries are regarded as an entity. Examples of entities with multiple words are book chapters, title of a paper, journal names and book names (Sarawagi 2007).

2.5.2 Automatic Training Approach

In this approach, the machine learning technique is used to generate rules or statistics from the large training corpora automatically (Manning and Schütze 1999).

In the rule based training approach, the training corpus should be annotated manually and based on the handcrafted annotation the new rules are generated. The two major types of algorithms for the rules creation and determination of the rules with the high precision are the top-down approaches (Soderland 1999) and bottom-up approaches (Califf and Mooney 2004).

In the statistical approaches, the system tries to find the hidden structure in the unlabelled corpus by using various statistical models such as Hidden Markov Models (HMM) (Agichtein and Ganti 2004), Maximum Entropy Markov Model (MEMM) (Ratnaparkhi 1999), Conditional Markov Models (CMM) (Malouf 2002) and Conditional Random Field (CRF) (Lafferty et al 2001). In order to determine a model, the unstructured text is decomposed in to smaller parts based on the predefined separators such as space, comma and full-stop. These smaller parts are then labelled using various training and inference algorithms (Tsochantaridis et al 2005, Vishwanathan et al 2006).

2.5.3 Information Extraction Frameworks

Among the IE frameworks, the three main frameworks that are used in many semantic annotation tools are Annotea (Kahan et al 2001), CREAM (Handschuh and Staab 2002) and Amilcare (Ciravegna et al 2003). Similarly, The General Architecture for Text Engineering (GATE) (Cunningham et al 2002) has been popular in the recent decade because it provides a platform to test NLP related tools and frameworks.

Annotea is designed for collaborative working on mainly HTML and XML documents. It employs XPointer, a W3C recommended method for locating annotations in the text. The annotated text is converted into the RDF triples, other W3C recommendations for data

interchange on the web. Various tools have embraced this framework such as Amaya, Annozilla and Vannotea.

The CREAM framework was developed in the University of Karlsruhe (Handschuh and Staab 2002). Similar to Annotea, it also follows the W3C recommended technologies such as RDF, OWL and XPointers. It goes beyond the scope of Annotea in that it also covers the semantic annotation of the deep web pages. In other words, it can generate the annotations from the databases from where the web pages are generated. Furthermore, it can generate annotation for the relation, which is essential for the knowledgebase of semantic services. Among the tools that support this framework are S-CREAM and M-OntoMat-Annotizer.

The Amilcare framework was developed by a team at the University of Sheffield (Ciravegna et al 2003). This is a pattern-discovery approach, which can be used in various kinds of documents including the web pages and the text documents. The framework comprises three main phases namely training, testing and production. In the training phase, an ontology and a training corpus are set as input and a set of rules are generated as its output. In the testing phase, the newly generated rules are executed against a new unlabelled corpus. The measurements of accuracy, such as precision, recall and f-measures are calculated and analysed in this phase. In the production phase, the training processes can be enhanced by a combination of system annotation and user annotations. This is particularly useful when the application is released and lesser user annotation is expected. Examples of the tools using these frameworks are Armadillo, AeroSWARM and Melita.

GATE is a toolkit in the field of NLP (Cunningham et al 2002) and is used as an IE tool in various domains such as bioinformatics, health and safety. GATE was designed for organising tasks such as data storage, data visualization, location and loading of components and executions of processes from the data structures and algorithms that process natural languages. The four main components in the GATE development environments are Language Resources (LR), Processing Resources (PR), Applications and the Data Stores. Among them, the most

crucial component is the PR. All the IE and text annotation tools are integrated in the PR. The LR represents the documents in different formats. The Application is a battery of processing resources to be run through the language resources, and the purpose of the Data Store is to store the annotated documents. A summary of the semantic annotation tools is depicted in Table 2-1. This table displays names of the IE frameworks, their extraction techniques and automation. A comprehensive review, analysis and comparison of these tools are available in Uren et al. (2006).

Table 2-1. Comparison of various extraction tools

Extraction Tools	Automation	Manual Rules	Pattern	Wrapper	Based On
AeroDAML	✓	✓	--	--	
KIM	✓	✓			
MUSE		✓	✓		
MnM	✓			✓	Amilcare
Ont-O-Mat: Amilcare	✓			✓	Amilcare
SemTag	✓			✓	
Ont-O-Mat: PANKOW	✓		✓		

Armadillo			✓		Amilcare
Amaya		✓			Annotea
Annozilla		✓			Annotea
Mangrove		✓			Annotea
Vannotea		✓			Annotea
OntoMat Annotizer		✓			CREAM
S-CREAM		✓			CREAM
M-OntoMat - Annotizer		✓			CREAM
OntoAnnotate		✓			CREAM
SHOE Knowledge Annotator		✓			
Running SHOE		✓		✓	
SMORE		✓			
Open Ontology Forge		✓			
COHSE Annotator		✓			Annotea
Lixto	✓			✓	
Melita	✓		✓		Amilcare

Cafetiere	✓	✓			
KnowItAll	✓		✓		Amilcare
SmartWeb	✓			✓	
AeroSWARM	✓		✓		Amilcare
Rainbow Project	✓	✓			
h-TechSight	✓	✓			
AktiveDoc	✓	✓	✓	✓	Amilcare
WickOffice	✓	✓			
OntoOffice	✓	✓			
Magpie		✓			
Thresher	✓	✓		✓	
GATE	✓	✓	✓		

.

2.5.4 Relevancy of the Information Extraction Frameworks

Between the two IE approaches, the rule based knowledge engineering approach is found suitable for the current research because the automatic training approaches need preparation of a large amount of training data, which exceeds the time required for the manual annotation of

the target regulation. Besides, creation of rules after analysing the structure of a regulation-document saves expertise time. Some rules are generic, which can be reused without modification.

Considering as a sub-domain of IE, SA can help in the extraction and representation of regulatory guidelines in a semantic format. Converting the regulation from raw-text to ontological individuals helps in inferring the hidden knowledge, which ultimately helps in semantic compliance checking. Using the ontology for the extraction of regulatory entities and populating an ontology can be considered as SA.

GATE is found suitable for IE in this research because it has the following advantages:

- It provides reusable java libraries for each NLP related task
- It allows us to work with WordNet the lexical ontology
- It provides interfaces and APIs to work with ontologies
- It is written in Java, which is platform independent.
- It is available freely under the GNU Library General Public Licence.
- It is considered as a standard platform and a tool to implement NLP related tasks, specifically information extraction. It supports the whole cycle of processes in IE. The tasks at the beginning such as collecting and organising the documents in a corpus, the tasks at the processing stage such as annotating, amending, deleting and exporting and the tasks at the end such as evaluation are all facilitated by the system.

In this section, technologies and frameworks for IE has been described, which can be used to extract regulatory entities from regulatory guidelines. The next section discusses various similarity approaches, which can be adapted to relate organisational processes with regulatory guidelines.

2.6 Semantic Similarity

This research aims to relate regulatory guidelines with organisational processes based on the existing similarity measures. Relating the guidelines with the processes requires identifying the similarity between them. Therefore, this section reviews the approaches to the similarity measures.

Semantic similarity is a comparison of two words in order to find out how much similar their meanings are. Semantic similarity is a useful component to improve various processes such as IR techniques (Lee 2005), ontology mapping (Euzenat and Valtchev 2004), word-sense disambiguation (Soler and Montoyo 2002) and reasoning (Rissland 2006). The similarity measure between two terms is evaluated as a numerical score. The score is computed with the help of some information sources such as WordNet lexical ontology (Hao et al 2011, Pedersen et al 2004, Richardson et al 1994), search engine (Bollegala et al 2007), Wikipedia (Gabrilovich and Markovitch 2006, Ponzetto and Strube 2007) and corpus (Jiang and Conrath 1997). There are several different approaches for similarity measure, which can be categorised based on their usage of the information sources. Some of the similarity approaches are described below.

2.6.1 Information Theoretic Approaches

Information theoretic approaches are based on the notion of Information Content (IC). IC of a concept is the probability of finding the concept in large corpora and can be exploited to measure the amount of information a concept expresses. As one goes deeper in a lexical taxonomy and chooses a concept, the concept will have less probability of its occurrence in corpora. Likewise, if one goes shallower in the lexical ontology and select a concept, the concept will likely to have more probability of its distribution in large corpora. Furthermore, the probability of finding the concept is cumulatively added to the super-concept from its sub-concept as one goes from specific-concepts to generic-concepts. The IC value is calculated by

considering negative of the log likelihood of the probability of the distribution of a concept in large corpora, and according to Resnik (1995), the formula is modelled as in Equation (1):

$$\boldsymbol{ic_{res}(c) = -\log p(c)} \quad (1)$$

In this equation, ic_{res} is the information content value computed using Resnik's methodology, c is a concept in a lexical ontology such as WordNet, $p(c)$ is the probability of encountering an instance of concept c in given corpora.

The formula is designed such a way that the IC of a concept monotonically decreases as one goes from a leaf concept towards its root concept in a lexical ontology. The purpose behind using negative likelihood is that the more frequently a concept appears in large corpora, the less information it expresses; thus making the specific-concepts more informative as compared to the abstract one. If there is a unique top node in the taxonomical hierarchy, then its probability value will be exact 1; hence its value of information content will be 0.

Example: Consider there are two concepts "carnivore" and "cat". Carnivore is more generic than cat. The probability of finding carnivore in a large corpus is higher than that of cat. This makes that carnivore less informative than cat. In other words, the IC value of carnivore is less than that of cat. In WordNet[11] similarity (see Figure 2-1), the IC values of carnivore and cat are 7.25 and 8.63 respectively.

In information theoretic concepts, the IC value of the two concepts considered are compared in order to compute the similarity metric between them. The three popular similarity measures, which employ IC, are Resnik (1995) similarity, Jiang & Conrath (1997) similarity and Lin (1998) similarity.

Resnik (1995) similarity is a semantic-similarity measure, which employs IC values for its similarity value calculation. In other words, it considers how much information the compared

[11] http://wordnet.princeton.edu/

concepts share each other. The shared information is the information given by the Least Commons Subsumer (LCS) that subsumes both concepts. It is also referred to as Most Specific Common Abstraction (MCSA). On its strength side, it has the simplicity in computing IC, which is deduced by counting the words in large corpora. However, measuring the occurrences in large corpora is a time intensive task and the IC value depends on the considered corpora. Another drawback of this approach is that the similarity calculation only considers the IC value of the LCS and ignores the IC value of the compared concepts.

Jiang & Conrath's approach is an information theoretic similarity approach which extends the Resnik's similarity in that it also based on the computing the IC values from considered corpora (Jiang and Conrath 1997). It also considers the IC values of the compared concepts, and measures the distance between the compared two concepts. The semantic distance measure is derived from the edge-based notion of distance with the addition of the IC as a decision factor

2.6.1.1 Lin Similarity

Lin (1998) similarity can be considered as an extension of Resnik's similarity. It is also based on the IC value calculation for the similarity measure. It considers not only the IC value of the lowest common subsumer, but also that of the concepts compared. Lin has formulated the similarity formula as in Equation (2).

$$sim_{lin}(c_1, c_2) = \frac{2 \times log(p(lcs(c_1, c_2)))}{log\, p(c_1) + log\, p(c_2)} \quad (2)$$

In this equation, $p(c_1)$ is the probability of finding the concept c_1, $p(c_2)$ is the probability of finding the concept c_2 and $p(lcs(c_1, c_2))$ is the probability of finding the LCS of the given concepts c_1 and c_2 in the given corpora.

Example: Consider, two concepts "cat" and "dog" are compared for Lin similarity. They are subsumed by concepts "carnivore", "whole" and "root". The LCS for cat and dog is carnivore.

Figure 2-1 displays some concepts and their IC values in WordNet based Lin similarity. Now, the IC values of cat, dog and carnivore is computed and the similarity between cat and dog is computed as $sim(cat, dog) = 2 \times ic(animal)/ (ic(cat) + ic(dog))$. The WordNet [12]similarity based on Lin similarity gives the similarity score 0.89. Whereas the LCS of cat and car is whole and the similarity score is 0.17.

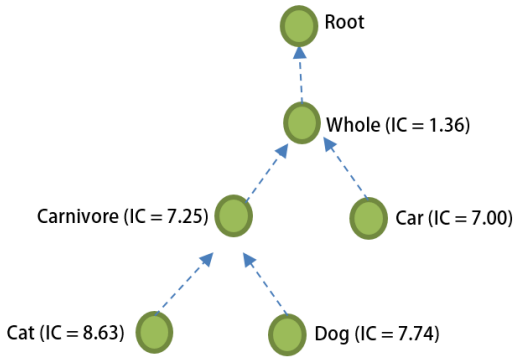


Figure 2-1 Comparison among cat, dog and car in Lin similarity

On its positive side, Lin similarity claims that it is not tied to a particular form of knowledge representation, and it is universal in application. Additionally, similarity between the ordinal values can also be measured. The advantage over Resnik's similarity is that it also considers the IC values of the compared concepts. However, in common with Resnik's similarity, requirement of the time intensive analysis of the applied corpora remains as a drawback. Furthermore, the similarity measure also depends on the considered corpora.

2.6.2 Edge-Counting Approaches

This approach is also called the shortest path method and is influenced by the geometric model in Cognitive Psychology, where the stimuli and response interaction is quicker if the distance between the origin of the stimuli and the response unit is shorter (Yang and Powers 2005). These measures consider the structure of concepts in an ontology. However, as a contrast to the information theoretic models, they do not take account of the IC values. Among these

[12] http://maraca.d.umn.edu/cgi-bin/similarity/similarity.cgi

measures, the work of Rada *et al*. (1989), Budanitsky & Hirst (2001) and Leacock & Chodorow (1998) are the most cited ones.

Rada *et al*. (1989) semantic similarity measure is similar to Resnik's measure in that they both took an account for the LCS for the metrics calculation. This means Rada *et al*. also consider the `is-a` hierarchy of the Mesh, a lexical ontology as the main and only relation for the similarity value computation. The concepts are represented as nodes in the ontology, and the path, also known as edge between the nodes, are considered for the similarity metrics. Prior to calculating the similarity measure, the number of the path between two concepts is computed. The number of the path is considered as the distance between the two concepts. Then the similarity is computed based on the distance in such a way that the lower the distance between the concepts, the higher the similarity between the concepts. The maximum similarity score is obtained if the compared words are synonyms and likewise the minimum similarity is deduced if the compared concepts are antonyms.

Budanitsky & Hirst (2001) similarity measure has gone a bit further by considering various kinds of relations in a lexical ontology. The conceptual similarity between two concepts in a lexical taxonomy depends on how far these concepts are from each other and how deep their LCS is from the taxonomical root node (Wu and Palmer 1994). Leacock and Chodorow (1998) considered the shortest edge-length between the compared words, as well as the depth of the concepts, and the lowest common subsumer from the root of the lexical taxonomy. Hao *et al*. (2011) approach has recently revitalised the edge based semantic similarity by adding the further assumption on it. This method considers the human intuition of comparing the similarity and differences between the concepts.

2.6.3 Gloss Overlap (Vector) Based Approaches

In contrast to ontology based approaches and information based approaches, the vector based approaches take advantage of relating the concepts when they are not represented with a semantic relation in a lexical taxonomy. These approaches are particularly helpful to determine

the semantic relatedness between the two compared concepts rather than deciding the semantic similarity; hence, it is used most often in word sense disambiguation.

The basic intuition behind these approaches is that they consider the glosses of the each compared words. A gloss is a description of a word in a dictionary, which interprets the meaning of the words. In these approaches, all the words in a gloss except the stop words are put into a vector as a bag of words. In this way, two separate vectors of gloss words are created and the shared words between these two vectors are counted. If the two vectors have some shared words within them, the compared words are regarded as related. In other words, the higher the number of shared words between the vectors, the more related the two compared words are. Some of these approaches are described below.

Lesk (1986) and Banerjee and Pedersen (2002) considered vector of words related to compared words for their relatedness metric computation. Their aim was to disambiguate the senses of the words. The basic intuition behind the vector-based approaches is that, if two words appear in the same sentence, their senses are also related. In particular, they looked into dictionary glosses of the each concept compared, created vector of the gloss words for each concept and computed the number of words overlapped between these two vectors. The higher the overlap, the more related their sense is. Lesk originally implemented this approach in some standard dictionaries, which had no semantic linking between the words. However, the vector-based approach heavily depends on the dictionary definition of the compared words and the definition is not adequately containing the expected words. In order to overcome the drawback of Lesk's algorithm, Banerjee and Pedersen (2002) extended the glosses with the glosses of the related words and expanded the gloss vectors. This expansion helped in giving clearer semantics to the words thus making the metric calculation more accurate.

2.6.4 Mixed Approaches

The other approaches employ various sources of information and either edge-based approaches or node-based approaches or combination of them to determine similarity between the given

two words. They are called mixed approaches. In recent researches, the information from the web content such as Wikipedia, search engines and social networking sites are also considered for the similarity measures, which can be categorised as mixed approaches. Some of the mixed approaches are described below.

Derivatives of Information Content

The derivation of information content approaches is found useful in recent works. The method of Seco *et al*. (2004) utilise the methods of information content approaches. However, the information content of the compared concepts and that of their LCS is computed from the lexical ontology, WordNet instead of analysing through various corpora. The intuition behind this approach is that the hierarchical relation between the concepts in the lexical ontology expresses semantic values among the concepts. Furthermore, when the concepts are arranged in `is-a` hierarchies, the concepts at the higher level in the hierarchy convey less information than that in the lower level. In other words, the leaf nodes express the highest information as compared to any other nodes above it. Similarly, in the work of Li *et al*. (2003), the similarity metric is determined by considering the edge-length between the compared words, the depth of their LCS from the root and the densities of the words compared in a lexical taxonomy.

Web Based Similarity

In order to overcome the individual web page processing, Cilibrasi and Vitanyi (2007) argue that the analysis of the search engine results is as strong as analysing all the pages in the web. In their semantic distance computation, they have considered the number of search results returned by the compared terms separately and that of their combination. Bollegala *et al*. (2007) proposed similar method with extension to consider the context in which the compared terms are placed in the web page. Strub and Ponzetto (20076; 2007) have argued that the information in the web is richer than that in the dictionary and lexical ontologies. The semantic similarity computation can be made closer to the human judgement by considering the knowledge in the web pages.

Ontology Mapping

Ontology mapping (OM) is also referred to as ontology matching or ontology alignment. OM is a process which aims to find semantic correspondences between similar elements of different ontologies (Noy 2004, Wick et al 2008). The examples of the similar elements are the concepts, properties and individuals. OM has utilised the well-known paradigm of the Schema Matching (Noy 2004) where elements of relational database or XML schemas are related.

OM is an active field of research, and survey and classification of OM approaches can be found in Noy (2004), Shvaiko & Euzenat (2005) and Kalfoglou & Schorlemmer (2003). The approaches of OM can be categorized into two groups: centralised and decentralised. In centralised approaches, the ontologies to be mapped have a common upper ontology. While in the decentralised approaches, there is no known common upper ontology. The decentralised approaches can be categorised into five major approaches (Mao 2008): (i) rule based (Ehrig and Staab 2004, Noy et al 2000) , (ii) machine learning based (Doan et al 2003), (iii) graph-based (Noy and Musen 2001), (iv) probabilistic (Mitra et al 2005) and (v) reasoning and theorem proving (Giunchiglia et al 2004). The rule based approaches use lexical information such as name, label and description and structural information such as hierarchical relations and properties in order to find corresponding entities (Ehrig and Staab 2004, McGuinness et al 2000, Noy et al 2000). In OM, combination of different similarity measures is used in order to compute the similarity scores between the corresponding entities.

2.6.5 Relevancy of the Similarity Frameworks

From the review of the similarity measures, it is difficult to select a single measure for this research since there is no measure, which can be used as such. It requires adapting existing measures with some modification, and there is more than one suitable candidate.

This research needs to adapt some methods in the regulatory domain in order to relate the concepts in the regulation ontology with that in the process ontology. In this case, the OM approaches seem to be viable solutions. However, the OM approaches cannot be applied as

such, which determine the similarity of the ontological concepts based on their meanings and relations. A regulation ontology and a process ontology are not similar ontologies in terms of their entities, structure and meanings. Therefore, the individuals representing regulatory guidelines in the regulation ontology and individuals representing organisational processes in the process ontology need to be related by adapting the OM approaches. In particular, adapting OM approaches considering the special nature of the regulatory guidelines and organisational process can be a good solution in order to relate the regulatory guidelines with organisational processes.

When computing the similarity between names, labels and descriptions, the Lin similarity algorithm is found the most appropriate since it exploits the conceptual hierarchy in a lexical ontology and the weight of a term with respect to their occurrence in general corpora. However, some of the concepts in general lexical ontology and a domain ontology are not similar in the same extent, and the similarity between them need to be processed differently. Adapting this measure, in order to compute the similarity between two words, needs to consider the context of the compared words.

In order to address the mapping challenge, the techniques used to compute semantic relatedness among entities such as sentence, word and concept, are useful. In particular, sentence to sentence mapping (Agirre et al 2012, Barzilay and Elhadad 2003, McCarthy et al 2012, Mohler et al 2011), concept to concept mapping (Chen et al 2010, Ge and Qiu 2008, Hawalah and Fasli 2011) and word to word mapping (Pedersen et al 2004) can be regarded as solutions to the problem to some extent. However, in the process domain where the processes are modelled in to an ontology, it requires a mapping process, which considers the structure of the processes, and the semantics embedded in the regulatory guidelines. Therefore, the word, concept and sentence similarity measures as such cannot be implemented, and their adaptation can be a candidate for the mapping between regulatory guidelines and organisational processes.

Hence, the combination and adaptation of (i) OM (Choi et al 2006, Kalfoglou and Schorlemmer 2003, Shvaiko and Euzenat 2005), (ii) word, concept and sentence similarity (Barzilay and Elhadad 2003, Ge and Qiu 2008, Yu and Zhou 2009) and (iii) Lin (1998) similarity are found to be useful to relate the regulatory guidelines and organisational processes.

2.7 Summary

This chapter has provided current approaches to the RCM and explained how the knowledge management help in the RCM. In addition, the approaches to the DSA, the regulatory IE, and similarity measures are analysed.

The aim of this thesis is to help in the process of the RCM. It is identified that the overall automation in the RCM systems can be improved by automating two important processes: (i) extracting regulatory entities from regulatory guidelines and (ii) mapping regulatory guidelines with organisational processes (El Kharbili et al 2008, Sesen et al 2010). Although IE has gained popularity in the recent research works (Reeves 2006, Sarawagi 2007), adaptation of this technology in the regulatory domain has not been explored as expected (Gao et al 2011, Kiyavitskaya et al 2008, Mu et al 2009). The semantic similarity and ontology mapping are well explored field of research (Kalfoglou and Schorlemmer 2003, Dekang Lin 1998, Pedersen et al 2004). However, these technologies have not been sufficiently exploited in the RCM in order relate regulatory guidelines with organisational processes (Sesen et al 2010). The next chapter will explain how this research aims to overcome some of the shortcomings described in this chapter such as extracting regulatory entities and relating regulatory guidelines with organisational processes.

3 The RegCMantic Framework

3.1 Introduction

This chapter proposes a new framework that aims to tackle some of the crucial research issues regarding RCM as identified in Chapter 2. One of the challenges mentioned in recent RCM frameworks (El Kharbili et al 2008, Sesen et al 2010) is to update RCM with changes in regulatory guidelines automatically. An organisation has to update RCM when there are some changes in existing regulatory guidelines, or new regulatory guidelines are introduced. Besides, if an organisation wishes to extend its business in a different country, the processes need to follow the new regulatory guidelines in the country. The compliance manager in the organisation has to determine which organisational processes are affected by the changes. If there is a large number of regulatory guidelines and organisational processes, it will take a huge amount of time and effort to determine the affected processes. This thesis aims to systematically automate the process of defining a relationship (or mapping) between the new regulatory guidelines and the existing organisational processes in two main stages: (i) extracting meaningful entities from regulatory guidelines and (ii) relating the regulatory guidelines with the organisational processes.

As described in Chapter 2, there are various approaches for IE and similarity measures, which have not been applied, to RCM. The proposed framework aims to adapt these technologies in RCM. The IE will be adapted in the regulatory domain in two steps: (i) DSA and (ii) semantic IE. In addition to the generic approaches to identify document-structure, this framework utilises the specific features of the regulatory text such as extensive use of model verbs and indicators as well as ontological concepts and parsers. Once the document-structure is identified, the meaningful entities embedded in the regulatory guidelines are extracted. The

identification of meaningful regulatory entities helps on relating regulatory guidelines and organisational processes.

The rest of the chapter is organised as follows. An overview of the proposed framework is presented in the Section 3.2. The extraction part of the framework is described in Section 3.3. Similarly, the mapping part of the framework is described in Section 3.4. The salient features of the framework are highlighted in the Section 3.5, and the summary is presented in Section 3.6.

3.2 Overview of the Proposed Framework: RegCMantic

A general overview of the proposed framework is given as follows:

1) **Document Structure Analysis (Identification of Document-Component in a Regulatory Document):** In addition to the generic approaches of identifying document-components, the special nature of the regulatory text- such as use of indicator terms and model verbs - can be exploited to identify the key document-components. The regulation-documents have a combination of different document-components such as topics, titles, paragraphs and footnotes. This thesis analyses the variations in the different regulation-documents to identify the document-components, which contain required regulatory entities.

2) **Information Extraction (Automatic Extraction of Regulatory Information):** The proposed framework enables the automation in the extraction of regulatory entities from regulatory text. The regulatory entities embedded in the regulation-documents are analysed, annotated, extracted and modelled into a computer interpretable knowledge base. The proposed framework uses four components to identify the entities: parser, definition terms, ontological concepts and rules.

3) **Semantic Similarity (Automatic Definition of Relationships between Regulatory Guidelines and Organisational Processes):** In order to determine whether an organisational process is complying with the relevant regulations, the process should be related with the regulations first. This framework aims to help identify the relationship between regulation and organisational process

The proposed framework, RegCMantic (see Figure 3-1) is divided into two main parts: (1) extraction part, and (2) mapping part. In the first part, the regulatory guidelines in different document formats such as pdf, rtf and doc, are converted into a uniform XML document structure format and is described as "DSA". In the XML document, the regulatory guidelines and the regulatory entities are annotated and this process is described as "Regulatory Entity Annotation". Finally, in the first part, the annotated entities are extracted and represented in an ontology, which is described as "Regulation Ontology Population". In the second part, each regulation statement is compared with organisational processes in order to determine the level of relationship or similarity. The comparison depends on three types of similarities: (i) topic similarity, (ii) core similarity and (iii) aux similarity depending on the topic, core and aux entities of a regulation-statement respectively.

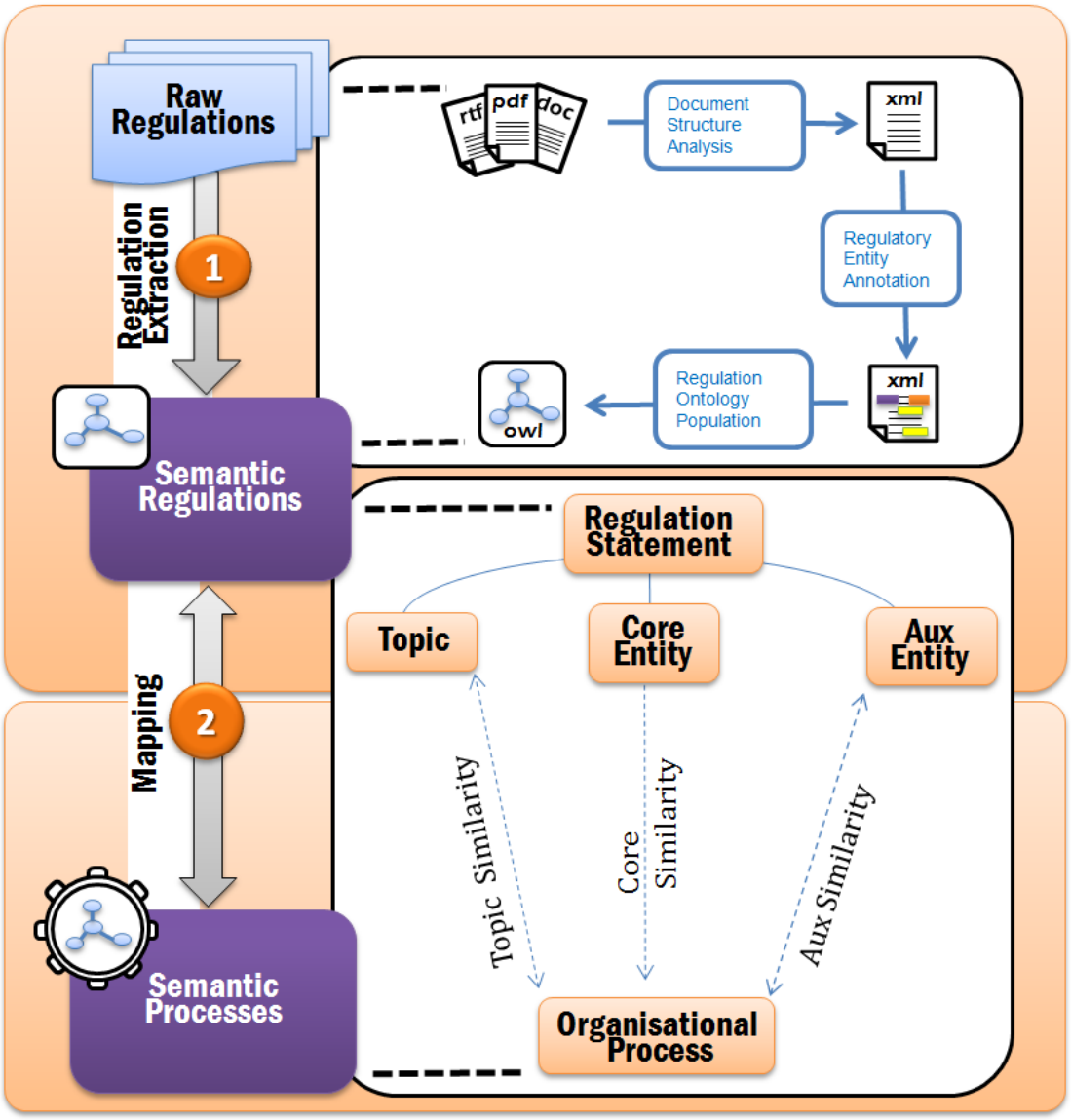


Figure 3-1. The RegCMantic framework

3.3 Regulatory Entity Extraction

The first phase in the RegCMantic framework is the extraction of regulatory entities. It includes representing the regulatory guidelines in an XML format and extracting meaningful entities from the text (see Figure 3-2), which is described later.

A regulatory document contains various document-components such as headers, footers, page numbers, footnotes, comments, titles and paragraphs. In order to extract meaningful regulatory entities from the regulatory text, it is essential to identify the document-components that

contain regulatory guidelines. In particular, it needs to annotate the regulatory entities embedded in regulatory-paragraphs. The regulatory-paragraphs also referred to as regulations in this thesis, are the paragraphs, which impose some restrictions on organisational processes. The restrictions are usually imposed by using model verbs such as **must**, **should** and **may**. The regulatory entity extraction process is described in four steps: (1) Document Conversion, (2) DSA, (3) Regulatory Entity Extraction and (4) Semantic Representation of Regulatory Guidelines.

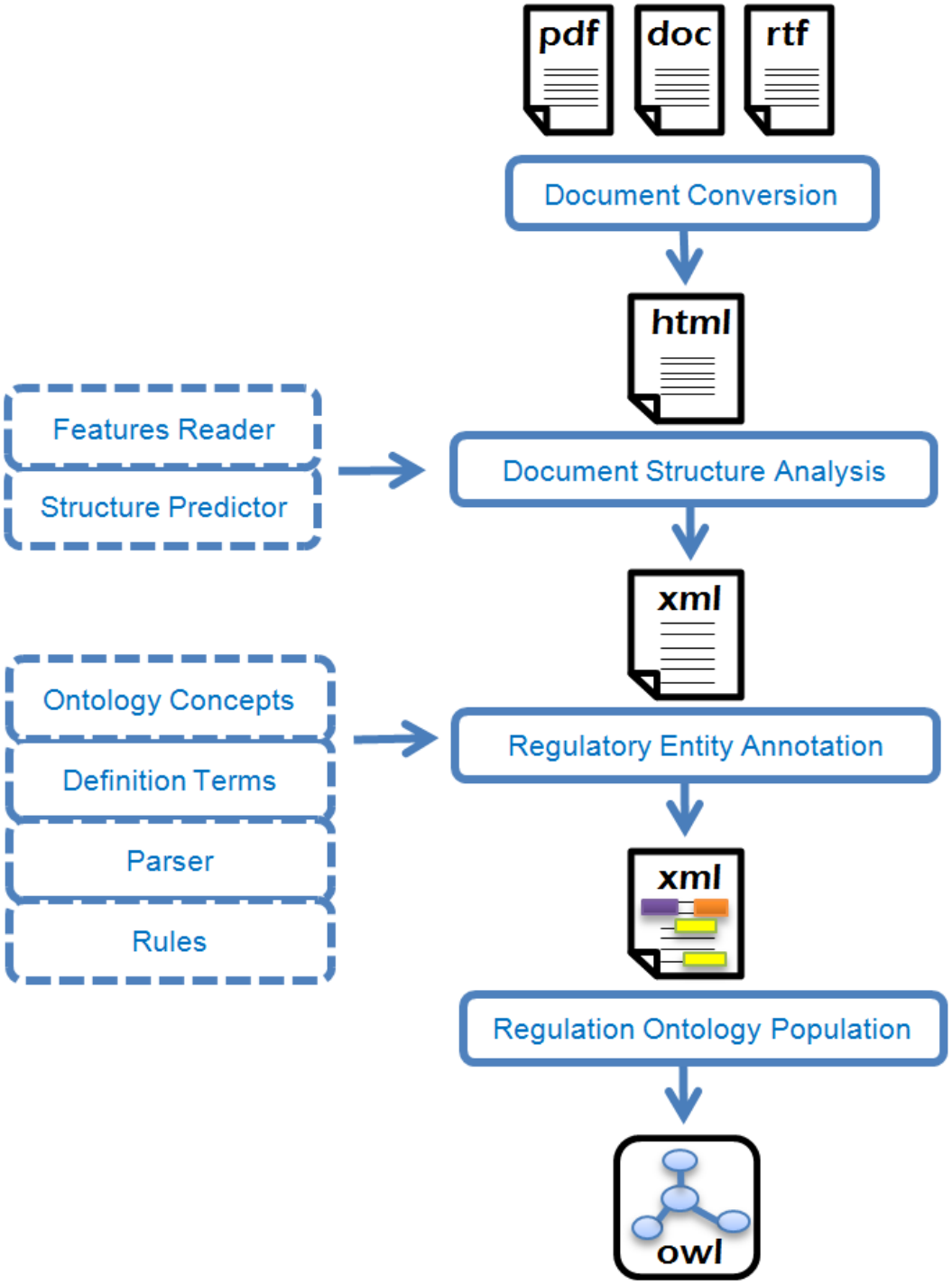
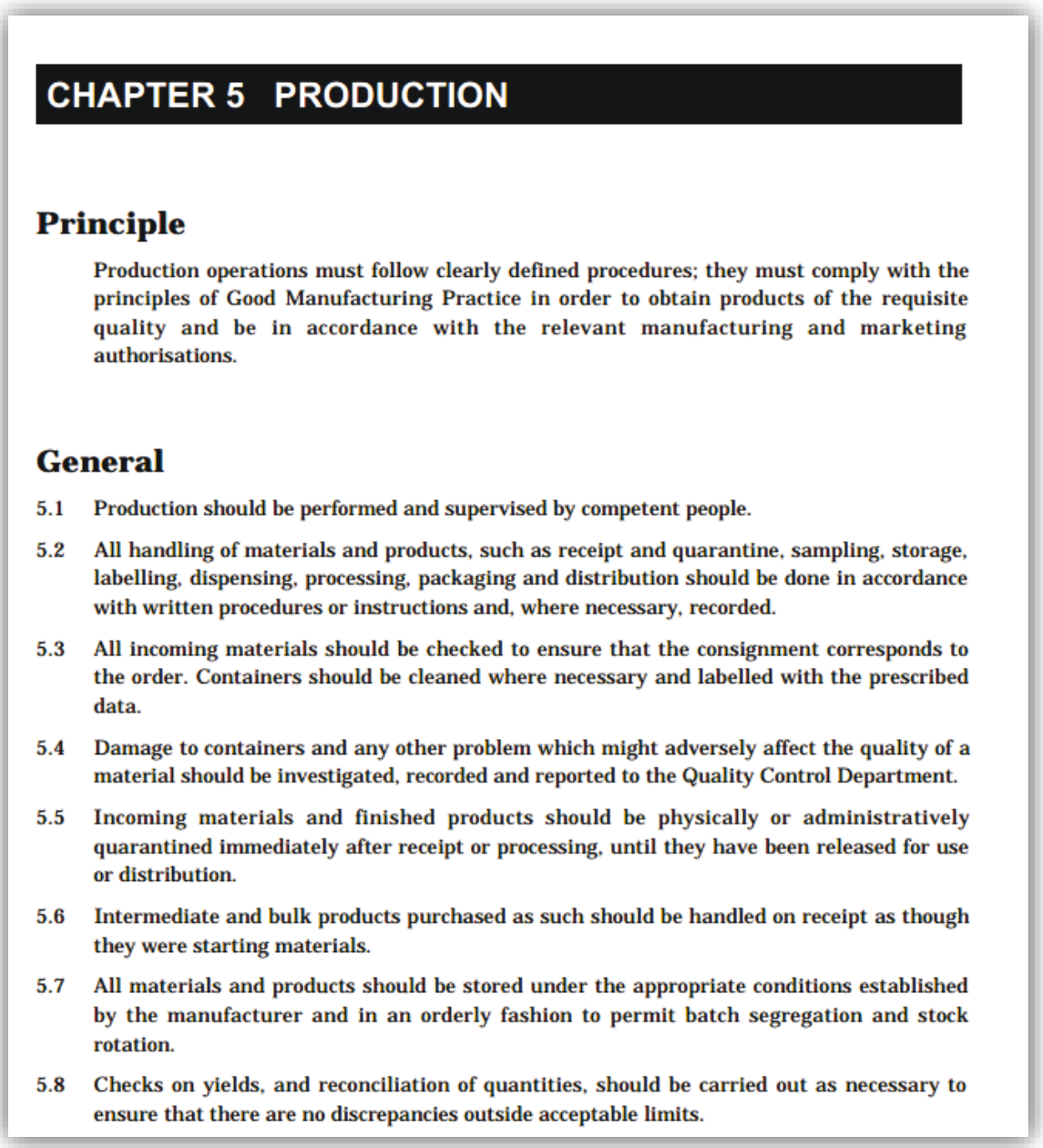


Figure 3-2. Regulatory entity extraction in the RegCMantic framework

3.3.1 Document Conversion

The regulatory guidelines are available in documents of various formats such as PDF, DOC, HTML and XML such as the UK (MHRA 2012), EU (Eudralex 2013) and USA (FDA 2012)

regulations for the Pharmaceutical industries. Instead of developing processors for each format, the RegCMantic approach is to convert them into a single uniform processing format: HTML. An example of converting regulatory guidelines from PDF file format to HTML file format is provided in Figure 3-3 and Figure 3-4. There are fair amount of tools, which convert documents into HTML format. There are also tools available that convert documents into XML formats. However, in the proposed framework (see Figure 3-2), the documents are first converted from various file formats to HTML and then to XML. They are not directly converted into XML because the direct conversion only converts the document into the XML file format; it does not identify the document-components. The proposed framework aims to represent the structure of a document explicitly, where each document-component is clearly labelled. Converting the files into HTML format preserves the original information such as font features and location of the text, which help in the processing the document to identify the different document-components, that can be represented in an explicit (and meaningful) format such as XML.



CHAPTER 5 PRODUCTION

Principle

Production operations must follow clearly defined procedures; they must comply with the principles of Good Manufacturing Practice in order to obtain products of the requisite quality and be in accordance with the relevant manufacturing and marketing authorisations.

General

5.1 Production should be performed and supervised by competent people.

5.2 All handling of materials and products, such as receipt and quarantine, sampling, storage, labelling, dispensing, processing, packaging and distribution should be done in accordance with written procedures or instructions and, where necessary, recorded.

5.3 All incoming materials should be checked to ensure that the consignment corresponds to the order. Containers should be cleaned where necessary and labelled with the prescribed data.

5.4 Damage to containers and any other problem which might adversely affect the quality of a material should be investigated, recorded and reported to the Quality Control Department.

5.5 Incoming materials and finished products should be physically or administratively quarantined immediately after receipt or processing, until they have been released for use or distribution.

5.6 Intermediate and bulk products purchased as such should be handled on receipt as though they were starting materials.

5.7 All materials and products should be stored under the appropriate conditions established by the manufacturer and in an orderly fashion to permit batch segregation and stock rotation.

5.8 Checks on yields, and reconciliation of quantities, should be carried out as necessary to ensure that there are no discrepancies outside acceptable limits.

Figure 3-3. Example regulatory guidelines in PDF file format

```
<html>
<head>
<title>pg_0001</title>
<style type="text/css">
.ft1{font-style:normal;font-weight:bold;font-size:23px;font-family:Arial;color:#ffffff;}
.ft2{font-style:normal;font-weight:bold;font-size:20px;font-family:Times New Roman;color:#000000;}
.ft3{font-style:normal;font-weight:normal;font-size:13px;font-family:Times New Roman;color:#000000;}
</style>
</head>
<body>
<span class="ft1">CHAPTER 5  PRODUCTION</span>
<span class="ft2">Principle</span></div>
<span class="ft3">Production operations must follow clearly defined procedures; they must comply with the</span>
<span class="ft3">principles of Good Manufacturing Practice in order to obtain products of the requisite</span><
<span class="ft3">quality and be in accordance with the relevant manufacturing and marketing</span>
<span class="ft3">authorisations.</span>
<span class="ft2"> General</span>
<span class="ft3">5.1 Production should be performed and supervised by competent people.</span>
<span class="ft3">5.2 All handling of materials and products, such as receipt and quarantine, sampling, storage,</span>
<span class="ft3">labelling, dispensing, processing, packaging and distribution should be done in accordance</span>
<span class="ft3">with written procedures or instructions and, where necessary, recorded.</span>
<span class="ft3">5.3 All incoming materials should be checked to ensure that the consignment corresponds to</span>
<span class="ft3">the order. Containers should be cleaned where necessary and labelled with the prescribed</span>
<span class="ft3">data.</span>
<span class="ft3">5.4 Damage to containers and any other problem which might adversely affect the quality of a</span>
<span class="ft3">material should be investigated, recorded and reported to the Quality Control Department.</span>
<span class="ft3">5.5 Incoming materials and finished products should be physically or administratively</span>
<span class="ft3">quarantined immediately after receipt or processing, until they have been released for use</span>
<span class="ft3">or distribution.</span>
<span class="ft3">5.6 Intermediate and bulk products purchased as such should be handled on receipt as though</span>
<span class="ft3">they were starting materials.</span>
<span class="ft3">5.7 All materials and products should be stored under the appropriate conditions established</span>
<span class="ft3">by the manufacturer and in an orderly fashion to permit batch segregation and stock</span>
<span class="ft3">rotation.</span>
```

Figure 3-4. Regulatory guidelines converted into HTML file format

3.3.2 Document-Structure Analysis (DSA)

In this step, the structure of the regulatory document is identified.

- A document contains different types of text in terms of their font-features such as font-size, font-style, font-strength and font-colour. In this framework, the type of the text is called **Text-Type**. A document contains a set of text-type: $T = \{t_1, t_2, .. , t_n\}$. For example, the font-size of the title of a document is bigger than that of the text in the body; therefore, they can be regarded as two different text-types.
- For each text-type, a score is computed considering all the font-features and is called **Feature-Score**. The main influencing factor for the feature-score is the font-size. This means the higher the font-size, the higher the feature-score. A document contains a set of feature-scores: $S = \{s1, s2, .. , sn\}$, a score for each text type.

- A level is defined for each text-type based on its feature-score, and is called **Text-Level**. This means, the higher the feature-score, the higher the text-level. A document contains a set of text-level: $L = \{l_1, l_2, .. , l_n\}$ for a set of text-type. In this set of the text-levels, $l_1 > l_2 > .. > l_n$.

Example: In the text in Figure 3-3, there are three text-types $t1$, $t2$ and $t3$ representing chapter, section and paragraph respectively. The first line of text "Chapter 5 Production" has the highest feature-score: $s1 = font\text{-}size \times 10 + font\text{-}weight = 23 \times 10 + 2 = 232$. The text in "Principal" and "General" has the second highest feature-score: $s2 = font\text{-}size \times 10 + font\text{-}bold = 20 \times 10 + 2 = 202$. The text in the paragraphs starting with some numbers has the feature-score lower than the above two: $s3 = font\text{-}size \times 10 + font\text{-}normal = 13 \times 10 + 0 = 130$. We have three types of feature-scores $s1$, $s2$ and $s3$ for three types of text-types $t1$, $t2$ and $t3$. Now we can assign levels: $l1$, $l2$ and $l3$ for $t1$, $t2$ and $t3$ respectively.

- Similarly, a document has a set of **Document-Components:** $C = \{c_1, c_2, .. , c_n\}$ such as chapter, section, sub-section, paragraph and page numbers. The document-components specify the structure of a document. Usually, they follow a hierarchical structure depending on the text-level of each text-type. In summary, each type-type is labelled with a text-level considering its feature-score, and each text-level is labelled with a document-component considering the document-component prediction algorithms. The document-component prediction algorithms are described below with examples.

When the document-components are identified, they are represented in an XML definition file, called **Document-Schema**. In order to create a document-schema, two processors are implemented: **Feature Reader** and **Structure Predictor** as shown in Figure 3-2.

The **Features Reader** identifies the document features such as font-style, font-weight, font-family, font-colour and text-content. Reading the sufficient amount of document features is helpful in processing the index for each document-component.

Based on the document features, the **Structure Predictor** infers the components of the document. The paragraph is the main document structure, which helps to determine the regulation. Therefore, among the document components, the paragraph is identified at first. Then, the other components are identified based on their preceding text or label. A series of algorithms is implemented in order to predict the structure of the document. The most important algorithms are presented in this chapter, and the rest are provided in Appendix C. Once the set of recommended document structures is presented in a user interface, the user verifies or modifies the suggested structure.

Paragraph Prediction

This process firstly creates a set of text-level $L = \{l_1, l_2, .. , l_n\}$ of the text in the document. In particular, a set of text-level is created on the basis of a set of a feature-score $S = \{s_1, s_2, .. , s_n\}$. Each text-level l determines how much text it contains; how many sentences it has; whether it contains any obligatory words such as **must** and **should**, and how far its font-size is from standard font-size of a paragraph text.

The prediction of a text as a paragraph depends on the paragraph index of the text. The paragraph index computation uses the indices of the sentence, text, obligation and deviation. The sentence index is the percentage of the sentence in a text-level. The text index of a text-level is the percentage of its text content. The obligation index of a text-level is the percentage of presence of obligatory words in the text. The deviation index of a text-level is the percentage of the distance of the text-level from the text-level of a standard paragraph. In general, the font-size of a paragraph is 12px, and not bold and not italic. The paragraph index prediction is the average value of the weighted values of these four indices. The text in the text-level having the highest paragraph index is regarded as the paragraph (see Algorithm 3-1).

Example: Following from the previous example, there are text-types in Figure 3-3: $t1$, $t2$ and $t3$. The feature-score of a typical paragraph is computed as $s_p = font\text{-}size \times 10 + font\text{-}normal = 12 \times 10 + 0 = 120$. In this case, the closest feature-score to the paragraph is that of

t3 (i.e. *130*). This suggests that *t3* is most likely to be a paragraph. Similarly, three other factors also suggest that *t3* is a paragraph: the amount of text in *t3* is the highest; *t3* has the highest number of sentences and there are more model verbs in *t3*.

Algorithm 3-1 Paragraph prediction

Input: L is a set of text-level in the document.

Output: L' is a new set of text- levels with the predicted text-level for the paragraph

function PREDICT-PARAGRAPH(*L)* **returns** *L'*

```
i = 0, lk = null
L = {l1, l2, .. , ln}
for each li ∈ L
    j = COMPUTE-PARA-PREDICTION-INDEX(li)
    if (j >i) then
        lk = li
        i = j
    end if
end for
lk.SET-COMPONENT(paragraph)
L' = L
return L'
```

Indicators Based Prediction

When the paragraph prediction is completed, the next process will predict the rest of the text-levels based on the preceding label or text also referred to as indicators. In many cases, the document components with higher text-level such as **part, chapter** and **section** are preceded with the relevant text such as "Chapter 5 Production" and "Section 5.3 Starting Materials". When a text-level with these preceding texts is found, the text-level is defined as that type of document component. For example, if the text in the text-level l_1 starts with the text

"Chapter", then the text-level l_1 is set as `chapter` for its document component (see Algorithm 3-2).

Example: Following from the previous example, the *t3* has been suggested as a paragraph in Figure 3-3 . Now, we need to identify the document-component of *t1* and *t2*. The text-type, *t1* is preceded with an indicator term "Chapter", which suggests that *t1* is a chapter.

Algorithm 3-2 Document-component prediction based on the indicator text

Input: C is a set of document-components (document-structure). L is a set of text- level in the document.
Output: L' is a new set of text- levels in the document with document structure values computed from the preceding text

function PREDICT-COMPONENT-WITH-INDICATOR(C, L) **returns** L'

```
C = { c1, c2, .. , cn }
L = { l1, l2, .. , ln }
for each li ∈ L
   ci = GET-COMPONENT(li)
   text = GET-INDICATOR-TEXT(li)
   if (ci = null) then
      for each cj ∈ C
         if (text = cj) then
            ci = cj
         end if
      end for
   end if
end for
L' = L
return L'
```

Prediction Based on Empirical Values

The prediction of the rest of the text-levels that have not been completed yet, are computed based on proximity and empirical values (see Algorithm 3-3). Based on proximity, the algorithm predicts the closest document component with respect to the already predicted document components. For an example, if a text-level l_1 is set to `chapter` as its document component, and l_3 is set to `paragraph` as its document component, then l_2 can be the remaining document-components in between the `chapter` and paragraph such as `section` and

`subsection`. These remaining components are obtained from an empirically created hierarchical component set $C = \{c1, c2, .. , cn\}$. When there is more than one possible document-component, the closest one to the highest predicted document-component is set. In the above example, value of l_2 is set as `section` – this is because the feature-score of the text-level l_1 (chapter) is closer to the feature-score of `section` as compared to the feature-score of `subsection`.

Algorithm 3-3 Predicting the remaining structure of a document

Input: C is a set of possible document-component s (document-structure). L is a set of text-levels in the document.
Output: L' is a new set of text-levels in the document with document structure values computed from the preceding text

function PREDICT-REMAINING-COMPONENT(C, L) **returns** L'

```
C = { c1, c2, .. , cn }
L = { l1, l2, .. , ln }
for each li ∈ L
        ci = GET-DOCUMENT-COMPONENT(li)
        ci+1 = GET-DOCUMENT-COMPONENT(li+1)
        if (ci = null) then
            c1 ∈ C
            ci= c1
        end if
        if (ci ≠ null or ci+1 = null) then
            for each cj ∈ C
               if (ci = cj) then
                  ci+1 = cj+1
               end if
            end for
        end if
end for
L' = L
return L'
```

Example: Following from the previous example, in Figure 3-3, $t1$ and $t3$ have been suggested as a chapter and a paragraph respectively. Now, we need to identify the document-component of $t2$. The empirical value suggests that there are document-components between a chapter and a paragraph such as section and subsection. In this case, the document-component closest

to the chapter is suggested as $t2$. The document-component closest to the chapter is the section. Therefore, it suggests that $t2$ is a section.

After completion of the prediction algorithms, the predicted document-structures are presented to users via a GUI. The users analyse, select and modify the suggested document-structures. After the selection, a schema defining the overall structure of the document is created.

XML Regulation

Based on the schema created in the earlier steps, the HTML document format is converted into XML document format (see Figure 3-5). The conversion from one document format to another document format is complicated as it identifies different document-components in a document, and represents the document-components in an explicit format. When the document-components are explicitly labelled or represented, it helps in the extraction of specific entities from specific document-components. Note that, in rare situation, if the regulators publish the documents in a standard and explicit format, the previous two steps may not be needed. However, this is not a common practice and those stages constitute an important part of the process.

The most important document-component represented in this format is the paragraph because the regulatory guidelines are represented in paragraphs. A regulation-document contains several paragraphs; however, not all the paragraphs are regulatory guidelines. In this thesis, a paragraph containing regulatory guidelines is called **regulation** or **regulation-paragraph.** Likewise, a sentence within in a regulation-paragraph is called **regulation-statement**.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<document>
  <meta>
    <name>Eudralex</name>
    <description>EU regulation for the pharmaceutical industry</description>
    <body>EMEA</body>
    <version>1.0</version>
    <published_on>2007</published_on>
  </meta>
  <content>
    <chapter title="CHAPTER 5  PRODUCTION">CHAPTER 5  PRODUCTION
      <section title="Principle">Principle
        <paragraph paraNum=""> operations must follow clearly defined procedures; they must comply with the principles of Good M
      </section>
      <section title=" General"> General
        <paragraph paraNum="5.1"> Production should be performed and supervised by competent people. </paragraph>
        <paragraph paraNum="5.2"> All handling of materials and products, such as receipt and quarantine, sampling, storage, labe
        <paragraph paraNum="5.3"> All incoming materials should be checked to ensure that the consignment corresponds to the o
        <paragraph paraNum="5.4"> Damage to containers and any other problem which might adversely affect the quality of a mat
        <paragraph paraNum="5.5"> Incoming materials and finished products should be physically or administratively quarantined i
        <paragraph paraNum="5.6"> Intermediate and bulk products purchased as such should be handled on receipt as though the
        <paragraph paraNum="5.7"> All materials and products should be stored under the appropriate conditions established by th
        <paragraph paraNum="5.8"> Checks on yields, and reconciliation of quantities, should be carried out as necessary to ensur
        <paragraph paraNum="5.9"> Operations on different products should not be carried out simultaneously or consecutively in th
        <paragraph paraNum="5.10"> At every stage of processing, products and materials should be protected from microbial and
        <paragraph paraNum="5.11"> When working with dry materials and products, special precautions should be taken to prever
        <paragraph paraNum="5.12"> At all times during processing, all materials, bulk containers, major items of equipment and w
        <paragraph paraNum="5.13"> Labels applied to containers, equipment or premises should be clear, unambiguous and in th
        <paragraph paraNum="5.14"> Checks should be carried out to ensure that pipelines and other pieces of equipment used fo
        <paragraph paraNum="5.15"> Any deviation from instructions or procedures should be avoided as far as possible. If a devia
        <paragraph paraNum="5.16"> Access to production premises should be restricted to authorised personnel. </paragraph>
        <paragraph paraNum="5.17"> Normally, the production of non-medicinal products should be avoided in areas and with the e
      </section>
```

Figure 3-5. An example of regulatory guidelines represented in XML representation format

3.3.3 Regulatory Entity Annotation

A regulation-statement contains important entities such as subject, obligation and action that help to express regulatory requirements. These entities are called **regulation-entities**. A **subject** is a regulation-entity, upon which the requirements are imposed. For example in a regulation-statement "Equipment should be cleaned after processing", the word **Equipment** is the subject. In a regulation- statement, a subject can be an equipment, a substance, a person, a document or a process. The text in a regulation document contains some model verbs such as **should, must** and **shall**. These model verbs are the means of expressing the requirements of a regulatory guideline and are called **obligations** .The strength of the obligations may also vary from soft and medium to strong such as **shall, should** and **must** are the soft, medium and strong obligations respectively. An **action** is a regulation-entity that represents the action to be performed in order to comply with the requirements and expectations. Usually an action is a verb; however, sometimes the verb may be modified to different grammatical forms such as nouns and adjectives. In the example described above, **cleaned** is the action. The subject,

obligation and action are called core-entities. Beside core-entities, there are entities that that express time, place, reason and quality, and are called **auxiliary-entities** or **aux-entities**.

In the annotation process, the system identifies the regulatory constraints for the organizational processes. The first task in this process is to identify the regulation-statements. In each regulation-statement, it annotates the regulation-entities. For the annotation, it uses four main components such as ontology concepts, definition terms, natural language parser, and IE rules.

Natural Language Parser

Natural language parsers interpret a sentence in terms of its grammatical structure. In particular, it identifies grammatical units and their relationship in a sentence such as subject, verb, object, preposition and determiners (see Table 3-1). Breaking down a regulation-statement into subject containing chunk, object-containing chunk, action containing chunk and complementary chunk helps in identifying the regulation-entities accurately. For example, if a concept or a term is identified in a regulation-statement, and the position of the concept or a term is within a subject-containing chunk, it verifies that it is a subject. In this step, a parser is used with some rules to identify the special chunks such as condition-chunk, subject-chunk, obligation-chuck, action-chunk, complement-chunk, where-chuck, when-chunk, why-chunk and how-chunk.

Table 3-1. Example of parsed text

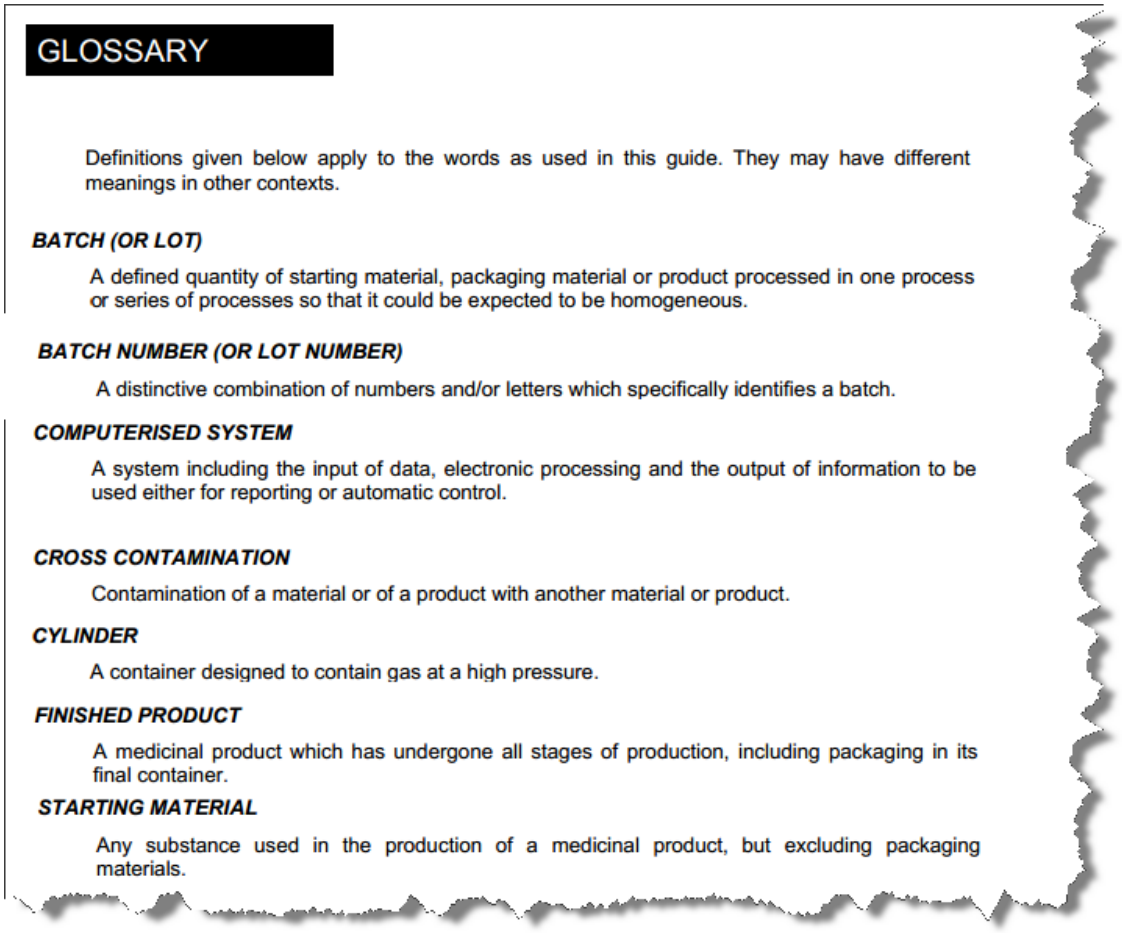
Natural Text	Parsed Text (Typed Dependencies)
Starting materials should only be purchased from approved suppliers named in the relevant specification and, where possible, directly from the producer.	amod(materials-2, Starting-1) nsubjpass(purchased-6, materials-2) aux(purchased-6, should-3) advmod(purchased-6, only-4) auxpass(purchased-6, be-5) root(ROOT-0, purchased-6) prep(purchased-6, from-7) amod(suppliers-9, approved-8) pobj(from-7, suppliers-9) partmod(suppliers-9, named-10) prep(named-10, in-11) det(specification-14, the-12) amod(specification-14, relevant-13) pobj(in-11, specification-14) cc(specification-14, and-15) dep(possible-18, where-17) dep(specification-14, possible-18) conj(specification-14, directly-20) prep(named-10, from-21) det(producer-23, the-22) pobj(from-21, producer-23)

Ontological Concepts

The ontological concepts defined in a domain are useful for IE. For example in the Pharmaceutical industry, some concepts in the process ontology are `Equipment`, `Substance` and `Filtering`. Using these concepts and their synonyms and hyponyms, the RegCMantic framework aims to identify meaningful entities in the regulatory guidelines. For this, a list of concept is created from the process ontology. The concepts and part of the concept names that can mislead the annotation process, should be removed. In this framework, these concepts are referred to as "Domain Specific Stop Words". Some examples of the domain specific stop words in the Pharmaceutical industry are `Action`, `Module`, `Entity and Domain` in `Equipment_Module`, `Physical_Entity`, `Abstract_Entity` and `Process_Domain`. The stops words are removed from the ontological concept list before annotation.

Definition Terms

Regulatory guidelines are usually provided with definition terms. The definition terms are also known as the introductory terms or the glossary in the regulatory documents, and are provided at the beginning of the documents. The terms are provided with their definition and context in which they are used in the rest of the document (see Figure 3-6). These terms contain the semantics of the regulatory guidelines, and help on annotating the text. Similar to the ontological concept processing, for the annotation, a list of definition terms is created.



GLOSSARY

Definitions given below apply to the words as used in this guide. They may have different meanings in other contexts.

BATCH (OR LOT)

A defined quantity of starting material, packaging material or product processed in one process or series of processes so that it could be expected to be homogeneous.

BATCH NUMBER (OR LOT NUMBER)

A distinctive combination of numbers and/or letters which specifically identifies a batch.

COMPUTERISED SYSTEM

A system including the input of data, electronic processing and the output of information to be used either for reporting or automatic control.

CROSS CONTAMINATION

Contamination of a material or of a product with another material or product.

CYLINDER

A container designed to contain gas at a high pressure.

FINISHED PRODUCT

A medicinal product which has undergone all stages of production, including packaging in its final container.

STARTING MATERIAL

Any substance used in the production of a medicinal product, but excluding packaging materials.

Figure 3-6. Example of definition terms

Information Extraction Rules

Application of pattern matching rules is regarded as an established IE method (Sarawagi 2007). Advancement on the regular expression, some special kind of rule specification languages are being used as state of art tools such as Common Pattern Specification Language (CPSL) (Douglas E Appelt and Onyshkevych 1998). Java Annotation Pattern Engine (JAPE) (Thakker et al 2009) is an example of implementation of the CPSL (see Figure 3-7). These rules typically have patterns as a condition on the left-hand-side (LHS) and actions to be performed on the right-hand-side (RHS). A typical example of the actions on the RHS is the annotation. Therefore, application of these rules helps to annotate the text if a specified pattern matching condition is met. In this step, the rules incorporate all the above three annotations and create a new set of annotation and/or confirms the existing annotations.

In Figure 3-7, line 4 indicates that it takes input the annotation called “action_container”. Line 5, determines what type of option is applied to the rule. Line 9 defines the rule name and line 10 defines the priority of the rule. In this example, it takes action_container as the annotations to process from the LHS. Once in the RHS the rule can be modified using Java programming language to process further annotations. Lines 15-16 accept the annotations passed from the LHS. Similarly, lines 18-22 define the names of the annotations that need to be processed. Finally, lines 26 – 43 process the annotations and output the results.

```
1  /*
2  * converts original markups ( annotation from the xml file) to starndard gate annotations.
3  */
4  Phase: action_final
5  Input:  action_container
6  Options: control = appelt
7
8  /* rule */
9  Rule: ActionRefiner
10 Priority:90
11 ({action_container}):ann
12 -->
13 {
14 // obtains the annotation
15 gate.AnnotationSet containerSet = (gate.AnnotationSet)bindings.get("ann");
16 gate.AnnotationSet containedSet = inputAS.getContained(containerSet.firstNode()
17                                   .getOffset(), containerSet.lastNode().getOffset());
18 Set selectedSet = new HashSet();
19 selectedSet.add("rule_action");
20 selectedSet.add("definition_term");
21 selectedSet.add("extracted_term");
22 selectedSet.add("concept_ontology");
23 Iterator annIter = containedSet.get(selectedSet).iterator();
24
25 /*  */
26 while (annIter.hasNext()){
27     gate.Annotation ann = (Annotation) annIter.next();
28
29     // get features from the annotation
30     String startNode = ann.getFeatures().get("startNode").toString();
31     String endNode = ann.getFeatures().get("endNode").toString();
32     String rule = ann.getFeatures().get("rule").toString();
33     String text = ann.getFeatures().get("text").toString();
34
35     // creating new annotation
36     gate.FeatureMap features = Factory.newFeatureMap();
37     features.put("rule","ActionRefiner");
38     int sNode = Integer.valueOf(startNode);
39     int eNode = Integer.valueOf(endNode);
40     features.put("startNode",sNode);
41     features.put("endNode", eNode);
42     features.put("text", text);
43     outputAS.add(ann.getStartNode(), ann.getEndNode(), "_ACTION",features);
44 }
45 }
```

Figure 3-7. An example of JAPE rule

The ontological concepts help to identify the synonyms and hyponyms of the concepts in the regulatory guidelines. The rules such as JAPE (Thakker et al 2009) help in specifying the grammar for pattern matching and incorporating the entities identified by the ontological concepts. Similar to ontological concepts, the definition terms provided by the regulatory document creators can help in identification of the terms, synonyms and hyponyms. The lexical parser can be used to separate different grammatical units in a sentence; this helps in identification of important chunks in a sentence such as subject containing chunk and action containing chunk.

3.3.4 Semantic Representation of Regulatory Guidelines

The semantic representation is the population of regulatory ontology with the extracted regulatory entities such as subject, action, obligation and modifiers. Representing regulatory guideline in semantic models such as ontology helps in the automation of RCM. For the population, an ontology with appropriate concepts is required. The ontology creation and population processes are described below.

Regulation Ontology Creation

In order to represent the regulatory guidelines semantically, a regulatory ontology called SemReg is created. It is recommended (Gómez-Pérez et al 2007) that the ontology engineering should utilise concepts of the existing ontologies in the similar domain and that of the upper ontologies. Therefore, an ontology called LKIF-Core (Hoekstra et al 2007) in the regulation domain is considered for the SemReg engineering. The research reviews, in Section 2.3, has shown that LKIF ontology (Hoekstra et al 2007) is the recent development in legal ontologies and has defined appropriate level of concepts. These concepts are extended to application level concepts and populated with the extracted entities. Although it is a core ontology, in order to adapt the concepts in the pharmaceutical domain, further concepts are created. Among the concepts created are `Subject`, `Obligation`, `Action`, `Regulation`, `Statement`, `Time`, `Place`, `Intention` and `Evaluative Expression`. Figure 3-8 shows the adaption of the LKIF-Core

concepts in the SemReg. In this figure, big boxes with dark borders are the extended concepts and the other boxes are the concepts in LKIF-Core ontology.

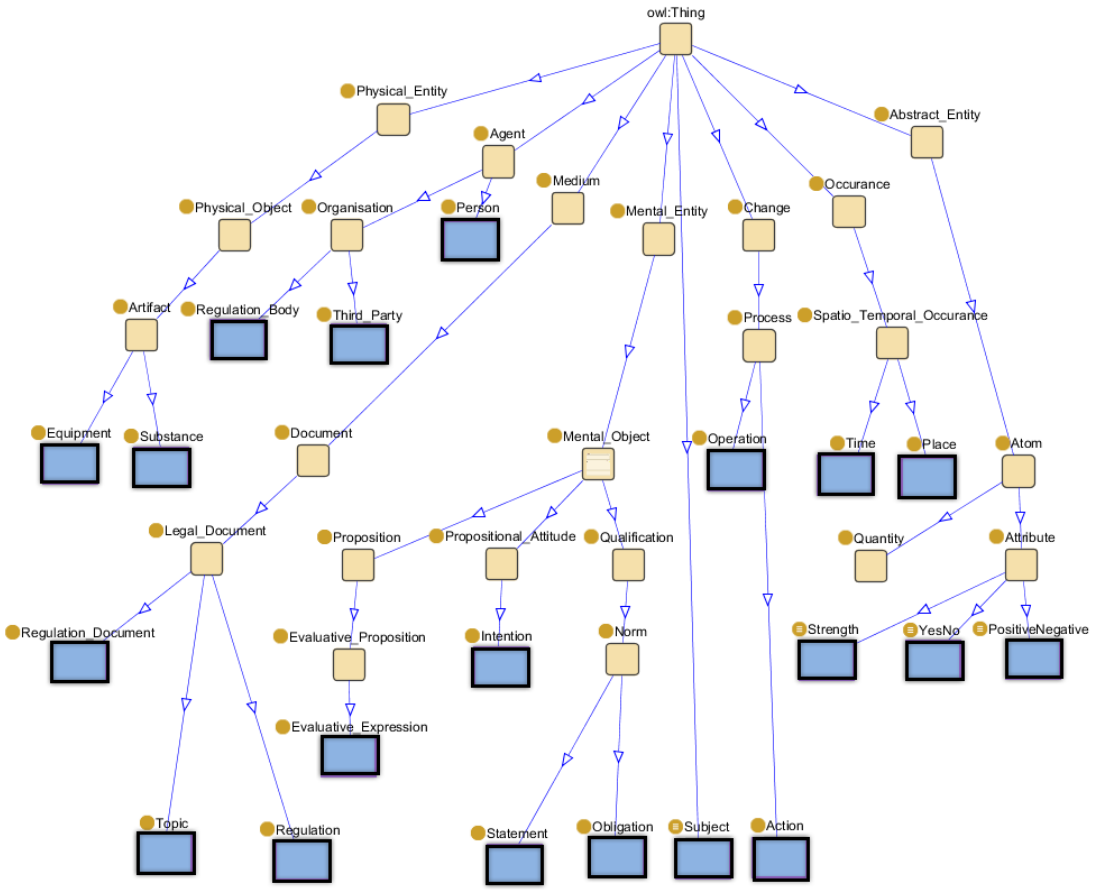


Figure 3-8. Concepts in SemReg ontology

`Subject`, `Obligation` and `Action` are the core regulation-entities as mentioned in the definition of the terminologies.

- `Subject` is a concept, which represents the entity to which the regulatory restriction applies. In the process ontology, OntoReg (Sesen et al 2010), the subject was defined as union of `Document`, `Equipment`, `Operation`, `Substance` and `Third Party`. In order to preserve the same interpretation, `Subject` is added as a separate entity in the LKIF-Core.
- The class `Document` is placed under its super class, `Medium`. The classes `Equipment` and `Substance` are kept under the class `Artifact` in LKIF-Core. `Artifact` is placed

under `Physical_Entity` and `Physical_Object` from top to bottom in a subsumption hierarchy.

- `Obligation` is defined as a kind of norm and placed under the LKIF-Core concept, `Norm`. The `Norm` is defined under the concepts `Mental_Entity`, `Mental_Object` and `Qualification` from top to bottom. The two important properties associated with `Obligation` are the `hasType` and `hasStrength`. The `hasType` property specifies whether the obligation is `positive` or `negative`. The `hasStrength` property specifies the strength of the obligation as `soft`, `medium` or `strong`.
- The class `Strength` is placed under the LKIF-Core class `Attribute`, which is placed under `Abstract_Entity` and Atom. `Strength` is defined as an enumerated class of the individuals `moderate`, `soft`, `strong` and `unknown`.
- Another subsumption of the class `Attribute` is the class `PositiveNegative`, which comprises the enumerated set of `negative`, `positive` or `unknown` individuals.
- The action is the verb part of a sentence, which is performed in order to meet the imposed obligation. The concept `Action` is defined under the LKIF-Core concept `Process`, which is situated under `Change`.
- The regulation-entities are the parts of a regulation-statement and are connected by the horizontal relations such as `isSubjectOf`, `isObligationOf` and `isActionOf`.
- A **regulation-paragraph** contains one or more regulation-statements. The regulation-paragraphs are defined as `Regulation` under the LKIF-Core concept, `Legal_Document`, and the regulation-statement is defined under the LKIF-Core concept, `Norm`. The `Legal_Document` is placed under `Medium` and `Document` in LKIF-Core.
- A regulation is a part of a topic, and the topic can be any higher document-structure such as **section**, **chapter** or **part**. The topic is associated with the regulation-document via object property `isTopicOf`. The topic and regulation-document are defined as `Topic` and `Regulation-Document` under the LKIF-Core concept `Legal_Document`.

- The `Legal_Document` is produced by a regulatory body, which is defined as `Regulation_Body` under `Organisation`. The `Organisation` is placed under `Agent` and connected with `Regulation_Document` via a property, `isDocumentOf`.

Besides the core regulation-entities, there are other helping entities called auxiliary regulation-entities. These entities are the concepts which represent the answers to the questions starting with **where, when, why** and **how**.

- The first two are represented by the concepts called `Place` and `Time` respectively and placed under the LKIF-Core concept `Spatio_Temporal_Occurance`, which is a subclass of `Occurrence`. The instances of the concept `Place` specify where the event should take place, and that of the concept `Time` specify when the event should take place.
- The next two entities are represented by the concepts called `Intention` and `Evaluative_Expression` respectively. The instances of the class `Intention` justify the purpose of the event, and that of the class `Evaluative_Expression` specify how the event or the process should be carried out.

The SemReg ontology Population

Ontology population is a process where ontological classes are populated with their instances. After the identification and annotation of the regulatory entities in the regulatory guidelines, they are converted into instances of the SemReg ontological classes (see Figure 3-9), and the regulatory guidelines are referred to as semantic regulations. In other words, the semantic regulations are the regulations represented in an ontology. Such a representation helps in processing the regulations with the least effort. The process of converting regulatory guidelines from text to semantic format has also been published in (Sapkota et al 2011). In this framework, the semantic regulations are needed for the mapping between regulatory guidelines and organisational processes. Figure 3-9 displays Protégé ontology engineering environment. On the left panel or class browser, it is showing hierarchies of classes preceded with circles.

On the middle panel or instance browser, it is enlisting the individuals of a class, which are indicated by purple diamonds. Similarly, on the right panel or individual editor, it is displaying the properties of an individual.

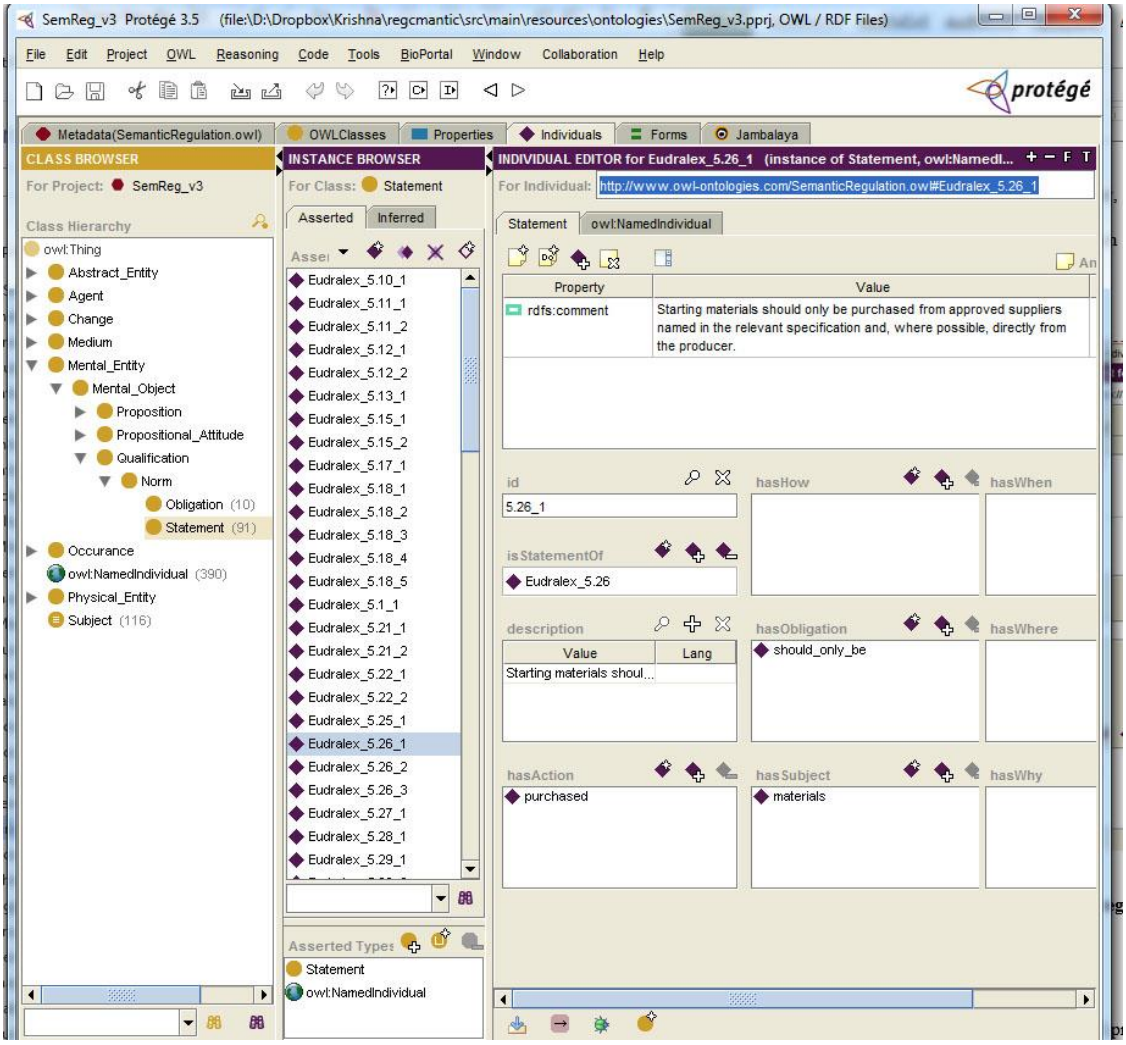


Figure 3-9. An example of population of regulatory ontology in Protégé

3.4 Mapping

In semantic compliance management, organisational processes are checked against regulatory guidelines. During this process, it is also essential to find out what regulatory guidelines an organisational process should confirm to. In other words, it is necessary to find out the related regulations for all the processes. In the RegCMantic framework, it provides automation on the process of relating the regulatory guidelines with the organisational processes.

In this thesis, the term mapping refers the relationship between the regulations with the organisational processes as shown in Figure 3-10. In this figure, list of regulations are shown on the left and that of processes are on the right. Some lines linking regulations and processes indicate that the mapping can be one to one, one to many or many to many. The regulations are the regulatory guidelines represented in a regulation-ontology, and the semantic processes are the processes represented in a process-ontology. In the mapping process, the relatedness between a regulation and a process is determined semantically. In particular, regulations and processes both have similar properties such as `hasSubject` and `hasAction`, and the value of these properties are analysed for the mapping process. If the values are similar in meaning, the regulation and the process are regarded as related (see Figure 3-11). This figure shows that mapping between a regulation-statement and an organisational process is determined by computing the similarities of subject and action of the statement and the process. The similarities between the values of the properties are measured in terms of similarity score. The similarity score helps to determine the level of relatedness between the regulation and process individuals. Furthermore, the higher the similarity score, the more related they are.

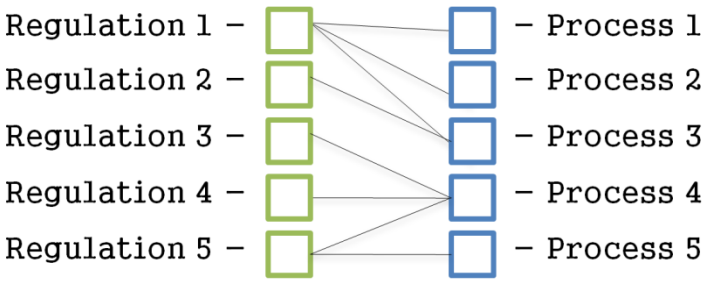


Figure 3-10. Mapping between regulations and processes

The similarity score is computed using the derivatives of the existing similarity algorithms. As mentioned in Section 2.6.6, the Lin similarity measure is found as an appropriate measure for this framework, which utilises the lexical relationships in the WordNet ontology, and the information content of the concepts in general corpora. When computation of the similarity score between a regulation and a process is completed, a set of mapping is created, and the mappings with high similarity scores are placed at the top. The compliance manager of the

organisation analyses the list of mapping and defines the acceptance of the similarity scores. The acceptable similarity score is a threshold, above which all the mappings are accepted, and below it, all the mappings are ignored.

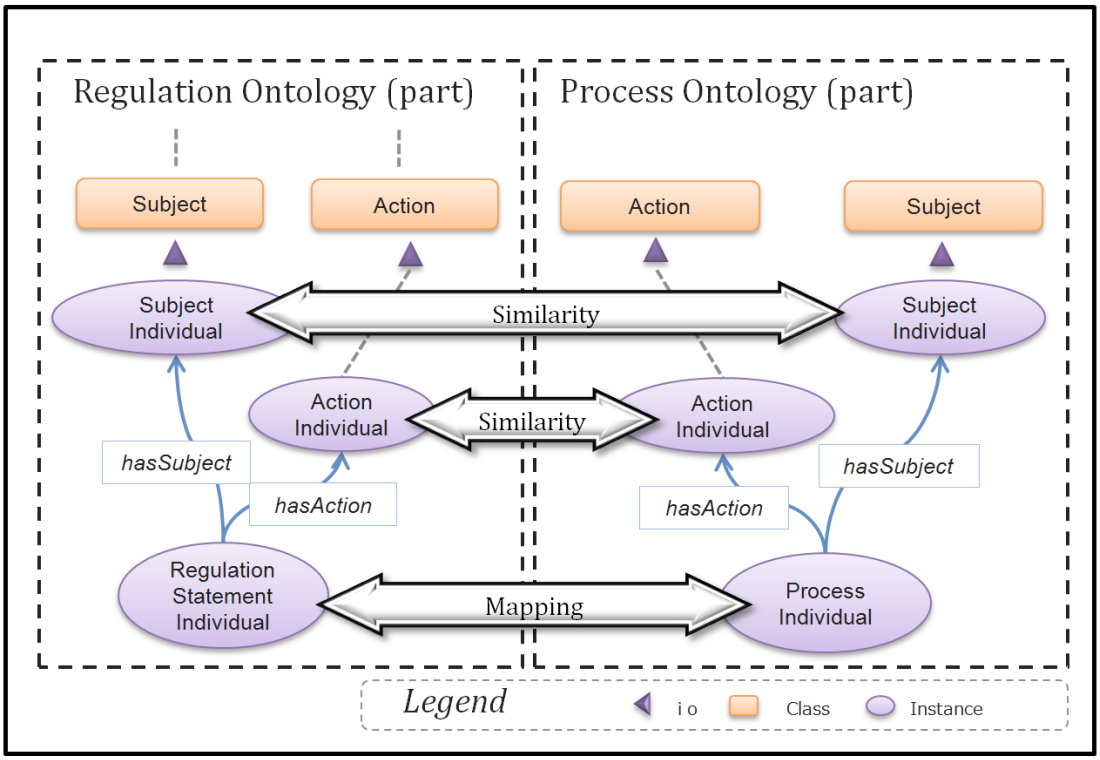


Figure 3-11. Mapping between a regulation-statement and a process based on subject and action similarities

The process of determining the mapping between a regulatory guideline and an organisational process is described in four steps: (i) identification of conceptual difference in an ontology, (ii) computation of three types of similarity scores, (iii) aggregation of three types of similarity scores and (iv) relating organisational processes with regulation-statements and regulation-paragraphs.

3.4.1 Conceptual Distance Computation

The similarity between two concepts is defined differently in different ontologies. In the similarity computation, the proposed framework determines whether a concept or an individual in the regulatory ontology is similar to a concept or individual in the process ontology.

Although, in general context, some concepts look like similar to each other, in a specific context (e.g. domain-ontology) they are different. For example, a lexical ontology such as WordNet defines that these concepts are similar; however, at the same time, a domain ontology such as process ontology defines that they are different from each other. In Figure 3-12, in WordNet ontology, the concepts `substance` and `equipment` are defined as similar to each other; where as in OntoReg ontology, they are defined as different from each other. Therefore, in order to define similarity between two concepts, it is necessary to determine their differences within a domain-ontology.

The distance between the two concepts in the process ontology is computed considering the axiom `disjointWith`. Currently, the value becomes 1 or 0 considering if they are disjoint with each other or not respectively; but in the future, we aim to consider the semantic distance computation algorithm (Ge and Qiu 2008) in conjunction with the current algorithm. After the conceptual difference computation, a table is created, and each row is represented by *<C1, C2, d>*, where *C1* and *C2* are two concepts in an ontology and d is the difference value.

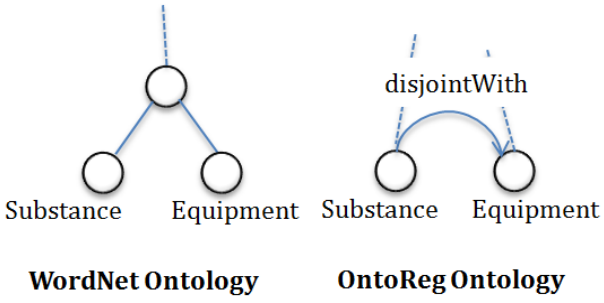


Figure 3-12. Different ontologies showing similarity and differences between the same concepts

3.4.2 Three Types of Similarity Score Computation

In the proposed framework, three types of similarities are computed: (i) Topic similarity (*Topic vs. Process*), (ii) Core-entity similarity (*Core vs. Process*) and (iii) Auxiliary-entity similarity (*Aux*

vs. Process). Topics intend to capture the overall semantics of a group of similar regulatory guidelines. Therefore, if the topic of a regulatory guideline is related to an organisational process, it can be considered that the regulatory guideline and the organisational process are also related. The core entities such as subject and action carry the semantics of a regulatory statement. Therefore, relating the core-entities with organisational processes can be considered as a semantic way of relating the guidelines with the processes. If the topic and core entities are represented implicitly and the similarity cannot be established, the words contained within a regulatory guideline may help to relate the regulatory guidelines with the processes. Figure 3-13 displays how the three similarities are computed.

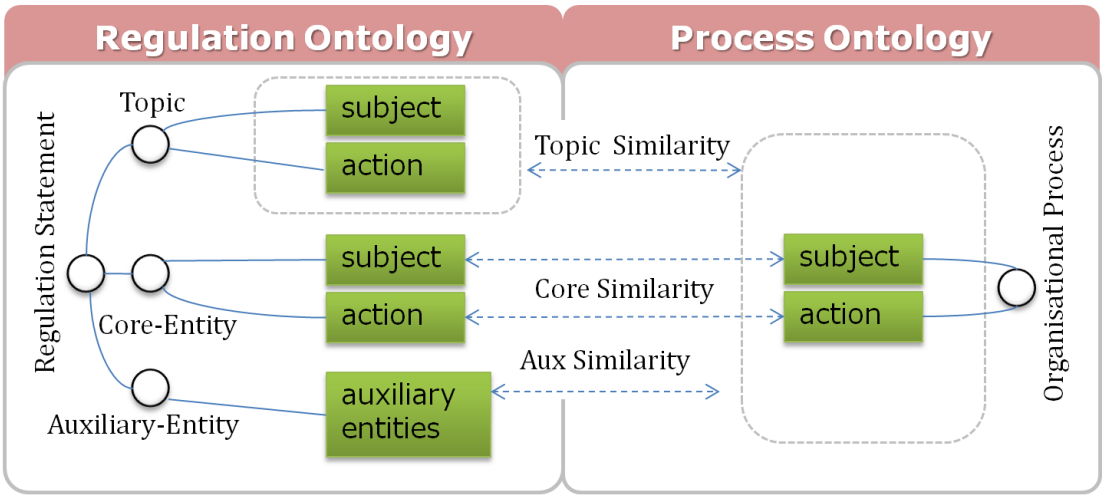


Figure 3-13. Three different types of similarity computations in the RegCMantic framework

In the core-entity similarity, average of the subject and action similarity is computed. In particular, the similarity scores are computed for subject and action separately. Algorithm 3-4 demonstrates the pseudocode to compute the similarity score between the subjects. This function computes a similarity score between the subjects of a regulation-statement and a process in regulation-ontology and process-ontology respectively. In this process, initially the score is set as *0*, which will be updated with the computed value. Consider there are two sets of subjects: *S1* from the regulation-statement and *S2* from the validation-task. A validation task is

the smallest unit of processes, which can be used to check the validation of the process. Now, each word in these sets are compared.

Algorithm 3-4. Similarity computation between subjects.

```
Input: r is a regulation and t is a validation task.
Output: score is the similarity score
function GET-SUBJECT-SCORE(r, t) returns score
    score = 0
    S1 = {s1 | s1 is_a_subject_in stmt}
    S2 = {s2 | s2 is_a_subject_in task}
    for each si ∈ S1
        for each sj ∈ S2
            d = GET-DIFFERENCE-VALUE(si, sj)
            if (d < θ ) then
                score' = SIMILARITY-SCORE(si, sj)
                if (score' >score) then
                    score = score'
                end if
            end if
        end for
    end for
    return score
```

First, the difference-value d is computed from the difference-table created from the process-ontology. A difference table represents two concepts in an ontology and their difference value. A difference value is the difference score between the concepts. It is computed by considering the special axioms in an ontology such as "owl:differentFrom", "owl:disjointWith" and "owl:allDifferent". If the two words are not defined as ontologically different in the process-ontology, d will be below the threshold θ and a similarity algorithm based on WordNet lexical ontology, Lin 91998) similarity is applied.

The Lin similarity considers the hierarchical structure of the terms in a lexical otology, WordNet (Pedersen et al 2004) and information content value *(IC)* of the terms from large corpora. This identifies the lowest common subsumer *(LCS)* between two compared words, computes the depth of the *LCS* from the root, measures the distance between the two compared terms via the *LCS*, and applies the *IC* values obtained from large corpora to compute the

similarity measure. It results into a set of similarity-scores, from which the highest similarity-score is selected as the final similarity-score of the subjects between the regulation-statement and the process.

Similarly, the action similarity is computed by comparing the action words associated with a regulation-statement and a process. When subject and action similarity scores are computed, an average of these two score is determined as core-similarity as shown below.

$$SIM_{core} = \frac{SIM_{subject} + SIM_{action}}{2} \quad (3)$$

Computation of topic-similarity and auxiliary-similarity is similar. In the topic similarity, the similarity score is computed between topic words and process words. The process words are the combination of subject and action in the process. The similarity between a topic and a process is determined by the percentage of the words in a topic that is similar to the process. Similarly, the auxiliary-similarity is computed by comparing the words in the regulation-statement and the process.

In order to increase the accuracy of the mappings, the following pre-processing is carried out.

1) The names of individuals and concepts are split. These names are usually made of words combined by using camel case and underline. They should be split in order to receive proper words.
2) The labels of individuals and concepts should be collected. Sometimes the individual and concept names are not composed of meaningful words such as individual names with abbreviations. In this case, they should be provided with the proper names in their labels, and if available, the labels should be used. For example, `T101CleaningTask` can be labelled as `Tank 101 Cleaning Task` as the important information about `Tank` was missing in the individual name.

3) The stop words in the text such as articles, prepositions, conjunctions and common words are removed. The removal of the stop words improves the accuracy and completeness of the result since they are less important in the text.
4) In all the similarity computation processes above, the two compared terms are checked against the "Difference Table" computed in the previous step in order to determine whether they are defined as different in the process ontology. It also requires setting a threshold θ; if a difference value exceeds the threshold, the compared terms are regarded as different from each other, and they will not be further processed for the similarity score. Figure 3-14 depicts a similarity computation process between two words, which states the similarity scores is only computed when the difference value d is smaller than the threshold, θ.

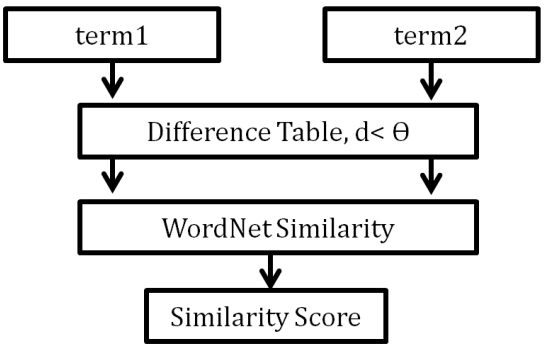


Figure 3-14. The similarity computation process showing consideration of difference table

3.4.3 Aggregating Three Similarity Scores

After computing the three types of similarity scores, we need to compute the aggregate similarity. The similarity aggregation algorithm (see Algorithm 3-5) emphasises the importance of the topic-similarity and the core-similarity, as these similarities are more meaningful as compared to the aux-similarity. The aux-similarity considers every annotated word in the regulatory text, which can be sometimes misleading such as the annotations within exceptions. Please note that the three types of similarity scores are computed between a validation task and a regulation-statement, not a regulation-paragraph.

In the aggregation algorithm, the maximum score between topic-score and core-score is chosen as the aggregate score. However, if the aux-score is the highest of all, the highest of the topic-score and the core-score is computed. Then, the average between the highest score and the aux-score is regarded as the aggregate score.

Algorithm 3-5 Computing aggregate score between a statement and a validation-task

Input: S_{topic}, S_{core} and is S_{aux} are topic-score, core-score and aux-score respectively
Output: S_{agg} is the aggregate similarity score of the three scores.
function GET-AGGREGATE-SCORE(S_{topic}, S_{core}, S_{aux}**) returns** S_{agg}

$S_{agg} = 0$
$S_{tc} = \text{MAX}(S_{topic}, S_{core})$
if $(S_{tc} \geq S_{aux})$ **then**
  $S_{agg} = S_{tc}$
else
  $S_{agg} = (S_{tc} + S_{aux})/2$
end if
return S_{agg}

3.4.4 Process-Statement Similarity to Process-Regulation Similarity Computation

As mentioned earlier, each regulation is composed of one or more statements. The similarity score, computed in the above steps, is the score between a statement and a validation-task. If a regulation contains multiple statements, it also contains a set of similarity scores. The highest score in the set is regarded as the similarity score between the regulation-paragraph and the validation-task, i.e. $Sim_{Reg} = MAX(Sim_{s1}, Sim_{s2}, ..., Sim_{Sn})$.

After the completion of similarity computation, the mappings with high similarity scores are placed at the top and presented to the user. The user then defines the acceptable level of the similarity score. The acceptable similarity score is the threshold, above which all the mappings are accepted, and below which all the mappings are ignored (see Algorithm 3-6).

Algorithm 3-6 Relating regulations with validation tasks

Input: *reg* is a regulation and *task* is a validation task.
Output: *true and false* are indications of related or not related respectively.
function RELATED(*reg, task***) returns** true **or** *false*

```
t = GET-ACCEPTABLE-SIM-SCORE-FROM-USER()
sim = GET-SIM-SCORE (reg, task)
if (sim≥t) then
  return true
end if
return false
```

The mapping algorithm was also evolved from baseline framework to extended framework. In the mapping part of the baseline framework, subject and action of a regulation-statement were considered for the mapping between regulatory guidelines and organisational processes. The subject and action, of a regulation-statement, are referred to as core-entities. In the extended framework, the mapping is performed considering three types of regulatory entities: (i) topic, (ii) core-entity and (iii) auxiliary-entity. Comparison of the core-entities of a regulation-statement and an organisation process is a meaningful way to compute the similarity score between them. However, some times the core entities are complex and represented implicitly. In this situation, it will be useful to consider the topic and auxiliary entities of the regulatory guideline in order to compare with the organisational process and compute similarity scores among them.

3.5 Features of the Proposed Framework

The RegCMantic is a generic framework, and likely to be applied to any domain where organisational processes are represented in an ontology. Application of this framework has the following benefits.

1) The creation of semantic regulation or representation of regulation in ontology is semi-automatic and saves expertise, time, efforts and money.

2) One of the important parts of the Semantic RCM is the mapping between the regulatory guidelines and organisational processes. In the RegCMantic framework, the mapping is semi-automatic, which also saves expertise, time and efforts.

3.6 Summary

In this chapter, a framework for RCM, RegCMantic is proposed and designed. It has described each step and process of the framework. The two main processes in the framework are extraction of regulatory guidelines and mapping between the regulatory guidelines and organisational processes. Therefore, the potential contributions of the proposed framework are (i) automatic extraction of the regulatory entities from regulatory guidelines and (ii) relating the regulatory guidelines with the organisational processes. The extraction process contains four components: (i) natural language parser, (ii) ontological concepts, (iii) definition terms and (iv) rules. Likewise, the mapping process considers three types of similarity scores: (i) topic-score, (ii) core-score and (iii) aux-score.

The RegCMantic aims to be a generic framework and is likely to be applied to various domains. Application of this framework benefits the RCM with enhanced reusability and makes the upgrading process smoother. The next chapter will provide the validation of the proposed framework in terms of implementation, design and development of a prototype. This will be followed by a chapter to evaluate the framework with some experimental results.

4 Implementation and Validation of the Framework

4.1 Introduction

This chapter describes validation phase of the proposed research. The validation is performed through the implementation of the RegCMantic framework in a case study. The description of the RegCMantic framework is provided in Chapter 3. This chapter establishes the rationale for the selection of the case study and uses the case study to examine each phase of the framework.

The RegCMantic framework aims to be a generic framework and is likely to be applicable to organisations, which have to relate regulatory guidelines with organisational processes for RCM. RCM is particularly important to the organisations which may pose a risk to the general public Yip et al. (2007). Therefore, these organisations are heavily regulated in order to protect the public from potential danger. Some of these organisations are healthcare industries, financial institutions and information managing companies.

The rest of the chapter is organised as follows. The selection of a domain and a scenario is explained in Sections 4.2 and 4.3 respectively. Likewise, the natures of regulations and processes is described in Section 4.4 and 4.5. The implementations of extraction and mapping processes of the RegCMantic framework are described in Sections 4.6 and 4.7 respectively. Section 4.8 has summarised this chapter highlighting its salient features.

4.2 Selection of the Case-Study

The implementation of the proposed framework in a case-study comprises two separate processes: (1) selection of the domain and (2) selection of a scenario in the domain. Selection of a domain is the process of choosing an industry or area for the application of the framework. Similarly, the selection of a scenario is the process of selecting a process and a set of related

regulations in order to validate the framework. In the following subsections, these two processes are explained separately.

4.2.1 Selection of the Domain

A case study in the Pharmaceutical industry is chosen after considering the following facts:

1) The Pharmaceutical industry is one of the heavily regulated industries since production and handling of drugs directly affects the health and safety of a large number of people (Haider 2002). In order to implement the proposed framework, a domain with strict regulations, is required. The Pharmaceutical industries in the UK are governed by the UK (MHRA 2012), and European regulations (EMEA 2012), and in order for them to sell the drugs in the USA, they also conform to the regulations in the USA (FDA 2012). The strictness and amount of regulations imposed on the domain is one of the considerations of the selection of this industry.
2) Another consideration for the case study is the selection of the domain, which comprises sufficient amount of complexity in order to apply the extraction and mapping processes. RCM in the Pharmaceutical industry is equipped with some tools and approaches, and improvement on the tools and approaches is highly desired. The extraction and mapping processes in the Pharmaceutical industry also possesses sufficient amount of complexity.
3) In order to validate the results of the framework, various methods can be used. One of the most appropriate methods is to perform validation through the experts in the domain. This research aims to involve the Chemical Engineers at the University of Oxford in the validation process.

4.2.2 Selection of a Scenario in the Domain

Selection of the regulatory guidelines and processes is required in order to validate the proposed framework in a selected domain. The processes involved in the aspirin production

are selected because the aspirin production has a comparatively clear and simple structure. The purpose of this study is not to offer a better representation solution for the complex processes, but rather to prove that the framework helps to improve automation in the overall compliance management system.

Having considered the aspirin production processes, we have the only option to select the regulatory guidelines that regulates the aspirin production processes. Therefore, the Chapter 5 of the EU Guidance on Good Manufacturing Practice 2007 (Eudralex) is selected because the UK pharmaceutical industries have to follow European regulations, and the guidelines in Chapter 5 govern the manufacturing processes.

4.3 Nature of Regulations

4.3.1 Background

Compliance Management in the Pharmaceutical industry is regarded as a very important process. Getting approval from the regulatory bodies for a drug is very expensive, and after the approval, the regulatory bodies start monitoring the processes to ensure it is effective and safe.

There are three main regulatory bodies, which affect the production and marketing of pharmaceutical products in the UK. One of them is the UK regulatory body, the Medicines and Healthcare products Regulatory Agency (MHRA). The other is the EU regulatory body, the European Medicines Agency (EMEA). The third one is the US regulation, the Food and Drug Administration (FDA). All these regulatory bodies have a common principle to protect the health of the public by ensuring safer medicines and healthcare products (MCA 2007). The pharmaceutical process in the UK is directly regulated by the MHRA and the EMEA and is subject to FDA regulations if the medicinal products are supplied in the USA.

The principles and guidelines provided by these regulatory bodies are very generic and in high level. Each body has put forward a set of principles and guidelines, some of them intersect

with each other. Some examples of such guidelines are the Good Manufacturing Practice (GMP), the Current Good Manufacturing Practice (cGMP), and the Good Laboratory Practice (GLP). In order to allow various implementations of the guidelines, the guidance's are written in generic forms. However, one has to translate them into a specific scenario in order to implement the guidelines.

A fair amount of literature is available for the Compliance Management systems in the Pharmaceutical industry. One such example is the "Rules and Guidance for Pharmaceutical Manufacturers and Distributers 2007", also known as the **Orange Book** (MHRA 2012). This book collates the regulatory information for the industry in the UK. Another documentation that comprises the European regulations for the industry is **Eudralex**. Two of the popular books having instructions to conform to FDA are the "Pharmaceutical Master Validation Plan: The Ultimate Guide to FDA, GMP, and GLP Compliance (Haider 2002)" and "Validation standard operating procedures: a step-by-step guide for achieving compliance in the pharmaceutical, medical device, and biotech industries**"** (Haider 2006). The former explains the instructions for validations steps, and the latter explains the validation and monitoring of the previously defined steps. However, as mentioned earlier, the instructions are presented in generic forms and one has to pay careful attention to recreate the instructions to fit into a particular case. RCM can only be made useful by creating low-level validation plans, monitoring, and reporting of these plans.

4.3.2 Regulations

The Pharmaceutical processes in the UK have to conform to the two main regulations, and they have different document-structures. The guidelines recommended in the MHRA, EMEA and FDA do not follow a uniform document-structure. This makes the identifications of the regulation-paragraph difficult. The guidelines from the same regulatory body are also presented in different document-structures from time to time. However, there is a common practise to represent some specific document-component such as **part, chapter** and **section**

among them. These high-level document-components are usually preceded with the relevant text. Some examples of regulatory guidelines are shown in Figure 3-3 and Figure 4-1. In this example, the guidelines are represented in the document-components proceeding with some numbers. The regulatory guidelines are generally preceded with some numbers, and these numbers are related to chapters or sections. However, this practice is not always followed. This makes the identification of regulation-paragraph challenging. Application of the first part of the framework identifies the regulation-paragraph, and the result is presented to the user for verification.

EudraLex
The Rules Governing Medicinal Products in the European Union

Volume 4
EU Guidelines to
Good Manufacturing Practice
Medicinal Products for Human and Veterinary Use

Part II
Basic Requirements for Active Substances used as Starting Materials

..... (details hidden...)

5 Process Equipment

.....(details hidden...)

5.2 Equipment Maintenance and Cleaning

.....(details hidden...)

5.22 Equipment and utensils should be cleaned, stored, and, where appropriate, sanitized or sterilized to prevent contamination or carry-over of a material that would alter the quality of the intermediate or API beyond the official or other established specifications.

Figure 4-1. An example of the regulatory text in Eudralex

4.4 Nature of Processes

The processes involved in the production of Aspirin are selected for the case study as these processes are represented in a process ontology called OntoReg (Sesen et al 2010). The description of the processes for Aspirin production is modelled as an instance of a concept called Process in OntoReg ontology.

The ontological representation of a validation-task is depicted in Figure 4-2. In the process ontology, OntoReg, a validation-task is associated with a `subject` via a property `hasPatient`, for which we have created an equivalent property called `hasSubject` for clarity. Similarly, an `action` is indirectly associated with a validation-task, which can be determined by traversing some properties and individuals. In the `FilterCleaningTask`, the subject is `Filter101`, which is an individual of a class `Filter`. The class `Filter` is subsumed by the classes `ProcessingEquipment` and `Equipment`. The `action` for the `FilterCleaningTask` is defined implicitly. Having traversed through the property `isReponsibilityOf` and `performs`, it is determined that `CleaningIndividual` is an individual of a class `Cleaning`. The class `Cleaning` is subsumed by its superclass `Action`.

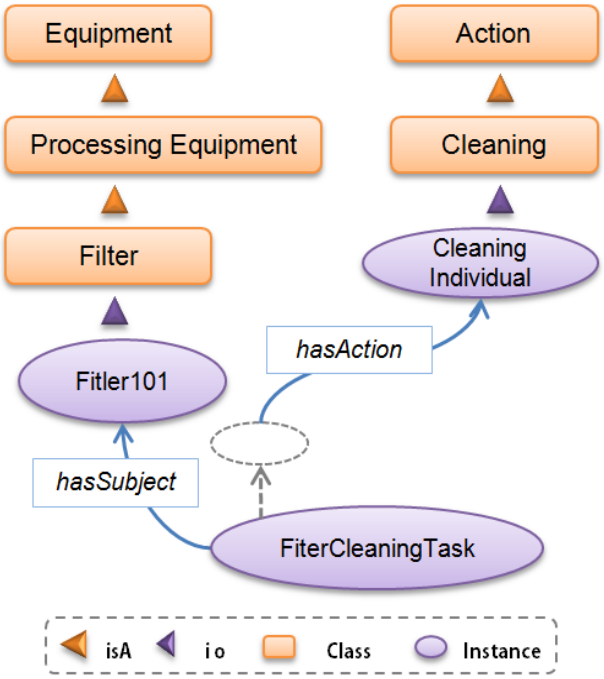


Figure 4-2. Filter Cleaning Task represented in OntoReg ontology

4.5 Regulatory Entity Extraction

The regulations selected for the case study are the regulations defined in the Eudralex 2007, an excerpt of which is shown in Figure 4-1. These regulations are provided in PDF formats. The

chapter in the document is represented by its preceding text and the chapter number. The example shows that the name of the chapter is "Production", and the chapter number is "5". The regulation-paragraphs in the document are preceded with some numbers corresponding to the chapter numbers. In the example, numbers are the decimals of the chapter numbers such as 5.21, 5.45 and 5.61. Other interesting features of all these three regulations are that they are all written in passive voice. In order to extract regulatory entities from this document, the following five steps for extraction are used.

4.5.1 Pre-Processing

In this step, the text in the PDF document format is converted into the HTML document format. The conversion process is based on existing state of art tools. In this case, VeryPDF[13] is selected as it is freely available and provides the necessary features. For instance, the organisation of Cascading Style Sheet (CSS) and the HTML tags is simple, clear and reads all the important font-features of the text. There also exist various versions of commercial software, which deal with a huge amount of PDF documents efficiently, and can be used in a commercial scale. However, for the given case study, VeryPDF can sufficiently process the required number of documents.

The converted HTML document is depicted in Figure 3-4. In this document, the head section contains the CSS definition of the different text-types in the document, and the body section contains the **div-sections** with their **class-id**. As in a typical web page, the style for each div-class is defined in the CSS, in the head section. Such a representation can help to utilise the style-information in the head section in order to determine the font-features of each **text-style** in the body. In this example, in the body section, we can see that the text containing "Chapter 5…" is assigned a CSS class "ft01". In the head section, this class is defined with font-weight bold and font-size 23. The information in the font such as font-weight and font-size helps to determine the text-type of the text.

[13] http://www.verypdf.com/app/html-converter/

Although the conversion of PDF document to HTML format is useful for processing the document, there are some challenges produced by the conversion. For instance, each line in the PDF document is identified as a paragraph. A line can be just a part of a sentence or some fragments of different sentences. Such an organisation of lines has posed a challenge to identify the completeness of sentences and paragraphs and in this thesis, it is solved by application of **Spanning-Paragraphs** algorithm (See Appendix C), which identifies the fragments of a sentence and combines them.

4.5.2 Schema-Generation

The Eudralex regulations in HTML format are processed with two important components: (1) **feature-reader** and (2) **structure-predictor** as described in Section 3.3.2. The first one has identified various important features related to the text and the later has predicted the possible **document-structure** of each **text-level**. After completion of the computation of the prediction, the result is presented in a panel (see Figure 4-3).

In the panel, a list of **text-levels** is presented along its right side. Each **text-level** is associated with its possible **document-structure** and is presented to the right side. The provided **document-structure** has a suggestion in a dropdown list of all the applicable **document-structures**. This feature enables a user to select appropriate document-structure if the suggestion has to be modified. Some examples of the text with application of the features in each **text-level** have also been provided on the right hand side of the panel. The examples help a user to verify the nature of the text at the **text-level**. There are also three important buttons at the bottom of the panel: (1) **'Process Document'** (2) '**Create Schema'** and (3) '**Convert to XML'**. The first button processes the document with application of feature-reader and structure predictor. The second button saves the document-schema about the document-structure. The third button creates XML representation of the document.

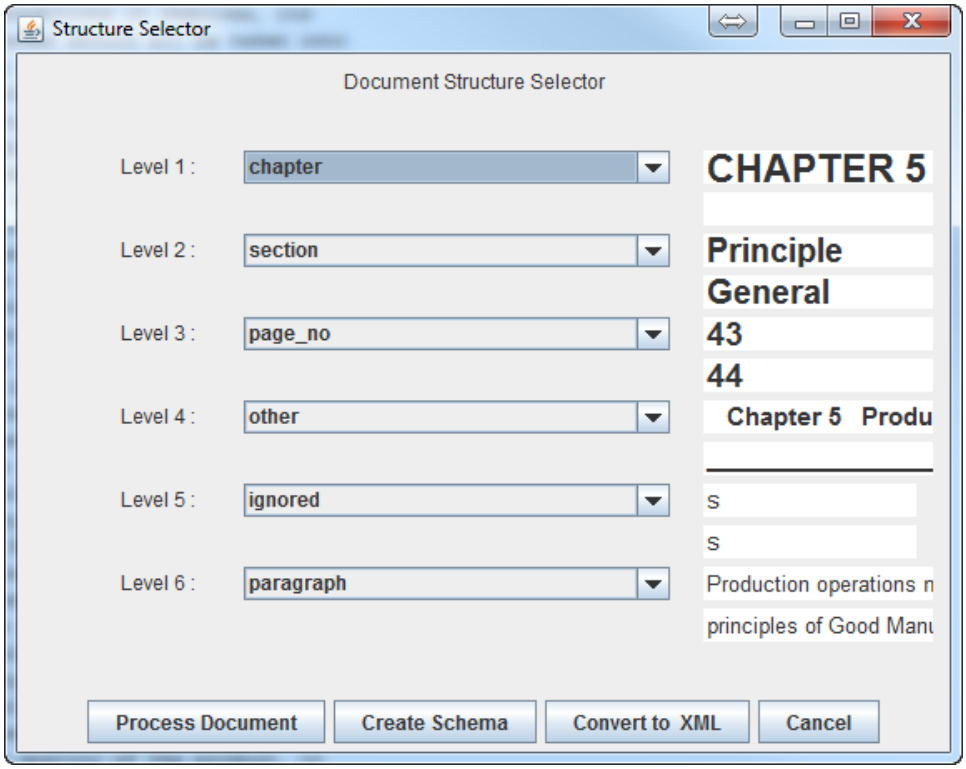


Figure 4-3. Predicted document-structure presented to users for verification

4.5.3 XML Regulation (Regulation with a Standard Structure)

As mentioned in the description of the framework, the creation of XML- regulation is not just the conversion of the HTML document format to the XML document format. It is a succinct representation of the document-components, which makes the processing of the information easier. Figure 3-5 depicts an excerpt of the Eudralex 2007 regulation in XML document format. It is generated using the document-schema created in Section 4.5.2.

The XML document format has two basic sections: (1) meta section and (2) content section. The meta section comprises meta information about the regulatory document such as name and description of the regulatory document, name of the regulatory body which produced the regulatory-document, version of the document and the date on which the document was published. The content section represents the document-structures of the document such as chapter, section and paragraph. The document-structures having a text-level higher than that of paragraph also contain titles. In Figure 3-5, chapter and section have attributes called **title**.

Similarly, the paragraphs are provided with **paraNum**, as their attributes. The paraNum specifies the number representing the paragraph in the regulatory document.

4.5.4 Regulatory Entity Annotation

This step has mainly focused on processing the regulation-paragraphs. However, the chapter and the section have also been used to provide more semantics to the process. The annotation process has used four basic components as described in Chapter 3: (1) **sentence parser**, (2) **definition terms**, (3) **ontological concepts** and (4) **annotation rules**. The annotation process is implemented in General Architecture for Text Annotation (GATE) platform (Cunningham et al 2002). The GATE has an API to work with Java, called GATE-Embedded. Lists of ontological concepts are provided in Appendix D. The annotation rules are created using Java Annotation Patterns Engine (JAPE) (Thakker et al 2009). Some examples of important rules are provided in Appendix E.

The parsers are the natural language parsers, which identify the relation among the grammatical component of a sentence. For instance, they identify a verb and its relationship with a subject, object and modifiers in a sentence. The parsers help to identify the chunks of the sentence where different regulation-entities are embedded such as **subject-chunk**, **obligation-chunk** and **action-chunk**. An excerpt of parsed sentence, converted into meaningful chunks for this framework, is presented in Figure 4-4. The full list of the parsed sentences is provided in Appendix F.

```
<parsed_sentence count="38">
    <subject> Records</subject>
    <obligation> should be</obligation>
    <action> maintained</action>
    <object> </object>
    <condition></condition>
    <modifier> stating the name , address , qualifications , and type of service provided<
</parsed_sentence>
<parsed_sentence count="39">
    <subject> Buildings and facilities used</subject>
    <obligation> should be</obligation>
    <action> located , designed , and constructed</action>
    <object> </object>
    <condition></condition>
    <modifier> to facilitate cleaning , maintenance , and operations as appropriate</modif
</parsed_sentence>
<parsed_sentence count="40">
    <subject> Facilities</subject>
    <obligation> should</obligation>
    <action> also be designed</action>
    <object> </object>
    <condition></condition>
    <modifier> to minimize potential contamination</modifier>
</parsed_sentence>
<parsed_sentence count="41">
    <subject> facilities</subject>
    <obligation> should</obligation>
    <action> also be designed</action>
    <object> </object>
    <condition> Where microbiological specifications have been established for the interme
    <modifier> to limit exposure to objectionable microbiological contaminants</modifier>
</parsed_sentence>
<parsed_sentence count="42">
    <subject> Buildings and facilities</subject>
    <obligation> should</obligation>
    <action> have</action>
    <object> adequate space</object>
    <condition></condition>
    <modifier> for the orderly placement of equipment and materials</modifier>
</parsed_sentence>
```

Figure 4-4. Parsed regulation-sentences divided into different chunks

The definition terms are domain specific terms used in a document and are generally provided by the regulators for each regulatory document. A gazetteer is a list of words or terms. A gazetteer terms is created for the annotation. Figure 3-6 shows an example of the definition terms in Eudralex regulatory document.

Figure 3-7 depicts an example of the JAPE rule, and the rest of the rules are provided in Appendix E. In this example, it shows how the actions are annotated using the chunk, `action_container`, and the annotators such as `rule_action`, `definition_term`, `extracted_term` and `concept_ontology`. Typically, a rule contains a left hand side (LHS) and a right hand side (RHS). In LHS, it accepts preconditions and input annotations and in the RHS, new annotations are created based on the preconditions and the input annotations.

Figure 4-5 presents an example of the annotation process, which depicts how the system identifies the entities in a regulatory paragraph. In this figure, the regulation 5.20 is annotated with the regulation-entities. The three core regulation-entities in this example are "Schedules and procedures", "should be" and "established", which are **subject, obligation** and **action** respectively. Similarly, an auxiliary-entity, **purpose** has also been identified in the example.

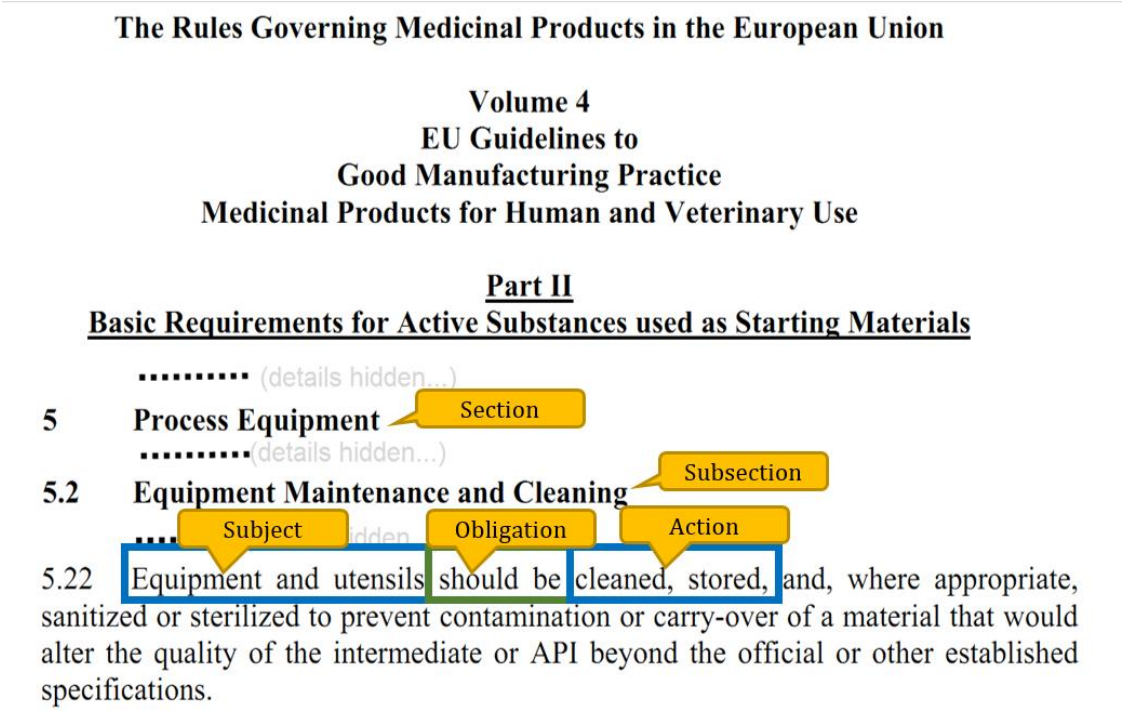


Figure 4-5. An example of annotated regulatory text

4.5.5 The SemReg ontology Population

The creation of semantic regulation involves two processes: (1) the engineering of an ontology to represent the regulations and (2) the populations of the ontology from the regulation text. The first process is described in Section 3.3. The implementation of the second process is described below.

The SemReg ontology population has used the annotations created in Section 4.5.4. An example of the ontological concepts, instances and their relationships are illustrated in Figure 4-6. In particular, the relationships among the instances of the concepts `Topic`, `Regulation`, `Statement`, `Equipment`, `Obligation` and `Action` in regulation, `Eudralex_5.22` are presented. The centre of this example is the statement, `Eudralex_5.22_1` that is connected to the regulation, `Eudralex_5.22` via inverse of an object property, `hasStatement`. The regulation, `Eudralex_5.22` is specified as regulation of a topic, `Eudralex_5.2` via inverse of an object property, `hasRegulation`. The topic, regulation and statement have also been provided with their text content via a data-type property, `description`. The most important properties of a statement are `hasSubject`, `hasObligation` and `hasAction`. The subjects, `Equipments` and `utensils` are connected to the statement via the object property `hasSubject`. Similarly, the obligation, `ObligationIndividual1` is linked with the statement via, the object property, `hasObligation`. The obligation is specified as `positive` and `strong` via the object properties, `hasType` and `hasStrength` respectively. The actions `cleaned` and `stored` are associated with the statement via the object property `hasAction`.

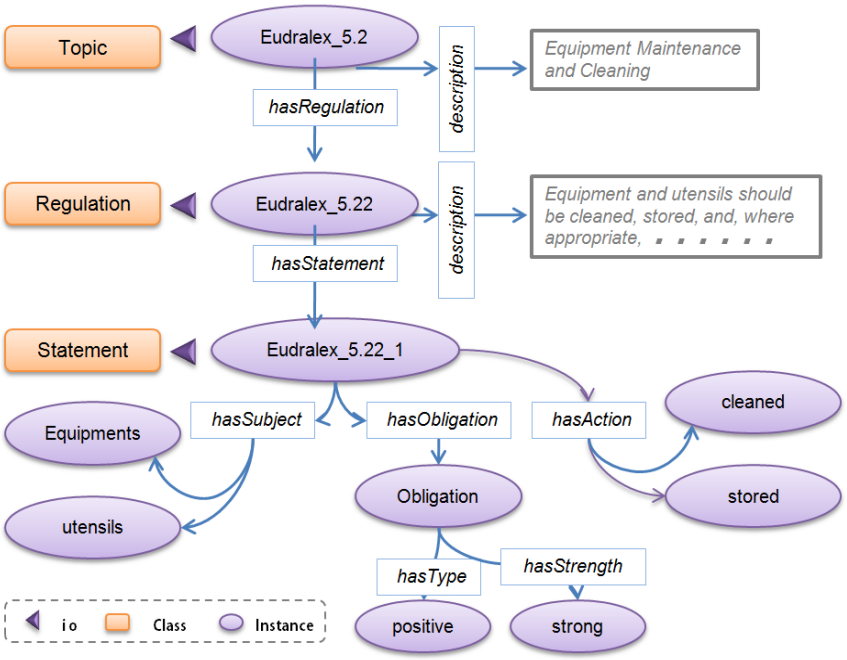


Figure 4-6. Eudralex 5.22 regulation represented in SemReg ontology

4.5.6 Challenges

While going through the regulatory entity extraction process, the author has come across the following challenges.

1) The PDF to HTML translation tool represented each line in the document as a new paragraph, which made the document structure analysis harder. An algorithm called "spanning paragraphs" has been applied to combine the related lines into a paragraph.
2) The regulatory entities are extracted from a sentence. Therefore identifying a sentence is essential. The sentence splitters are not 100% accurate, for example, ANNIE Sentence Splitter misses the sentence boundary if they start with number. Adaptation of the sentence splitter considering the nature of the text solved the challenge.
3) While representing the subject entities in an ontology, it is difficult to identify the category of they fall in. For example, in a process ontology the subject is defined as union of concepts such as Document, Equipment, Substance and Operation. There are equivalent concepts in the regulatory ontology. However, it is difficult to determine the concepts they should fall in and requires developing algorithms or frameworks,

which is recommended for the future work. In this research, it has been solved by creating a separate concept called Subject.

4.6 Mapping Regulations with Organisational Processes

In the mapping process, as described in Section 3.4, the `Subject` and `Action` concepts related to the regulations and the organisational processes are compared. A similarity-score between two compared concepts is computed in order to determine whether a regulation and an organisational process are related. Consider an example of a regulation-statement, `Eudralex_5.22_1` and a validation-task `T101CleaningTask`. The regulation states, "*Equipment and utensils should be cleaned, stored, and,….*". The core regulation-entities in this regulation are `Equipments` and the `Cleaned` and are defined under the concepts `Subject` and `Action` respectively in the SemReg ontology. The validation-task `T101CleaningTask` in the OntoReg ontology is associated with a subject, `Tank101` and an action, `Cleaning101`. From the structure of the OntoReg ontology, it can be deduced that the `Tank101` is a `Storage-Tank`. A `Storage-Tank` is a `Storage-Equipment`, and a `Storage-Equipment` is an `Equipment`. Likewise, since, the `Cleaning101` is specified as an instance of its class, `Cleaning`, it can be inferred that `Cleaning101` is a `Cleaning`.

The similarity-score between the subject terms, **tank** in the regulation and **equipment** in the process are computed using a WordNet[14] based similarity algorithm, Lin (1998) similarity. The result of the similarity-score computation is found 1.00. Similarly, the action terms, **cleaned** in the regulation and **cleaning** in the process are compared using the same algorithm, and their similarity-score has also been found to be 1.00. In the similarity computation, the scores, 0.00 and 1.00 are considered as lowest and highest similarity respectively. Having the high similarity-scores in both subject and action terms indicate that the regulation and the process are more closely related. However, the overall similarity between a regulation and a process is

[14] http://wordnet.princeton.edu/

determined by computing three types of similarity scores and their aggregation as described in Chapter 3.

4.6.1 Three types of similarity scores computation

The three types of similarity-scores, which are needed to determine the overall similarity as explained in Chapter 3, are **topic-similarity, core-similarity** and **aux-similarity**. For the three types of similarity score computation, three types of entities are selected from regulations, and they are the `topic`, `core` and `aux` entities. Figure 4-7 depicts an XML representation of the three types of entities in the regulation, `Eudralex_5.22`. Similarly, the entities required from the process domain are `subject`, `action` and `annotation`. Figure 4-8 depicts the collection of `subject`, `action` and `annotation` of `FilterCleaningTask`. The subjects are identified by the names and labels of the subject individual and its classes and superclass. Likewise, the actions are determined by the names and labels of the action individuals, their classes and super classes. The annotations are created by combining subjects and actions. A bag of words (bow) is the collection of words in a sentence or a phrase. The `bow` in the `topic` and `aux` of the regulation, `Eudralex_5.22` are created using the distinct terms in the text.

```
<regulation id="Eudralex_5.22">
  <statement id="Eudralex_5.22_1">
    <topic>
      <text>Chapter 5 Production, Process Equipement, Equipment Maintainance and Cleaning</text>
      <annotation>Process,Equipment,Cleaning</annotation>
      <bow>Equipement,Maintainance,Process,Equipment,Cleaning</bow>
    </topic>
    <core>
      <subject>equipment, utentils</subject>
      <action>cleaned,stored, sanitized, sterilized</action>
    </core>
    <aux>
      <text>5.22 Equipment and utensils should be cleaned, stored, and, where appropriate, sanitized or … </text>
      <annotation>API,quality,material,Equipment,other established specfications,contamination</annotation>
      <bow>utensils,sanitized,sterilized,prevent,alter,intermediate,official,API,quality,material,Equipmen … </bow>
    </aux>
  </statement>
</regulation>
```

Figure 4-7. Three types of entities in Eudralex 5.22 regulation

```
<task id="FilterCleaningTask">
  <subject>filter,processing equipment, equipment</subject>
  <action>cleaning</action>
  <annotation>filter, processing equipment, equipment, cleaning</annotation>
</task>
```

Figure 4-8. Subject, action and annotations in Filter Cleaning Task

For the core-score computation,

- The `subject` and `action` in the regulation-statement, `Eudralex_5.22_1` are compared with the `subject` and `action` of the validation-task, `FilterCleaningTask` respectively. In particular, the terms in regulatory subject, "*equipment, and utensils*" are compared with the terms in the process subject, "*filter, processing equipment, equipment*". This comparison has produced a set of similarity between these two subjects.

After the two separate comparisons, it produces two sets of scores such as subject-score set (see Table 4-1) and action-score set (see

Table 4-2).

Table 4-1. An example of similarity scores computation between regulatory and process subjects

Regulatory Subject	Process Subject	Similarity Score
Equipment	Filter	0.42
Equipment	Processing Equipment	0.54
Equipment	Equipment	1.00
Utensils	Filter	0.32
Utensils	Processing Equipment	0.27
Utensils	Equipment	0.48
	Highest Similarity Score	1.00

Table 4-2. An example of similarity scores computation between regulatory and process actions

Regulatory Action	Process Action	Similarity Score
Cleaned	Cleaning	1.00
Stored	Cleaning	0.00
Sanitized	Cleaning	0.00
Sterilized	Cleaning	0.84

Highest Similarity Score	1.00

- In the `subject-score` set, *{ 0.42, 0.54, 1.00, 0.32, 0.27, 0.48 }* the highest score is determined as *1.00*. Therefore, 1.00 is set as the similarity score between the sets of subjects in the regulation-statement, `Eudralex_5.22_1` and the process, `FilterCleaningTask`.
- Similarly, in the `action-score` set, *{ 1.00, 0.00, 0.00, 0.84, 1.00 }* the highest score is found as 1.00. Therefore, the similarity score between the sets of actions in the regulation-statement, `Eudralex_5.22_1` and the process, `FilterCleaningTask`.
- Then, an average score between the `subject-score` and `action-score`, 1.00 is determined as the `core-score`.

The computations of `topic-score` and `aux-score` are similar. In the `topic-score` computation, the terms, "*Equipment, Maintenance, Process, Equipment, Cleaning*" in the `bow` of `topic` in the regulation-statement, `Eudralex_5.22_1` are compared with the terms, "*filter, processing equipment, equipment, cleaning*" in the `annotation` of `FilterCleaningTask` (see Table 4-3). The highest similarity score between the term "*Equipment*" in regulation and the terms "*filter, processing equipment, equipment, cleaning*" in the process is found as 1.00. Similarly, the highest similarity scores of "*Maintenance", "Process", "Equipment"* and *"Cleaning"* with respect to their comparison with the terms in process annotations are found as 0.73, 0.56, 1.00 and 1.00 respectively. Then, the average of these scores, 0.86 is determined as the `topic-score` between the regulation-statement, `Eudralex_5.22_1` and the process, `FilterCleaningTask`.

In the `aux-score` computation, the terms, "*utensils, sanitized, sterilized, prevent, alter, intermediate, official, API, quality, material, equipment...*" in the `bow` of `aux` in the regulation-statement, `Eudralex_5.22_1` are compared with the terms, "*filter, processing equipment,*

equipment, cleaning" in the `annotation` of `FilterCleaningTask`. It has also carried out the highest similarity score computation and the average of the highest similarity score computation. Then, the `aux-score` between the regulation-statement, `Eudralex_5.22_1` and the process, `FilterCleaningTask` is computed as 0.42. A part of an XML file representing the three scores computed between the regulation, Eudralex_5.22 and the process, FilterCleaningTask is provided in Figure 4-9.

Table 4-3. An example of similarity scores computation between regulatory topic and process

Regulatory Topic	Process Annotation	Similarity Score
Equipment	Filter	0.42
Equipment	Processing Equipment	0.54
Equipment	Equipment	1.00
Equipment	Cleaning	0.06
	Highest Similarity Score	**1.00**
Maintenance	Filter	0.00
Maintenance	Processing Equipment	0.12
Maintenance	Equipment	0.00
Maintenance	Cleaning	0.73
	Highest Similarity Score	**0.73**
Process	Filter	0.08
Process	Processing Equipment	0.56
Process	Equipment	0.12
Process	Cleaning	0.40
	Highest Similarity Score	**0.56**
Equipment	Filter	0.42
Equipment	Processing Equipment	0.54
Equipment	Equipment	1.00
Equipment	Cleaning	0.06
	Highest Similarity Score	**1.00**
Cleaning	Filter	0.00
Cleaning	Processing Equipment	0.00
Cleaning	Equipment	0.00

Cleaning	Cleaning	1.00
	Highest Similarity Score	**1.00**
	Average of the Highest Similarity Scores	**0.86**

4.6.2 Aggregate Similarity Computation

In the examples in 4.6.1 the `topic-score`, `core-score` and `aux-score` are computed as 0.86, 1.00 and 0.42 respectively. In the aggregation algorithm, the maximum score between `topic-score` and `core-score` is computed as:

$S_{tc} = MAX(S_{topic}, S_{core}) = MAX(0.86, 1.00) = 1.00$

Where, S_{tc} is the maximum score between `topic-score`, S_{topic} and `core-score`, S_{core}. In this case, the S_{tc} is greater than the `aux-score`, S_{aux}. Hence, the final similarity score between the regulation-statement, `Eudraxlex_5.22_1` and the validation-task, `FilterCleaningTask` is determined as 1.00, which is represented as the `final-score`. Then, an XML file, containing all the three scores and the aggregate score between regulation-statements and processes, is generated (see Figure 4-9.).

```
2  <mapping mapping_id = "mid_133">
3      <reg_id>Eudralex_5.22</reg_id>
4      <stmt_id>Eudralex_5.22_1</stmt_id>
5      <task_id>FilterCleaningTask</task_id>
6      <topic_score>0.86</topic_score>
7      <core_score>1.00</core_score>
8      <aux_score>0.42</aux_score>
9      <final_score>1.00</final_score>
10 </mapping>
```

Figure 4-9. Three types of similarity scores computed between Eudralex_5.22_1 and FilterCleaningTask

4.6.3 Process-Statement Similarity to Process-Regulation Similarity Computation

The similarity scores computed in Section 4.6.3 are the similarity scores between a regulation-statement, `Eudraxlex_5.22_1` in SemReg ontology with an organisational process, `FilterCleaningTask` in OntoReg ontology. In this case, the regulation, `Eudralex_5.22` contains only one regulation-statement, `Eudralex_5.22_1`. Therefore, the similarity score between the regulation, `Eudraxlex_5.22` and the validation-task, `FilterCleaningTask` is set as 1.00. If the regulation had two or more regulation-statements, the highest score among them would be regarded as the similarity score between the regulation and the process. After the similarity scores computations, a list of mapping between all the regulations and processes with their similarity score is generated (Figure 4-10). The highlighted line 8 shows the similarity score between `Eudralex_5.22` and `FilterCleaningTask` is 1.0.

```
1  mapping_id, reg_id, task_id, score,accuracy
2  mid_1, Eudralex_5.21, FilterCleaningTask, 1.0,
3  mid_2, Eudralex_5.21, T101CleaningTask, 1.0,
4  mid_3, Eudralex_5.21, T102CleaningTask, 1.0,
5  mid_4, Eudralex_5.21, FilterCleanlinessTestTask, 1.0,
6  mid_5, Eudralex_5.21, T101CleanlinessTestTask, 1.0,
7  mid_6, Eudralex_5.21, T102CleanlinessTestTask, 1.0,
8  mid_133, Eudralex_5.22, FilterCleaningTask, 1.0,
9  mid_134, Eudralex_5.22, T101CleaningTask, 0.97562,
10 mid_135, Eudralex_5.22, T102CleaningTask, 0.97562,
11 mid_153, Eudralex_5.26, StartingMaterialPurchase_1, 0.92955,
12 mid_136, Eudralex_5.22, FilterCleanlinessTestTask, 0.92357,
13 mid_137, Eudralex_5.22, T101CleanlinessTestTask, 0.89919,
14 mid_138, Eudralex_5.22, T102CleanlinessTestTask, 0.89919,
15 mid_165, Eudralex_5.31, StartingMaterialTestTask_7, 0.86649,
```

Figure 4-10. An excerpt of computed mapping between regulations and validation-tasks

4.6.4 Challenges

While implementing the mapping part of the RegCMantic framework in the case study, the following challenges have been found.

1) Computing similarity between a substance and an equipment by using a lexical ontology such as WordNet generates a high similarity score. This is true in many

cases. However, in some domain they are treated as completely different from each other. In this thesis, the process ontology defines them as different concepts. Initially, the RegCMantic framework generated several spurious mappings because of the similarities of these concepts. Later, it was solved by defining a difference table, which stores the difference score between two concepts in a domain ontology.

2) The topics in the regulatory guidelines and the concepts in a process ontology are considered for computing the similarity score between a regulation-statement and an organisational process. If the topics and the concepts are at the higher level, they express generic meaning, which may lead to generate false positive mappings. Initially, the framework generated several spurious mappings with higher scores because of considering the topics and the concepts. This challenge was tackled by setting an acceptable level in the topic and concepts in the ontology.

4.7 Summary

This chapter described the implementation and validation of the RegCMantic framework in a case study. The case study is carried out in the Pharmaceutical industry, as it is one of the heavily regulated industries. In particular, Eudralex, the EU regulation governing the Good Manufacturing Guidelines (GMP) in the Pharmaceutical industry is selected. The regulation ontology, SemReg is created extending the LKIF-Core ontology. Prior to the regulatory entity extraction, the framework has identified the important document-components such as **topic** and **regulation-paragraphs**. The core-entities such as **subject, obligation** and **action** and the aux-entities such as **place, time** and **reason** are extracted from the regulation-paragraphs. The extraction process is followed by the SemReg ontology population with the extracted regulation-entities.

In the mapping process, a regulation-statement, `Eudralex_5.22_1` in SemReg ontology and a process `FilterCleaningTask` in OntoReg ontology are compared. Mapping is the process of determining whether a regulation-statement is relevant to a process. Three types of similarity

scores are computed between the regulation-statement and the process: **topic-score, core-score** and **aux-score**. An algorithm to compute the aggregate similarity score from the three similarity scores is applied, which has provided the similarity score between the regulation-statement and the process. Since a regulation-paragraph has one or many regulation-statements, the highest similarity score among the regulation-statements is selected as the similarity score between the regulation-paragraph, `Eudralex_5.22` and the process, `FilterCleaningTask`.

The examples presented in this chapter have demonstrated how the RegCMantic framework is likely to be implemented in a real scenario and in a specific domain. Although the case study is carried out in the Pharmaceutical industry, the framework is generic and is likely to be applied to other domains. The overall results of the framework, analysis and evaluations of the results are discussed in Chapter 5.

5 Results and Evaluations

5.1 Introduction

This chapter describes the results and evaluations of the RegCMantic framework. The description of the framework is provided in Chapter 3, and its validation in a case study is described in Chapter 4. The case study chosen for the framework is the regulatory guidelines governing the aspirin production process in the Pharmaceutical industry. This chapter describes and analyses the results of extracting regulatory entities and mapping regulatory guidelines with organisational processes. The results of the extraction are analysed to determine how accurately it identifies the regulatory entities embedded in the regulatory text. Likewise, the results of the mapping are examined to find out how efficiently the proposed framework related regulatory guidelines with organisational processes.

The mapping between regulatory guidelines and organisational processes is necessary when new regulations are introduced, or there are changes or updates on the existing regulatory guidelines. This framework aims to relate the guidelines and the processes automatically. Therefore, it is important to evaluate the accuracy and completeness of the mappings generated by the framework. Likewise, in order to relate the guidelines with the processes, the important entities in the guidelines have to be identified accurately. In other words, the accuracy and completeness of the mapping part of the framework is directly affected by the accuracy and completeness of the extraction part of the framework. Therefore, both parts of the framework: extraction and mapping are considered for evaluation.

This chapter describes the evaluation criteria, evaluations for extraction and mapping, main factors of the two evaluations, relevancy of the evaluation techniques with similar works and reasons for three types of annotations and mappings.

The evaluation is divided into two parts: evaluation of the automatic extraction of guidelines, which is discussed in Section 5.2; and the evaluation of the mapping between regulatory guidelines and organisational process, which is described in Section 5.3. An estimated time saved by using the RegCMantic framework is analysed in Section 5.4.

5.2 Extraction of Regulatory Entities

5.2.1 Evaluation Criteria

As described in Chapter 4, the Pharmaceutical industry in the EU is selected as a case study to evaluate the proposed framework (Eudralex 2013). For the process domain, a process-ontology for the Pharmaceutical processes, OntoReg is used (see Section 4.4). For the regulatory domain, a regulatory-ontology, SemReg is created by extending LKIF-Core ontology (See Section 3.3.4). The concepts of SemReg are populated with the extracted regulatory-entities. In particular, the first 50 regulation-paragraphs in the Eduralex-5 document are chosen for the evaluation of the extraction part of the framework.

Since there is no annotation benchmark to compare with system-generated annotations, the annotations are compared with annotations created by experts in the area. The comparison is carried out in order to determine the number of annotations that are correctly recognized; those incorrectly identified, and those that are missed. It is important to identify the correctness of the annotations since the more correct the annotations, the more accurate the mapping between regulatory guidelines and organisational processes. Similarly, identification of incorrect and missing annotations helps in the evaluation of completeness of the extraction part of the framework.

Precision, recall and f-measure are regarded as standard techniques to evaluate the result in IR and IE systems. Therefore, they are considered for the evaluation of this framework. Precision computes the accuracy of the result-set. In order to compute precision and recall, it is needed to collect the correct annotations or true positive (TP), the incorrect annotations or false

positive (FP) and missing annotations or false negative (FN). The precision, recall and f-measure are computed (Baeza-Yates and Ribeiro-Neto 1999) as:

$$precision = \frac{Correct\ Annotations\ (TP)}{Correct\ Annotations\ (TP) + Incorrect\ Annotations\ (FP)}$$

$$recall = \frac{Correct\ Annotations\ (TP)}{Correct\ Annotations\ (TP) + Missing\ Annotations\ (FN)}$$

$$f = \frac{2\ (precision\ \times\ recall)}{precision + recall}$$

During the evaluation process of the baseline framework, it was found if boundaries were set for subject and action annotations, the identification of subject and action could be improved. Then, the next objective was set to identify boundaries such as subject-chunk, obligation-chunk, action-chunk and modifier-chunks. In order to identify chunks containing specific type of regulatory entity, use of a parser was the potential solution. A parser identifies grammatical units in a sentence and their relationship. Similarly, use of definition terms provided for a regulatory document can increase the correct identification of annotations. Hence, the framework evolved to the current version incorporating the two more components. The results of the baseline framework and the extended framework have also been provided in the evaluation.

The result and evaluation of the extraction phase of the RegCMantic framework is provided in two tables. Table 5-1 provides the result, and Table 5-2 presents the evaluation of the result. In Table 5-1, the first columns contain the types of the annotations such as subject, obligation and action. The second column presents the number of annotations that are created by the user. Likewise, the third column provides total annotations of each type in the first column, which are generated by the RegCMantic framework. This column is divided into two sub columns: annotations generated by the baseline framework (BF) and that by the extended framework (EF). The BF and EF sub columns are provided in the fourth column (Correct), the fifth

column (Incorrect) and the sixth column (Missing). Similarly, in Table 5-2, precision, recall and f-measure computed for each annotation type by the baseline framework (BF) and extended framework (EF) are provided. This can be further clarified by explaining the first row.

Table 5-1. Accuracy of different types of annotations

Number of Annotations	Total Manual	Total System		Correct (TP)		Incorrect (FP)		Missing (FN)	
Annotation Types		BF	EF	BF	EF	BF	EF	BF	EF
Subject	51	45	46	40	44	5	2	11	7
Obligation	52	52	52	52	52	0	0	0	0
Action	94	97	97	85	93	12	4	9	1
Object	7	2	6	2	6	0	0	5	1
Modifier	41	19	25	11	22	8	3	30	19
Condition	9	4	6	2	6	2	0	7	3
Total	254	219	232	192	223	27	9	62	31

The first row in Table 5-1 shows that the manual annotation has identified 51 subjects. The baseline framework has identified 45 subjects; among them, 40 subject annotations are correct and 5 subject annotations are incorrect. Furthermore, the baseline framework has missed 11 subject annotations. The extended framework has annotated 46 subjects; among them, 44 are correct subjects and 2 incorrect. It has missed 7 subject annotations.

Analysis of result of the baseline and extended frameworks is presented in Table 5-2. The precision of the baseline framework and extended framework are determined as 0.89 and 0.96 respectively. Similarly, the recall of the baseline framework and extended framework is found 0.78 and 0.86 respectively. The f-measure of the baseline and extended framework is computed as 0.83 and 0.91 respectively. This means that the extended framework has performed better than the baseline framework in terms of identification of subjects.

The first three rows in these tables present information about subject, obligation and action, which are described as the core-entities in this framework. The core-entities play a more important role as compared to other or auxiliary-entities (see Chapter 3). The both frameworks have identified all 52 obligations. This is because the framework has created an exhaustive list of obligatory words such as "should be", "must" and "can be". About action, the extended framework has shown a good f-measure, 0.97. Identification of an object, a modifier and a condition has not performed as well as that of the core-entities because the RegCMantic framework focuses on identification of the core-entities. A comprehensive algorithm to identify the auxiliary-entities remains recommended for the future-work of this research (see Chapter 6).

A comparison between the extended framework (EF) and the baseline framework (BF) presented that the current version has outperformed the initial version. Although there is no change on identification of obligations, there is improvement in identification of other core-entities: subject and action. On the extraction of auxiliary entities such as object, modifier and condition, it has shown better improvement in the extended framework.

Table 5-2. Evaluation of different types of annotations

Evaluation Measures	**Precision**		**Recall**		**F-Measure**	
Annotation Types	BF	EF	BF	EF	BF	EF
Subject	0.89	0.96	0.78	0.86	0.83	0.91
Obligation	1.00	1.00	1.00	1.00	1.00	1.00
Action	0.88	0.96	0.90	0.99	0.89	0.97
Object	1.00	1.00	0.29	0.86	0.44	0.92
Modifier	0.58	0.88	0.27	0.54	0.37	0.67
Condition	0.50	1.00	0.22	0.67	0.31	0.80

5.2.2 Comparing the Extraction Result with Other Frameworks

There are regulatory entity extraction frameworks and generic entity extraction frameworks and both these techniques are related to the current research. The evaluation techniques used in the extraction result are precision, recall and f-measure. Some of the popular and relevant entity extraction frameworks, their dataset, and results in terms of precision, recall and f-measure are presented in Table 5-3. Among them, the frameworks which extract regulatory entities are Gaius T framework (Kiyavitskaya et al 2009), Mu *et al*. (2009), Gao *et al*. (2011) and Cleland-Huang *et al*. (2006). Details of these frameworks are provided in Chapter 2.

In the frameworks mentioned in Table 5-3, the quality of the annotations were evaluated by comparing the results with manual annotations. The documents used for the annotations were found similar to the case study of this research as they were also provided in textual format. In the application of the Gaius T framework (Kiyavitskaya et al 2009), US Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule and the Italian accessibility law, the Stanca Act were used. The documents were in a structured text format called **legalese**. Legalese is a guideline to write legal documents in the U.S.A. For the evaluation, three types of entities: (i) rights, (ii) obligation and (iii) constraints are annotated. The f-measure was found as 90.7. In Mu et al. (2009) software-requirement specification documents in text format were processed to extract various entities such as agentive, action, objective and constraints. There were 249 sentences with 5,669 words. The overall f-measure was computed as 72.6. Similarly, in Gao et al. (2011), exceptions were extracted in 2,647 contracts from Onecle repository. There were seven types of contracts namely licensing, consulting, outsourcing, supply, manufacturing, purchase and stock options. The f-measure was found as 90.0. In Cleland-Huang et al. (2006), 15 requirement specifications developed by MS students at DePaul University were considered for the evaluation. In particular, the system had to identify non-functional requirements (NFR) from the specifications, which contained 326 NFRs. Its recall was found better than its precision.

The web document extraction frameworks such as Armadillo (Dingli et al 2003), KIM (Popov et al 2003), Ont-O-Mat:Pankow (Cimiano et al 2004) and SemTag (Dill et al 2003) are also related to the current research since this research also requires to convert the documents in HTML format before processing. Comparison of the RegCMantic framework with these frameworks has also shown that the result is very close. In particular, application of the proposed framework to extract the core entity has performed better than the application of the framework for all the regulatory entities. It is because this framework has focused on the extraction of the core entities.

Table 5-3. Comparison of extraction result with existing frameworks

Framework	Precision	Recall	F-Measure	Data Sets
Armadillo	91.0	74.0	81.6	Web sites of Computer Science Department
KIM	86.0	82.0	84.0	Web document
Ont-O-Mat: Pankow	65.0	28.2	39.3	Web document
SemTag	82.0	n/a	n/a	Web document
Gaius T Framework	90.6	90.8	90.7	Academic papers, legislations
Mu et al	66.2	80.3	72.6	Paragraphs, text
Gao et al	90.0	90.0	90.0	Text
Cleland-Huang et al	29.9	59.9	39.9	Text, 15 requirement specification developed by MS students
RegCMantic (all-entities)	*94.2*	*82.0*	*87.7*	*PDF, Eudralex legislation*
RegCMantic (core-entities)	*97.0*	*95.0*	*96.0*	*PDF, Eudralex legislation*

5.3 Mapping Regulations with Organisational Processes

5.3.1 Evaluation Criteria

As mentioned in Chapter 3, this framework requires two ontologies; one for the regulatory domain called SemReg and the other for the process domain called OntoReg. Since there is no standard for comparing the system-generated mappings, the author has decided to compare the mapping with the manual mappings. The manual mappings were created by expertise in the Engineering Science Department of the University of Oxford.

Among the evaluation techniques for the similarity results are Pearson correlation, Spearman's rank correlation, precision and recall. In order to compute Pearson correlation index (Pearson 1904), experts have to score each pair of entities with a value between 1 and 100. Then, the similarity scores generated by the application of a framework and that by human judgement are compared to compute the correlation index. In Spearman's rank correlation (Spearman 1904), two ranks are created from the similarity scores generated by the system and that by human experts. These two ranks are compared in order to compute the rank correlation index. These techniques are particularly important when a search engine has to rank pages in IR. In order to adopt this technique, all the possible mappings between the regulatory guidelines and organisational processes should be scored by the domain experts. The experts, working on the process ontology in the University of Oxford, are consulted for viability of this option, and it is found that time and resources were not sufficient to implement it. In addition, the rankings are very important for search engines; however, in this research it is important to evaluate the accuracy and completeness of the result.

Although there are some frameworks, which have various similarity components similar to the current work, the author has not found a framework that investigates the same issue. In sentence to sentence alignment algorithms (Barzilay and Elhadad 2003, McCarthy et al 2012, Mohler et al 2011), similarity between the description of two sentences was presented, which is similar to the topic-similarity and aux-similarity of this research. However, the core-

similarity in the current research also uses ontological structures, which was not provided in these frameworks. Therefore, they cannot be compared with the current framework. Similarly, the ontology based similarity algorithms (Chen et al 2010, Ge and Qiu 2008, Hawalah and Fasli 2011, Yu and Zhou 2009) presented the similarity between the concepts of an ontology or two ontologies with the similar meaning. In this research, the ontology mapping has not been applied; instead, relation between a regulatory guideline and an organisational process is determined. The regulation and process are not the similar concept; instead, they are two different concepts. These frameworks did not provide solutions to relate two different concepts. Therefore, these frameworks have also been found as unsuitable for comparison.

The mapping accuracy of the extended framework is only compared with that of the baseline framework and manual processes. The extended framework is evolved from the baseline framework, where the mapping between a regulatory guideline and an organisational process was determined solely by core-score. In the extended framework, it is extended with two other types of similarity scores: topic-score and aux-score. The result of the extended framework is presented as a comparison with the baseline framework. Figure 5-2 shows the accuracy and completeness of the proposed framework.

5.3.2 Evaluating the Mapping Result

The OntoReg ontology is developed by experts (Sesen et al 2010) in the domain, which contains a set of mapping between Eudralex regulations and validation-tasks. In particular, each validation-task is associated with one or more regulations, and each regulation is related with one or more validation-tasks, called **manual mapping**. A subset of manual mapping collected from the OntoReg is depicted in Table 5-4, where mapping id, mid_133 indicates that there is a mapping between the regulation `Eudralex_5.22` and the validation-task, `FilterCleaningTask`. This list is created by using the values of the object-property, `isRegulationOf` of individuals under the concept, `Regulation`.

The determination of the mapping between a regulation and a validation-task with the application of the proposed framework is referred to as **computed mapping**. A subset of computed mappings collected from the result of the similarity computation is shown in Table 5-5. The mapping id, em_1 indicates that there is a mapping between the regulation, `Eudralex_5.22` and the validation-task, `FilterCleaningTask`. The selection of the mappings also needs to define the minimum threshold, t to accept the mappings. In this case, the value of t is set to 0.85 because from the repeated observations, the value of t as 0.85 is found producing the best value for the f-measure. The mappings with the score 0.85 or above are selected as the accepted mapping.

Table 5-4. An excerpt of the manual mapping

Mapping ID	Reg ID	Task ID
em_1	Eudralex_5.22	FilterCleaningTask
em_2	Eudralex_5.22	T102CleaningTask
em_3	Eudralex_5.22	T101CleaningTask
em_4	Eudralex_5.22	T101CleanlinessTestTask
em_5	Eudralex_5.22	T102CleanlinessTestTask
em_6	Eudralex_5.22	FilterCleanlinessTestTask
em_7	Eudralex_8.14	ReactionYieldTestTask_1
em_8	Eudralex_8.14	InvestigationTask_1
em_9	Eudralex_5.21	FilterCleaningTask
em_10	Eudralex_5.21	FilterCleanlinessTestTask
em_11	Eudralex_5.21	T101CleanlinessTestTask
em_12	Eudralex_5.21	T102CleanlinessTestTask
em_13	Eudralex_5.21	T101CleaningTask
em_14	Eudralex_5.21	T102CleaningTask
em_15	Eudralex_5.31	StartingMaterialTestTask_7
em_16	Eudralex_5.26	PharmaSupplierAssess_1
em_17	Eudralex_5.26	StartingMaterialPurchase_1

Table 5-5. An excerpt of mapping created by the application of the RegCMantic framework

Mapping ID	Reg ID	Task ID	Similarity Score
mid_1	Eudralex_5.21	FilterCleaningTask	1.00
mid_4	Eudralex_5.21	FilterCleanlinessTestTask	1.00
mid_133	Eudralex_5.22	FilterCleaningTask	1.00
mid_134	Eudralex_5.22	T101CleaningTask	1.00
mid_135	Eudralex_5.22	T102CleaningTask	1.00
mid_136	Eudralex_5.22	FilterCleanlinessTestTask	1.00
mid_153	Eudralex_5.26	StartingMaterialPurchase_1	1.00
mid_2	Eudralex_5.21	T101CleaningTask	0.98
mid_3	Eudralex_5.21	T102CleaningTask	0.98
mid_5	Eudralex_5.21	T101CleanlinessTestTask	0.98
mid_6	Eudralex_5.21	T102CleanlinessTestTask	0.98
mid_137	Eudralex_5.22	T101CleanlinessTestTask	0.95
mid_138	Eudralex_5.22	T102CleanlinessTestTask	0.95
mid_154	Eudralex_5.26	StartingMaterialTestTask_7	0.90
mid_165	Eudralex_5.31	StartingMaterialTestTask_7	0.90
mid_172	Eudralex_8.14	ReactionYieldTestTask_1	0.87
mid_164	Eudralex_5.31	StartingMaterialPurchase_1	0.86

Among the mappings generated by the system, most of them are found as correct mapping. Table 5-6 and Figure 5-1 show the number of correct, incorrect and missing mappings. Similarly, Table 5-7 and Figure 5-2 show the precision, recall and f-measure of the mapping result. From the comparison between the baseline framework and the extended framework, it can be seen that the extended framework has performed better than the baseline framework in terms of its precision, recall and f-measure. The reasons behind the improvement are the consideration of topic-score, which is the similarity of the topic of a regulation-statement with a validation-task. Although, aux-score has not contributed significantly in the current result, it has produced some similarity scores, which are higher than the other two and are found meaningful. Hence, the consideration of topic-score and aux-score can be considered as improving factors in computing the similarity between a regulation-paragraph and a validation-task.

Table 5-6. Comparison of correct, incorrect and missing mappings

Number of Mappings / Mapping Frameworks	Total Existing	Total System	Correct (TP)	Incorrect (FP)	Missing (FN)
Baseline Framework (BF)	192	188	152	36	40
Extended Framework (EF)	192	177	171	6	21

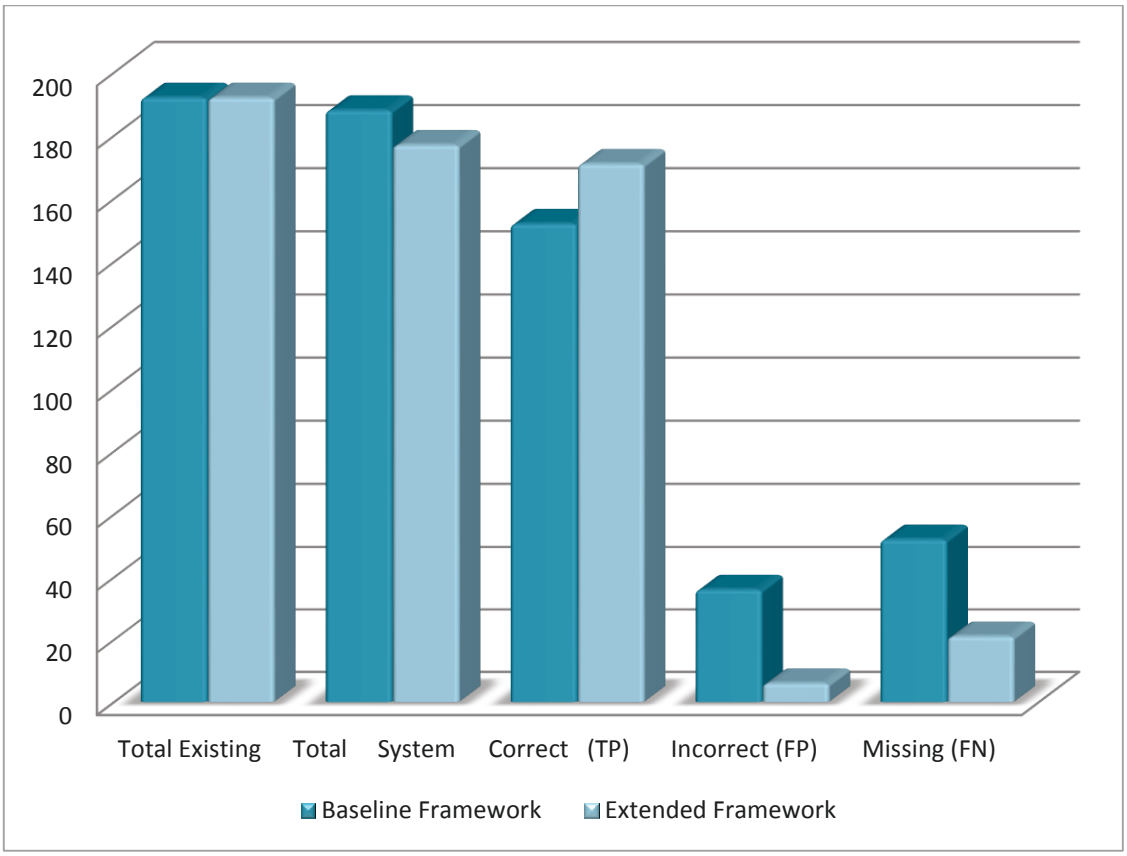


Figure 5-1 Graphical representation of correct, incorrect and missing mappings

Table 5-7. Comparison and evaluation of mapping result

Evaluation Measures / Mapping Frameworks	Precision	Recall	F-Measure
Baseline Framework (BF)	0.81	0.75	0.78
Extended Framework (EF)	0.97	0.89	0.93

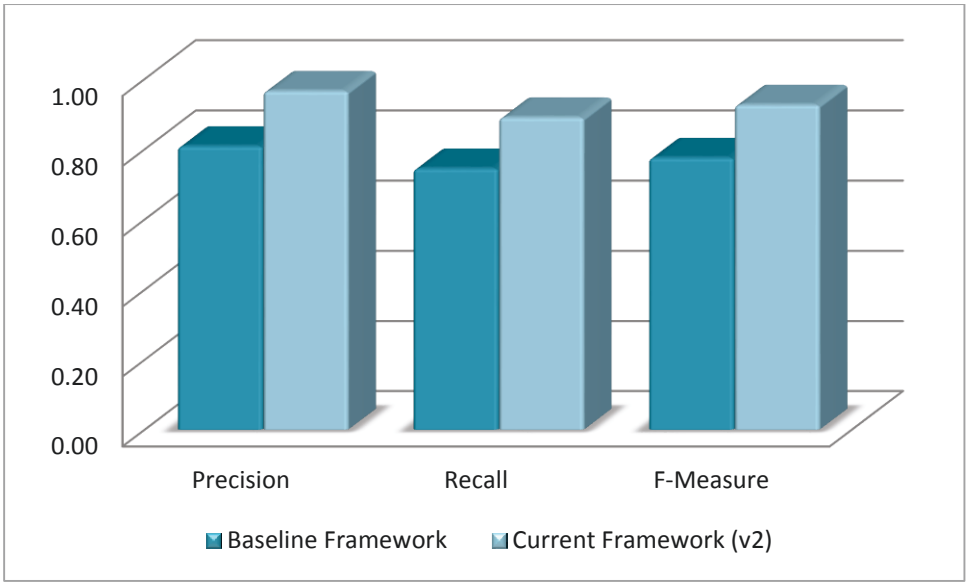
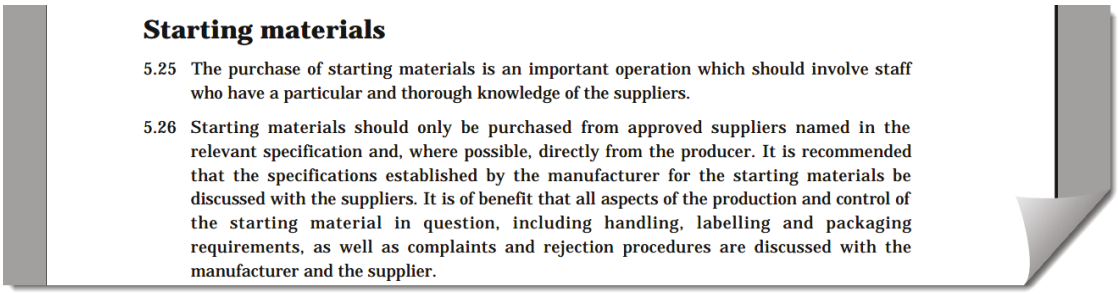


Figure 5-2. Comparison evaluation of mapping result

5.3.3 Analysis of Incorrect and Missing Mappings

Although the mapping results are encouraging, there are some incorrect and missing mappings. Some of the incorrect and missing mappings are analysed in order to determine why they were not correctly represented.



Starting materials

5.25 The purchase of starting materials is an important operation which should involve staff who have a particular and thorough knowledge of the suppliers.

5.26 Starting materials should only be purchased from approved suppliers named in the relevant specification and, where possible, directly from the producer. It is recommended that the specifications established by the manufacturer for the starting materials be discussed with the suppliers. It is of benefit that all aspects of the production and control of the starting material in question, including handling, labelling and packaging requirements, as well as complaints and rejection procedures are discussed with the manufacturer and the supplier.

Figure 5-3. Regulatory guidelines in Eudralex 5.26

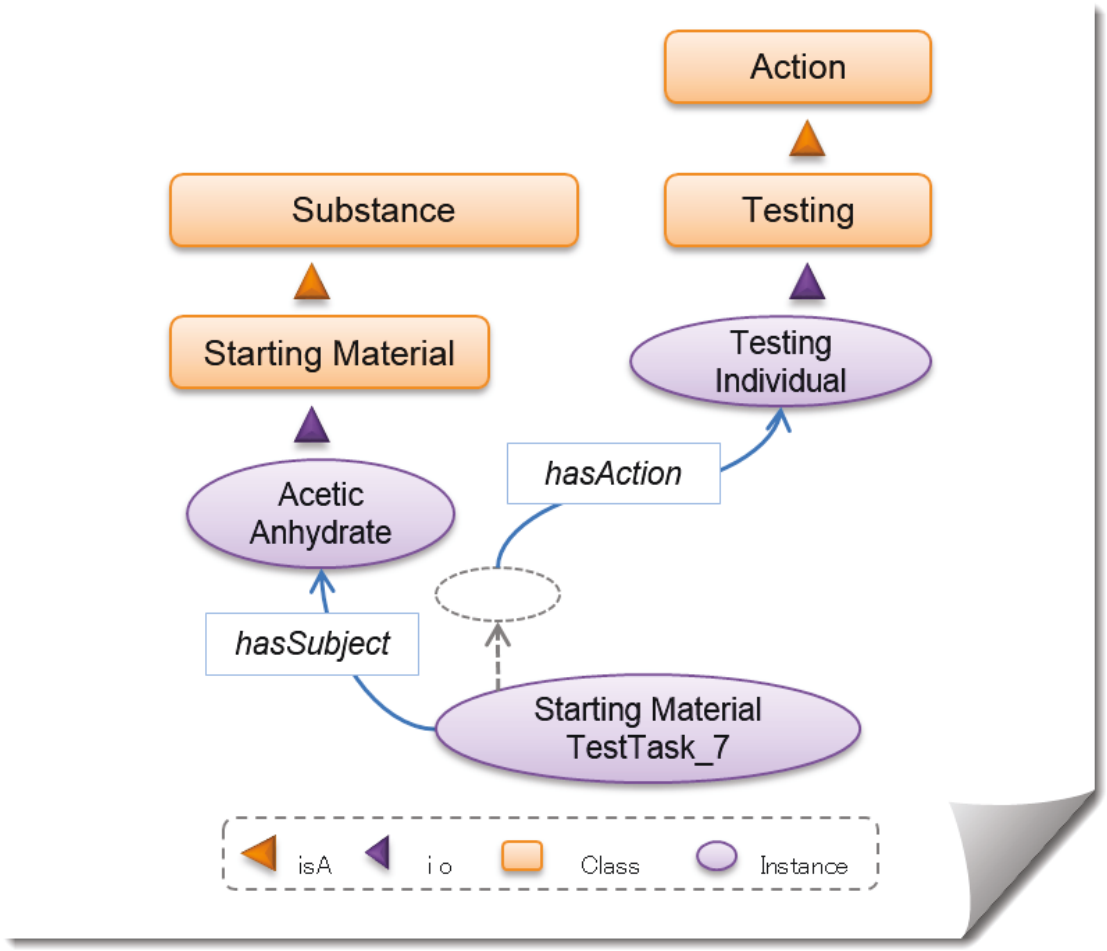


Figure 5-4. Ontological representation of StartingMaterialTestTask_7

Among the incorrectly identified mappings, the selected mapping for analysis is the mapping between a regulation-paragraph, `Eudralex_5.26` (see Figure 5-3) and a validation-task, `StartingMaterialTestTask_7` (see Figure 5-4). From the analysis of the three scores computation (see Figure 5-5), it is found that the incorrect mapping was determined by the topic-score. The topic of the regulation-paragraph contains a phrase, `Starting Material`, which is also present in the validation-task, `StartingMaterialTestTask_7`. Therefore, the topic-score was higher, and the mapping was present in the top accepted mapping sets. However, the regulation-paragraph states "`Starting materials should only be purchased from approved suppliers ...`" and the validation-task means, "starting material assessment". In this example, the author also considered that the mapping was correct since the regulatory guidelines and the organisational processes both are concerned about the

starting material. However, in the mappings experts' consideration, they should not be mapped since they are two different entities: supplier assessment and starting material assessment. The presence of incorrect mapping owes to its inability to understand the whole sentence and to interpret its abstract meaning, which is a difficult task and is not the scope of this research.

```
<mapping mapping_id="mid_154">
    <reg_id>Eudralex_5.26</reg_id>
    <stmt_id>Eudralex_5.26_1</stmt_id>
    <task_id>StartingMaterialTestTask_7</task_id>
    <topic_score>0.89652</topic_score>
    <core_score>0.5</core_score>
    <aux_score>0.75946</aux_score>
    <final_score>0.89652</final_score>
</mapping>
```

Figure 5-5. Incorrect mapping between Eudralex_5.26 and StartingMaterialTestTask_7

The mapping between a regulation-paragraph, `Eudralex_5.26` and a validation-task, `PharmaSupplierAssess_1` is one of the missing mappings in the result. From the analysis of three-scores generated between the regulation-paragraph and the validation-task (see Figure 5-6), it is found that the highest score among them was the core-score, and it was far below the acceptable level. The process engineers created a mapping between them since the regulation-paragraph implicitly means assessment of suppliers, which is also the intension of creating the validation-task, `PharmaSupplierAsses_1`. In this case, the implicitly represented regulation-paragraph is found as the system's inability to recognise the mapping between them.

```
<mapping mapping_id="mid_152">
    <reg_id>Eudralex_5.26</reg_id>
    <stmt_id>Eudralex_5.26_1</stmt_id>
    <task_id>PharmaSupplierAssess_1</task_id>
    <topic_score>0.42429</topic_score>
    <core_score>0.65776</core_score>
    <aux_score>0.33764</aux_score>
    <final_score>0.65776</final_score>
</mapping>
```

Figure 5-6. Missing mapping between Eudralex_5.26 and PharmaSupplierAssess_1

The analysis of the above two examples has showed that the incorrect and missing mappings are mainly because of the system's inability to process the statements, which express their

semantics implicitly. The processing the semantics of implicitly stated statements still remains as an open issue (Sarawagi 2007) and is out of the scope of the current research. Furthermore, automatic mapping process cannot be 100% accurate and requires users to validate the mappings.

5.3.4 Comparing the Mapping Result with Other Frameworks

Although the author has not come across a framework that has the same purpose as the RegCMantic framework has, the components of some frameworks are similar to that of the RegCMantic framework. The frameworks having the similar components are described and compared with the RegCMantic framework below.

There are many similarity algorithms dedicated to different purposes (Barzilay and Elhadad 2003, Budanitsky and Hirst 2006, Cilibrasi and Vitányi 2007, Hawalah and Fasli 2011, Yu and Zhou 2009). The author has not come across a framework whose purpose exactly matches the purpose of the RegCMantic framework. Furthermore, the evaluation techniques used in the existing similarity approaches also are found different (Chen et al 2010, Gabrilovich and Markovitch 2006, Ge and Qiu 2008, McCarthy et al 2012, Mohler et al 2011, Pirró 2009, Ponzetto and Strube 2007). Since the result the mapping is evaluated using precision, recall and f-measures, this thesis compares the results with the similarity frameworks that uses these evaluation techniques

Some similarity frameworks, their domain of interest, evaluation methods and results are presented in Table 5-8. The description of these frameworks is provided in Chapter 2. In Budanitsky & Hirst (2006), the results of some of the WordNet based similarity algorithms are evaluated using precision, recall and f-measure. For the evaluations, 500 articles from Wall Street Journal corpus were selected, and some words were replaced by malapropism words, which were identified by the evaluating algorithms. In some works (Barzilay and Elhadad 2003, Yu and Zhou 2009), the precision was compared with recall and plotted a graph to show the results. In Barzilay & Elhadad (2003), 92 pairs of sentences were chosen from

Encyclopaedia Britannica and Britannica Elementary about children. The results of the similarity between the pair of sentences were evaluated by plotting a graph using precision against recall. Similar techniques were used in Yu & Zhou (2009), where 500 movie metadata were selected in order to identify TV content similarity from the program description.

Similar to the precision, some authors (Cilibrasi and Vitányi 2007, Hawalah and Fasli 2011) have computed the accuracy as correct result divided by the total number of result. In Hawalah & Fasli (2011), Miller and Charles benchmark was selected, which contained 30 pairs of nouns extracted from WordNet. The results of the WordNet based similarity between the pairs is evaluated using the accuracy. The similar approach was presented in Cilibrasi & Vitányi (2007) while evaluating Google similarity algorithm against manual and WordNet based similarity algorithms. For the training, 50-labelled samples were selected, and the experiment was carried out in 100 samples.

Table 5-8. Comparison of mapping result with other frameworks

Framework	Similarity About	Evaluation Method
Budanitsky and Hirst	Evaluating WordNet based measures	Precision, Recall, *f*-Measure
Barzilay et al	Sentence Alignment	Precision vs. Recall
Yu et al	TV content similarity	Precision vs. Recall
Hawalah and Fasli	Semantic Relatedness	Accuracy
Cilibracy and Vitanyi	Google Similarity Distance	Accuracy
RegCMantic	*Regulatory guidelines and organisational processes*	*Precision, Recall f-Measure*

The similarity results (Chen et al 2010, Gabrilovich and Markovitch 2006, Ge and Qiu 2008, McCarthy et al 2012, Mohler et al 2011, Pirró 2009, Ponzetto and Strube 2007) were also evaluated using either Pearson (1904) correlations or Spearman (1905) rank correlation. In order to compute Pearson correlation index, experts have to score each pair of entities with a value between 1 and 100. Then, the similarity scores generated by the application of a

framework and that by human judgement are compared to compute the correlation index. In Spearman's rank correlation, two ranks are created from the similarity scores generated by the system and that by human experts. These two ranks are compared in order to compute the rank correlation index. These techniques are important particularly when a search engine has to rank pages in IR. Likewise, comparison of a framework with its own previous result (Chen et al 2010) or application in different domains (Ge and Qiu 2008) has also been considered for the evaluation of the frameworks. Some of the similarity frameworks, their domain of interest and evaluation methods are presented in Table 5-9. Description of these frameworks is provided in Chapter 2.

In Mohler et al. (2011), 80 questions and their correct answers were created by graduate tutors; then, 31 undergraduate students were asked to answer these questions. The answers were compared with the correct answers and scored between 0 and 5. The scores generated by the framework and that assigned by the tutors were considered to compute the Pearson correlation and Root Mean Square Error (RMSE).

In WikiRelate (Strube and Ponzetto 2006), 30 pairs of nouns, 65 pairs of word synonyms, 353 pairs words from three different types of datasets: M&C (Miller and Charles 1991), R&G (Rubenstein and Goodenough 1965) and 353-T (Finkelstein et al 2002) were considered for Pearson correlation. In Pirro (2009), 65 word pairs from R&G dataset were selected, and intrinsic information content was used. The evaluation was carried out by computing Pearson correlation with human judgement.

In McCarthy et al. (2012), sentence similarities were carried out in various datasets using align heuristics, average and wordism, and was evaluated by using Spearman's correlation. In Ge & Qiu (2008), concept similarity was computed based on semantic distance. The selected concepts were **person, creator, author, illustrator** and **writer**. The scores of the framework were compared with similarity scores produced by synonymy similarity and gloss-overlap similarity and evaluated using Spearman's correlation. In Chen et al. (2010), similarity of a

new object with an existing object was computed by considering two ontologies: (i) object ontology and (ii) danger ontology. A new object **nail** is compared with the existing objects such as **button, nipper, needle, knife** and **scoop**. The result was evaluated by comparing with previous research.

Table 5-9. Evaluation methods used in different similarity computation approaches

Framework	Similarity About	Evaluation Method
Mohler et al	Grading answers	Pearson Correlation RMSE, Median RMSE
Strube and Ponzetto	Wiki Relate	Pearson Correlation
Gabrilovich and Markovitch	Semantic Relatedness in Wikipedia	Pearson Correlation
Pirro	Intrinsic Information Content	Pearson Correlation
McCarthy and Gella	Text Similarity	Spearman's Correlation Align Heuristics, Average, Wordism
Ge and Qiu	Concept Similarity	Comparison of Similarity Score Among 5 terms
Chen et al	Similarity of Danger Objects	Previous vs. Current Research

5.4 Approximation of Time Saved Using the RegCMantic Framework

This section analyses the approximate amount of time saved by using this framework. These figures are only approximation and the author recommends the actual analysis of time saved for future work.

This can be described by analysing the time saved in each component of the RegCMantic framework. The key components developed in the RegCMantic framework are:

(i) Regulatory guidelines identification (Generic)

(ii) Parser (Generic)

(iii) Concepts (Domain Specific)

(iv) Term (Domain Specific)

(v) Rules (Generic and Domain Specific)

(vi) Three scores computation (Generic)

(vii) Mapping regulatory guidelines with organisational processes (Generic)

Apart from the concepts, terms and rules all the other components are generic. Rules are partly generic and partly domain specific.

Consider a RCM has come across around 200 regulatory guidelines in around 5 pages regulatory document. There are more than 500 processes in the organisation. Among them, 50 regulatory guidelines affect 50 organisational process. Therefore, the objective is to find around the 200 mappings. The amount of saved time using the RegCMantic framework can be analysed into the following two parts.

5.4.1 In the Same Domain

Consider the RegCMantic framework is already in place in the RCM.

Manual Mapping: A compliance manager has to compare each regulatory guideline with organisational process. Therefore, she has to go through 200 X 500 = 10000 comparisons carefully. If we consider that each comparison takes minimum minute, the whole process will take minimum 10000 minutes or minimum 167 hours.

RegCMantic Mapping: In order to deal with the new regulatory guidelines the compliance manager has to follow the following steps:

(i) Converts the regulatory guidelines into HTML format (less than 5 minute)

(ii) verifies document structure (less than 15 minutes)

(iii) populates ontology (less than 10 minute) and

(iv) Verification of ontology (around 30 minutes)

(v) Verifies the suggested mappings (around 200 minutes).

In total, it takes around 230 minutes.

Saved Time: In this case, by using the RegCMantic framework, the compliance manger has to work around forty times (1000/230 = 43.47) less time than the time for her manual mapping.

5.4.2 In a Different Domain

Consider this is the first time the RCM is implementing the RegCMantic framework. The manual time remains the same (i.e, 10000 minutes).

RegCMantic Mapping: In order to deal with the new regulatory guidelines, in addition to the "RegCMantic Mappings" steps described in Section 5.4.1, the compliance manager has to follow the following steps:

(i) collect terms: around 20 terms at the beginning of the regulatory document (less than 5 minute)

(ii) collect concepts: the RegCmantic framework has a module to collect ontological concepts automatically (less than 5 minutes)

(iii) adjust rules (around 8 hours = 480 minutes)

In total, it takes around 230 + 5 + 5 + 480 = 720 minutes

Saved Time: In this case, by using the RegCMantic framework, the compliance manger has to work around ten times (1000/900 = 13.88) less time than the time for her manual mapping.

5.5 Summary

This chapter presented the result and evaluation of the application of the RegCMantic framework in the Pharmaceutical industry. There are two separate results and evaluation sections for extraction and mapping part of the RegCMantic framework. The results of the framework are compared with the results of manual processes. In addition, the extended framework has also been compared with the baseline framework.

For the extraction, three types of annotations are analysed: correct, incorrect and missing annotations. Similarly, for the mapping evaluation, three types of mappings are examined: correct, incorrect and missing mappings. The three types of annotations and mappings are used to compute precision, recall and f-measures. With compared to the manual process, the results are found encouraging, and as compared to the baseline framework, the extended framework has performed better. Likewise, the comparison of the RegCMantic framework with other frameworks has shown an encouraging result. The summary of the improvements in the results of the RegCMantic framework is presented below.

1) **Improvement on the extraction of core-entities from regulatory guidelines:** The core-entities such as subject, obligation and action play an important role on relating regulatory guidelines with organisational processes. While extracting these entities, the RegCMantic framework has shown better result as compared to other frameworks (Breaux et al 2008, Gao et al 2011, Kiyavitskaya et al 2007, 2008, 2009, Mu et al 2009) in terms of precision, recall and f-measure.

2) **Improvement on the mapping between regulatory guidelines and organisational processes:** Relating the regulatory guidelines with organisational processes is important when a new regulatory guideline comes in effect, or there are some changes to the existing guidelines. The author has not found a framework that is similar to the proposed framework in terms of relating regulatory guidelines with organisational processes. The ontology mapping algorithms, the sentence similarity and WordNet based similarity algorithms cannot be directly applied to the mapping process. The RegCMantic framework has proposed a hybrid technique to use all these types of similarity computations. The result of the mapping process in RegCMantic framework as compared to the other mapping processes (Barzilay and Elhadad 2003, Chen et al 2010, Ge and Qiu 2008, Hawalah and Fasli 2011, McCarthy et al 2012, Mohler et al 2011, Yu and Zhou 2009)is found encouraging.

The next chapter summarises the research and findings of this thesis, concludes the findings of the research and provides future direction of the research.

6 Conclusions and Future work

6.1 Summary of the Thesis

This chapter summarises the findings of this research, highlights the contributions of the RegCMantic framework and presents the directions for future work.

The central focus of the research conducted in this thesis was to automate the compliance management process. The thesis provided a detailed review and analysis of existing models and techniques, and identified major research challenges involved in the automation of compliance management. Two processes are identified as crucial towards automation of compliance management processes. These included: (i) extraction of meaningful entities from the regulatory text and (ii) mapping regulatory guidelines with organisational processes. Taking into account the essence of these two processes in the automation of compliance management, this thesis has designed and developed a new a framework, called RegCMantic. These two processes are summarised as follows:

The first part of the framework dealt with the extraction of regulatory entities from the text in regulatory guidelines. The application of the framework extracted various regulatory entities with the help of: (i) parser, (ii) definition terms, (iii) ontological concepts and (iv) rules. The lexical parser has identified the parts of a sentence where relevant entities are embedded. Similarly, the use of definition terms and ontological concepts, their synonyms and hyponyms helped in the extraction of the entities. The entities extracted are (i) the core-entities such as subject, action and obligation, and (ii) the aux-entities such as the entities representing time, place, purpose, procedure and condition. A case study is carried out in the Pharmaceutical industry in order to validate the framework. The results showed significant improvement in the accuracy of the extraction of regulatory entities from regulatory guidelines (Sapkota et al

2012) as compared to other similar frameworks (Breaux et al 2006, Gao et al 2011, Kiyavitskaya et al 2008, Mu et al 2009).

The second part of the framework explained the mapping between a regulatory guideline and an organisational process. The application of proposed framework has used three different types of entities in the regulatory guidelines: regulation-topic, core-entities and aux-entities. Similarly, in the process domain, it has used the concepts related to subject and action. It has computed three types of similarity scores: topic-score, core-score and aux-score and determined the aggregate similarity score between the regulatory guideline and the organisational process. The mapping part of the proposed framework has also produced improved accuracy in terms of mapping between the regulatory guidelines and the organisational processes.

The development of the proposed framework was a complex process, and has encountered various challenges such as: (i) ambiguity and complexity of the regulatory text, (ii) implicit information in the description of organizational processes, and (iii) absence of a standard framework to map regulations and processes. Ontological concepts, definition terms, parser and rules are used to extract the regulatory information from the complex regulatory text. An ontology based mapping process is implemented in the RegCMantic framework.

The contributions of the proposed framework are listed in Section 6.2. The critical evaluation of the framework is provided in Section 6.3. The directions to the future work are emphasized in Section 6.4.

6.2 Contributions

The main contributions presented in this thesis are summarised in the following four main areas:

1) **Document-Component and Document Structure Identification:** The framework systematically illustrated the process of how document-component identification can

be achieved in the regulatory domain. It has exploited the unique nature of the regulatory documents, which has not been done previously. The identification of special document-component (such as regulation-paragraph) is carried out more accurately as compared to existing approaches. The framework employed special nature of the regulatory guidelines such as use of model verbs, preceded with indicators and use of passive voice.

2) **Automation in Extraction of Regulatory Information:** The novel aspects of this part were to analyse, annotate, extract and represent the regulatory entities embedded in the regulation-documents to ensure automation of the compliance management. In contrast to the existing framework to extract regulatory entities, the RegCMantic proposed a holistic approach to extract regulatory information from an unstructured document. It identified the document-component and applied four extraction components: parser, definition terms, ontological concepts and rules. The results showed that the extraction of core-entities from the regulatory guidelines in the proposed framework (Sapkota et al 2012) have outperformed other frameworks (Breaux et al 2006, Gao et al 2011, Kiyavitskaya et al 2008, Mu et al 2009) in terms of the degree of accuracy.

3) **Automation in Mapping the Regulations with Processes:** These contributions were to determine whether an organisational process was related to a regulatory guideline. Though there exist algorithms for similarity and mapping (Agirre et al 2012, Euzenat and Valtchev 2004, Kalfoglou and Schorlemmer 2003, Rissland 2006) they fall short of appropriately mapping regulatory guidelines with organisational processes (Sesen et al 2010, Zhao et al 2003). The process designed in the proposed framework addressed this issue and automatically mapped regulatory guidelines with organisational processes.

4) **Possible Generic Framework:** The proposed framework presented a holistic approach to the mapping process. It explained the identification of document-components, extraction of regulatory entities and relation of regulatory guidelines and organisational processes. The framework was applied to and validated through a Pharmaceutical industry case study, which showed meeting the research aims and objectives such as document-structure identification, regulatory entities extraction and mapping between regulatory guidelines and organisational processes. The framework is likely to be generic and is likely to be applicable to other domains with little or no efforts. For instance, the framework components such as parsers, definition terms and

rules are reusable. However, it needs to have a domain specific ontology in order to provide domain specific concepts for the extraction from regulatory guidelines.

6.3 Critical Evaluation

This section provides a critical analysis and observation of the research work. The first observation is that due to the limited time and resources, the case study is carried out in one domain of the Pharmaceutical industry. In order to validate the level of reusability of the framework, it could are evaluated in more than one domain. The resource limitation is due to the lack of the ontological representation of the processes. Second, the extraction of auxiliary entities has not performed as well as that of core-entities. This is because the more time is dedicated to defining the concepts and creating the whole framework rather than focusing on the practicalities. Furthermore, the identification of aux-entities needs more time and effort as compared to that of core-entities because the aux-entities are often provided in an implicit and complex manner. Third, there were few number of processes represented as special concepts called validation-tasks in the process ontology. Therefore, a limited number of regulatory guidelines are selected for the validation of the mapping part of the framework.

6.4 Directions for Future Work

The research carried out in this thesis has addressed some of the issues in automating the Compliance Management processes. The extraction of regulatory entities and the mapping of processes with regulations are the focus of this thesis. However, these processes can be improved with the following extensions.

1) **Document Structure Analysis (DSA):** The DSA conducted in this research has helped in identification of important document-components such as a regulation-paragraph, topic and sub-topic. Identification of such components has allowed the extraction tools to focus on the document-component of their interest. For example, extraction of core entities needs to identify regulation-paragraph and regulation-

statement. However, because of the time limitation, the evaluation of the DSA technique adopted in this framework has not carried out and can be a direction of future work.

2) **Application of Mapping in More Processes:** When the case study of this research was carried out in the Pharmaceutical industry, there were a limited number of processes represented in a special kind of concepts called, `validation-task`. Furthermore, the ontology was equipped with few manual mappings between organisational processes and regulatory guidelines. This has led the author to select a limited number of regulatory guidelines to compare with the existing mappings. Exploring the implementation of the mapping part of the framework in more processes, which are designed as ontological concepts, can be a future work.

3) **Auxiliary Entity Extraction:** The core-entities in a regulatory-statement comprise subject, obligation and action. The auxiliary entities represent the entities that make the core-entities more meaningful. The examples of the auxiliary entities are the entities that represent place, time, purpose, condition and procedure. In this research, a brief overview of aux-entities is provided. In future work, it is planned to have the aux-entities extensively explored. In particular, the future investigation needs to deal with complex situations since the aux-entities are often represented implicitly.

4) **Mapping Considering Time, Place, Procedure and Purpose:** The current framework uses subject and action from the process ontology for the similarity score computation. The subject and action in the process ontology are compared with topic, core-entity and aux-entities in regulatory guidelines in order to compute the similarity score between a regulatory guideline and an organisational process. However, other aspects of improving result that is more accurate can be considered. In particular, the similarity computation can be more meaningful by considering the concepts representing time, place, procedure and purpose of an organisational process. For

example, similarity between place and time of a process can be compared with that of a regulation. Therefore, an investigation towards this direction can be beneficial towards the automated semantic compliance management process.

5) **Implementation of the Framework in other Industries:** Implementation of the framework in other domains is another direction for future research work. For example, it can be implemented in (i) legislations: for mapping laws with cases and (ii) healthcare industry: for mapping healthcare protocols with treatment of patients.

6) **Compliance Checking:** Another direction of future work can be implementing the framework in a RCM and keep track of the compliance tasks by using the ontological information about the regulatory guidelines.

6.5 Summary

The automation in RCM processes helps in streamlining the RCM system. The two crucial processes in RCM are extracting the regulatory entities and mapping regulatory guidelines with organisational processes. These two processes make the updating RCM automatic when the regulatory guidelines change. From the observation in a case study, it can be concluded that the proposed, RegCMantic framework has a considerable amount of automation and accuracy in the extraction and mapping processes.

References

Agichtein E and Ganti V (2004) Mining reference tables for automatic text segmentation. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining KDD 04*. ACM, 20. Available at: http://portal.acm.org/citation.cfm?doid=1014052.1014058.

Agirre E, Cer D, Diab M and Gonzalez-Agirre A (2012) SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), in Conjunction with the First Joint Conference on Lexical and Computational Semantics (SEM 2012)*. Montreal,Canada: ACL Press, 385–393. Available at: http://acl.eldoc.ub.rug.nl/mirror/S/S12/S12-1051.pdf (accessed 01/09/12).

Agrawal R, Johnson C, Kiernan J and Leymann F (2006) Taming Compliance with Sarbanes-Oxley Internal Controls Using Database Technology. *Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*. Atlanta: IEEE Computer Society, 92–92. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1617460.

Anjewierden A (2001) AIDAS: Incremental Logical Structure Discovery in PDF Documents. *Proceedings of Sixth International Conference on Document Analysis and Recognition (ICDAR' 01)*. Seattle, USA: IEEE Comput. Soc, 374–378. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=953816.

Appelt D, Hobbs J, Bear J, Israel D and Tyson M (1993) FASTUS: A finite-state processor for information extraction from real-world text. *International Joint Conference on Artificial Intelligence*. Chambery, France: Citeseer, 1172–1178. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.4634&rep=rep1&type=pdf.

Appelt DE and Onyshkevych B (1998) The common pattern specification language. *TIPSTER Worskhop (TIPSTER '98)*. Baltimore, Maryland: Association for Computational Linguistics, 23–30. Available at: http://portal.acm.org/citation.cfm?doid=1119089.1119095.

Baeza-Yates R and Ribeiro-Neto B (1999) *Modern Information Retrieval* (1st edition). Harlow, UK: Addison Wesley, 544.

Baird HS, Jones SE and Fortune SJ (1990) Image segmentation by shape-directed covers. *Proceedings of International Conference on Pattern Recognition*. Atlantic City, USA: IEEE Computer Society Press, 820–825.

Barzilay R and Elhadad N (2003) Sentence Alignment for Monolingual Comparable Corpora. *Proceedings of the 2003 Conference on Empirical Methods in Natural Lnguage Processing (EMNLP'03)*. Stroudsburg, PA, USA: ACL Press, 25–32. Available at: http://dl.acm.org/citation.cfm?id=1119359 (accessed 27/06/12).

Bollegala D, Matsuo Y and Ishizuka M (2007) Measuring semantic similarity between words using web search engines. *The 16th International Conference on World Wide Web*. Banff, Alberta, Canada: ACM Press, 757–766. Available at: http://dl.acm.org/citation.cfm?id=1242675 (accessed 12/04/12).

Borst WN (1997) Construction of Engineering Ontologies for Knowledge Sharing and Reuse. *Technology*. Universiteit Twente, 243. Available at: http://doc.utwente.nl/17864/.

Breaux TD, Antón AI and Doyle J (2008) Semantic Parameterization: A Process for Modeling Domain Descriptions. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 18(2): 1–27. Available at: http://dl.acm.org/citation.cfm?id=1416565 (accessed 30/08/12).

Breaux TD, Vail MW and Antón AI (2006) Towards Regulatory Compliance: Extracting Rights and Obligations to Align Requirements with Regulations. *Proceedings of 14th IEEE International Requirements Engineering Conference (RE'06)*. Minneapolis: IEEE Computer Society, 49–58. Available at: http://portal.acm.org/citation.cfm?id=1173696.1174062 (accessed 25/02/11).

Breuker J and Hoekstra R (2004) Core concepts of law: taking common-sense seriously. *International Conference on Formal Ontology in Information Systems (FOIS '04)*. Torino, Italy: IOS-Press, 210–221. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.76.3075 (accessed 28/02/11).

Budanitsky A and Hirst G (2006) Evaluating wordnet-based measures of lexical semantic relatedness. *Journal of Computational Linguistics* 32(1): 13–47. Available at: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2006.32.1.13 (accessed 30/11/12).

Califf ME and Mooney RJ (2004) Bottom-Up Relational Learning of Pattern Matching Rules for Information Extraction. *Journal of Machine Learning Research*. JMLR. org 4(2): 177–210. Available at: http://www.crossref.org/jmlr_DOI.html.

Castillo J a. R, Silvescu A, Caragea D, Pathak J and Honavar VG (2003) Information Extraction and Integration from Heterogeneous, Distributed, Autonomous Information Sources – A Federated Ontology-Driven Query-Centric Approach. *Proceedings of the 2003 IEEE International Conference on Information Reuse and Integration (IRI '03)*. Las Vegas, NV, USA: IEEE Computer Society, 183–191. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1251412.

Chen Y, Wang J, Cheng Z, Jing L and Zhou Y (2010) An Algorithm to Compute Similarity between Danger Objects Based on Ontology for Danger-Aware Systems. *Proceedings of the 2nd International Symposium on Aware Computing (ISAC'10)*. Tainan, Taiwan: IEEE Computer Society, 128–135. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5670467 (accessed 03/09/12).

Choi N, Song I-Y and Han H (2006) A Survey on Ontology Mapping. *ACM SIGMOD Record* 35(3): 34–41.

Cilibrasi R and Vitányi P (2007) The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering*. Published by the IEEE Computer Society 19(3): 370–383. Available at: http://eprints.pascal-network.org/archive/00002784/.

Cimiano P, Handschuh S and Staab S (2004) Towards the Self-Annotating Web. *The 13th International Conference on World Wide Web (WWW '04)*. Manhattan, New York: ACM Press, 462–471.

Ciravegna F, Wilks Y and Petrelli D (2003) Adaptive information extraction for document annotation in amilcare. In: Handschuh S and Staab S (eds) *The 25th annual international ACM SIGIR conference on Research and development in information retrieval*. Tampere, Finland: ACM Press, 451–451. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.66.33&rep=rep1&type=url&i=0.

Conley J (2000) *Automated Regulatory Compliance*. *Scientific Computing & Instrumentation, http://www.scimag.com*. Available at: http://www.scimag.com (accessed 06/10/10).

Conway A (1993) Page grammars and page parsing: A syntatic approach to document layout recognition. *The Second International Conference on Document Analysis and Recognition (ICDR '93)*. Tsukuba Science City, Japan: IEEE Computer Society Press, 761–764. Available at: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=395626.

Cunningham H, Maynard D, Bontcheva K and Tablan V (2002) GATE: A framework and graphical development environment for robust NLP tools and applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania: ACL Press, 168–175. Available at: http://eprints.aktors.org/90/.

Dill S, Eiron N, Gibson D, Gruhl D, Guha R, Jhingran A, Kanungo T, Mccurley KS, Rajagopalan S, Tomkins A, Tomlin JA, Zien JY, Road H and Jose S (2003) A Case for Automated Large Scale Semantic Annotation. *Journal of Web Semantics* 1(1): 115–132.

Dingli A, Ciravegna F and Wilks Y (2003) Automatic Semantic Annotation using Unsupervised Information Extraction and Integration. *Workshop on Knowledge Markup and Semantic Annotation*. Sanibel Island, Florida, USA: ACM Press, 1–8.

Doan A, Madhavan J, Dhamankar R, Domingos P and Halevy A (2003) Learning to match ontologies on the Semantic Web. *The International Journal on Very Large Data Bases (VLDB)*. Springer 12(4): 303–319. Available at: http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s00778-003-0104-2.

Ehrig M and Staab S (2004) QOM - Quick Ontology Mapping. In: McIlraith SA, Plexousakis D and Van Harmelen F (eds) *3rd International Semantic Web Conference (ISWC2004)*. Hirosima, Japan: Springer, 1–28. Available at: http://www.springerlink.com/index/fj4075bm4231x35w.pdf.

EMEA (2012) *Mission Statement*. . Available at: http://www.emea.europa.eu/htms/aboutus/emeaoverview.htm# (accessed 03/05/12).

Esposito F, Ferilli S, Basile TMA and Mauro N Di (2008) Machine Learning for Digital Document Processing: from Layout Analysis to Metadata Extraction (90th edition). In: Marinai S and Fujisawa H (eds) *Machine Learning in Document Analysis and Recognition*. Springer Berlin Heidelberg, 105–138.

Eudralex (2013) *The Rules Governing Medicinal Products in the European Union*. *Online*. Available at: http://ec.europa.eu/health/documents/eudralex/cd/index_en.htm (accessed 18/05/13).

Euzenat J and Valtchev P (2004) Similarity-based ontology alignment in OWL-Lite. *Processing*. Ios Pr Inc 16(C): 333–337. Available at: http://www.iro.umontreal.ca/~owlola/pdf/align-ECAI04-FSub.pdf.

FDA (2012) *U.S. Food and Drug Administration: Protecting and Promoting Your Health*. *U.S. Department of Heatlh and Human Services*. Available at: http://www.fda.gov/.

Finkelstein L, Gabrilovich E, Matias Y, Rivlin E, Solan Z, Wolfman G and Ruppin E (2002) Placing Ssearch in Context: The Concept Revisited. *ACM Transactions on Information Systems* 20(1): 116–131.

Gabrilovich E and Markovitch S (2006) Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *The 20th international joint conference on Artifical intelligence (IJCAI '07)*. Hyderabad, India: Morgan Kaufmann Publishers Inc, 1606–1611. Available at: http://dl.acm.org/citation.cfm?id=1625275.1625535.

Gangemi A, Guarino N, Masolo C, Oltramari A and Schneider L (2002) Sweetening ontologies with DOLCE. *The 13th International Conference on Knowledge Engineering and Knowledge Management. (EKAW '02)*. Singuenza, Spain: Springer-Verlag, 166–181. Available at: http://www.springerlink.com/index/5p86jk323x0tjktc.pdf (accessed 27/07/11).

Gao X, Singh MP and Mehra P (2011) Mining Business Contract for Service Exceptions. *IEEE Transactions on Services Computing* 5(3): 333–344. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5708127 (accessed 30/08/12).

Ge J and Qiu Y (2008) Concept Similarity Matching Based on Semantic Distance. *Proceedings of the Fourth International Conference on Semantics, Knowledge and Grid (SKG'08)*. Beijing, China: IEEE Computer Society, 380–383. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4725943 (accessed 27/06/12).

George D (2010) Examining the Application of Modular and Contextualised Ontology in Query Expansions for Information Retrieval - A PhD Thesis. *Sciences-New York*. University of Central Lancashire.

Ghanavati S, Amyot D and Peyton L (2007) Towards a Framework for Tracking Legal Compliance in Healthcare. *Proceedings of the 19th International Conference on Advanced Information Systems Engineering (CAiSE'07)*. Trondheim, Norway: Springer-Verlag Berlin, 218–232. Available at: http://portal.acm.org/citation.cfm?id=1768051 (accessed 06/06/12).

Giunchiglia F, Shvaiko P and Yatskevich M (2004) S-Match: an algorithm and an implementation of semantic matching. In: Bussler CJ, Davies J, Fensel D and Studer R (eds) *1st European Semantic Web Symposium (ESWC '04)*. Heraklion, Greece: Springer, Heidelberg, 61–75.

Gómez-Pérez A, Fernández-López M and Corcho O (2007) *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web.* (2nd edition). Secaucus, NJ, USA: pringer-Verlag, 505. Available at: http://www.amazon.com/dp/1846283965.

Governatori G, Hoffmann J, Sadiq S and Weber I (2009) Detecting regulatory compliance for business process models through semantic annotations. *Business Process Management Workshops*. Springer 17(1): 5–17. Available at: http://www.springerlink.com/index/l1434406632903u3.pdf (accessed 01/12/10).

Gruber TR (1995) Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies* 43(5-6): 907–928. Available at: http://linkinghub.elsevier.com/retrieve/doi/10.1006/ijhc.1995.1081.

Guarino N (1998) Formal Ontology and Information Systems. *Proceedings of FOIS98*. IOS Press 46(June): 3–15. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.1776&rep=rep1&type=pdf.

Haider SI (2002) *Pharmaceutical master validation plan: the ultimate guide to FDA, GMP, and GLP compliance* (1st edition). *Florida: CRC Press*. New York: Informa Healthcare, 208. Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Pharmaceutical+Master+Validation+Plan:+The+Ultimate+Guide+to+FDA,+GMP,+and+GLP+Compliance#0 (accessed 13/12/10).

Haider SI (2006) *Validation standard operating procedures* (2nd edition). New York: Informa Healthcare, 1144. Available at: http://books.google.co.uk/books?id=uLroiosj41MC.

Handschuh S and Staab S (2002) Authoring and annotation of web pages in CREAM. *Proceedings of the eleventh international conference on World Wide Web (WWW '02)*. Honolulu, Hawaii: ACM Press, 462–473. Available at: http://portal.acm.org/citation.cfm?doid=511446.511506.

Hao D, Zuo W, Peng T and He F (2011) An Approach for Calculating Semantic Similarity between Words Using WordNet. *Proceedings of the Second International Conference on Digital Manufacturing & Automation*. Hong Kong: IEEE Computer Society Press, 177–180. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6051913 (accessed 16/12/11).

Hawalah A and Fasli M (2011) A Graph-Based Approach to Measuring Semantic Relatedness in Ontologies. *Proceedings of the International Conference on Web Intelligence, Mining and Semantics (WIMS'11)*. New York: ACM Press, 12. Available at: http://dl.acm.org/citation.cfm?id=1988722 (accessed 27/06/12).

Hoekstra R, Breuker J, Di Bello M and Boer A (2007) The LKIF Core Ontology of Basic Legal Concepts. In: Casanovas P, Biasiotti MA, Francesconi E and Sagri MT (eds) *Proceedings of the 2nd Workshop on Legal Ontologies and Artificial Intelligence Techniques(LOAIT'07)*. Stanford, California, USA: CEUR-WS.org, 43–63. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.9650&rep=rep1&type=pdf#page=43 (accessed 29/07/11).

Horrocks I (2003) Description Logic Programs: Combining Logic Programs with Description Logic. *The Twelfth International World Wide Web Conference (WWW 2003)*. Budapest, Hungary: ACM Press, 48–57. Available at: http://www.kde.cs.uni-kassel.de/conf/iccs05/horrocks_iccs05.pdf (accessed 11/04/12).

Ishitani Y (1999) Logical structure analysis of document images based on emergent computation. *Proceedings of International Conference on Document Analysis and Recognition*. Bangalore, India: IEEE Computer Society Press, 189–192.

Jayram TS, Krishnamurthy R, Raghavan S, Vaithyanathan S and Zhu H (2006) Avatar Information Extraction System. *Computational Linguistics*. Citeseer 29(1): 1–9. Available at: http://dblp.uni-trier.de/rec/bibtex/journals/debu/JayramKRVZ06.

Jiang J and Conrath D (1997) Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proceedings of the International Conference on Research in Computational Linguistics*. Taipei, Taiwan: MIT Press, 19–33. Available at: http://arxiv.org/abs/cmp-lg/9709008 (accessed 31/08/12).

Kahan J, Koivunen M, Prud'Hommeaux E and Swick R (2001) Annotea: An Open RDF Infraestructure for Shared Web Annotations. *Proceedings of the 10th international conference on World Wide Web*. Hong Kong: ACM Press, 623–632. Available at: http://portal.acm.org/ft_gateway.cfm?id=372166&type=pdf&coll=GUIDE&dl=GUIDE&CFID=92736257&CFTOKEN=71919243.

Kalfoglou Y and Schorlemmer M (2003) Ontology mapping: the state of the art. *The Knowledge Engineering Review* 18(1): 1–31. Available at: http://www.journals.cambridge.org/abstract_S0269888903000651.

El Kharbili M (2012) Business process regulatory compliance management solution frameworks: A comparative evaluation. *The Eighth Asia-Pacific Conference on Conceptual Modelling (APCCM '12)*. Melbourne, Australlia: APCCM Press, 1–10. Available at: http://marco.gforge.uni.lu/papers/APCCM12.pdf (accessed 27/12/12).

El Kharbili M, Stein S, Markovic I and Pulvermüller E (2008) Towards a framework for semantic business process compliance management. *Proceedings of the Workshop on Governance, Risk and Compliance for Information Systems (GRCIS ' 08)*. Milan, Italy: Citeseer, 1–15. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.142.9939&rep=rep1&type=pdf (accessed 30/11/10).

Kim J, Le DX and Thoma GR (2001) Automated labeling in document images. *Proceedings of SPIE Conference on Document Recognition and Retrieval VIII*. San Jose, CA: SPIE Digital Library, 111–122.

Kise K, Sato A and Iwata M (1998) Segmentation of page images using the area Voronoi diagram. *Computer Vision and Image Understanding* 70(1): 370–390.

Kiyavitskaya N, Zeni N and Breaux TD (2008) Automating the Extraction of Rights and Obligations for Regulatory Compliance. *Proceedings of 27th International Conference on Conceptual Modeling (ER'08)*. Barcelona, Spain: Springer-Verlag Berlin, 154–168. Available at: http://www.springerlink.com/index/dn7240937t72lx46.pdf (accessed 06/06/12).

Kiyavitskaya N, Zeni N, Breaux TD, Cordy JR and Mylopoulos J (2007) Extracting Rights and Obligations from Regulations : Toward a Tool-Supported Process. *Proceedings of the Twenty-Second IEEE/ACM International Conference on Automated Software Engineering (ASE '07)*. Atlanta, Georgia, USA: ACM Press, 429–423. Available at: http://doi.acm.org/10.1145/1321631.1321701.

Kiyavitskaya N, Zeni N, Cordy JR and Mich L (2009) Cerno : Light-Weight Tool Support for Semantic Annotation of Textual Documents Introduction : The Semantic Annotation Challenge. *Data & Knowledge Engineering* 68(12): 1470–1492. Available at: http://research.cs.queensu.ca/~cordy/Papers/KZCMM_DKE_Cerno.pdf.

Klink S, Dengel A and Kieninger T (1999) Document Structure Analysis Based on Layout and Textual Features. *Proc. of International Workshop on Document Analysis Systems, (DAS '00)*. Manchester: IAPR, 99–111.

Krishnamoorthy M, Nagy G, Seth S and Viswanathan M (1993) Syntactic segmentation and labeling of digitized pages from technical journals. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15(1): 737–747.

Lafferty J, McCallum A and Pereira F (2001) Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. *Computer*. Citeseer pages(2): 282–289. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.23.9849&rep=rep1&type=pdf.

Lee C (2005) Ontology-based information retrieval and extraction. *The Third International Conference on INformation Technology: Research and Education (ITRE '05)*. Hsinchu, Taiwan: IEEE Computer Society Press, 265–269. Available at: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1503119.

Lin CC, Niwa Y and Narita S (1997) Logical structure analysis of book document images using contents information. *Proceedings of International Conference on Document Analysis and Recognition (ICDAR '97)*. Ulm, Germany: IEEE Computer Society Press, 1048–1054.

Lin D (1998) An Information-Theoretic Definition of Similarity. In: Shavlik JW (ed) *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 296–304. Available at: http://webdocs.cs.ualberta.ca/~lindek/papers/sim.pdf (accessed 31/08/12).

Lin M, Tapamo J and Ndovie B (2006) A Texture-based Method for Document Segmentation and Classificatio. *South African Computer Journal* 36(1): 49–56.

Liu Y, Muller S and Xu K (2007) A Static Compliance-Checking Framework For Business Process Models. *IBM Systems Journal* 46(2): 335–361. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5386614.

Luong M-T, Nguyen TD and Kan M-Y (2010) Logical Structure Recovery in Scholarly Articles with Rich Document Features. *IInternational Journal of Digital Library Systems (IJDLS)* 1(4): 1–25. Available at: http://www.igi-global.com/article/international-journal-digital-library-systems/48200 (accessed 08/01/13).

Malouf R (2002) Markov models for language-independent named entity recognition. *proceeding of the 6th conference on Natural language learning COLING02*. Association for Computational Linguistics 1(2): 1–4. Available at: http://portal.acm.org/citation.cfm?doid=1118853.1118872.

Manning CD and Schütze H (1999) *Foundations of Statistical Natural Language Processing*. *Reading*. Cambridge: MIT Press. Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Foundations+of+Natural+Language+Processing#3.

Mao M (2008) Ontology Mapping: Towards Semantic Interoperability in Distributed and Heterogeneous Environments. University of Pittsburgh, 162.

Mao S, Rosenfeld A and Kanungo T (2003) Document Structure Analysis Algorithms: A Literature Survey. *The SPIE Electronic Imaging Conference*. Santa Clara, California, USA: SPIE Digital Library, 197–207. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1.6218&rep=rep1&type=pdf (accessed 06/06/11).

MCA (2007) *Rules and guidance for pharmaceutical manufacturers and distributors 2002* (6th edition). London: Pharmaceutical Press, 432. Available at: http://www.amazon.co.uk/dp/0853697191 (accessed 02/05/12).

McCarthy D, Gella S and Reddy S (2012) DSS: Text Similarity Using Lexical Alignments of Form, Distributional Semantics and Grammatical Relations. *Proceedings of the First Joint Conference on Lexical and Computational Semantics (SEM'12)*. Montreal, Canada: ACL Press, 557–564. Available at: http://www.aclweb.org/anthology-new/S/S12/S12-1081.pdf (accessed 27/06/12).

McCarty LT (1989) A Language for Legal Discourse. *Proceedings of the 2nd international conference on Artificial intelligence and law (ICAIL'89)*. New York, NY, USA: ACM Press, 180–189. Available at: http://logic.stanford.edu/classes/cs204/mccarty.pdf.

McGuinness DL, Fikes R, Rice J and Wilder S (2000) An Environment for Merging and Testing Large Ontologies. In: Cohn AG, Giunchiglia F and Selman B (eds) *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR '00)*. Breckenridge, Colorado, USA: Citeseer, 483–493. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.109.1812&rep=rep1&type=pdf.

MHRA (2012) *Good Manufacturing Practice*. . Available at: http://www.mhra.gov.uk/home/idcplg?IdcService=SS_GET_PAGE&nodeId=613 (accessed 03/05/12).

Miller GA and Charles WG (1991) Contextual Correlates of Semantic JSimilarity. *Language and Cognitive Processes* 6(1): 1–28.

Mitra P, Noy NF and Jaiswal AR (2005) OMEN : A Probabilistic Ontology Mapping Tool. *The 4th International Semantic Web Conference (ISWC '05)*. Galway: Springer-Verlag Berlin, 1–12. Available at: http://www.springerlink.com/index/36535wm17876447h.pdf.

Mohler MAG, Bunescu R and Mihalcea R (2011) Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT'11)*. Stroudsburg, PA, USA: ACM Press, 752–762. Available at: http://ace.cs.ohiou.edu/~razvan/papers/acl11.pdf (accessed 27/06/12).

Mu Y, Wang Y and Guo J (2009) Extracting Software Functional Requirements from Free Text Documents. *Proceedings of International Conference on Information and Multimedia Technology, 2009. ICIMT '09.* Jeju Island, Republic of Korea: IEEE Computer Society Press, 194–198. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5381217 (accessed 06/06/12).

Muhammed A (2007) *Information security: The route to compliance*. *Computer Weekly*. Available at: http://www.computerweekly.com/Articles/2007/04/23/223385/Information-security-The-route-to-compliance.htm (accessed 07/12/10).

Müller H-M, Kenny EE and Sternberg PW (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS biology* 2(11): e309. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=517822&tool=pmcentrez&rendertype=abstract (accessed 17/06/11).

Nagy G, Seth S and Viswanathan M (1992) A prototype document image analysis system for technical journals. *Computer* 25(1): 10–22.

Namboodiri A and Jain A (2007) Document structure and layout analysis. In: Chaudhuri BB (ed) *Digital Document Processing*. London: Springer London, 29–48. Available at: http://www.springerlink.com/index/l7p043wp802814g2.pdf (accessed 12/10/12).

Namiri K and Stojanovic N (2007) A formal approach for internal controls compliance in business processes. *8th Workshop on Business Process Modeling, Development, and Support (BPMDS 2007), In conjunction with CAiSE*. Trondheim: Tapir Academic Press, 1–9. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.122.4201&rep=rep1&type=pdf (accessed 30/11/10).

Noy NF (2004) Semantic Integration: A Survey Of Ontology-Based Approaches. *ACM SIGMOD Record* 33(4): 65–67. Available at: http://ufesmestradoanabringuente.googlecode.com/svn/trunk/2per/ed/fichamentos/09_10_02_Semantic Integration - A Survey Of Ontology-Based Approaches - Noy 2004.pdf (accessed 02/01/13).

Noy NF and Musen MA (2001) Anchor-PROMPT: Using Non-Local Context for Semantic Matching. *The Workshop on Ontologies and Information Sharing at the Seventeenth International Joint Conference on Artificial Intelligence IJCAI-2001*. Seattle, USA: AAAI Publication, 63–70. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.13.3504&rep=rep1&type=pdf#page=65.

Noy NF, Musen MA and Informatics SM (2000) PROMPT: Algorithm and tool for automated ontology merging and alignment. *The 17th National Conference on Artificial Intelligence (AAAI ' 00)*. Austin, Texas, USA: AAAI Press, 450–455. Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Algorithm+and+Tool+for+Automated+Ontology+Merging+and+Alignment#0.

O'Gorman L (1993) The document spectrum for page layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15(1): 1162–1173.

Paaß G and Konya I (2012) Machine Learning for Document Structure Recognition. In: Mehler A, Kühnberger K-U, Lobin H, Lüngen H, Storrer A and Witt A (eds) *Modeling, Learning, and Processing of Text Technological Data Structures*. Springer-Verlag Berlin, 221–247. Available at: http://www.springerlink.com/index/X729712016643J45.pdf (accessed 18/03/13).

Pavlidis T and Zhou J (1992) Page segmentation and classification. *CVGIP: Graphical Models and Image Processing* 54(6): 484–496. Available at: http://www.sciencedirect.com/science/article/pii/1049965292900689.

Pearson K (1904) *On the Theory of Contingency and Its Relation to Association and Normal Correlation* (1st edition). *Biometric*. London: Dulau and Co., 35.

Pedersen T, Patwardhan S and Michelizzi J (2004) WordNet:: Similarity: Measuring the Relatedness of Concepts. *Proceeding of the Demonstration Papers at HLT-NAACL 2004 (HLT-NAACL--Demonstrations '04)*. Stroudsburg, PA, USA: ACL Press, 38–41. Available at: http://dl.acm.org/citation.cfm?id=1614025.1614037 (accessed 31/08/12).

Pirró G (2009) A Semantic Similarity Metric Combining Features and Intrinsic Information Content. *Data & Knowledge Engineering*. Elsevier 68(11): 1289–1308. Available at: http://www.sciencedirect.com/science/article/pii/S0169023X09000986 (accessed 19/02/12).

Ponzetto S and Strube M (2007) Knowledge derived from Wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research* 30(1): 181–212. Available at: http://www.aaai.org/Papers/JAIR/Vol30/JAIR-3005.pdf (accessed 18/03/13).

Popov B, Kiryako A, Kirilov A, Manov D, Ognyanoff D and Goranov M (2003) KIM - Semantic Annotation Platform. *2nd International Semantic Web Conference (ISWC2003)*. Florida, USA: Sirma Group Company, 834– 849. Available at: http://www.ontotext.com/sites/default/files/publications/KIM_SAP_ISWC168.pdf.

Ratnaparkhi A (1999) Learning to Parse Natural Language with Maximum Entropy Models. *Machine Learning*. Kluwer Academic Publishers, Boston 34(1-3): 151–175. Available at: http://www.springerlink.com/index/q2u2rhlg2r118445.pdf.

Reeves S (2006) The Code Document ' s Structure and Analysis. *TeamEthno-Online* June(2): 34–51. Available at: http://eprints.nottingham.ac.uk/408/1/paper-revised.pdf.

Richardson R, Smeaton A and Murphy J (1994) *Using WordNet as a Knowledge Base forMeasuring Semantic Similarity between Words*. *Proceedings of AICS*. Dublin, 1–15. Available at: http://www.compapp.dcu.ie/wpapers/1994/1294.pdf (accessed 18/03/13).

Rissland EL (2006) AI and Similarity. *IEEE Intelligent Systems*. IEEE Computer Society Press 21(3): 39–49. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1637349.

Rubenstein H and Goodenough JB (1965) Contextual Correlates of Synonymy. *Communications of the ACM* 8(10): 627–633.

Sapkota K, Aldea A, Younas M, Duce DA and Banares-Alcantara R (2011) Semantic-ART: a framework for semantic annotation of regulatory text. *Proceedings of the Fourth Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR '11)*. Glasgow: ACM Press, 23–24.

Sapkota K, Aldea A, Younas M, Duce DA and Banares-Alcantara R (2012) Extracting Meaningful Entities from Regulatory Text. *Proceedings of the Fifth International Workshop on Requirements Engineering and Law (RELAW '12)*. Chicago: IEEE Computer Society Press, 29–32. Available at: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6347798&url=http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6347798.

Sarawagi S (2007) Information extraction. *Communications of the ACM* 1(3): 261–377. Available at: http://portal.acm.org/citation.cfm?id=234209 (accessed 04/07/11).

Sarawagi S and Agichtein E (2006) Scalable Information Extraction and Integration. *Proceedings of the 13th International Conference on Management of Data (COMAD'06)*. Delhi: Tata McGraw-Hill Publishing Company Limited, 249. Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Scalable+Information+Extraction+and+Integration#0 (accessed 03/09/12).

Sesen MB, Suresh P, Banares-Alcantara R and Venkatasubramanian V (2010) An Ontological Framework for Automated Regulatory Compliance in Pharmaceutical Manufacturing. *Computers & Chemical Engineering* 34(7): 1155–1169. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0098135409002336 (accessed 03/08/10).

Shen W, Doan A, Naughton JF and Ramakrishnan R (2007) Declarative Information Extraction Using Datalog with Embedded Extraction Predicates. In: ` (ed) *Proceedings of the 33rd international conference on Very large data bases*. Vienna, Austria: VLDB Endowment, 1033–1044. Available at: http://portal.acm.org/citation.cfm?id=1325968&dl=GUIDE,.

Shvaiko P and Euzenat J (2005) A Survey of Schema-Based Matching Approaches. *Journal on Data Semantics IV* 3730(1): 146–171.

Slimani T, Yagahlane B and Mellouli K (2006) A New Similarity Measure based on Edge Counting. *World Academy of Science, Engineering and Technology* 23(1): 34–38. Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:A+New+Similarity+Measure+based+on+Edge+Counting#0 (accessed 18/03/13).

Soderland S (1999) Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning*. Springer 34(1): 233–272. Available at: http://www.springerlink.com/index/M23N8197VG924T51.pdf.

Soler S and Montoyo A (2002) A Proposal for WSD Using Semantic Similarity. *Computational Linguistics and Intelligent Text Processing*. Springer Berlin Heidelberg 2276(1): 51–67. Available at: http://dx.doi.org/10.1007/3-540-45715-1_14.

Spearman C (1904) The Proof and Measurement of Association Between Two Things. *The American Journal of Psychology* 15(1): 72–101.

Stoffel A, Spretke D, Kinnemann H and Keim DA (2010) Enhancing Document Structure Analysis using Visual Analytics. *Proceedings of the 2010 ACM Symposium on Applied Computing (SAC'10)*. New York: ACM Press, 8–12. Available at: http://dl.acm.org/citation.cfm?id=1774091 (accessed 08/01/13).

Strube M and Ponzetto SP (2006) WikiRelate ! Computing Semantic Relatedness Using Wikipedia. *The 21st national conference on Artificial intelligence (AAAI' 06)*. Boston, MA, USA: AAAI Press, 1419–1424.

Tateisi Y and Itoh N (1994) Using stochastic syntactic analysis for extracting a logical structure from a document image. *Proceedings of International Conference on Pattern Recognition*. Jerusalem, Israel: IEEE Computer Society Press, 391–394.

Thakker D, Osman T and Lakin P (2009) GATE JAPE Grammar Tutorial. Nottingham, UK, 1–38. Available at: http://gate.ac.uk/sale/thakker-jape-tutorial/GATE JAPE manual.pdf.

Tsochantaridis I, Joachims T, Hofmann T and Altun Y (2005) Large Margin Methods for Structured and Interdependent Output Variables. *Journal of Machine Learning Research*. Citeseer 6(2): 1453–1484. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.6373&rep=rep1&type=pdf.

Uszok A, Bradshaw JM and Jeffers R (2004) Applying KAoS Services to Ensure Policy Compliance for Semantic Web Services Workflow Composition and Enactment. *The Semantic Web – ISWC 2004*. Hiroshima, Japan: Springer-Verlag Berlin, 425–440. Available at: http://www.springerlink.com/index/PUBDXQ2VJ19L1802.pdf (accessed 18/03/13).

Vishwanathan SVN, Schraudolph NN, Schmidt MW and Murphy KP (2006) Accelerated training of conditional random fields with stochastic gradient methods. *Proceedings of the 23rd international conference on Machine learning ICML 06*. ACM Press 148(1): 969–976. Available at: http://portal.acm.org/citation.cfm?doid=1143844.1143966.

Watts P (2006) Compliance management in the corporate world. *Computer Fraud & Security* 2006(12): 19–20. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1361372306704547.

Wick M, Mccallum A and Doan A (2008) A Discriminative Approach to Ontology Mapping. *The International Workshop on New Trends in Information Integration, NTII' 08*. Auckland, New Zealand,, 16–19. Available at:

http://people.cs.umass.edu/~mccallum/papers/mwick08discriminative.pdf (accessed 19/03/13).

Wu Z and Palmer M (1994) Verb Semantics and Lexical Selection. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA: ACL Press, 6. Available at: http://arxiv.org/abs/cmp-lg/9406033.

Yang D and Powers DMW (2005) Measuring semantic similarity in the taxonomy of WordNet. *Reproduction*. Australian Computer Society, Inc. 315–322. Available at: http://portal.acm.org/citation.cfm?id=1082196.

Yang D and Powers DMW (2007) Word similarity on the taxonomy of WordNet. *Proceedings of the Twenty-eighth Australasian conference on Computer Science (ACSC'05)*. Newcastle, NSW, Australia: Australian Computer Society Inc., 315–322. Available at: http://david.wardpowers.info/Research/AI/papers/200603-ACL+CoLing-WordSimWN.pdf (accessed 12/04/12).

Yu Z and Zhou X (2009) Combining Vector Space Model and Category Hierarchy Model for TV Content Similarity Measure. *Proceedings of the Third International Conference on Multimedia and Ubiquitous Engineering (MUE'09)*. Qingdao, China: IEEE Computer Society, 130–136. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5319034 (accessed 27/06/12).

Zhang IX (2007) Economic Consequences of the Sarbanes-Oxley Act of 2002. *Journal of Accounting and Economics*. Elsevier 44(1-2): 74–115. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0165410107000213 (accessed 10/12/10).

Zhao C, Bhushan M and Venkatasubramanian V (2003) Roles of ontology in automated process safety analysis. *Computer Aided Chemical Engineering*. Elsevier 14(1): 341–346. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1570794603801384 (accessed 30/11/10).

APPENDIX - A

Regulatory Documents

Eudralex - The rules governing medicinal products in the European Union

CHAPTER 5 PRODUCTION

Principle

Production operations must follow clearly defined procedures; they must comply with the principles of Good Manufacturing Practice in order to obtain products of the requisite quality and be in accordance with the relevant manufacturing and marketing authorisations.

General

5.1 Production should be performed and supervised by competent people.

5.2 All handling of materials and products, such as receipt and quarantine, sampling, storage, labelling, dispensing, processing, packaging and distribution should be done in accordance with written procedures or instructions and, where necessary, recorded.

5.3 All incoming materials should be checked to ensure that the consignment corresponds to the order. Containers should be cleaned where necessary and labelled with the prescribed data.

5.4 Damage to containers and any other problem which might adversely affect the quality of a material should be investigated, recorded and reported to the Quality Control Department.

5.5 Incoming materials and finished products should be physically or administratively quarantined immediately after receipt or processing, until they have been released for use or distribution.

5.6 Intermediate and bulk products purchased as such should be handled on receipt as though they were starting materials.

5.7 All materials and products should be stored under the appropriate conditions established by the manufacturer and in an orderly fashion to permit batch segregation and stock rotation.

5.8 Checks on yields, and reconciliation of quantities, should be carried out as necessary to ensure that there are no discrepancies outside acceptable limits.

5.9 Operations on different products should not be carried out simultaneously or consecutively in the same room unless there is no risk of mix-up or cross-contamination.

5.10 At every stage of processing, products and materials should be protected from microbial and other contamination.

5.11 When working with dry materials and products, special precautions should be taken to prevent the generation and dissemination of dust. This applies particularly to the handling of highly active or sensitising materials.

5.12 At all times during processing, all materials, bulk containers, major items of equipment and where appropriate rooms used should be labelled or otherwise identified with an indication of the product or material being processed, its strength (where applicable) and

43

batch number. Where applicable, this indication should also mention the stage of production.

5.13 Labels applied to containers, equipment or premises should be clear, unambiguous and in the company's agreed format. It is often helpful in addition to the wording on the labels to use colours to indicate status (for example, quarantined, accepted, rejected, clean, ...).

5.14 Checks should be carried out to ensure that pipelines and other pieces of equipment used for the transportation of products from one area to another are connected in a correct manner.

5.15 Any deviation from instructions or procedures should be avoided as far as possible. If a deviation occurs, it should be approved in writing by a competent person, with the involvement of the Quality Control Department when appropriate.

5.16 Access to production premises should be restricted to authorised personnel.

5.17 Normally, the production of non-medicinal products should be avoided in areas and with the equipment destined for the production of medicinal products.

Prevention of cross-contamination in production

5.18 Contamination of a starting material or of a product by another material or product must be avoided. This risk of accidental cross-contamination arises from the uncontrolled release of dust, gases, vapours, sprays or organisms from materials and products in process, from residues on equipment, and from operators' clothing. The significance of this risk varies with the type of contaminant and of product being contaminated. Amongst the most hazardous contaminants are highly sensitising materials, biological preparations containing living organisms, certain hormones, cytotoxics, and other highly active materials. Products in which contamination is likely to be most significant are those administered by injection, those given in large doses and/or over a long time.

5.19 Cross-contamination should be avoided by appropriate technical or organisational measures, for example:

a) production in segregated areas (required for products such as penicillins, live vaccines, live bacterial preparations and some other biologicals), or by campaign (separation in time) followed by appropriate cleaning;

b) providing appropriate air-locks and air extraction;

c) minimising the risk of contamination caused by recirculation or re-entry of untreated or insufficiently treated air;

d) keeping protective clothing inside areas where products with special risk of cross-contamination are processed;

e) using cleaning and decontamination procedures of known effectiveness, as ineffective cleaning of equipment is a common source of cross-contamination;

f) using "closed systems" of production;

g) testing for residues and use of cleaning status labels on equipment.

5.20 Measures to prevent cross-contamination and their effectiveness should be checked periodically according to set procedures.

Validation

5.21 Validation studies should reinforce Good Manufacturing Practice and be conducted in accordance with defined procedures. Results and conclusions should be recorded.

5.22 When any new manufacturing formula or method of preparation is adopted, steps should be taken to demonstrate its suitability for routine processing. The defined process, using the materials and equipment specified, should be shown to yield a product consistently of the required quality.

5.23 Significant amendments to the manufacturing process, including any change in equipment or materials, which may affect product quality and/or the reproducibility of the process should be validated.

5.24 Processes and procedures should undergo periodic critical re-validation to ensure that they remain capable of achieving the intended results.

Starting materials

5.25 The purchase of starting materials is an important operation which should involve staff who have a particular and thorough knowledge of the suppliers.

5.26 Starting materials should only be purchased from approved suppliers named in the relevant specification and, where possible, directly from the producer. It is recommended that the specifications established by the manufacturer for the starting materials be discussed with the suppliers. It is of benefit that all aspects of the production and control of the starting material in question, including handling, labelling and packaging requirements, as well as complaints and rejection procedures are discussed with the manufacturer and the supplier.

5.27 For each delivery, the containers should be checked for integrity of package and seal and for correspondence between the delivery note and the supplier's labels.

5.28 If one material delivery is made up of different batches, each batch must be considered as separate for sampling, testing and release.

5.29 Starting materials in the storage area should be appropriately labelled (see Chapter 5, item 13). Labels should bear at least the following information:

— the designated name of the product and the internal code reference where applicable;

— a batch number given at receipt;

— where appropriate, the status of the contents (e.g. in quarantine, on test, released, rejected);

— where appropriate, an expiry date or a date beyond which retesting is necessary.

When fully computerised storage systems are used, all the above information need not necessarily be in a legible form on the label.

5.30 There should be appropriate procedures or measures to assure the identity of the contents of each container of starting material. Bulk containers from which samples have been drawn should be identified (see Chapter 6, item 13).

5.31 Only starting materials which have been released by the Quality Control Department and which are within their shelf life should be used.

5.32 Starting materials should only be dispensed by designated persons, following a written procedure, to ensure that the correct materials are accurately weighed or measured into clean and properly labelled containers.

5.33 Each dispensed material and its weight or volume should be independently checked and the check recorded.

5.34 Materials dispensed for each batch should be kept together and conspicuously labelled as such.

Processing operations: intermediate and bulk products

5.35 Before any processing operation is started, steps should be taken to ensure that the work area and equipment are clean and free from any starting materials, products, product residues or documents not required for the current operation.

5.36 Intermediate and bulk products should be kept under appropriate conditions.

5.37 Critical processes should be validated (see "VALIDATION" in this Chapter).

5.38 Any necessary in-process controls and environmental controls should be carried out and recorded.

5.39 Any significant deviation from the expected yield should be recorded and investigated.

Packaging materials

5.40 The purchase, handling and control of primary and printed packaging materials shall be accorded attention similar to that given to starting materials.

5.41 Particular attention should be paid to printed materials. They should be stored in adequately secure conditions such as to exclude unauthorised access. Cut labels and other loose printed materials should be stored and transported in separate closed containers so as to avoid mix-ups. Packaging materials should be issued for use only by authorised personnel following an approved and documented procedure.

5.42 Each delivery or batch of printed or primary packaging material should be given a specific reference number or identification mark.

5.43 Outdated or obsolete primary packaging material or printed packaging material should be destroyed and this disposal recorded.

Packaging operations

5.44 When setting up a programme for the packaging operations, particular attention should be given to minimising the risk of cross-contamination, mix-ups or substitutions. Different products should not be packaged in close proximity unless there is physical segregation.

5.45 Before packaging operations are begun, steps should be taken to ensure that the work area, packaging lines, printing machines and other equipment are clean and free from any products, materials or documents previously used, if these are not required for the current operation. The line-clearance should be performed according to an appropriate check-list.

5.46 The name and batch number of the product being handled should be displayed at each packaging station or line.

5.47 All products and packaging materials to be used should be checked on delivery to the packaging department for quantity, identity and conformity with the Packaging Instructions.

5.48 Containers for filling should be clean before filling. Attention should be given to avoiding and removing any contaminants such as glass fragments and metal particles.

5.49 Normally, filling and sealing should be followed as quickly as possible by labelling. If it is not the case, appropriate procedures should be applied to ensure that no mix-ups or mislabelling can occur.

5.50 The correct performance of any printing operation (for example code numbers, expiry dates) to be done separately or in the course of the packaging should be checked and recorded. Attention should be paid to printing by hand which should be re-checked at regular intervals.

5.51 Special care should be taken when using cut-labels and when over-printing is carried out off-line. Roll-feed labels are normally preferable to cut-labels, in helping to avoid mix-ups.

5.52 Checks should be made to ensure that any electronic code readers, label counters or similar devices are operating correctly.

5.53 Printed and embossed information on packaging materials should be distinct and resistant to fading or erasing.

5.54 On-line control of the product during packaging should include at least checking the following:

a) general appearance of the packages;

b) whether the packages are complete;

c) whether the correct products and packaging materials are used;

d) whether any over-printing is correct;

e) correct functioning of line monitors.

Samples taken away from the packaging line should not be returned.

5.55 Products which have been involved in an unusual event should only be reintroduced into the process after special inspection, investigation and approval by authorised personnel. Detailed record should be kept of this operation.

5.56 Any significant or unusual discrepancy observed during reconciliation of the amount of bulk product and printed packaging materials and the number of units produced should be investigated and satisfactorily accounted for before release.

5.57 Upon completion of a packaging operation, any unused batch-coded packaging materials should be destroyed and the destruction recorded. A documented procedure should be followed if uncoded printed materials are returned to stock.

Finished products

5.58 Finished products should be held in quarantine until their final release under conditions established by the manufacturer.

5.59 The evaluation of finished products and documentation which is necessary before release of product for sale are described in Chapter 6 (Quality Control).

5.60 After release, finished products should be stored as usable stock under conditions established by the manufacturer.

Rejected, recovered and returned materials

5.61 Rejected materials and products should be clearly marked as such and stored separately in restricted areas. They should either be returned to the suppliers or, where appropriate, reprocessed or destroyed. Whatever action is taken should be approved and recorded by authorised personnel.

5.62 The reprocessing of rejected products should be exceptional. It is only permitted if the quality of the final product is not affected, if the specifications are met and if it is done in accordance with a defined and authorised procedure after evaluation of the risks involved. Record should be kept of the reprocessing.

5.63 The recovery of all or part of earlier batches which conform to the required quality by incorporation into a batch of the same product at a defined stage of manufacture should be authorised beforehand. This recovery should be carried out in accordance with a defined procedure after evaluation of the risks involved, including any possible effect on shelf life. The recovery should be recorded.

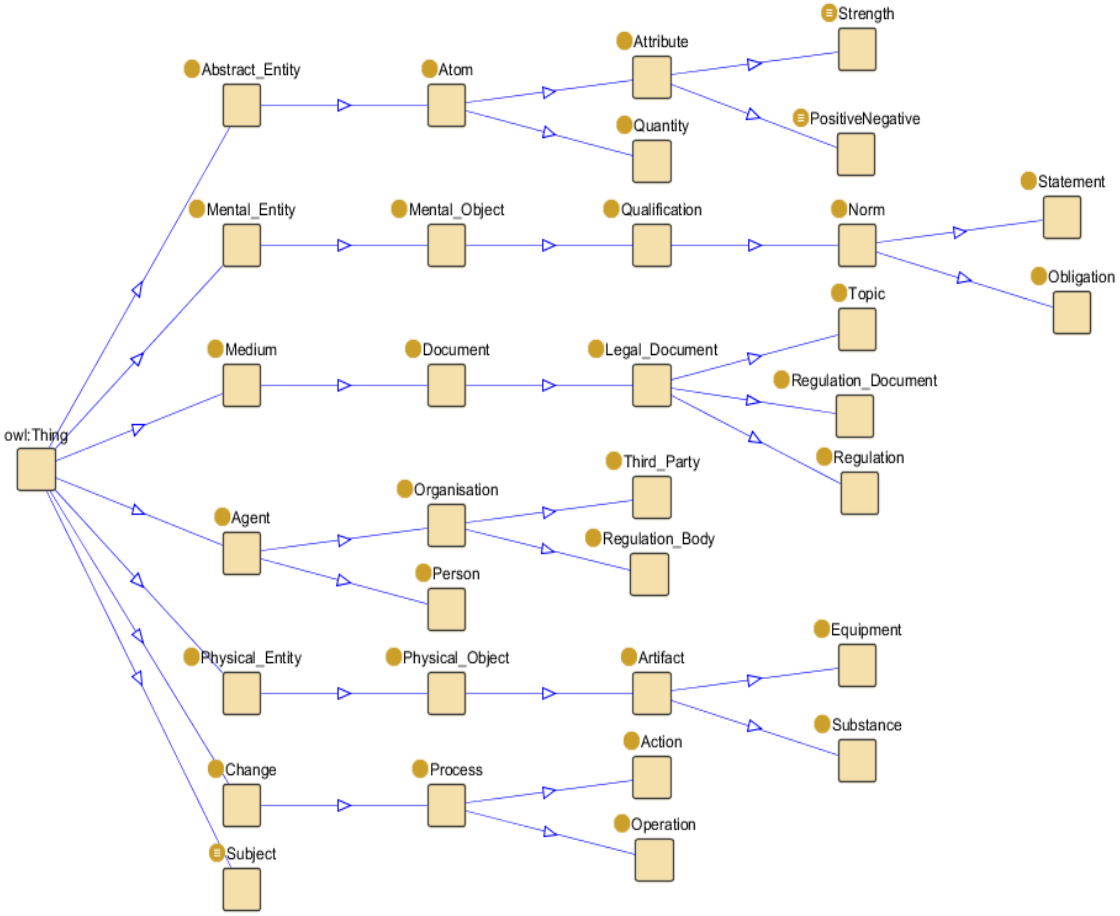
5.64 The need for additional testing of any finished product which has been reprocessed, or into which a recovered product has been incorporated, should be considered by the Quality Control Department.

5.65 Products returned from the market and which have left the control of the manufacturer should be destroyed unless without doubt their quality is satisfactory; they may be considered for re-sale, re-labelling or recovery in a subsequent batch only after they have been critically assessed by the Quality Control Department in accordance with a written procedure. The nature of the product, any special storage conditions it requires, its condition and history, and the time elapsed since it was issued should all be taken into account in this assessment. Where any doubt arises over the quality of the product, it should not be considered suitable for re-issue or re-use, although basic chemical re-processing to recover active ingredient may be possible. Any action taken should be appropriately recorded.

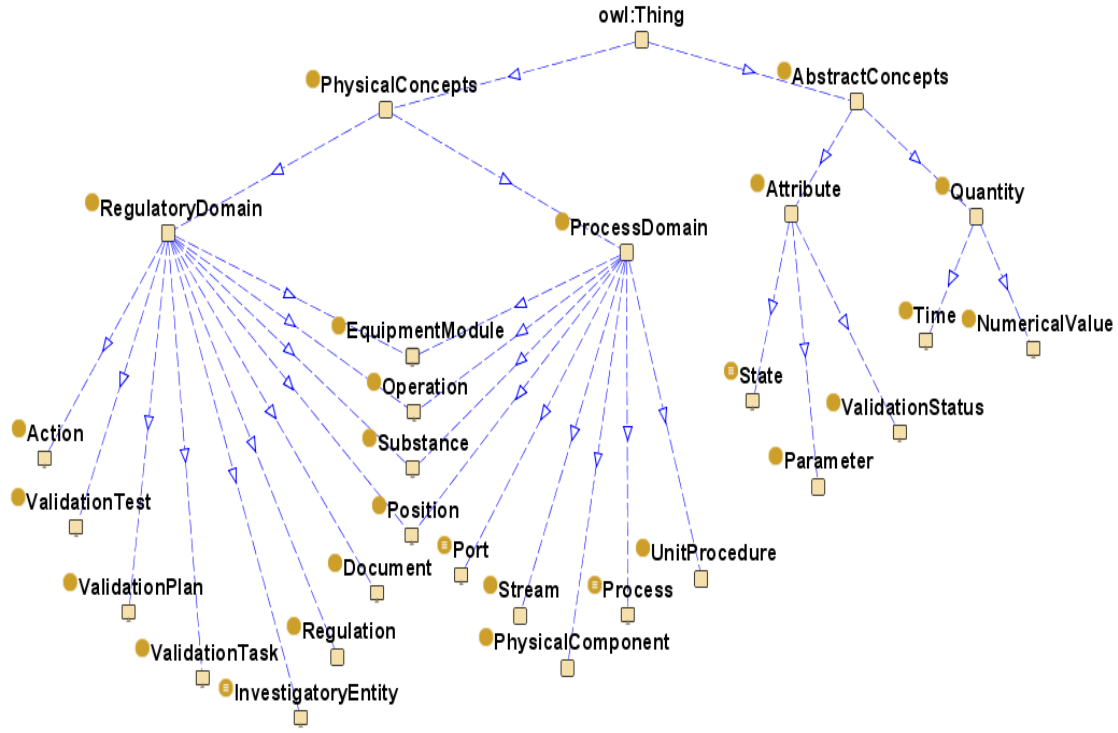
APPENDIX - B

Ontologies

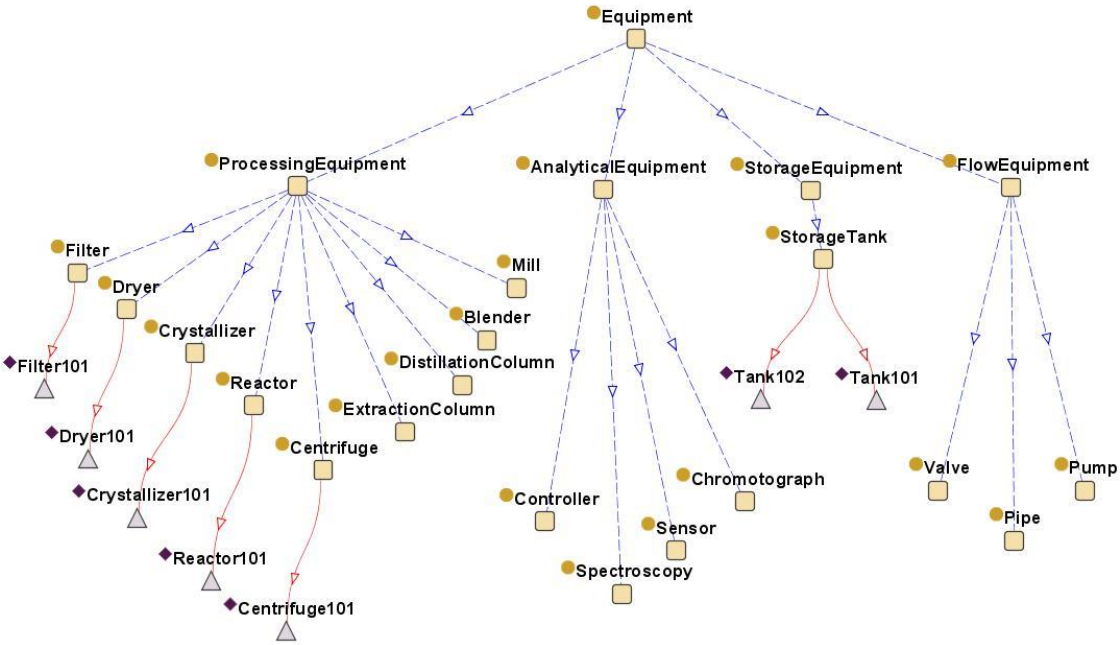
SemReg ontology representing the regulatory guidelines in the Eudralex



OntoReg ontology (Sesen et al, 2010) representing organisational processes in the Pharmaceutical industry



A concept "Equipment" in OntoReg is shown in detail below.



APPENDIX - C

Algorithms

Various algorithms to identify document structures and relating regulatory guidelines with organisational processes.

Mapping Regulation with Tasks

Description: This algorithm defines the relationship between the regulations in the regulation ontology with the validation tasks in the process ontology. It adopts an existing WordNet similarity algorithm in order to compute similarity score between two words.

function RELATED(*reg, task*) **returns** true or false
    $score_1$ ← GET-ACCEPTABLE-SUBJECT-SCORE-FROM-USER()
    $score_2$ ← GET-ACCEPTABLE-ACTION-SCORE-FROM-USER()
    subject_score ← GET-SUBJECT-SCORE (*reg, task*)
    action_score ← GET-ACTION-SCORE (*reg, task*)
    if (*subject_score* ≥ $score_1$ **and** *action_score* ≥ $score_2$) **then**
      return true
    end if
    return false

function GET-SUBJECT-SCORE(*reg, task*) **returns** *subject_score*
    subject_score = 0
    S_1 = {s_1 | s_1 is_a_subject_in *reg*}
    S_2 = {s_2 | s_2 is_a_subject_in *task*}
    f or each $s_i \in S_1$
      for each $s_j \in S_2$
        new_subject_score ← COMPUTE-WORDNET-SIMILARITY-SCORE(s_i, s_j)
        if (*new_subject_score* > *subject_score*) **then**
            subject_score ← *new_subject_score*
        end if
      end for
    end for
    return *subject_score*

function GET-ACTION-SCORE(*reg, task*) **returns** *action_score*
action_score = 0
A_1 = {a_1 | a_1 is_an_action_in *reg*}
A_2 = {a_2 | a_2 is_an_action_in *task*}
for each $a_i \in A_1$
  for each $a_j \in A_2$
    new_action_score ← COMPUTE-WORDNET-SIMILARITY-SCORE(a_i, a_j)

```
    if (new_action_score > action_score) then
        action_score ← new_action_score
    end if
  end for
end for
return action_score
```

Spanning Level of Style (Joining sentences)

Description: This algorithm helps to combine the text with the same level of style. When pdf file is converted into html pages, each line is considered as a different paragraph, which breaks down a sentence or a paragraph in illogical fragments. This algorithm helps to combine the illogically fragmented pieces of lines and paragraphs to make them intact.

Input: T is a set of text in the corpus.
Output: T' is the modified (processed) text.

```
function JOIN-TEXT(T) returns T'
    T = {t1, t2, .. , tn}
    repeat = true
    while (repeat) do
    repeat = false
    for each ti ∈ T
        l1 ← GET-STYLE_LEVEL(ti)
        l2 ← GET-STYLE-LEVEL(ti+1)
        if (l1 == l2) then
            ti ← ti +ti+1
            repeat = true
        end if
    end for
    end while
    T' = T
    return T'
```

Structure Prediction (Paragraph)

Description: When pdf document is converted into html, we get different levels of text style. They may have different font size, font-weight, and font-style and so on. This algorithm will compute the best possible level of style to be the candidate for paragraph text.

Input: S is a set of sentences and l is a level of style
Output: *percent* is the percentage of the given level in the set of sentences.

```
function COMPUTE-SENTENCE-LEVEL-PERCENT(S, l) returns percent
```

```
S = {s1, s2, .. , sn}
count = 0, total = 0, percent = 0
for each si ∈ S
    total = total + 1
    li← GET-STYLE_LEVEL(si)
    if (l == li) then
        count = count + 1
    end if
end for
percent = (count/total) * 100
return percent
```

Input: *text* is set of the text in the corpus and *l* is a level of style
Output: *percent* is the percentage of the given level in the set of text.

```
function COMPUTE-TEXT-LEVEL-PERCENT(T, l) returns percent
    T = {t1, t2, .. , tn}
    count = 0, total = 0, percent = 0
    for each ti ∈ T
        total = total + 1
        li← GET-STYLE_LEVEL(ti)
        if (l == li) then
            count = count + 1
        end if
    end for
    percent = (count/total) * 100
    return percent
```

Input: *O* is a set of obligation words in the corpus and *l* is a level of style
Output: *percent* is the percentage of the given level in the set of obligation.

```
function COMPUTE-OBLIGATION-LEVEL-PERCENT (O , l) returns percent
    O = {o1, o2, .. , on}
    count = 0, total = 0, percent = 0
    for each oi ∈ O
        total = total + 1
        li← GET-STYLE_LEVEL(oi)
        if (l == li) then
            count = count + 1
        end if
    end for
    percent = (count/total) * 100
    return percent
```

Input: *s standard* size of a paragraph in general convention and *l* is a level of style
Output: *percent* is the percentage of deviation of the level from the standard size of a paragraph .

function COMPUTE-SIZE-DEVIATION-LEVEL-PERCENT(*s , l)* **returns** *percent*
    s_1 is_the_size_of *l*
    $d = \pm (s - s_1)$
    $percent = (d/s) * 100$
    return *percent*

Input: *l* is a style level.
Output: *index* is the paragraph prediction index for the level *l* .

function COMPUTE-PARA-PREDICTION-INDEX(*l)* **returns** *index*
    p_1 is_percentage_of_sentences_in *l*
        p_2 is_percentage_of_content_in *l*
        p_3 is_percentage_of_obligation_in *l*
        p_4 is_percentage_of_deviation_in *l*
    w_1 is_weight_of p_1
    w_2 is_weight_of p_2
    w_3 is_weight_of p_3
    w_4 is_weight_of p_4
    $index = getAverage(p_1 * w_1 + p_2 * w_2 + p_3 * w_3 + p_4 * w_4)$
    return *index*

Description: The level with the highest level of predicted index is the paragraph.

Input: *L* is a set of style level in the document.
Output: *L'* is a new set of style levels with paragraph level predicted

function PREDICT-PARAGRAPH(*L)* **returns** *L'*
    $i = 0, paragraph = null$
    $L = \{l_1, l_2, .. , l_n\}$
    for each $l_i \in L$
        $i_1 \leftarrow$ COMPUTE-PARA-PREDICTION-INDEX(l_i)
        if ($i_1 > i$) **then**
            $paragraph \leftarrow$ GET-STRUCTURE(l_i)
            $paragraph \leftarrow l_1$
            $i \leftarrow i_1$
        end if
    end for
    $L' = L$
    return *L'*

Structure Prediction (Others, based on preceding text)

Description: This algorithm is applied after the paragraph prediction algorithm is applied. If a list of possible document structures (components) are provided and some of the component is preceded in the style level text, then the preceding text is set as its structure (style, component).

Input: C is a set of possible document structure component . L is a set of style level in the document.
Output: L' is a new set of style levels in the document with document structure values computed from the preceding text

function PREDICT-PRECEDED-STRUCTURE(C, L) **returns** L'

$C \leftarrow \{ c_1, c_2, .. , c_n\}$
$L \leftarrow \{ l_1, l_2, .. , l_n\}$
for each $l_i \in L$
$c_i \leftarrow$ GET-STRUCTURE(l_i)
$p \leftarrow$ GET-TEXT-PRECEDING(l_i)
if ($c_i = null$) **then**
for each $c_j \in C$
if ($p = c_j$) **then**
$c_i = c_j$
end if
end for
end if
end for
$L' = L$
return L'

Structure Prediction (Filling the Rest)

Description: This algorithm is applied after the paragraph and other component prediction algorithms are applied. It just takes a list used last time, and fills the empty places from the biggest value towards lowest.

Input: C is a set of possible document structure component . L is a set of style level in the document.
Output: L' is a new set of style levels in the document with document structure values computed from the preceding text

function PREDICT-REMAINING-STRUCTURE(C, L) **returns** L'

$C \leftarrow \{ c_1, c_2, .. , c_n\}$
$L \leftarrow \{ l_1, l_2, .. , l_n\}$
for each $l_i \in L$
$c_i \leftarrow$ GET-STRUCTURE(l_i)
$c_{i+1} \leftarrow$ GET-STRUCTURE(l_{i+1})

if (c_i = *null* **or** c_i = *null)* **then**
for each $c_j \in C$
if ($ci = c_j$) **then**
. . . //todo
end if
end for
end if
end for
$L' = L$
return L'

Extraction (Style Head & Style Body)

Description: This algorithm helps style body to correspond with a style head. An HTML document has a css definition for each div tag in its body, which we referred as style head. A div tag in the body section, has its name attached to the css style definition and we called it as style body.

Input: H is a set of style head . B is a set of style body.
Output: B' is a new set of style body with style levels assigned.

function COMPUTE-BODY-LEVEL(H, B) **returns** B'
$H \leftarrow \{ h_1, h_2, .. , h_n\}$
$B \leftarrow \{ b_1, b_2, .. , b_n\}$
for each $h_i \in H$
for each $b_j \in B$
$n_i \leftarrow$ GET-NAME(h_i)
$n_j \leftarrow$ GET-NAME(b_j)
if ($n_i = n_j$) **then**
$l_i \leftarrow$ GET-LEVEL(h_i)
$l_j \leftarrow$ GET-LEVEL(b_j)
$l_j \leftarrow l_i$
end if
end for
end for
$B' = B$
return B'

Input: H is a set of style head .
Output: H' is a new set of style head with style levels assigned.

function COMPUTE-HEAD-LEVEL(H) **returns** H'
$H \leftarrow \{ h_1, h_2, .. , h_n\}$
for each $h_i \in H$

$s_i \leftarrow$ GET-LEVEL-SCORE(h_i)
$s \leftarrow$ COMPUTE-LEVEL-SCORE(h_i)
$s_i \leftarrow$ s
. . . //todo score to level
end for
H' = SORT-HEAD(H)
return H'

Input: H is a set of style head .
Output: H' is a new set of style head sorted by style levels.

function SORT-HEAD(*H)* **returns** H'
$H \leftarrow \{ h_1, h_2, .. , h_n\}, H \leftarrow \varnothing$
for each $h_i \in H$
. . . //todo sort, and assign level
end for
return H'

Style Score Calculation

Description: This algorithm helps to calculate the score of a style level based on the font features. It considers the standard font values from the java.awt.Font for font-size, font-weight and font-style and adds all with their added weights.

Input: l is a style level.
Output: *score is* the score of the level computed considering font-size, font-weight and font-style

function COMPUTE-LEVEL-SCORE(*l)* **returns** *score*
s_1 is_the_font_size_in l
s_2 is_the_font_weight_in l
s_3 is__the_font_style_in l
w_1 is_weight_of *s1*
w_2 is_weight_of *s2*
w_3 is_weight_of *s3*
*score = s1 * w1 + s2 * w2 + s3 *w3*
return *score*

APPENDIX - D

Gazetteers

Some gazetteers used in RegCMantic framework

extracted_term.lst

```
api
production
process
quality
materials
apis
equipment
batch
intermediates
use
validation
control
procedures
records
material
testing
specifications
packaging
batches
contamination
manufacturer
cleaning
manufacturing
system
steps
controls
cell
processing
date
number
systems
unit
storage
test
changes
laboratory
processes
product
stability
data
facilities
operations
sampling
results
```

definition_term.lst

```
air-lock
batch
lot
batch number
lot number
biogenerator
biological agents
bulk product
cell bank
cell culture
clean area
clean area
contained
containment
contained area
controlled area
computerised system
cross contamination
crude plant
vegetable drug
cryogenic vessel
cylinder
exotic organism
finished product
herbal medicinal product
infected
in-process control
intermediate product
liquifiable gases
manifold
manufacture
manufacturer
medicinal plant
medicinal product
packaging
packaging material
procedures
production
qualification
quality control
quarantine
radiopharmaceutical
reconciliation
…
```

Ontology_concept.lst

```
a auditor
a coordinator
a inspector
active pharmaceutical
ingredient
actuating equipment
administration department
administrations department
allowing
allowing task
analytical equipment
approved supplier
assess
assess task
assessing
assessment report
assistant
assistant purchasing officer
assurance auditor
assurance coordinator
assurance department
assurance inspector
auditor
batch number
blender
c coordinator
c officer
centrifuge
centrifuging
charging
check
checking task
chromatograph
chromotograph
classifying
clean
cleaniness test task
cleaning
cleaning task
cleaning validation plan
cleaning validation report
cleanliness report
cleanliness test
cleanliness test task
clerical assistant
column
comply
complying
component
computer validation plan
computer validation report
…
```

APPENDIX - E

Rules

Rules to read CSS styles and interpret the corresponding HTML tags

style_head2.jape | obligation.jape | rule_action.jape | rule_subject.jape

```
/*
 *
 */
Phase: style
Input: Token  SpaceToken Split Style
Options: control = appelt

//Macros
Macro: DOT ({Token.string=="."})
Macro: CURLY_START ({Token.string=="{"})
Macro: CURLY_END ({Token.string=="}"})
Macro: FT({Token.string=="ft"})
Macro: NUM ({Token.category==CD})
Macro: CONTROL ({SpaceToken.kind==control})
Macro: SPACE ({SpaceToken.kind==space})

// defines the details of the styles in  a page
Rule: StyleDetails
Priority:90
(
      DOT (FT NUM ):name CURLY_START
      {Token.string=="font-style"} {Token.string==":"}({Token}):fs  {Token.string==";"}
      {Token.string=="font-weight"} {Token.string==":"}({Token}):fw {Token.string==";"}
      {Token.string=="font-size"} {Token.string==":"}({Token}):size {Token.string=="px"}{Token.string==";"}
      {Token.string=="font-family"} {Token.string==":"}({Token}| SPACE)[1,10]:family {Token.string==";"}
      {Token.string=="color"} {Token.string==":"}({Token.string=="#"} {Token}):color
      ({Token}|SPACE)* CURLY_END CONTROL
):curly
-->
{
// obtains the annotation for whole style
gate.AnnotationSet curlySet = (gate.AnnotationSet)bindings.get("curly");
// adding features to the annotation
gate.FeatureMap features = Factory.newFeatureMap();
features.put("rule","StyleDetails");
// adding string of the annotation to the fetures. It needs start and end node offsets as int values
int sNodeCurly = curlySet.firstNode().getOffset().intValue();
int eNodeCurly = curlySet.lastNode().getOffset().intValue();
features.put("startNode",sNodeCurly);
features.put("endNode", eNodeCurly);
String wholeString = doc.getContent().toString().substring(sNodeCurly, eNodeCurly);
features.put("string",wholeString);
    // ------ obtains the annotation for font name ------------------
    gate.AnnotationSet nameSet = (gate.AnnotationSet)bindings.get("name");
    // adding string of the annotation to the fetures. It needs start and end node offsets as int values
    int sNodeName = nameSet.firstNode().getOffset().intValue();
    int eNodeName = nameSet.lastNode().getOffset().intValue();
    String fontName = doc.getContent().toString().substring(sNodeName, eNodeName);
    features.put("name", fontName);

    // ------ obtains the annotation for font style ------------------
    gate.AnnotationSet fsSet = (gate.AnnotationSet)bindings.get("fs");
    // adding string of the annotation to the fetures. It needs start and end node offsets as int values
    int sNodeFS = fsSet.firstNode().getOffset().intValue();
    int eNodeFS = fsSet.lastNode().getOffset().intValue();
    String fontStyle = doc.getContent().toString().substring(sNodeFS, eNodeFS);
    features.put("font-style", fontStyle);

    // ------ obtains the annotation for font weight ------------------
    gate.AnnotationSet fwSet = (gate.AnnotationSet)bindings.get("fw");
    // adding string of the annotation to the fetures. It needs start and end node offsets as int values
    int sNodeFW = fwSet.firstNode().getOffset().intValue();
    int eNodeFW = fwSet.lastNode().getOffset().intValue();
```

Rules to identify the obligations

style_head2.jape | obligation.jape | rule_action.jape | rule_subject.jape

```
/**
 * Identifies obligations within a sentence.
 * uses look up (gazetteer) for the purpose.
 */
Phase:obligation
Input:  Lookup  Token
Options: control = appelt

/* basic macros */
Macro: CC ( {Token.category == CC}) //CC - coordinating conjunction: 'and', 'but', 'nor', 'or', 'yet', plus, minus, less,
that').
Macro: COMMA ( {Token.category == ","}) //, - literal comma
Macro: RB ( {Token.category == RB}) //RB - adverb: most words ending in '-ly'. Also 'quite', 'too', 'very', 'enough',
Macro: RBR ( {Token.category == RBR}) //RBR - adverb - comparative: adverbs ending with '-er' with a comparat
Macro: RBS ( {Token.category == RBS}) //RBS - adverb - superlative

/*  obligation from the parsed sentencec*/
//MACRO: PARSED ({parsed_obligation})

/* complex macros */
Macro: CONJ (
 (CC) | (COMMA)
)
Macro: ADVERB (
 (RB)| (RBR)|(RBS)
)

/* obligation related macros */
Macro: OBL_STRONG ({Lookup.majorType == obligation,Lookup.minorType==strong})
Macro: OBL_STRONGNEG ({Lookup.majorType == obligation,Lookup.minorType==strong_neg})
Macro: OBL_MOD ({Lookup.majorType == obligation,Lookup.minorType==moderate})
Macro: OBL_MODNEG ({Lookup.majorType == obligation,Lookup.minorType==moderate_neg})
Macro: OBL_SOFT ({Lookup.majorType == obligation,Lookup.minorType==soft})
Macro: OBL_SOFTNEG ({Lookup.majorType == obligation,Lookup.minorType==soft_neg})


/* rule:  Strong Positive */
Rule: Strongobligation
Priority:100
(
    (OBL_STRONG)
):obligation
-->
{
// obtains the annotation
gate.AnnotationSet team = (gate.AnnotationSet)bindings.get("obligation");
gate.Annotation teamAnn = (gate.Annotation)team.iterator().next();

// adding features to the annotation
gate.FeatureMap features = Factory.newFeatureMap();
features.put("rule","Strongobligation");

// adding string of the annotation to the fetures. It needs start and end node offsets as int values
int sNode = teamAnn.getStartNode().getOffset().intValue();
int eNode = teamAnn.getEndNode().getOffset().intValue();
features.put("startNode",sNode);
features.put("endNode", eNode);
String theWord = doc.getContent().toString().substring(sNode, eNode);
features.put("text",theWord);
features.put("strength","Strong");
features.put("type", "Positive");

// finally adding the feature with the annotation.
outputAS.add(team.firstNode(), team.lastNode(), "obligation",features);
}
```

Rules to identify actions

style_head2.jape | obligation.jape | rule_action.jape | rule_subject.jape

```
/**
 * Identifies actions.
 */
Phase:action
Input:  Token obligation
Options: control =  brill

/* simple macros */
Macro: CC ( {Token.category == CC}) //CC - coordinating conjunction: 'and', 'but', 'nor', 'or', 'yet', plus, minus, less, times (multiplication), over (division). .
that').
Macro: RB ( {Token.category == RB}) //RB - adverb: most words ending in '-ly'. Also 'quite', 'too', 'very', 'enough', 'indeed', 'not', '-n't', and 'never'.
Macro: RBR ( {Token.category == RBR}) //RBR - adverb - comparative: adverbs ending with '-er' with a comparative meaning.
Macro: RBS ( {Token.category == RBS}) //RBS - adverb - superlative
Macro: VBD ( {Token.category == VBD}) //VBD - verb - past tense: includes conditional form of the verb 'to be'; 'If I were/VBD rich...'.
Macro: VBG ( {Token.category == VBG}) //VBG - verb - gerund or present participle
Macro: VBN ( {Token.category == VBN}) //VBN - verb - past participle
Macro: VBP ( {Token.category == VBP}) //VBP - verb - non-3rd person singular present
Macro: VB ( {Token.category == VB}) //VB - verb - base form: subsumes imperatives, infinitives and subjunctives.
Macro: VBZ ( {Token.category == VBZ}) //VBZ - verb - 3rd person singular present
Macro: COMMA ( {Token.category == ","}) //, - literal comma
Macro: PREP ({Token.category == IN})
Macro: OBL ({obligation})

/* complex macros */
Macro: CONJ (CC | COMMA)

Macro: ADVERB (RB | RBR | RBS)

Macro: VERB (VBN |  VB |    VBP |   VBD |   VBG |   VBZ )
Macro: VERB_PHRASE(VERB PREP)    // prepostion

// should be (appropriately) cleaned
Macro: ONE_ACTION(  OBL (ADVERB)?   VERB )

//, appropriately desined
Macro: ACT_PART ((CONJ)+(ADVERB)? VERB)

// should be (appropriately) cleaned
// should be (appropriately) cleaned and (appropriately) stored
//should be (appropriately) cleaned ,(appropriately) stored, (carefully) handled and (appropriately) delivered.
Macro: ACTIONS(
    (   OBL(ADVERB)?(VERB ):action)   |
        ONE_ACTION(CONJ)+(ADVERB)?(VERB):action      |
        ONE_ACTION  (ACT_PART)+(CONJ)+(ADVERB)?(VERB):action

)

/* rule: */
Rule: ActionFinder
Priority:90
(
    ACTIONS
)
-->
{
// obtains the annotation
gate.AnnotationSet actionSet = (gate.AnnotationSet)bindings.get("action");
gate.Annotation action = (gate.Annotation)actionSet.iterator().next();

// adding features to the annotation
gate.FeatureMap features = Factory.newFeatureMap();
features.put("rule","ActionFinder");

// adding string of the annotation to the fetures. It needs start and end node offsets as int values
int sNode = action.getStartNode().getOffset().intValue();
int eNode = action.getEndNode().getOffset().intValue();
features.put("startNode",sNode);
features.put("endNode", eNode);
String theWord = doc.getContent().toString().substring(sNode, eNode);
features.put("text",theWord);

```

APPENDIX - F

Parsed Sentences

Sentences are parsed into various chunks using Stanford Parser.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<parsed_sentences size="49">
    <parsed_sentence count="1">
        <subject> Quality</subject>
        <obligation> should be</obligation>
        <action> the responsibility</action>
        <object> </object>
        <condition></condition>
        <modifier> of all persons involved</modifier>
    </parsed_sentence>
    <parsed_sentence count="2">
        <subject> Each manufacturer</subject>
        <obligation> should</obligation>
        <action> establish , document , and implement</action>
        <object> </object>
        <condition></condition>
        <modifier> </modifier>
    </parsed_sentence>
    <parsed_sentence count="3">
        <subject> The system for managing quality</subject>
        <obligation> should</obligation>
        <action> encompass</action>
        <object> the organisational structure , procedures , processes and resources ,
as well as activities</object>
        <condition></condition>
        <modifier> </modifier>
    </parsed_sentence>
    <parsed_sentence count="4">
        <subject> All quality related activities</subject>
        <obligation> should be</obligation>
        <action> defined and documented</action>
        <object> </object>
        <condition></condition>
        <modifier> </modifier>
    </parsed_sentence>
    <parsed_sentence count="5">
        <subject> </subject>
        <obligation> should be</obligation>
        <action> a quality unit -LRB- s -RRB- that is independent of production and
that fulfills</action>
        <object> both quality assurance -LRB- QA -RRB- and quality control -LRB- QC -
RRB- responsibilities</object>
        <condition></condition>
        <modifier> </modifier>
    </parsed_sentence>
    <parsed_sentence count="6">
        <subject> This</subject>
        <obligation> can</obligation>
        <action> be</action>
        <object> </object>
        <condition></condition>
        <modifier> in in the form of separate QA and QC units or a single individual or
group</modifier>
    </parsed_sentence>
    <parsed_sentence count="7">
        <subject> The persons authorised</subject>
        <obligation> should be</obligation>
        <action> specified</action>
        <object> </object>
        <condition></condition>
        <modifier> </modifier>
    </parsed_sentence>
    <parsed_sentence count="8">
        <subject> All quality related activities</subject>
        <obligation> should be</obligation>
        <action> recorded</action>
        <object> </object>
```

```
        <condition></condition>
        <modifier> at the time they are performed</modifier>
    </parsed_sentence>
    <parsed_sentence count="9">
        <subject> Any deviation from established procedures</subject>
        <obligation> should be</obligation>
        <action> documented and explained</action>
        <object> </object>
        <condition></condition>
        <modifier> </modifier>
    </parsed_sentence>
    <parsed_sentence count="10">
        <subject> Critical deviations</subject>
        <obligation> should be</obligation>
        <action> investigated , and the investigation and its conclusions should be
documented</action>
        <object> </object>
        <condition></condition>
        <modifier> </modifier>
    </parsed_sentence>
    <parsed_sentence count="11">
        <subject> No materials</subject>
        <obligation> should be</obligation>
        <action> released or used</action>
        <object> </object>
        <condition></condition>
        <modifier> </modifier>
    </parsed_sentence>
    <parsed_sentence count="12">
        <subject> Procedures</subject>
        <obligation> should</obligation>
        <action> exist</action>
        <object> </object>
        <condition></condition>
        <modifier> notifying responsible management in a timely manner of regulatory
inspections , serious GMP deficiencies , product defects and related actions -LRB- e.g.
quality</modifier>
    </parsed_sentence>
    <parsed_sentence count="13">
        <subject> The quality unit -LRB- s</subject>
        <obligation> should be</obligation>
        <action> involved</action>
        <object> </object>
        <condition></condition>
        <modifier> in all quality-related matters</modifier>
    </parsed_sentence>
    <parsed_sentence count="14">
        <subject> The quality unit -LRB- s</subject>
        <obligation> should</obligation>
        <action> review and approve</action>
        <object> </object>
        <condition></condition>
        <modifier> </modifier>
    </parsed_sentence>
    <parsed_sentence count="15">
        <subject> The main responsibilities of the independent quality unit</subject>
        <obligation> should not be</obligation>
        <action> delegated</action>
        <object> </object>
        <condition></condition>
        <modifier> </modifier>
    </parsed_sentence>
    <parsed_sentence count="16">
        <subject> regular internal audits</subject>
        <obligation> should be</obligation>
        <action> performed</action>
        <object> </object>
        <condition>   In order to verify compliance with the principles of GMP for
APIs</condition>
        <modifier> in with accordance</modifier>
    </parsed_sentence>
    <parsed_sentence count="17">
        <subject> Audit findings and corrective actions</subject>
        <obligation> should be</obligation>
        <action> documented and brought</action>
        <object> </object>
        <condition></condition>
```

```
        <modifier> </modifier>
    </parsed_sentence>
    <parsed_sentence count="18">
        <subject> Agreed corrective actions</subject>
        <obligation> should be</obligation>
        <action> completed</action>
        <object> </object>
        <condition></condition>
        <modifier> in a timely and effective manner</modifier>
    </parsed_sentence>
    <parsed_sentence count="19">
        <subject> Regular quality reviews of APIs</subject>
        <obligation> should be</obligation>
        <action> conducted</action>
        <object> </object>
        <condition></condition>
        <modifier> with the objective of verifying</modifier>
    </parsed_sentence>
    <parsed_sentence count="20">
        <subject> Such reviews</subject>
        <obligation> should</obligation>
        <action>  normally  be  conducted  and  documented  annually  and  should
include</action>
        <object> </object>
        <condition></condition>
        <modifier> </modifier>
    </parsed_sentence>
    <parsed_sentence count="21">
        <subject> </subject>
        <obligation></obligation>
        <action> A review</action>
        <object> </object>
        <condition></condition>
        <modifier> of critical in-process control and critical API test results ; - A
review of all batches that failed to meet established specification -LRB- s -RRB- ; - A
review of all critical deviations or non-conformances and related investigations ; - A
review of any changes carried out to the processes or analytical methods ; - A review
of  results  of  the  stability  monitoring  program  ;  -  A  review  of  all  quality-related
returns , complaints and recalls ; and - A review</modifier>
    </parsed_sentence>
    <parsed_sentence count="22">
        <subject> The results of this review</subject>
        <obligation> should be</obligation>
        <action> evaluated and an assessment made</action>
        <object> </object>
        <condition></condition>
        <modifier>  whether  corrective  action  or  any  revalidation  should  be
undertaken</modifier>
    </parsed_sentence>
    <parsed_sentence count="23">
        <subject> Reasons for such corrective action</subject>
        <obligation> should be</obligation>
        <action> documented</action>
        <object> </object>
        <condition></condition>
        <modifier> </modifier>
    </parsed_sentence>
    <parsed_sentence count="24">
        <subject> Agreed corrective actions</subject>
        <obligation> should be</obligation>
        <action> completed</action>
        <object> </object>
        <condition></condition>
        <modifier> in a timely and effective manner</modifier>
    </parsed_sentence>
    <parsed_sentence count="25">
        <subject> </subject>
        <obligation> should be</obligation>
        <action> an adequate number of personnel qualified by appropriate education ,
training and\/or experience</action>
        <object> </object>
        <condition></condition>
        <modifier>  There  should  be  an  adequate  number  of  personnel  qualified  by
appropriate  education  ,  training  and\/or  experience  to  perform  and  supervise  the
manufacture</modifier>
    </parsed_sentence>
    <parsed_sentence count="26">
```

```
        <subject> The responsibilities of all personnel</subject>
        <obligation> should be</obligation>
        <action> specified</action>
        <object> </object>
        <condition></condition>
        <modifier> writing</modifier>
    </parsed_sentence>
    <parsed_sentence count="27">
        <subject> Training</subject>
        <obligation> should be</obligation>
        <action> regularly conducted by qualified individuals and should cover</action>
        <object> </object>
        <condition></condition>
        <modifier> at a minimum , the particular operations</modifier>
    </parsed_sentence>
    <parsed_sentence count="28">
        <subject> Records of training</subject>
        <obligation> should be</obligation>
        <action> maintained</action>
        <object> </object>
        <condition></condition>
        <modifier> </modifier>
    </parsed_sentence>
    <parsed_sentence count="29">
        <subject> Training</subject>
        <obligation> should be</obligation>
        <action> periodically assessed</action>
        <object> </object>
        <condition></condition>
        <modifier> </modifier>
    </parsed_sentence>
    <parsed_sentence count="30">
        <subject> Personnel</subject>
        <obligation> should</obligation>
        <action> practice</action>
        <object> good sanitation and health habits</object>
        <condition></condition>
        <modifier> </modifier>
    </parsed_sentence>
    <parsed_sentence count="31">
        <subject> Personnel</subject>
        <obligation> should</obligation>
        <action> wear clean clothing suitable for the manufacturing activity with which
they are involved and this clothing should be changed</action>
        <object> </object>
        <condition></condition>
        <modifier> </modifier>
    </parsed_sentence>
    <parsed_sentence count="32">
        <subject> Additional protective apparel , such as head , face , hand , and arm
coverings</subject>
        <obligation> should be</obligation>
        <action> worn</action>
        <object> </object>
        <condition></condition>
        <modifier> to protect intermediates and APIs from contamination</modifier>
    </parsed_sentence>
    <parsed_sentence count="33">
        <subject> Personnel</subject>
        <obligation> should</obligation>
        <action> avoid</action>
        <object> direct contact</object>
        <condition></condition>
        <modifier> with with intermediates or APIs</modifier>
    </parsed_sentence>
    <parsed_sentence count="34">
        <subject> eating , drinking , chewing and the storage of food</subject>
        <obligation> should be</obligation>
        <action> restricted</action>
        <object> </object>
        <condition>  Smoking</condition>
        <modifier> to certain designated areas separate</modifier>
    </parsed_sentence>
    <parsed_sentence count="35">
        <subject> Personnel suffering from an infectious disease or having</subject>
        <obligation> should not</obligation>
        <action> engage</action>
```

```
        <object> </object>
        <condition></condition>
        <modifier> in activities that could result</modifier>
    </parsed_sentence>
    <parsed_sentence count="36">
        <subject> Any person shown</subject>
        <obligation> should be</obligation>
        <action> excluded</action>
        <object> </object>
        <condition></condition>
        <modifier> from activities where the health condition could adversely
affect</modifier>
    </parsed_sentence>
    <parsed_sentence count="37">
        <subject> Consultants advising</subject>
        <obligation> should</obligation>
        <action> have</action>
        <object> </object>
        <condition></condition>
        <modifier> education , training , and experience , or any combination thereof ,
to advise on the subject</modifier>
    </parsed_sentence>
    <parsed_sentence count="38">
        <subject> Records</subject>
        <obligation> should be</obligation>
        <action> maintained</action>
        <object> </object>
        <condition></condition>
        <modifier> stating the name , address , qualifications , and type of service
provided</modifier>
    </parsed_sentence>
    <parsed_sentence count="39">
        <subject> Buildings and facilities used</subject>
        <obligation> should be</obligation>
        <action> located , designed , and constructed</action>
        <object> </object>
        <condition></condition>
        <modifier> to facilitate cleaning , maintenance , and operations as
appropriate</modifier>
    </parsed_sentence>
    <parsed_sentence count="40">
        <subject> Facilities</subject>
        <obligation> should</obligation>
        <action> also be designed</action>
        <object> </object>
        <condition></condition>
        <modifier> to minimize potential contamination</modifier>
    </parsed_sentence>
    <parsed_sentence count="41">
        <subject> facilities</subject>
        <obligation> should</obligation>
        <action> also be designed</action>
        <object> </object>
        <condition> Where microbiological specifications have been established for the
intermediate or API</condition>
        <modifier> to limit exposure to objectionable microbiological
contaminants</modifier>
    </parsed_sentence>
    <parsed_sentence count="42">
        <subject> Buildings and facilities</subject>
        <obligation> should</obligation>
        <action> have</action>
        <object> adequate space</object>
        <condition></condition>
        <modifier> for the orderly placement of equipment and materials</modifier>
    </parsed_sentence>
    <parsed_sentence count="43">
        <subject> such equipment</subject>
        <obligation> can be</obligation>
        <action> located</action>
        <object> outdoors</object>
        <condition> Where the equipment itself -LRB- e.g. , closed or contained systems
-RRB- provides adequate protection of the material</condition>
        <modifier> </modifier>
    </parsed_sentence>
    <parsed_sentence count="44">
        <subject> The flow of materials and personnel</subject>
```

```
        <obligation> should be</obligation>
        <action> designed</action>
        <object> </object>
        <condition></condition>
        <modifier> to prevent mix-ups or contamination</modifier>
    </parsed_sentence>
    <parsed_sentence count="45">
        <subject> areas or other control systems</subject>
        <obligation> should be</obligation>
        <action> defined</action>
        <object> areas or other control systems</object>
        <condition>  There should be</condition>
        <modifier> for the following activities</modifier>
    </parsed_sentence>
    <parsed_sentence count="46">
        <subject> </subject>
        <obligation></obligation>
        <action> Receipt , identification , sampling</action>
        <object> </object>
        <condition></condition>
        <modifier> </modifier>
    </parsed_sentence>
    <parsed_sentence count="47">
        <subject> clean washing and toilet facilities</subject>
        <obligation></obligation>
        <action>  Adequate  ,  clean  washing  and  toilet  facilities  should  be
provided</action>
        <object> </object>
        <condition>  Adequate</condition>
        <modifier> for personnel</modifier>
    </parsed_sentence>
    <parsed_sentence count="48">
        <subject> These washing facilities</subject>
        <obligation> should be</obligation>
        <action> equipped</action>
        <object> </object>
        <condition></condition>
        <modifier> with with with hot and cold water as appropriate , 13 soap or
detergent , air driers or single service towels</modifier>
    </parsed_sentence>
    <parsed_sentence count="49">
        <subject> The washing and toilet facilities</subject>
        <obligation> should be</obligation>
        <action> separate</action>
        <object> </object>
        <condition></condition>
        <modifier> from accessible to , manufacturing areas</modifier>
    </parsed_sentence>
</parsed_sentences>
```