Universidad Carlos III de Madrid

PhD Dissertation

Semantic Resources in Pharmacovigilance: A Corpus and an Ontology for Drug-Drug Interactions

Author: María Herrero Zazo

Directors: Dr. Isabel Segura Bedmar
Dr. Paloma Martínez Fernández

PhD Program in Computer Science and Technology
Computer Science Department

Leganés, 2015

TESIS DOCTORAL

Semantic Resources in Pharmacovigilance: a Corpus and an Ontology for Drug-Drug Interactions

Autor: MARÍA HERRERO ZAZO

Directoras: ISABEL SEGURA BEDMAR

PALOMA MARTÍNEZ FERNÁNDEZ

EL TRIBUNAL

Presidente: _____

Vocal: _____

Secretario: _____

CALIFICACIÓN:

VOCAL

SECRETARIO PRESIDENTE

«Estoy seguro de que los tiempos nuevos tienen que traer, entre otras cosas, previstas o no, en la Ciencia, y muy especialmente en la Medicina, una profunda modificación de la literatura. La situación actual no puede seguir, entre otras razones, porque seguir equivale a empeorar. El turbión de revistas y libros, la mayoría de unas y de otros escritos sin responsabilidad, ahoga al lector más voluntarioso. No puede pretenderse – y esto desde hace ya muchos años – reunir la bibliografía completa de un problema médico. Algunos autores, hace unos decenios, lo intentaron, y el empeño les abrumó. Recuerdo al profesor Arturo Bield que, con titánica paciencia y rodeado de un ejército de buenos auxiliares, logró reunir y publicar, en 1913, una bibliografía de glándulas de secreción interna con aspiraciones, casi logradas, de ser completa. Pocos años después la reedición de su libro se había hecho imposible. En los años que siguieron a la edición inicial, los trabajos se habían multiplicado disparatadamente, aquí, en Europa; y surgió, por contera, el fenómeno americano con su avalancha de citas nuevas, que nos llegaba, ya de la parte Norte, ya de la parte Sur del Nuevo Continente […]»

Dr. Gregorio Marañón
Toledo, 1949

Agradecimientos

Por mucho que lo intentase, esta sección estaría siempre incompleta si tratase de recoger en ella a todos los que han contribuido a este trabajo. Desde los que me animaron y me apoyaron al empezar esta nueva etapa, pasando por todos los que me han preguntado alguna vez *¿de qué trata tu tesis?* y todos los que, cada vez que nos hemos visto durante este tiempo, me han dedicado un sincero *¿cómo lo llevas?* o me han dado ánimos para terminar. Ellos no saben el impulso que supone ese apoyo en un camino tan largo como este. Y es que han sido muchos los que, sabiendo poco o a veces nada de lo que contamos en las siguientes hojas, no han dejado de interesarse por todos los pasos que he ido dando para llegar hasta aquí.

Los primeros se los tengo que agradecer a mis directoras, Isabel y Paloma, por haberme dado la oportunidad de trabajar con ellas, en un principio, y por confiar en que una farmacéutica podría conseguir un Doctorado en Informática, después. Siempre he podido contar con su apoyo, sus consejos y sus respuestas. Especialmente quiero darle las gracias a Isa por ser a la vez tutora y compañera, por todo lo que me ha enseñado en este tiempo y, sobre todo, por emocionarse con los avances de esta tesis tanto como yo. Supongo que no es lo habitual que un tutor se involucre en la tesis de su doctorando como lo ha hecho ella, y le estaré siempre agradecida por ello.

Quiero dar las gracias también a Mercedes Martínez, mi tutora durante los cursos de Doctorado en la Universidad de Valladolid, porque sin su apoyo en aquellos momentos iniciales esta tesis no habría sido posible. Moreover, I would like to show my gratitude to Christoph Steinbeck and Janna Hastings for giving me the opportunity to join their research group. This has been one of the most special experiences in this thesis. Also, I want to say thank you to all my colleagues at the UC3M and the EBI for the time we have spent together. Y, como no, tengo que agradecer a todos los que, como Minerva, han leído o escuchado parte de este trabajo y han contribuido con sus consejos y sugerencias.

No puedo olvidar tampoco a todas esas personas que, fuera de la Universidad, han sido también imprescindibles en este proyecto. A mi padre le agradezco las llamadas mientras estaba sola en la biblioteca y los *cómo vas con eso que estás haciendo*. A mi madre, le agradezco todo. A Periquín, haber encontrado la cita que introduce esta tesis. A mis hermanas, las risas cuando, después de llevar un año matriculada de la tesis, se enteraron de que iba a ser Doctora en Informática. También quiero dar las gracias al resto de mi familia (a los de siempre y a los que vienen) y a mis amigos, por haber estado en esta y en todas las etapas de mi vida.

Y si hay alguien que ha vivido este proyecto día a día desde que comenzó es Gerar. A él le doy las gracias por enseñarme que lo importante es el camino, y por hacerlo conmigo. Ahora sólo pido que el camino sea largo…

María

Resumen

Hoy en día ha habido un notable aumento del número de pacientes polimedicados que reciben simultáneamente varios fármacos para el tratamiento de una o varias enfermedades. Esta situación proporciona el escenario ideal para la prescripción de combinaciones de fármacos que no han sido estudiadas previamente en ensayos clínicos, y puede dar lugar a un aumento de interacciones farmacológicas (DDIs por sus siglas en inglés). Las interacciones entre fármacos son un tipo de reacción adversa que supone no sólo un riesgo para los pacientes, sino también una importante causa de aumento del gasto sanitario. Por lo tanto, su detección temprana es crucial en la práctica clínica. En la actualidad existen diversos recursos y bases de datos que pueden ayudar a los profesionales sanitarios en la detección de posibles interacciones farmacológicas. Sin embargo, la calidad de su información varía considerablemente de unos a otros, y la consistencia de sus contenidos es limitada. Además, la actualización de estos recursos es difícil debido al aumento que ha experimentado la literatura farmacológica en los últimos años. De hecho, mucha información sobre DDIs se encuentra dispersa en artículos, revistas científicas, libros o informes técnicos, lo que ha hecho que la mayoría de los profesionales sanitarios se hayan visto abrumados al intentar mantenerse actualizados en el dominio de las interacciones farmacológicas.

La ingeniería informática puede representar un papel fundamental en este campo permitiendo la identificación, explicación y predicción de DDIs, ya que puede ayudar a recopilar, analizar y manipular grandes cantidades de datos biológicos y farmacológicos. En concreto, las técnicas del procesamiento del lenguaje natural (PLN) pueden ayudar a recuperar y extraer información sobre DDIs de textos farmacológicos, ayudando a los investigadores y profesionales sanitarios en la complicada tarea de buscar esta información en diversas fuentes. Sin embargo, el desarrollo de estos métodos depende de la disponibilidad de recursos específicos que proporcionen el conocimiento del dominio, como bases de datos, vocabularios terminológicos, corpora u ontologías, entre otros, que son necesarios para desarrollar las tareas de extracción de información (EI).

En el marco de esta tesis hemos desarrollado dos recursos semánticos en el dominio de las interacciones farmacológicas que suponen una importante contribución a la investigación y al desarrollo de sistemas de EI sobre DDIs. En primer lugar hemos revisado y analizado los corpora y ontologías existentes relevantes para el dominio y, en

base a sus potenciales y limitaciones, hemos desarrollado el corpus DDI y la ontología para interacciones farmacológicas DINTO. El corpus DDI ha demostrado cumplir con las características de un estándar de oro de gran calidad, así como su utilidad para el entrenamiento y evaluación de distintos sistemas en la tarea de extracción de información *SemEval-2013 DDIExtraction Task*. Por su parte, DINTO ha sido utilizada y evaluada en dos aplicaciones diferentes. En primer lugar, hemos demostrado que esta ontología puede ser utilizada para inferir interacciones entre fármacos y los mecanismos por los que ocurren. En segundo lugar, hemos obtenido una primera prueba de concepto de la contribución de DINTO al área del PLN al proporcionar el conocimiento del dominio necesario para ser explotado por un prototipo de un sistema de EI. En vista de estos resultados, creemos que estos dos recursos semánticos pueden estimular la investigación en el desarrollo de métodos computaciones para la detección temprana de DDIs.

Este trabajo ha sido financiado parcialmente por el Gobierno Regional de Madrid a través de la red de investigación MA2VICMR [S2009/TIC-1542], por el Ministerio de Educación Español, a través del proyecto MULTIMEDICA [TIN2010-20644-C03-01], y por el Séptimo Programa Macro de la Comisión Europea a través del proyecto TrendMiner [FP7-ICT287863].

Palabras clave: Corpus, Ontología, Representación del Conocimiento, Farmacovigilancia, Interacción Farmacológica, Procesamiento del Lenguaje Natural, Extracción de Información, Inferencia.

Abstract

Nowadays, with the increasing use of several drugs for the treatment of one or more different diseases (polytherapy) in large populations, the risk for drugs combinations that have not been studied in pre-authorization clinical trials has increased. This provides a favourable setting for the occurrence of drug-drug interactions (DDIs), a common adverse drug reaction (ADR) representing an important risk to patients safety, and an increase in healthcare costs. Their early detection is, therefore, a main concern in the clinical setting. Although there are different databases supporting healthcare professionals in the detection of DDIs, the quality of these databases is very uneven, and the consistency of their content is limited. Furthermore, these databases do not scale well to the large and growing number of pharmacovigilance literature in recent years. In addition, large amounts of current and valuable information are hidden in published articles, scientific journals, books, and technical reports. Thus, the large number of DDI information sources has overwhelmed most healthcare professionals because it is not possible to remain up to date on everything published about DDIs.

Computational methods can play a key role in the identification, explanation, and prediction of DDIs on a large scale, since they can be used to collect, analyze and manipulate large amounts of biological and pharmacological data. Natural language processing (NLP) techniques can be used to retrieve and extract DDI information from pharmacological texts, supporting researchers and healthcare professionals on the challenging task of searching DDI information among different and heterogeneous sources. However, these methods rely on the availability of specific resources providing the domain knowledge, such as databases, terminological vocabularies, corpora, ontologies, and so forth, which are necessary to address the Information Extraction (IE) tasks.

In this thesis, we have developed two semantic resources for the DDI domain that make an important contribution to the research and development of IE systems for DDIs. We have reviewed and analyzed the existing corpora and ontologies relevant to this domain, based on their strengths and weaknesses, we have developed the DDI corpus and the ontology for drug-drug interactions (named DINTO). The DDI corpus has proven to fulfil the characteristics of a high-quality gold-standard, and has demonstrated its usefulness as a benchmark for the training and testing of different IE systems in the

SemEval-2013 DDIExtraction shared task. Meanwhile, DINTO has been used and evaluated in two different applications. Firstly, it has been proven that the knowledge represented in the ontology can be used to infer DDIs and their different mechanisms. Secondly, we have provided a proof-of-concept of the contribution of DINTO to NLP, by providing the domain knowledge to be exploited by an IE pilot prototype. From these results, we believe that these two semantic resources will encourage further research into the application of computational methods to the early detection of DDIs.

This work has been partially supported by the Regional Government of Madrid under the Research Network MA2VICMR [S2009/TIC-1542], by the Spanish Ministry of Education under the project MULTIMEDICA [TIN2010-20644-C03-01] and by the European Commission Seventh Framework Programme under TrendMiner project [FP7-ICT287863].

Keywords: Corpus, Ontology, Knowledge Representation, Pharmacovigilance, Drug-drug interaction, Natural Language Processing, Information Extraction, Inference.

Contents

List of figures

List of tables

List of annexes

Chapter 1

Introduction

1.1 Motivation

Recent technological developments and advances in the field of biomedicine have brought an increasing knowledge of molecular and cellular physiology, genomics, proteomics, and pharmacology. This has led to the generation of large amounts of experimental and computational biomedical data along with new discoveries, which are generally described, in the first instance, in research biomedical publications. Only considering the bibliographic database MEDLINE, the number of published research articles is increasing between 10,000 and 20,000 articles per week (NLM, 2014). The process of reviewing all the literature related to a biomedical or pharmacological subject is very time-consuming. Natural Language Processing (NLP) techniques can provide an interesting way to reduce the time spent by healthcare professionals and scientific researches on reviewing biomedical literature, as well as a promising approach for new knowledge discovery (Mack & Hehenbergerb, 2002).

Recently, one of the areas that have attracted a great deal of attention by the NLP research community is pharmacovigilance. Pharmacovigilance is formally defined by the World Health Organization (WHO) as the science and activities relating to the detection, assessment, understanding, and prevention of adverse effects (AEs) or any other drug-related problem (WHO, 2002). A type of common and important adverse drug reaction (ADR), having a significant impact on patient safety and healthcare costs, is drug-drug interactions (DDIs) (Aronson, 2007; Jankel, McMillan, & Martin, 1994; Pirmohamed et al., 2004). A DDI is the process that occurs when one drug affects the levels or effects of

another drug in the body. Although there is a large quantity of pharmacological databases and semi-structured resources − such as DrugBank (Wishart et al., 2006), Stockley (Baxter, 2013), and Drug Interactions Facts (Tatro, 2010), among others − to assist healthcare professionals in the prevention of DDIs, the quality of these databases is very uneven and the consistency of their content is limited, so it is very difficult to assign a real clinical significance to each interaction (Paczynski, Alexander, Chinchilli, & Kruszewski, 2012; Rodríguez-Terol et al., 2009). On the other hand, despite the availability of these databases, a large proportion of the most current and valuable information on DDIs is unstructured, written in natural language and hidden in published articles. A simple search for the term "drug-drug interactions" in the web search engine Google Scholar® increased from 63,500 results in January 2014 to 73,400 results in December 2014, and, at the time of this writing, 139,984 documents are indexed in the online library MEDLINE with the Medical Subject Headings (MeSH) term "drug interaction".

Computational methods can play a key role in the identification, explanation, and prediction of DDIs on a large scale, since they can be used to collect, analyze, and manipulate large amounts of biological and pharmacological data (Percha & Altman, 2013). On the one hand, several recent NLP systems have shown promising results in extracting DDIs from biomedical literature (Chowdhury & Lavelli, 2013b; Segura-Bedmar, Martínez, & de Pablo-Sánchez, 2011b; Segura-Bedmar, Martínez, & Herrero-Zazo, 2013; Segura-Bedmar, 2010; Thomas, Neves, Rocktäschel, & Leser, 2013). The major bottleneck for advancing in this area is, however, that these systems rely on specific resources providing the domain knowledge (databases, terminological vocabularies, corpora, ontologies, etc.) necessary to address the Information Extraction (IE) tasks. Although there is a wealth of linguistic resources for NLP in the biomedical domain, including terminologies and corpora, most of them are too broad to support the development of IE approaches applied to pharmacovigilance.

Furthermore, the same problem is encountered in the development of other computational tools to assist in pharmacovigilance, such as clinical decision support systems (CDSS) or signal detection systems for the early detection of ADRs and DDIs. These methods rely on the availability of computable representations of the general domain knowledge that can be understood and exploited by information systems (Olivié, 2007). Ontologies can be powerful tools representing pharmacovigilance knowledge and, specifically, the DDI domain. They can be used to integrate, in a common and harmonised framework, information from different resources, and can be exploited by reasoning engines systems to infer new knowledge, such as, as we describe in this thesis, the inference of DDIs. However, ontologies developed so far in this domain do not provide a comprehensive representation of the DDI domain, since they focus on shallow and partial representations of the domain.

1.2 Context

The occurrence of ADRs is one of the main concerns in the clinical setting, and DDIs are known to be a risk factor for their development (Ganeva, Gancheva, Troeva, Kiriyak, & Hristakieva, 2013). For example, in a study involving more than two thousand

patients, it was observed that 1.4% of hospitalized patients in medical wards experienced potentially preventable adverse drug events (ADE), 11.7% of them due to DDIs (Otero-López et al., 2006). Although not all exposures to combinations of potentially interacting drugs lead to the occurrence of an adverse clinical manifestation, DDIs have been described to be the fifth cause involved in ADRs (Vargas et al., 1997). A prospective study in patients admitted to an internal medicine service showed that the 43% of the studied patients was exposed at least to one potential DDI, of which a 14% showed a relationship with an ADR (Ibáñez, Alcalá, García, & Puche, 2008). It was observed too, as in similar studies (Ganeva et al., 2013; Klarin, 2007; Obreli Neto et al., 2012), that the number of interactions did relate to the number of prescriptions.

Nowadays, with the increasing use of several drugs for the treatment of one or more different diseases (polytherapy) in large populations, the risk for drug combinations that have not been studied in pre-authorization clinical trials has increased (Back & Else, 2013). Moreover, it has been shown that genetic factors can lead to differences in a drug's effect between individuals (Martiny & Miteva, 2013). Therefore, the consequence of a DDI can differ from one patient to another. In this scenario, it is vital to increase the efforts devoted to the prediction and prevention of DDIs (Huang et al., 2008).

With this deeper knowledge about DDIs and their related factors, there is an increase in the publication of new aspects of known DDIs and in the discovery of new DDIs, too. Therefore, DDI information resources should be updated continuously in order to reflect all this new information. As mentioned before, currently known DDIs are described in different sources, such as compendia, databases, or approved-drug information, such as the Summary of Product Characteristics (SPC) or Package Insert (PI). However, deficiencies and inconsistencies between different DDI information resources have been reported by different authors (Aronson, 2004; Barillot, Sarrut, & Doreau, 1997; Bergk, Haefeli, Gasse, Brenner, & Martin-Facklam, 2005; Hansten, Horn, & Hazlet, 2001; Nikolić & Ilić, 2013; Paczynski et al., 2012). Thus, the development of automatic methods for collecting, maintaining, and interpreting the information about drugs is crucial to achieve a real improvement in the early detection of DDIs.

The computer science research community has worked, on the recent years, in different approaches regarding ADRs and DDIs. On the one hand, prediction of unknown DDIs is a very attractive subject. There have been approaches based on ontologies and taxonomies (Arikuma et al., 2008; Cami, Manzi, Arnold, & Reis, 2013) in conjunction with statistical comparison of similar characteristics (Vilar et al., 2012). Digital workbenches for the quantitative prediction of DDIs (Bonnabry, Sievering, Leemann, & Dayer, 1999), computational inference methods (Gottlieb, Stein, Oron, Ruppin, & Sharan, 2012), or database analysis (Ito, Brown, & Houston, 2004; van Puijenbroek, Egberts, Heerdink, & Leufkens, 2000) are other strategies in this domain. Data signalling for the early detection of DDIs has been studied, too (Tatonetti, Fernald, & Altman, 2011; Thakrar, Grundschober, & Doessegger, 2007), and adverse events reporting systems have been used for data mining of DDIs (Harpaz et al. 2010; Iyer et al. 2013).

On the other hand, NLP has become a very active research area. Several research groups have developed different systems for the extraction of ADRs from a diverse set of resources, such as FDA drug labels or PIs (Bisgin, Liu, Fang, Xu, & Tong, 2011; Boyce, Gardner, & Harkema, 2012), electronic health records (EHR) and clinical notes (Lependu et al., 2013), scientific literature (Karnik, Subhadarshini, Wang, Rocha, & Li, 2011), or social network websites (Nikfarjam & Gonzalez, 2011). Information retrieval (IR) of

DDI-related research articles has been studied, too (Duda, Aliferis, Miller, Statnikov, & Johnson, 2005). Segura-Bedmar (2010) carried out the deepest study on the research field of NLP applied to the DDI domain, leading to several relevant achievements, such as the study of anaphora resolution for DDI extraction (Segura-Bedmar, Crespo, de Pablo-Sánchez, & Martínez, 2010) or the study of different IE techniques for the extraction of DDI relations (Segura-Bedmar, Martínez, & de Pablo-Sánchez, 2011a, 2011b, 2010). Moreover, one of the main contributions of this work was the development of the DrugDDI corpus, a gold-standard used in the first edition of the *DDIExtraction 2011 challenge task* (Segura-Bedmar, Martínez, & Sánchez-Cisneros, 2011). It was the first evaluation task designed to provide a framework for comparing different approaches to extracting DDIs from texts, and encouraged the research in this domain (Chowdhury & Lavelli, 2013a; Karnik et al., 2011).

As demonstrated by the interest raised by the *DDIExtraction 2011 challenge*, the development of NLP systems for the DDI domain relies heavily on manually annotated corpora for training and testing purposes (Bada et al., 2012). Similarly, the development of other pharmacovigilance supporting systems relies on resources providing the knowledge of the domain, such as ontologies (Cimiano, Unger, & McCrae, 2014; Wimalasuriya, 2010). However, there is a lack of these specific and appropriate resources for the DDI domain.

In summary, the availability of proper resources providing a deep knowledge of the pharmacological domain is necessary for the encouragement and support of computer science research groups working in pharmacovigilance. The combination of computer science techniques and pharmacological domain knowledge becomes essential for the development of new systems and techniques for the prediction and early detection of DDIs in pharmacovigilance. For this purpose, we intend to provide, within the framework of this thesis, two semantic resources necessary for the progress of this research field. The first one is a manually annotated corpus for training and evaluation of IE systems, and the second one is an ontology that will be validated and proven useful in two application domains: NLP of pharmacological text and inference of DDI knowledge.

1.3 Objectives

The objective of this thesis is twofold: (1) contributing to the improvement on the early detection of DDIs from scientific literature through the development of two different resources, an annotated corpus and a comprehensive ontology, which will enable the development, training, and evaluation of automatic NLP systems for pharmacological texts in the field of DDIs, and (2) the application of such ontology to infer new knowledge, in particular, new DDIs that could not have been reported in biomedical publications.

Concerning the first objective, most NLP techniques heavily rely on annotated corpora to learn models that can be used to extract information from raw text. Annotated corpora are valuable resources as they provide a gold-standard data for the reproducible automatic training and evaluation of machine learning-based NLP techniques (van Mulligen et al., 2012). Most recent research has focused on biological entities and their

relationships (such as gene and protein interactions), mainly because of the availability of annotated corpora in the biological domain. However, the extraction of relations in the pharmacological domain, such as DDIs, requires a corpus created and annotated specifically for these purpose. Unfortunately, to date the number of corpora annotated with DDIs is very small (see **Chapter 2** for a review of corpora annotated with pharmacological substances and DDIs). Moreover, they have been annotated considering a unique type of DDIs: pharmacokinetic DDIs (PK DDIs), excluding the annotation of pharmacodynamic DDIs (PD DDIs). These two types of DDIs relate to the type of mechanism preceding them. A pharmacokinetic (PK) mechanism occurs when the concentration of one drug is altered by another one (e.g., *ciprofloxacin* increases the blood levels of *duloxetine* by impairing its elimination from the body). A pharmacodynamic (PD) mechanism leads to an alteration on the effect of one drug without a variation of its levels in the body (e.g., concomitant use of two different drugs, such as *alcohol* and *sedative* pills, with a depressive effect in the central nervous system, can produce the potentiation of their effects and increase related symptoms, such as somnolence). This kind of PD mechanism is often described in the biomedical literature. However, there is no corpus annotated with PD DDIs. Therefore, the creation of a new annotated corpus including all possible types of DDIs is necessary for the development of NLP systems suitable for the comprehensive extraction of DDI information. Therefore, one of the main goals of this thesis is the creation of a large corpus annotated specifically for its final application in DDI extraction and including both types of DDIs: PK and PD DDIs.

On the other hand, ontologies and controlled vocabularies have been commonly used in NLP techniques applied to pharmacology. The main application of ontologies has been to support the construction of conceptual dictionaries or lists of terms for different pharmacological semantic categories (Segura-Bedmar et al., 2013). The integration of different ontologies to this purpose has proven to be useful for the Named Entity Recognition (NER) task (Grego & Couto, 2013; Lamurias, Grego, & Couto, 2013). However, different common issues, such as ambiguity, polysemy, synonymy, and spelling variations, cannot be addressed through the application of these dictionary-based methods. On the other hand, several works have applied lexical Relation Extraction (RE) to ontology learning and population (Poesio, Barbu, Giuliano, & Romano, 2008). However, using ontologies for RE has not been sufficiently studied. Despite this, ontologies, through the formal representation of relationships between different concepts, can be exploited in NLP techniques for both, NER and RE tasks, since they provide a contextual framework and semantic knowledge base. Therefore, there has been an increasing interest in bringing together traditional approaches on NLP and recent developments in the Sematic Web and ontological engineering fields (Cimiano et al., 2014), the application of ontologies to IE and IR (Hassanpour, O'Connor, & Das, 2011; Müller, Kenny, & Sternberg, 2004; Wimalasuriya, 2010; Zhang, Hoffmann, & Weld, 2012), or the use of ontologies to normalize relations extracted by NLP techniques (Coulet et al., 2011; Percha & Altman, 2012). Application of ontologies to the extraction of DDIs is limited, however, by the lack of appropriate resources (a review of related knowledge resources for DDIs is provided in **Chapter 5**). The main challenge is that DDIs can be described in many different ways in text. For example, these three sentences describe the same DDI, providing different information about it:

« The effects of **duloxetine**, including serotonergic syndrome, are increased by **ciprofloxacin**. » (i)

« **Ciprofloxacin** decreases the metabolism of **duloxetine** by inhibiting the enzyme CYP 2D6. » (ii)

« Treatment with **ciprofloxacin** is contraindicated in patients taking **duloxetine**. » (iii)

As this example shows, a DDI relationship is not simple: sentence (i) describes the consequence of the DDI; sentence (ii) explains how the DDI occurs; sentence (iii) provides a recommendation to avoid the DDI. No existing ontology represents all these different types of relationships between two interacting drugs. Moreover, related concepts, such as '*metabolism'* (a pharmacokinetic process), '*CYP 2D6'* (a metabolic enzyme), or '*serotonergic syndrome'* (an adverse effect), are not collected in a specific ontology for the recognition of DDI-related concepts. Therefore, one on the main contribution of this thesis is the development of a specific and appropriate ontology that systematically organizes all DDI-related information, necessary to provide a knowledge base to be integrated in NLP systems devoted to IE in pharmacovigilance.

In addition to this, such a comprehensive ontology might be useful for the development of other tools supporting pharmacovigilance-related activities. Therefore, the second main objective of this thesis is the application of the ontology to the inference of new knowledge and, in particular, to the inference of DDIs and their mechanisms. These inference capabilities might represent the first step in the development of further CDSS or signal detection systems, providing new approaches for the detection of DDIs.

The specific objectives to be achieved in this thesis are:

Objective 1 To study the annotated corpora relevant to the DDI domain.

Objective 2 To create the DDI corpus, a manually annotated corpus that will be a benchmark for IE of DDIs. Specifically, the following topics will be addressed:

a. annotation of pharmacological substances at different levels of granularity.
b. annotation of different types of DDIs.
c. annotation of different types of documents.
d. creation of annotation guidelines.
e. measurement of inter-annotator agreement (IAA) to assess the quality of the corpus.

Objective 3 To validate the DDI corpus as gold-standard for training and evaluation of NLP systems devoted to the NER and RE tasks in the DDI domain.

Objective 4 To study the different linguistic phenomena in text describing DDIs.

Objective 5 To study the main semantic resources in the pharmacological domain.

Objective 6 To analyse and compare current modeling approaches in the DDI domain.

Objective 7 To create an ontology for the representation of all DDI-related knowledge, including formal representation of:

a. different types of DDIs, including all possible mechanisms.

b. possible effects or consequences of DDIs.

c. different recommendations for avoiding a DDI.

Objective 8 To evaluate the ontology in different tasks:

a. prediction or inference of DDIs.

b. information extraction of DDIs from text.

1.4 Document structure

The layout of this thesis is split into two main parts. The first one focuses on the creation of the DDI corpus.

Chapter 2 reviews and compares the different corpora relevant in the DDI domain, including those annotated with pharmacological substances and those annotated with DDIs, too.

Chapter 3 describes the construction of the new DDI corpus, including the collection process, the pre-processing stage, and the annotation process. The main issues encountered during this annotation process are also described and exemplified in this chapter. Quantitative features of the DDI corpus are provided, and the corpus is compared to the corpora reviewed in **Chapter 2**.

Chapter 4 provides a detailed description of the evaluation of the DDI corpus. Firstly, results of the IAA are provided and discussed. Secondly, we describe the *SemEval-2013 DDIExtraction shared task*, where the DDI corpus is used as a gold-standard for training and evaluation of different IE systems. Their results are described in this chapter, along to an error analysis of the results of the DDI extraction systems, which allows for the identification of further improvements in the DDI corpus.

The second part of this thesis describes the creation of an ontology for drug-drug interactions: DINTO.

Chapter 5 reviews existing semantic resources in the pharmacological domain covering relevant aspects related to DDIs. These include terminological resources for chemical and pharmacological substances, adverse drug reactions (ADRs), and DDI-related aspects.

Chapter 6 reviews the state of the art on current modeling efforts that have dealt with the representation of some aspect of DDIs. Their different conceptualizations are analysed, represented in a common representation framework, and compared to identify strengths and weaknesses of current conceptual models in the DDI domain.

Chapter 7 describes the construction of DINTO and the different development activities. The final ontology is described in detail, too, and a summary of the available files is provided at the end of this chapter.

Chapter 8 introduces ontology evaluation and describes the strategy followed to evaluate DINTO, which combines technical and application-based evaluations. In particular, in this chapter the ontology is technically evaluated in its form and content to assess its consistency and expressivity, and to detect possible errors.

In **Chapter 9**, the ontology is evaluated in an application scenario for the inference of DDIs and their mechanisms. Firstly, we review the related work on computational inference of DDIs. Then, we describe three different experiments designed and conducted to evaluate, in a comprehensive way, the inference capabilities of DINTO.

Chapter 10 describes the evaluation of DINTO in a different application scenario. To this purpose, we provide a proof-of-concept of the contribution of DINTO to different NLP tasks, by providing the domain knowledge to be exploited by an IE pilot prototype.

Finally, **Chapter 11** highlights the main conclusions of this thesis, discusses the achievement of the proposed objectives, the dissemination and publication of our work, and highlights the future work on both the DDI corpus and DINTO as directions for further related research.

Chapter 2

Corpora in the pharmacological domain

The first main contribution of this thesis is the development of an annotated corpus for drugs and their interactions, which will be a benchmark to train and test DDI extraction systems. Gold-standard annotated corpora are necessary resources when building and evaluating NLP systems, since they provide correct annotations for both the training and evaluation of automatic systems.

A gold-standard corpus in NLP has been defined as a collection of texts manually annotated with the instances and/or relationships relevant to the specific tasks by one or more annotators (Deleger et al., 2012; Klein, Riazanov, Hindle, & Baker, 2014; Neves & Leser, 2014; Neves, 2014; Wissler, Almashraee, Monett, & Paschke, 2014). However, a quality gold-standard requires not only manual annotation, but its subsequent checking and correction (Baker, Hardie, & McEnery, 2006). Although it is not possible to avoid the introduction of errors in almost any manual task, incorrect annotations in the training corpus are propagated to the final system (Wissler et al., 2014). Therefore, the number of mistakes should be reduced to their minimum level. The most common way to ensure quality of the annotations is the analysis of the inter-annotator agreement (IAA) – a measure of the degree of concordance between the annotations made by different annotators (Cohen, 1960) –, and the creation of annotation guidelines – the document describing in detail the annotation process and the concepts to be annotated (Klein et al., 2014).

In addition to these characteristics, to be useful to the NLP community, a gold-standard should be rich in information and include large variety of documents and annotated instances that represent the diversity of document types and instances of interest for a specific task (Deleger et al., 2014). Finally, a wide acceptance and use by the NLP community converts a corpus into a de facto gold-standard, such as the GENIA corpus (Kim, Ohta, Tateisi, & Tsujii, 2003), which has been described to have the highest usage rate among several corpus designed for biomedical NLP (Cohen, Ogren, Fox, & Hunter, 2005).

Therefore, from the analysis provided above, we can conclude that the characteristics of a gold-standard must be:

1. Manual annotation.
2. High quality, ensured by:
    a. the measurement of the IAA,
    b. the creation of annotation guidelines.
3. Usefulness, by providing:
    a. rich information,
    b. diversity of document types and annotated instances,
    c. large number of documents and annotations.
4. Wide acceptance and use by the NLP community.

In this chapter, we analyse existing corpora annotated with entities and relationships relevant to the DDI domain. **Table 2.1** summarizes all the mentioned corpora, with a brief description of their main purpose, and **Figure 2.1** shows the timeline of pharmacological corpora annotated with pharmacological entities and DDIs.

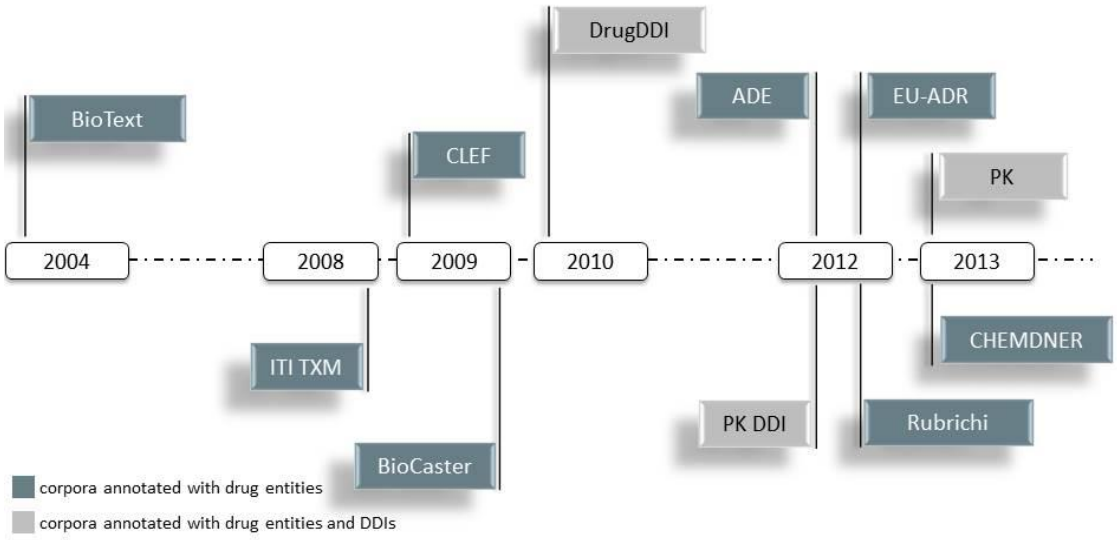

Figure 2.1. Timeline of pharmacological corpora included in this review

Corpora	Description
BioText (Rosario & Hearst, 2004)	BioText is a corpus created for the evaluation of relation extraction of treatment-disease relationships.
ITI TXM (Alex et al., 2008)	The Tissue Expressions and Protein–Protein Interactions (ITI TXM) corpus was created in the framework of the TXM project for development of tools for assisting in the curation of biomedical research papers. It was annotated with a broad range of biomedical entities and relationships between them.
CLEF (Roberts et al., 2009)	The Clinical E-Science Framework (CLEF) corpus was developed within the CLEF project to support extracting information from clinical patient reports.
BioCaster (Doan et al., 2009)	BioCaster is a manually annotated corpus created for the evaluation of a system aimed to detect outbreak diseases.
DrugDDI (Segura-Bedmar, 2010)	The DrugDDI corpus was created and used in the *DDIExtraction 2011 challenge* to promote research and provide a common framework for comparing the latest advances in information extraction techniques applied to the extraction of DDIs from biomedical texts.
ADE (Gurulingappa et al., 2012)	The Adverse Drug Effects (ADE) corpus was created to support the development and validation of methods for the automatic extraction of drug-related adverse effects from medical case reports.
PK DDI (Boyce et al., 2012)	The PK DDI corpus was created for the development of automated methods for identifying PK DDIs from drug package inserts.
Rubrichi (Rubrichi & Quaglini, 2012)	These authors created a small in-house annotated corpus for the automatic extraction of drug information conveyed in the Summary of Product Characteristics.
EU-ADR (van Mulligen et al., 2012)	The Exploring and Understanding Adverse Drug Reactions (EU-ADR) corpus was developed as part of the EU-ADR project. It contains annotations of multiple entities (drugs, diseases, and targets) and relationships between them.
PK (Wu et al., 2013)	The PK corpus was created to cover the domain of pharmacokinetic studies, including *in vivo* and *in vitro* PK DDI studies.
CHEMDNER (Krallinger et al., 2013)	The CHEMDNER corpus was created for the "*Chemical compound and drug name recognition*" *task* of the *BioCreative IV Challenge and Workshop*, which goal was to promote the implementation of systems to detect mentions of chemical compounds and drugs, in particular those chemical entity mentions that can subsequently be linked to a chemical structure.

Table 2.1. Summary of corpora annotated with pharmacological substances and corpora annotated with DDIs.

The identification of drug names is a preliminary and crucial step in many text mining tasks such as the detection of the outbreak of diseases (Doan et al., 2009), the extraction of medication-related information (Deléger, Grouin, & Zweigenbaum, 2010), the detection of adverse-drug events (Warrer, Hansen, Juhl-Jensen, & Aagaard, 2012), or the extraction of relationships such as drug-disease (Xu & Wang, 2013), drug-gene interactions (Sutton, Wojtulewicz, Mehta, & Gonzalez, 2012), or drug-drug interactions (Segura-Bedmar, Martínez, & de Pablo-Sánchez, 2011b), among many others. In fact, several corpora have been built for these purposes in recent years (**Figure 2.1**). Here, we review the main corpora annotated with drug entities (**Section 2.1**), giving a special focus on those corpora that also contain DDIs (**Section 2.2**). Then, in **Section 2.3**, we examine these resources with respect to the previously described characteristics of a gold-standard corpus. Finally, main conclusions are discussed in **Section 2.4**.

2.1 Corpora annotated with drug entities

Annotation of drug entities differs between corpora, especially from those created some years ago and the more recent ones. Since each corpus has been developed for a specific task, the definition of a drug entity varies significantly from corpus to corpus.

Thus, in **CLEF** (Roberts et al., 2009) and **BioText** (Rosario & Hearst, 2004) corpora drug names and therapeutic devices or interventions were annotated with the same entity type. Other corpora, such as **ADE** (Gurulingappa et al., 2012), **EU-ADR** (van Mulligen et al., 2012), or **ITI TXM** (Alex et al., 2008), used a single entity type to annotate both drugs and chemicals, while the **BioCaster** corpus (Doan et al., 2009) distinguished between substances for the treatment of diseases and chemicals not intended for therapeutic purposes.

In contrast, corpora such as **PK DDI** (Boyce et al., 2012) or that developed by **Rubrichi & Quaglini** (Rubrichi & Quaglini, 2012) proposed a more fine-grained classification of pharmacological substances. The annotation schema of the **PK DDI** corpus described three entity types to annotate pharmacological substances: ACTIVE INGREDIENT, DRUG PRODUCT and METABOLITE. Similarly, **Rubrichi & Quaglini**, proposed three different entity types: ACTIVE DRUG INGREDIENT, DRUG and DRUG CLASS. Similarly, the **CHEMDNER** corpus distinguished between eight subtypes of chemical named entities, the TRIVIAL type being the closest to drug mentions.

Finally, in the **DrugDDI** corpus (Segura-Bedmar, 2010), drugs were automatically recognized on the basis of drug-related Unified Medical Language System (UMLS) semantic types using the MetaMap Transfer tool (MMTx) (Aronson, 2001) and annotated as a unique entity DRUG. In a similar way, drug names and metabolites were assigned the same type DRUG in the **PK** corpus (Wu et al., 2013).

This diversity in the annotation of pharmacological substances among different corpora shows that there is not a consensus between annotators about which definition should be used for the drug entity, mainly because it is application dependent.

2.2 Corpora annotated with DDIs

To the best of our knowledge, only three works have addressed the annotation of DDIs: the **DrugDDI** corpus, the **PK DDI** corpus and the **PK** corpus. The first corpus annotated with pharmacological substances and DDIs was the **DrugDDI** corpus. It was developed in the framework of Dr. Segura-Bedmar's PhD thesis: '*Application of information extraction techniques to pharmacological domain*' (Segura-Bedmar, 2010). The objective of this work was the development and evaluation of IE techniques in biomedical documents, particularly for automatic detection of DDIs from unstructured text. The **DrugDDI** corpus was created and used in the *DDIExtraction 2011 challenge* (Segura-Bedmar, Martínez, & Sánchez-Cisneros, 2011). The goal of this task was to promote research and provide a common framework for comparing the latest advances in IE techniques applied to the extraction of DDIs from biomedical texts. This work was a starting point for research on such problems and has received much interest in the community of text mining, which is demonstrated by the number of papers that have appeared later and attempting to address the same problem. This work is the onset of this thesis, and can be considered as a preliminary version of the DDI corpus.

The **DrugDDI** corpus consisted of 579 documents describing drug interactions that were taken from the DrugBank database (Wishart et al., 2006). A total of 3,160 DDIs were manually annotated as relationships between two interacting drugs at a sentence level. However, this version presented important limitations. Firstly, drug names were automatically annotated without any manual intervention in the process. Secondly, no annotation guidelines were produced. Thirdly, the annotation was carried out by a single annotator, without pharmacological background. Finally, the quality of the corpus was not evaluated in terms of the IAA.

Another corpus created for the development of automated methods for identifying DDIs from texts is the **PK DDI** corpus (Boyce et al., 2012). It is a manually annotated corpus made up of FDA-approved drug package inserts (PIs), annotated with 592 DDIs. These documents were annotated only for a specific type of DDIs, PK DDIs. They were annotated as POSITIVES (if they asserted the existence of the DDI) or NEGATIVES (if the sentence denied the occurrence of the DDI) and as QUANTITATIVE (when the statement contained quantitative data) or QUALITATIVE (otherwise).

The last annotated corpus for DDIs is the **PK corpus** (Wu et al., 2013), which also focused on the annotation of PK DDIs. However, it consisted of MEDLINE abstracts that were manually annotated with a total of 1,333 DDIs. In this corpus, sentences describing DDIs were classified regarding their level of certainty, which could be mainly classified as types DDI, AMBIGUOUS DDI or NON-DDI.

2.3 Analysis of corpora features

2.3.1 Type of annotation procedure

As mentioned before, the first characteristic defining a gold-standard corpus is the type of annotation procedure: manual, automatic, semi-automatic or hybrid (Neves, 2014). In this section, we describe the different ways to create annotated corpora, and compare the processes adopted by the studied corpora.

Manual annotation is performed by annotators from scratch based on a pre-established annotation schema. Well-designed annotation processes, such as that in the annotation of the **ADE** corpus (Gurulingappa et al., 2012), establish a training period for annotators. During this period, they study a small set of texts from the corpus and contribute to the creation of a comprehensive annotation schema, adapted to the information and peculiarities that will be found in the text (Herrero-Zazo, Segura-Bedmar, & Martínez, 2013). However, it is a very time consuming task, and there might be a high number of missing annotations or other sources of error due to the human factor. The participation of several annotators in the annotation process contributes significantly to avoid these issues.

Corpus	Entity annotation	DDI annotation	Annotation guidelines	IAA
BioText	Manual	-	✓	×
ITI TXM	Manual	-	×*	✓
CLEF	Manual	-	✓	✓
BioCaster	Manual	-	×*	×
DrugDDI	Automatic	Manual	×	×
ADE	Manual	-	×*	✓
PK DDI	Semi-automatic	Manual	✓	✓
Rubrichi	Manual	-	×	×
EU-ADR	Automatic	-	×*	✓
PK	Automatic	Manual	×*	✓
CHEMDNER	Manual	-	✓	✓

Table 2.2. Annotation procedure, availability of annotation guidelines, and evaluation based on the IAA (* is used when authors mentioned the creation of specific annotation guidelines, but these are not available).

In contrast, automatic annotation of biomedical corpora relies on programs or tools that recognize and annotate biomedical terms by mapping them with pre-established vocabularies or dictionaries. The UMLS MetaMap Transfer tool (MMTx) (Aronson, 2001) is a highly configurable program to map biomedical text to the UMLS

Metathesaurus[1]. Moreover, MMTx enables the syntactic and semantic analysis of the documents by performing sentence splitting, tokenization, POS-tagging, shallow syntactic parsing, and linking of phrases with UMLS concepts. Those corpora completely derived from automated methods and never manually validated by experts are referred to as silver-standard corpora (Neves, 2014). Automatic annotation reduces the time required by the annotation process and those problems associated with the manual annotation. However, the quality of the annotation relies strongly on the characteristics of the vocabulary or dictionary used by the annotation tool. Therefore, problems such as incorrect annotation of ambiguous terms or missing annotation of spelling variations, abbreviations, or synonyms, are frequent. In fact, previous works have shown that the recognition of pharmacological substances cannot be performed properly by current automatic IE systems without human intervention (Jagannathan et al., 2009).

A common variant to manual annotation is semi-automatic approaches combining both manual and automatic annotation. First, documents are pre-annotated with an automatic tool, and then annotators review and annotate these pre-annotated documents. With this method, time consumed in the annotation process is reduced, as well as the number of missing annotations associated to manual annotation. If the reviewing process includes not only the validation of the automatic annotations, but is accompanied by a carefully reading of the text to identify missing ones, then this method provides the same quality that manual annotation.

Finally, some corpora combine automatic annotation for entities and manual annotation for relationships. They are known as hybrid corpora (Neves, 2014).The most important limitation of these resources is that they cannot be considered as gold-standards for the NER task, and that the quality of the annotation of relationships relies on the number of missing and incorrectly annotated entities.

Most corpora were manually annotated with pharmacological substances (**CLEF**, **BioText**, **BioCaster**, **ITI-TXM**, **ADE, CHEMDNER** or **Rubrichi and Quaglini**'s corpora), while automatic annotation was the method used in the **EU-ADR**, **PK** and **DrugDDI** corpora. Only the **PK DDI** corpus was annotated semi-automatically for pharmacological entities using a dictionary from the database DrugBank, and later manually reviewed. In contrast, all the DDI relationships were manually annotated (**Table 2.2**).

The type of annotations determines the suitability of the corpora for different IE tasks. On the one hand, corpora manually annotated with chemical entities could be suitable for chemical NER (**ITI-TXM**, **BioCaster**, **ADE,** or **CHEMDNER**), while those annotated specifically with pharmacological entities, such as **BioCaster**, **PK DDI**, **Rubrichi & Quaglini** or **CHEMDNER**, could be useful for drug NER. In addition to this, corpora with different annotation types for drug entities, such as the **PK DDI** or **Rubrichi & Quaglini's** corpora, could be suitable for drug named entity classification, too. However, as we explain in next sections, other aspects such as the quality and number of annotations should be considered when selecting a corpus for NER.

On the other hand, from the three resources annotated with DDIs, only the **PK DDI** and the **PK corpus** could be used in both DDI extraction and classification tasks. The former one only classified relationships as positives or negatives and as qualitative or

[1] http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

quantitative, while the latter one classified PK DDIs based on the level of certainty expressed in the sentence. These resources could be useful, therefore, for the development of systems training modality recognition (Aramaki et al., 2009). However, they are not suitable for classification of relationships based on DDI-related aspects, such as those describing the mechanism of a DDI or those describing the effect.

2.3.2 Corpora quality: annotation guidelines and IAA

Annotation guidelines are the documents defining the annotation task and the annotation conventions (Bird, Klein, & Loper, 2009). The extent and detail of these documents are related to the quality of the annotation process and the agreement between annotators. Moreover, the usefulness of the corpora depends on the quality of the annotation guidelines (Dipper, Götze, & Skopeteas, 2004). Therefore, it is necessary that the annotation guidelines be accessible to the final user of the corpus (Leech, 1993; Pustejovsky & Stubbs, 2012) and that these documents fulfil requirements such as explicitness and completeness (Dipper et al., 2004). However, some biomedical annotated corpora do not provide annotation guidelines or linguistic aspects are not described with sufficient level of detail (Lu, Bada, Ogren, Cohen, & Hunter, 2006). For example, the **CLEF corpus** was annotated based on exhaustive and well defined annotation guidelines. However, since the aim of this project was not the annotation of drugs, descriptions regarding their annotation are not detailed enough. In contrast, the **PK DDI corpus**' annotation guidelines are available to the research community and provide instruction for the annotation of drugs and PK DDIs. They include definitions, examples, and specific instructions regarding annotation aspects and use of the annotation tool. Although definitions and examples for all those entities describing drugs in their annotation schema are provided, drug names term variants are not explicitly described. Regarding DDIs annotation rules, however, explicit and clear instructions are provided (**Table 2.2**).

As mentioned before, most common way to ensure quality of the annotations is the analysis of the inter-annotator agreement (IAA), a measure of the degree of concordance between the annotations made by different annotators (C. Müller & Strube, 2006). It is commonly measured in terms of the standard Kappa statistic (Cohen, 1960), which takes into account that a certain degree of agreement between annotators can also be ascribed to chance. IAA enables the assessment of the quality and consistency of the corpus and the annotation guidelines, as well as the complexity of the annotation task.

The measurement of the IAA requires that more than one annotator annotates the same texts in the corpus. This is another important factor in the quality of the annotation process. Strategies of multi-annotation, in which all documents are annotated by more than one annotator, have been proposed as the best way to avoid problems related to manual annotation (Jagannathan et al., 2009; Wilbur, Rzhetsky, & Shatkay, 2006).

However, only some corpora provided IAA scores (**Table 2.2**). These included the **CLEF**, **ITI-TXM**, **EU-ADR**, **ADE**, **PK,** and **PK DDI** corpora. In general, the consistency in the annotation of drug entities was high (greater than 75%), while IAA scores reported for the **PK DDI** corpus (around 60%) suggested that annotation of DDIs is a more complex task than simply the identification of drug names.

2.3.3 Type of documents

Pharmacological information can be found in different texts, such as research articles, manually curated databases, patent documents, patient clinical records, clinical notes, and so forth. Manual annotation of these texts should be carried out by domain experts capable of understanding the information described on them. However, not all corpora are equally complex. Complexity varies depending on several facts, such as the source of the documents (e.g., manual curated databases or primary scientific literature), the type of study described in the text (e.g., clinical study or *in vitro* study), and content-related and linguistic aspects, such as the use of technical vocabulary, complex sentences, and so forth. The level of complexity of texts determines which annotators should be selected for the annotation task (e.g., a pharmacovigilance expert or a life science bachelor student), the length of training for annotators, and the explicit rules that should be described in the annotation guidelines. In summary, the complexity of the texts influences the manual annotation process and the subsequent IAA results. Similarly, the precision and recall obtained by automatic IE systems trained and tested in a corpus might vary considerably with the purpose and style of the texts used to train and/or test them (Jessop, Adams, Willighagen, Hawizy, & Murray-Rust, 2011).

Regarding the type of document, almost all of the aforementioned corpora were made up of MEDLINE abstracts (**BioText**, **ADE**, **EU-ADR**, **PK** and **CHEMDNER**), while full articles were used only in the **ITI TXM** corpora. Both the **PK DDI** corpus and that developed by **Rubrichi & Quaglini** consisted of texts taken from PIs, which are one of the most important sources of information for healthcare professionals and patients on the use of medicines. The **DrugDDI** corpus is the only corpus consisting of DrugBank documents (**Table 2.3**).

Corpus	Type of document	Size	Sentences	Drugs	DDIs
BioText	MEDLINE abstracts	*np*	*np*	*np*	-
ITI TXM	Full articles	400	*np*	18,000	-
CLEF	Patient records	150	*np*	197	-
BioCaster	Internet news	1,000 articles	*np*	1,022	-
DrugDDI	DrugBank documents	579	5,806	23,190	3,160
ADE	MEDLINE abstracts	*np*	4,272	5,063	-
PK DDI	Package Inserts	68	*np*	3,896	592
Rubrichi	Package Inserts	100 sections	*np*	*np*	-
EU-ADR	MEDLINE abstracts	300	*np*	1,753	-
PK	MEDLINE abstracts	428	5,026	*np*	1,333
CHEMDNER	MEDLINE abstracts	10,000	*np*	84,355	-

Table 2.3. Comparison of size, type of documents, and number of annotations for corpora annotated with drugs or chemicals and DDIs (we use *np* for those figures not provided by the authors).

2.3.4 Size and number of annotations

The size of corpora consisting of clinical texts is very small. For example, the **CLEF** corpus consisted of 150 patient records, which were annotated only with 197 drugs. Conversely, the **ITI-TXM** corpus were made up of 455 full text research articles and were annotated with almost 18,000 drug compounds. Usually, the size of the different corpora of MEDLINE abstracts never exceeded 500 abstracts, until the recent development of the **CHEMDNER** corpus, which included 10,000 abstracts annotated with 84,355 entities. Meanwhile, the **EU-ADR** corpus contained 300 MEDLINE abstracts annotated with 1,753 drugs, and the **ADE** corpus consisted of 4,272 sentences annotated with a total of 5,063 drugs. The size of the **PK** corpus was 428 MEDLINE abstracts divided into 5,026 sentences. In contrast, the **DrugDDI** corpus was made up of 5,806 sentences from 579 DrugBank documents. However, the latter contained 3,160 DDIs while the former 1,333. Meanwhile, in the **PK DDI** corpus, consisting of 208 multi-sentence sections from 64 PIs, which were annotated with a total of 3,986 drugs, the number of annotated DDIs was 592. A review of these metrics is shown in **Table 2.3**.

2.4 Discussion and conclusions

In this chapter, we have reviewed the primary corpora annotated with drugs, paying special attention to those annotated with DDIs. One of the purposes of this thesis is to create a corpus that can provide a benchmark framework for the automatic extraction of DDIs to cope with some of the limitations of the existing ones.

First, differences in the annotation of pharmacological substances among different corpora have been observed. The main reason is that each corpus has been developed for a specific task, leading to disparate definitions of the drug concept, which is annotated at different levels of detail. A preliminary analysis of texts describing DDIs showed that a DDI might be described between different types of drugs, such as generic drugs, brand drugs, groups of drugs or other substances not approved for human use (e.g., toxins or abused substances). References to different types of drugs are used in texts because they provide important information that can be used and interpreted by domain experts. For example, a healthcare professional knows that a DDI described for the drug *paracetamol* must be extrapolated to the brand drugs Gelocatil® and Efferalgan®, since they contain *paracetamol* as their active substance. He or she can infer, as well, that a document describing those different interactions for the drug product Gelocatil® would probably have been written by the manufacturer of this specific drug product. Moreover, a domain expert infers from a DDI described to occur between *duloxetine* and the group of drugs *CYP450 inhibitors* that the mechanism of the DDI is a reduction in the metabolism of *duloxetine.* In spite of this, none of the aforementioned corpora provided annotations for all these different types of drugs. One of our goals is, therefore, to create a corpus including specific annotations for all these different entity types.

Regarding DDI-annotated corpora, only three resources have been identified. The **DrugDDI** corpus proved to be useful as a training and evaluation resource in the *DDIExtraction 2011 challenge*. Moreover, its size and the number of annotated drugs and DDIs were higher than for the other related corpora. However, some aspects such as the

automatic annotation of drugs, the lack of annotation guidelines, or the evaluation through the IAA were shortcomings of this version to be a gold-standard. Other corpora, such as the **PK** corpus and the **PK DDI** corpus provided different types of annotated texts to those included in the **DrugDDI** corpus. However, they had a small size and only PK DDIs were annotated, thus excluding the equally important PD DDIs.

In general, we have observed that different corpora vary on the type of documents, size, and number of annotations. None of them addressed, however, the task of annotating different types of documents and studying their influence on training and testing machine learning systems. Moreover, none of the DDI-annotated corpora was annotated with different type of DDI information, such as effect, mechanism, or recommendation.

We try to overcome all these issues in this thesis. An important contribution of our work is to provide the first annotated corpus with PK and PD DDIs, and with a fine-grained annotation schema for pharmacological substances. In addition, our goal is that the corpus has a bigger size than those related corpora and provides different styles of texts, thus enabling training and development of NLP systems devoted to DDI extraction in different text sources. Finally, in order to assess the quality and consistence of the corpus, annotation guidelines should be created and the IAA should be measured.

To conclude this chapter, we provide a list of the specific limitations that will be addressed in this thesis.

1. There are significant differences in the annotation of drugs names among different corpora, since no existing work has focused on the study and annotation of drug classification and nomenclature variations, such as synonyms, abbreviations, acronyms, and so forth. For example, the same concept '*Nonsteroidal Anti-inflammatory drug'* can be expressed as well as '*Nonsteroidal Anti-inflammatory drugs'*, '*Nonsteroidal Anti-inflammatory agents*' or '*NSAID'*, among others. However, corpora annotated with pharmacological substances have not included these, or similar, aspects in the annotation.

2. Corpora annotated with pharmacological substances do not provide detailed guidelines with the different definitions of these substances and the annotation conventions or rules.

3. The study and identification of the issues affecting the manual annotation of pharmacological texts can be useful to anticipate the problems that will be encountered by NLP systems, since there is a relationship between the complexity of the manual annotation and the performance of automatic systems. However, there is not existing work studying the aspects influencing the annotation of pharmacological substances.

4. There is a lack of gold-standard corpora for the extraction of DDIs. Available corpora have been annotated only with a type of DDIs, PK DDIs. However, there are not corpora annotated with PD DDIs.

5. A DDI can be described in multiple ways regarding its mechanism, effect, or recommendation (see **Section 1.3** for an example). These aspects are crucial for the appropriate management of DDIs in the clinical setting (Bergk et al., 2005; Tatro, 2010) and, all together, provide the whole picture of a specific DDI. However, different sources of DDI information have been reported to have a lack of some of them (Bergk et al., 2005), leading to differences among different types of documents. In spite of this, there is not one annotated corpus classifying the type of information provided to describe each DDI.

6. Each of the existing DDI corpora consists of homogeneous documents (DrugBank documents, PIs, or MEDLINE abstracts). However, none of them includes different types of documents in the same corpus. This approach is interesting for the study of the influence of different texts in the annotation process and the performance of evaluated NLP systems.

7. As we explain in **Section 1.3**, DDIs are described in text in very different ways and there are multiple patterns used to describe a DDI relationship. Therefore, the number of annotations of a DDI corpus must be large enough to cover most of them in a comprehensive way. In addition to this, corpora are usually divided into two sets: one for training and one for testing purposes. In order to be useful for the development and evaluation of IE systems, both of them must contain a representative sample of these different annotated patterns. However, the limited size and number of annotations in the existing corpora annotated with DDIs might compromise their use as gold-standards for NLP purposes.

Chapter 3

The DDI corpus

As we have shown in the previous chapter, there is no existing corpus that can be considered a gold-standard for pharmacological substances and DDIs. The lack of an appropriate resource becomes a bottleneck to apply NLP techniques to the extraction of DDIs. To overcome this problem, we have developed a new corpus: the DDI corpus. This chapter describes its construction and annotation processes (**Section 3.1**, **Section 3.2** and **Section 3.3**). The main annotation issues identified during the annotation of the corpus are described in **Section 3.4**. Then, we outline the main characteristics of the corpus in terms of the frequency of entities and relationships in **Section 3.5**. Finally, in **Section 3.6** we compare the DDI corpus with previous corpora for DDIs reviewed in our related work, and highlight the main conclusions in **Section 3.7**.

3.1 Collecting the corpus

As mentioned before (**Section 2.2**), the origin of this thesis is the DrugDDI corpus (Segura-Bedmar, 2010), an antecedent version of the new DDI corpus that was used in the *DDIExtraction 2011 challenge* (Segura-Bedmar, Martínez, & Sánchez-Cisneros, 2011). The DrugDDI corpus consisted of 579 documents describing DDIs taken from the DrugBank database. To improve the quality and usefulness of the corpus, we have added 213 DrugBank new texts, and a second type of documents: abstracts from the MEDLINE database. We refer to the first subset of the corpus as the DDI-DrugBank dataset, while the latter one is referred to as the DDI-MEDLINE dataset.

3.1.1 The DDI-DrugBank dataset

The initial source of unstructured textual information on DDIs is the DrugBank database[2]. It is a free online resource created and maintained by the *Wishart Research Group* at University of Alberta (Canada), which contains information on a large number of pharmacological substances including small molecule drugs, biotech drugs, nutraceuticals, and experimental drugs. This database provides information oriented to biochemists and biologists regarding the nomenclature, structure, and physical properties of drugs and their drug targets. DrugBank also offers clinical information such as pharmacology, metabolism, and indications. Since its first release in 2006, it has become a very popular resource and has been widely used in several contexts, including drug repositioning (Pérez-Nueno, Karaboga, Souchet, & Ritchie, 2014), drug target discovery (Liu et al., 2014), and drug interaction prediction (Cheng & Zhao, 2014), among many other applications.

For each drug, DrugBank contains more than 100 data fields including drug synonyms, brand names, chemical formula and structure, drug categories, corresponding codes for the ATC and AHFS drug codes systems, mechanism of action, indication, dosage forms, and toxicity. In addition to this, DrugBank offers drug interaction information, which is manually curated by the DrugBank team. DDI information has changed along the different releases of the database. The version used in the creation of the corpus (DrugBank 2.1 (Wishart et al., 2008)) provides a detailed description of DDIs in unstructured text. Since this information has been manually curated, texts focus completely on the description of DDIs and use a language similar to that used in PIs (see example in **Section 3.6**). In contrast, most recent versions of DrugBank provide a semi-structured but briefer description of the DDIs. **Figure 3.1** and **Figure 3.2** show the description of the DDIs for the drug *heparin* in both versions DrugBank 2.1 and DrugBank 4.1, respectively.

Drug Interactions:

a. Drugs Enhancing Heparin Effect:
Oral anticoagulants: Heparin sodium may prolong the one-stage prothrombin time. Therefore, when heparin sodium is given with dicumarol or warfarin sodium, a period of at least 5 hours after the last intravenous dose or 24 hours after the last subcutaneous dose should elapse before blood is drawn if a valid prothrombin time is to be obtained.

Platelet inhibitors: Drugs such as acetylsalicylic acid, dextran, phenylbutazone, ibuprofen, indomethacin, dipyridamole, hydroxychloroquine and others that interfere with platelet-aggregation reactions (the main hemostatic defense of heparinized patients) may induce bleeding and should be used with caution in patients receiving heparin sodium.

The anticoagulant effect of heparin is enhanced by concurrent treatment with antithrombin III (human) in patients with hereditary antithrombin III deficiency. Thus in order to avoid bleeding, reduced dosage of heparin is recommended during treatment with antithrombin III (human).

b. Drugs Decreasing Heparin Effect:
Digitalis, tetracyclines, nicotine, or antihistamines may partially counteract the anticoagulant action of heparin sodium. Heparin Sodium Injection should not be mixed with doxorubicin, droperidol, ciprofloxacin, or mitoxantrone, since it has been reported that these drugs are incompatible with heparin and a precipitate may form.

Drug/ Laboratory Tests Interactions

Hyperaminotransferasemia: Significant elevations of aminotransferase (SGOT [S-AST] and SGPT [S-ALT]) levels have occurred in a high percentage of patients (and healthy subjects) who have received heparin sodium. Since aminotransferase determinations are important in the differential diagnosis of myocardial infarction, liver disease and pulmonary emboli, rises that might be caused by drugs (heparin sodium) should be interpreted with caution.

Figure 3.1. Description of the DDIs for the drug *heparin* in DrugBank version 2.1 (from (Segura-Bedmar, 2010))

[2] http://www.drugbank.ca/

Interacting Drug	Interaction Description	Possible Basis
Heparin		10 interactions
Acetylsalicylic acid	Increased risk of bleeding.	
Drospirenone	Heparin can increase risk of hyperkalemia for patients on drospirenone	
Ticlopidine	Increased bleeding risk. Monitor aPTT.	
Tobramycin	Increased risk of nephrotoxicity	
Treprostinil	The prostacyclin analogue, Treprostinil, increases the risk of bleeding when combined with the anticoagulant, Heparin. Monitor for increased bleeding during concomitant thearpy.	
Aprotinin	Aprotinin, in the presence of heparin, has been found to prolong the activated clotting time (ACT) as measured by a celite surface activation method. The kaolin activated clotting time appears to be much less affected.	

Figure 3.2. Description of the DDIs for the drug *heparin* in DrugBank version 4.1

The DDI-DrugBank dataset includes a total of 792 documents from DrugBank 2.1. From them, 579 made up the DrugDDI corpus, while the remaining 213 have been included in this new version. We use the Kapow's free RoboMaker screen-scraper[3] to download the interaction documents, which are then analysed by the UMLS MMTx (**Section 3.2**).

3.1.2 The DDI-MEDLINE dataset

The MEDLINE database is a free, publicly available service of the National Library of Medicine (NLM)[4]. It is a primary repository for biomedical peer-reviewed journal articles, which has become a valuable resource to the health care professionals and biomedical research community. These abstracts provide an appropriate type of text for training and testing of NLP systems, since they represent the scientific literature where new pharmacological discoveries are described. Therefore, mining these types of documents has a great interest for the biomedical community (Kim, Ohta, & Tsujii, 2008). In contrast to DrugBank, MEDLINE texts are usually written in a very scientific language, and the main topic of the scientific texts would not necessarily be on DDIs.

Document selection for the DDI-MEDLINE corpus has been carried out against PubMed[5], the search engine for the MEDLINE database. An initial set of documents is selected from PubMed using a query with ''drug interactions'' as MeSH term. This query return 116,919 citations (published between 1975 and 2011) of which 233 documents are randomly selected for annotation. Documents without an abstract section are discarded.

[3] http://openkapow.com/
[4] http://www.nlm.nih.gov/
[5] http://www.ncbi.nlm.nih.gov/pubmed

3.2 Processing the corpus

The aim of this activity is processing the documents to generate a corpus in Extensible Markup Language (XML) format and annotated with shallow syntactic and semantic information. The latter one refers to an initial annotation of drug entities in the corpus, which will be subsequently reviewed during the manual annotation. This semi-automatic approach assists annotators during the annotation process, reducing the task duration and the number of missing entities (Fort & Sagot, 2010). Here, we review the main NER tools available for processing of biomedical text, and describe the processes used to analyse the texts conforming the DDI corpus.

3.2.1 Review of the main NER tools for biomedical text

In order to choose the best option for our task, we make a detailed review of the main tools for NER task in the biomedical domain. Although current efforts, such as the *CHEMDNER task* at the *Fourth BioCreative challenge* (Krallinger et al., 2013), have encourage research into the development of chemical NER systems, the number of freely available tools in the past years was small. Indeed, to the best of our knowledge, only five tools were available when we started the annotation of the DDI corpus. They are summarized in **Table 3.1** and described below. In order to provide an updated description of available tools for chemical NER, we also include the recently developed **CheNER** tool.

In this review, we have observed that one of the most important differences among chemical NER tools is the type of named chemical entities that they recognize. Chemical named entities can be divided into two main groups: systematic and common names. Systematic nomenclatures use a set of rules to generate systematic names for chemical compounds, being the most frequently used worldwide the IUPAC (International Union of Pure and Applied Chemistry) nomenclature (McNaught & Wilkinson, 1997). Authors specialized in chemical NER assert that IUPAC names are usually complex multiword terms including punctuation marks, sequences of numbers separated by commas, and so forth (e.g., *N-(4-hydroxyphenyl)acetamide*) calling for a classification-based tool, whereas common names follow hardly any rule (e.g., *acetaminophen*) and are best captured by an exhaustive dictionary (Rocktäschel, Weidlich, & Leser, 2012; Usié, Alves, Solsona, Vázquez, & Valencia, 2014).

Therefore, dictionary-based approaches are used to identify drug names, abbreviations, trivial names, molecular formulas, and family names, whereas machine learning techniques are applied for NER of IUPAC chemical entities. In our case, the DDI corpus focuses on the former type of entities, being less common those mentions to IUPAC-like entities.

To the best of our knowledge, the first freely available chemical NER tool was the **Open-Source Chemistry Analysis Routines (OSCAR),** whose latest version is OSCAR4 (Jessop et al., 2011). It is a system for the recognition of chemical entities, which has been developed since 2002. It offers different methods to recognize chemical

entities, including dictionary-based approaches, predetermined regular-expression, or machine learning in the form of a Maximum Entropy Markov Model (MEMM).

Tool	Purpose	Method	Syntactic analysis
OSCAR	NER in chemistry publications	MEMMs	×
Jochem	Identification of small molecules and drugs in text	Dictionary-based approach combining information from different sources	×
ChemSpot	Identification of chemical mentions in natural language texts, including trivial names, drugs, abbreviations, molecular formulas, and IUPAC entities	CRF for IUPAC named recognition Dictionary-based approach for drug named recognition (ChemIDPlus)	×
MetaMap	To provide access to the concepts in the UMLS Metathesaurus from biomedical text	UMLS mapping	✓
DrugNER	Identification and classification of drug named entities	MetaMap + INN rule-based system	✓ (MetaMap)
CheNER	Recognition of IUPAC chemical names	Machine learning based on CRF	×

Table 3.1. Comparison of text processing tools

Although not a NER tool, the **Joint Chemical Dictionary** (**Jochem**) should be mentioned in this review. It is a dictionary for the identification of small molecules and drugs in text, which combines data extracted and from different resources: UMLS, MeSH, ChEBI, DrugBank, KEGG, HMDB, and ChemIDplus. Hettne et al. (2009) used the dictionary and their concept recognition software Peregrine (Schuemie, 2007) combined with disambiguation rules to extract chemical entities from SCAI (Kolárik, Klinger, Friedrich, Hofmann-Apitius, & Fluck, 2008), a corpus annotated with chemical entities, and compared the performance for each individual dictionary alone and for Jochem. Although results obtained with the combined resource outperformed those obtained for individual resources, the performance of a dictionary based on ChemIDplus alone was comparable to the performance of the combined dictionary.

ChemSpot is a hybrid system combining Conditional Random Fields (CRF) for IUPAC named recognition, and a dictionary built from ChemIDplus for drug names (Rocktäschel et al., 2012). The ChemSpot dictionary-matching component uses the previously mentioned ChemIDplus dictionary processed by Hettne et al., (2009). However, ChemSpot includes a different post-processing architecture based on the

LINNAEUS software (Gerner, Nenadic, & Bergman, 2010) and match expansion instead of the dictionary-matching software Peregrine.

A well-known and very popular tool that has been broadly used for NER task in the biomedical domain is the **MetaMap Transfer tool (MMTx)** (Aronson & Lang, 2010). This is a NLP engine created and maintained by the NLM that maps free text to biomedical concepts in the UMLS Metathesaurus (Bodenreider, 2004). In addition to this, MMTx performs lexical and syntactic analysis of the text, including tokenization and sentence boundary, acronym/abbreviation identification, part-of-speech tagging or lexical lookup of input words, which are the pre-processing steps that must be performed prior to any named recognition step (Eltyeb & Salim, 2014).

In an attempt to increase the performance of MMTx for drug named entity recognition, Segura-Bedmar et al. (2008) developed **DrugNER**, a drug name recognition and classification system that combines the information obtained by the MMTx program and a set of nomenclature rules recommended by the WHO International Nonproprietary Names (WHOINN) Program (WHO, 2006).

Finally, to the best of our knowledge the most recently developed chemical NER tool is **CheNER**, a machine learning application based on CRFs, which has been created and trained for the recognition of IUPAC chemical entities in text. The authors reported that CheNER performed better than **OSCAR4** and **ChemSpot** in identifying IUPAC names. However, results for non-IUPAC names were lower.

While **CheNER** has been created specifically for IUPAC named recognition, **OSCAR** does not differentiate between entity types. However, Hette et al. (2009) and Rocktäschel et al. (2012) showed in two different studies that a dictionary-based system employing **Jochem**, and another one based on the dictionary-component of **ChemSpot** alone, respectivelly, outperformed **OSCAR** for chemical NER on the SCAI corpus. On the other hand, the performance of the combined dictionary **Jochem** was comparable to those obtained by the ChemIDplus dictionary alone, while being the latter one substantially smaller. In addition to this, **ChemSpot**, which uses as dictionary the same ChemIDplus dictionary, obtained better results than the **Jochem**-based system. This ChemIDplus dictionary was created from concepts extracted from the NLM ChemIDplus database[6]. In contrast, **MMTx** relies on a rich and updated terminological resource, the UMLS Metathesaurus, which covers not only drugs, but chemical names, too.

UMLS is a set of resources developed by the NLM, whose main objective is to assist in the developing of natural language technology for biomedical texts. UMLS has three major knowledge sources: the Metathesaurus, the Semantic Network and the Specialist Lexicon. The Metathesaurus is a large vocabulary containing concepts, concept names, and other attributes from more than 100 terminologies, classifications, and thesauri. These include MeSH (Lipscomb, 2000), RxNorm (Nelson, Zeng, Kilbourne, Powell, & Moore, 2011), SNOMED CT (Stearns, Price, Spackman, & Wang, 2001), and the ATC classification system (WHO, n.d.), among many others. All concepts in the Metathesaurus are assigned to at least one semantic type from the UMLS Semantic Network, providing a consistent categorization of all concepts represented in the UMLS Metathesaurus. The Semantic Network contains 135 semantic types such as 'Pharmaceutical substance', 'Amino Acid, Peptide, or Protein', 'Disease or Syndrome' or

[6] http://chem.sis.nlm.nih.gov/chemidplus/

'Gene or Genome'. Therefore, in addition to drugs and chemical entities, UMLS includes other concepts related to the DDI domains, such as diseases, signs and symptoms, proteins, and so forth. **MMTx** is highly configurable and allows the selection of the vocabularies or data models, including selection by sematic types, to be used. This is an interesting characteristic for our project, since it would enable a further annotation of relevant concepts (such as ADRs or proteins) in future work. Finally, the Specialist Lexicon is a biomedical lexicon with syntactic, morphological, and orthographic information.

Another advantage of **MMTx** over other tools is that it performs shallow syntactic analysis, such as sentence splitting. Moreover, at least one enhanced release is launched every year, and has been widely used for text mining applications in the biomedical domain (Bashyam, Divita, Bennet, Browne, & Taira, 2007; Meystre & Haug, 2006; Yip, Mete, Topaloglu, & Kockara, 2010).

For all these reasons, we have selected MMTx as our text processing tool for the syntactic and semantic analyses of the documents in the corpus.

3.2.2 Analysing the texts

This section focuses on the description of the processes used to analyse the texts. The main processes perform by MMTx are represented in **Figure 3.3**.

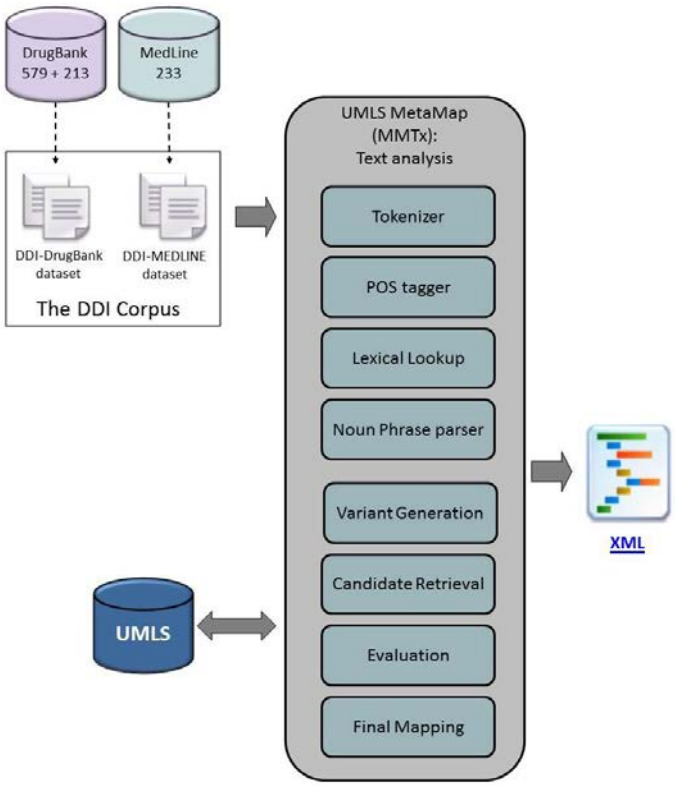

Figure 3.3. MMTx processes

During the syntactic analysis, the first activity consists in tokenization, sentence boundary determination, and acronym/abbreviation (AA) identification. Then, during the part-of-speech (POS) tagging, phrases in each sentence are identified and classified. The different types of phrases that can be assigned by MMTx are shown in **Table 3.2**. In case the type could not be determined, the phrase is annotated with the label UNK, which indicates that the corresponding type is "unknown".

Type of phrase	Examples
Noun phrase (NP)	Drug Interactions, the cytochrome P450 3A4 enzyme system
Prepositional phrase (PP)	with drugs, of azole antimycotics, of orally administered
Verbal phrase (VP)	administered, inhibit, decrease
Adjectival phrase (ADJ)	hypersensitive
Adverbial phrase (ADV)	concurrently, no, to significantly
Conjunctions (CONJ)	and, or, since

Table 3.2. Types of phrases identified by MMTx

As shown in **Figure 3.4**, MMTx annotates each phrase with its type, the number of tokens, the text, and an identifier in the XML document. The next process is a lexical lookup of input words in the SPECIALIST lexicon (McCray, Aronson, Browne, & Rindflesch, 1993), and the assignment of the POS tags to the tokens. In case that a token has several tags in the lexicon, MMTx uses the Xerox part-of-speech tagger (Cutting, Kupiec, Pedersen, & Sibun, 1992) to select the correct one. The final syntactic analysis consists of a shallow parse in which phrases and their lexical heads are identified by the SPECIALIST minimal commitment parser (Mccray, Srinivasan, & Browne, 1994). In this way, each token is annotated with its POS tag, its word, and a boolean value indicating if it is the head of the phrase (ISHEAD). In addition, the starting and ending offsets of each token within the text are stored in the attributes start and end, respectively. These character offsets allow mapping from the annotation to the raw text easily.

Once the shallow syntactic parsing has been performed, MMTx looks for the phrases in the UMLS Metathesaurus. For each phrase, a set of variants is generated using the SPECIALIST lexicon and linguistic techniques. The set of variants consists of the text of the phrase and its acronyms, abbreviations, synonyms, and derivational, inflectional, and spelling variants. MMTx looks these variants up in the Metathesaurus and retrieves those concepts containing at least one of them, which are considered candidates. Each candidate is evaluated against the text of the phrase using several linguistic metrics to determine its similarity. Finally, those concepts with a highest similarity are selected as the final mapping. For each concept in the final mapping set, MMTx provides the concept unique identifier (CUI), the concept name, and the semantic types.

```
- <document id="DDI-DrugBank.d505">
  - <sentence id="DDI-DrugBank.d505.s0" text="No formal drug/drug interaction studies with Plenaxis were performed.">
      <entity id="DDI-DrugBank.d505.s0.e0" charOffset="45-52" type="brand" text="Plenaxis"/>
    </sentence>
  - <sentence id="DDI-DrugBank.d505.s1" text="Cytochrome P-450 is not known to be involved in the metabolism of
    Plenaxis.">
      <entity id="DDI-DrugBank.d505.s1.e0" charOffset="66-73" type="brand" text="Plenaxis"/>
    </sentence>
  - <sentence id="DDI-DrugBank.d505.s2" text="Plenaxis is highly bound to plasma proteins (96 to 99%).">
      <entity id="DDI-DrugBank.d505.s2.e0" charOffset="0-7" type="brand" text="Plenaxis"/>
    </sentence>
  - <sentence id="DDI-DrugBank.d505.s3" text="Laboratory Tests Response to Plenaxis should be monitored by measuring
    serum total testosterone concentrations just prior to administration on Day 29 and every 8 weeks thereafter.">
      <entity id="DDI-DrugBank.d505.s3.e0" charOffset="29-36" type="brand" text="Plenaxis"/>
      <entity id="DDI-DrugBank.d505.s3.e1" charOffset="83-94" type="drug" text="testosterone"/>
    </sentence>
  - <sentence id="DDI-DrugBank.d505.s4" text="Serum transaminase levels should be obtained before starting treatment with
    Plenaxis and periodically during treatment.">
      <entity id="DDI-DrugBank.d505.s4.e0" charOffset="76-83" type="brand" text="Plenaxis"/>
    </sentence>
    <sentence id="DDI-DrugBank.d505.s5" text="Periodic measurement of serum PSA levels may also be considered."/>
  </document>
```

Figure 3.4. Example of a document processed by MMTx

Finally, the output of MMTx is transformed into XML format, which follows the Document Type Defintion (DTD) shown in **Figure 3.5**.

```
<!ELEMENTdocument          (sentence*) >
<!ELEMENTsentence  (entity*,ddi*) >
<!ELEMENTentity     EMPTY>
<!ELEMENTddi                EMPTY>

<!ATTLISTdocument
          id                ID        #REQUIRED>

<!ATTLISTsentence
          id                ID        #REQUIRED
          text       CDATA  #IMPLIED>

<!ATTLISTentity
          id                ID        #REQUIRED
          charOffset  CDATA  #IMPLIED
          type        CDATA  #IMPLIED
          text        CDATA  #IMPLIED>

<!ATTLISTddi
          id                ID        #REQUIRED
          e1         CDATA  #IMPLIED
          e2         CDATA  #IMPLIED
          typeCDATA         #IMPLIED>
```

Figure 3.5. DTD for the XML files in the DDI corpus

As shown in the Figure above, the root element is the <document> element with the following attribute:

- id: an unique id that is composed by the name of the corpus (DDI-DrugBank or DDI-MEDLINE) and an identifier beginning with "d" and followed by a number.

The next element is the <sentence> element. Each one of them has the following attributes:

- id: a unique id which is composed by the name of the corpus (DDI-DrugBank or DDI-MEDLINE), the id of the document (d505), and an id beginning with "s" and followed by the index of the sentence (the index of the first sentence should be 0).
- text: contains the text of the sentence.

Within the <sentence> element, there are the following elements: <entity> and <ddi>. Elements of type <entity> correspond to all annotated pharmacological substances, while elements of type <ddi> correspond to all annotated drug-drug interactions. Each <entity> element has the following attributes:

- id: a unique id that is composed by the name of the corpus (DDI-DrugBank or DDI-MEDLINE) , the id of the document (d505), the id of the sentence, and an id beginning with "e" and followed by the index of the entity in the sentence (the first entity of the sentence should have the index 0).
- charOffsets: contains the start and end positions, separated by a dash, of the mention in the sentence. When the mention is as discontinuous name, it will contain the start and end positions of all parts of the mention separated by semicolon.
- text: stores the text in the mention.
- type: stores the type of the pharmacological substance (`drug`, `brand`, `group` or `drug_n`)[7].

Similarly, each <ddi> element has the following attributes:

- id: a unique id that is composed by the name of the corpus (DDI-DrugBank or DDI-MEDLINE), the id of the document (d505), the id of the sentence, and an id beginning with "d" and followed by the index of the ddi in the sentence (the first ddi of the sentence should have the index 0).
- e1: stores the id of the first interacting entity.
- e2: stores the id of the second interacting entity.
- type: stores the type of the drug-drug interaction (`advice`, `effect`, `mechanism`, `int`)[7].

The XML format provides maximum flexibility for the use of the DDI corpus. In addition, the corpus is distributed in a standoff annotation format that involves storing annotation and text separately (Leech, 1993). An advantage of the standoff annotation format is that the original texts can be immediately retrieved without need of recovering

[7] These types are described in Section 3.1.1.

it from the annotations. Furthermore, this format preserves useful information about the structure of the texts.

3.3 Annotating the corpus

As we have described in the previous section, we carry out an automatic pre-annotation of drug entities with MMTx. This tool allows for the recognition of a variety of biomedical entities occurring in texts, which correspond to different semantic types of the UMLS Metathesaurus. In prior work the UMLS Semantic Network was reviewed, and those semantic types including drugs were selected (Segura-Bedmar, 2010). However, automatic annotation leads to certain issues, such as incorrect annotation of ambiguous terms or missing entities. Therefore, the annotation process for the DDI corpus includes a manual review of the texts and their automatic annotations, plus the manual annotation of DDI relationships. All the decisions made during the annotation process, as well as the description of the annotation process and the annotated entities and relations, are documented in form of annotation guidelines (Cohen et al., 2005).

3.3.1 Annotation guidelines

Annotation guidelines are the documents defining the annotation task and the annotation conventions (Bird et al., 2009). Their extent and detail are related to the quality of the annotation process, and the final agreement between annotators (Corbett, Batchelor, House, & Teufel, 2007). Moreover, the usefulness of the corpora has been defined to be dependent on the quality of the annotation guidelines, which should fulfil requirements such as explicitness and completeness (Dipper et al., 2004). Since they provide a detailed description of the annotations and, thus, of the corpus itself, it is important to make them available along with the annotated corpora (Leech, 1993; Pustejovsky & Stubbs, 2012). Therefore, the annotation guidelines for the DDI corpus are publicly available and can be downloaded from http://www.cs.york.ac.uk/SemEval-2013/task9/data/uploads/annotation_guidelines_ddi_corpus.pdf.

These annotation guidelines provide clear and accurate definitions for all those relevant entities and relationships described in the annotation schema (**Figure 3.6**). This document also contains the rules and conventions on how the annotation task should be carried out as well as providing examples clarifying their use.

As shown in **Figure 3.6**, four entity types are proposed to annotate pharmacological substances: `drug`, `brand`, `group` and `drug_n`.

- `drug`: The `drug` type is used to annotate human medicines known by a generic name (e.g., *ciprofloxacin*, *acetaminophen*).

- `brand`: Drugs described by a trade or brand name are annotated as `brand` entities (e.g., *Adacip*®, *Gelocatil*®). A drug medication frequently has several

brand names since different companies can market it. The use of a brand-name drug instead of its generic name might be related to a higher risk of ADEs (Hochman, Hochman, Bor, & McCormick, 2008; Steinman, Chren, & Landefeld, 2007). The use of either generic or brand names depends on the drug information source. Thus, while generic names are used in medical and pharmacological textbooks and scientific medical journals, brand names are to be used in drug product labels (SPCs and PIs).

- `group`: Since the descriptions of DDIs involving groups of drugs are very common in texts, we decide to include the `group` type to annotate groups of drugs (e.g., *quinolones, NSAIDs*). Extrapolation of drug interactions involving a specific compound to interactions involving its group is a common procedure in some DDI information sources. However some authors have established that this procedure is wrong because this generalization is not true for all drugs (Aronson, 2004; Bergk et al., 2005).

- `drug_n`: The last entity type refers to active substances not approved for human use, such as, and among others, toxins or pesticides (e.g., *picrotoxin, MPTP*). This type is included because interactions between drugs and substances not approved for human use are frequently reported in MEDLINE documents.

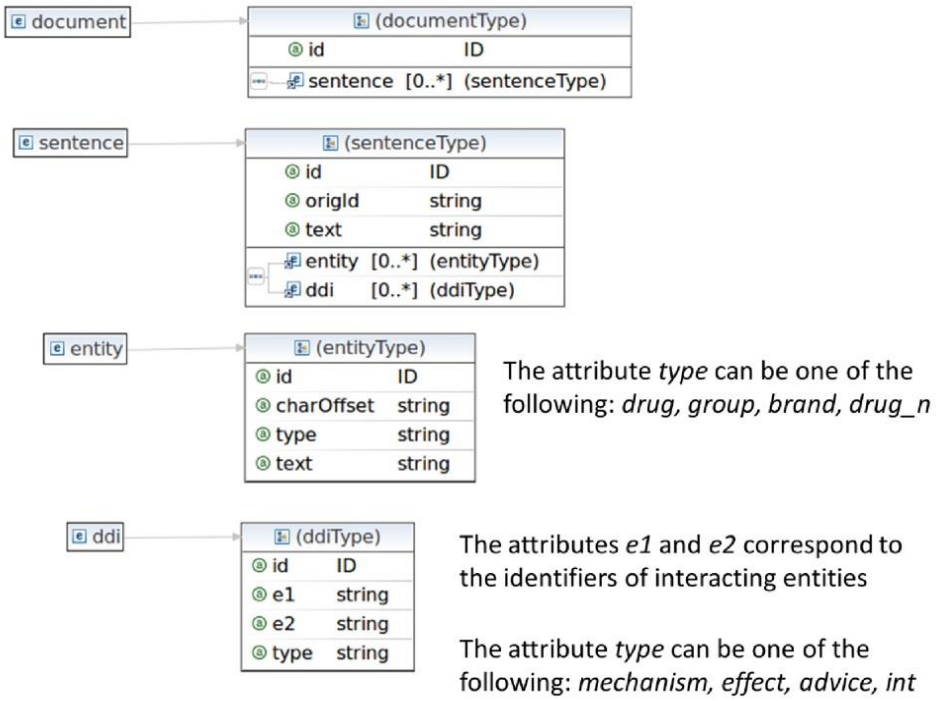

Figure 3.6. Annotation schema in the DDI corpus

Concerning the relationships, four different types of DDI relationships are proposed:

- `mechanism`: This type is used to annotate DDIs that are described by their PK mechanism (e.g., *Grepafloxacin may inhibit the metabolism of theobromine*).

- `effect`: This type is used to annotate DDIs describing an effect (e.g., *In uninfected volunteers, 46% developed rash while receiving SUSTIVA and clarithromycin*) or a PD mechanism (e.g., *Chlorthalidone may potentiate the action of other antihypertensive drugs*).

- `advice`: This type is used when a recommendation or advice regarding a DDI is given (e.g., *UROXATRAL should not be used in combination with other alpha-blockers*).

- `int`: This type is used when a DDI appears in the text without providing any additional information (e.g., *The interaction of omeprazole and ketoconazole has been established*).

Figure 3.7 and **Figure 3.8** show sentences describing DDIs. In **Figure 3.7**, the first sentence describes two interactions: `effect` and `mechanism`, and the last one also describe a DDI of `effect` type. In **Figure 3.8**, DDIs of `effect` type are described between *fenfluramine* and a group of drugs, *antihypertensive drugs*, as well as with some of its members (*guanethidine*, *methyldopa*, *reserpine*). The last sentence gives an advice to avoid a DDI.

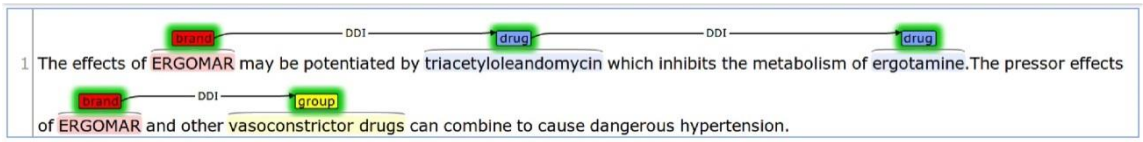

Figure 3.7. Examples of DDIs: `effect` and `mechanism` types (from the stav text annotation visualizer[8] (Stenetorp & Topi, 2011))

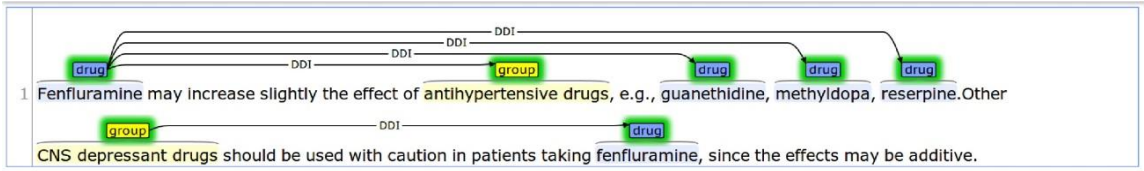

Figure 3.8. Examples of DDIs: `effect` and `advice` types (from the stav text annotation visualiser[8] (Stenetorp & Topi, 2011))

The proposed classification of DDIs is consistent with the information requirements established by pharmacology experts for an appropriate management of DDIs in the

[8] http://http://corpora.informatik.hu-berlin.de/

clinical setting (Aronson, 2004; Bergk et al., 2005). For this purpose, healthcare professionals should be provided with information on how the interaction occurs (mechanism), what consequences can be expected (effect) and how it can be managed to avoid or reduce the associated risk (advice).

Furthermore, this classification is useful to reflect the type of information provided from different sources. Thus, drug product labels provide little advice on how to minimize the risk of an interaction, whereas PK descriptions are very common in these documents (Bergk et al., 2005). On the other hand, DDI compendia (such as Stockley's Drug Interactions (Baxter, 2013) or Drug Interaction Facts (Tatro, 2010)) also contain considerable information on advice regarding drug interactions.

Although the principal aim of the annotation task is annotation at the semantic level, linguistic phenomena usually arise during the annotation process. The reason is that grammar and meaning are intertwined and, therefore, most annotation efforts should combine the two (Simpson & Demner-Fushman, 2012). Therefore, quality annotated corpora should be annotated taking into account both semantic and grammatical aspects. During the annotation of the DDI corpus, we have identified linguistic phenomena that complicate the manual annotation of drug named entities, and different syntactic phenomena that should be considered during the annotation of DDIs. We provide a detailed description of them in **Section 3.4**.

3.3.2 Annotation process

This section describes the process followed in the annotation of drugs and their interactions in the DDI corpus. Four people participate in this activity: two expert pharmacists with a substantial background in pharmacovigilance and two text miners involved in the creation of the DrugDDI corpus. The process is divided in the following steps: training, manual annotation, IAA measurement and harmonization. **Figure 3.9** shows a representation of these phases, which are described in detail below.

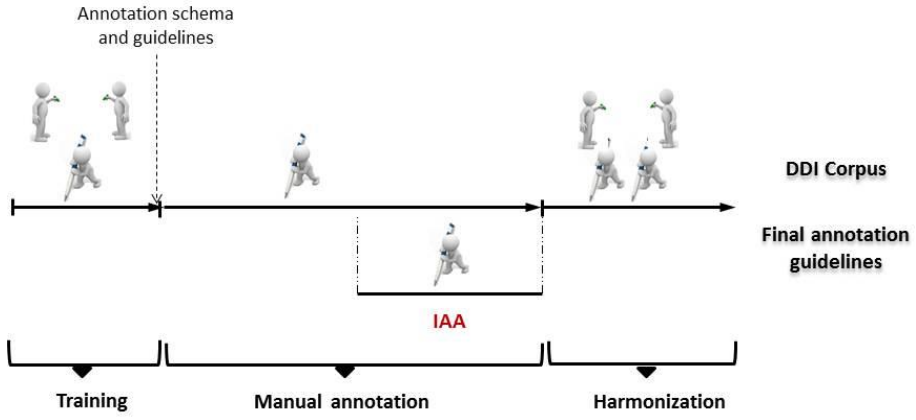

Figure 3.9. Annotation process

As mentioned before, the annotation process relies on annotation guidelines, which are created in an iterative process. In the training stage, the first annotator studies and annotates a set of 30 documents from DrugBank and 10 MEDLINE abstracts. During this annotation, a previously defined annotation schema and initial annotation guidelines are discussed between the pharmacist and the two text mining experts, until the creation of a defined schema and annotation guidelines, whose main points have been presented in the previous sections.

The first annotator marks up then the whole corpus, while the second one annotates a total of 1600 randomly selected sentences from the DDI-DrugBank dataset and 400 ones from the DDI-MEDLINE dataset. Since the documents are pre-annotated with pharmacological substance entities by the MMTx tool and labelled with `drug` type, the annotators have to review manually these labels, adding missing entities, removing erroneous ones, and modifying spans or type of entities when necessary. All mentions of pharmacological substances are annotated, even those that are not involved in a DDI. Finally, DDIs are manually annotated at a sentence level. During this process, one of the text miners assists annotators in technical aspects, such as the use of the annotation tool. Afterward, the double annotation is used to measure the IAA to assess the consistency and quality of the corpus.

An important aspect in the creation of a gold-standard corpus is the selection of the annotation tool that will support the manual annotation. Neves & Leser (2014) provided a recent and detailed review of annotation tools for the biomedical literature, which involved almost 30 tools, of which 13 were selected and compared using predefined criteria and hands-on experiences whenever possible. Therefore, a new review of these tools is out of the scope of this thesis, and we refer the interested reader to this publication for more details about available annotation tools.

From their review, the authors concluded that there is no single tool supporting all use cases with equal robustness, and that their suitability for an annotation project depends on their individual pros and cons and the characteristics of the annotation task (Neves & Leser, 2014). In our case, there are two main requirements influencing the selection of the annotation tool. Firstly, the DDI corpus is built from a preliminary corpus for the *DDIExtraction 2011 task*, rather than raw texts, which already contained drug and DDI annotations. Therefore, we required a tool supporting the adequate integration of pre-annotations in XML format. Secondly, the annotation process of the DDI corpus needs an annotation tool ensuring the annotation of relationships in an easy and effective manner. Although three annotation tools support the first requirement (Bionotate (Cano, Monaghan, Blanco, Wall, & Peshkin, 2009), WordFreak (Morton & LaCivita, 2003), and XConc (Kim et al., 2003)), none of them fulfil both conditions.

For these reasons, we select the open-source XML editor XML Notepad, published by Microsoft® (CodePlex, 2014). XML Notepad provides a simple intuitive user interface for browsing and editing large XML documents, and has real time XML schema validation. Moreover, the editor features incremental search in both tree and text views, drag/drop support, IntelliSense, find/replace with regular expressions and XPath expressions, and support for XInclude.

Finally, during the harmonization process the text mining experts check and review those sentences containing disagreements between the two annotators, and classify them according to the main reason for discrepancy (e.g., missed entity annotation, partial

matching, different entity type assigned, missed DDI annotation, different DDI type assigned). These cases are studied and discussed between the two annotators and the two text mining experts, who help to achieve consensus on the final corpus. Changes in the corpus are made accordingly to the consensus-driven decisions. In the same way, the annotation guidelines are modified to include new rules and examples.

3.4 Annotation issues in pharmacological texts

In this section, we review the main sources of annotation problems that affect in general the manual annotation process, the linguistic phenomena that complicate the manual annotation of drug named entities, and the different syntactic phenomena that should be considered during the annotation of DDIs.

3.4.1 Main sources of annotation problems

In this section, we describe the main sources of annotation problems identified during the annotation process of the DDI corpus. These issues have been addressed in the framework of a pharmacological corpus. However, due to their general nature, they could be extrapolated to domains others than pharmacology, such as medicine or chemistry.

- **Tokenization problems**

Corpus texts need to be segmented into words and sentences through tokenization before any further processing can be done. Different tokenizers have been developed for this purpose (He & Kayaalp, 2006). However, pharmacological texts, and specifically chemical and drug names, are a source of ambiguous numbers, punctuation marks, and parentheses (e.g., '*N-methyl-D-aspartic acid*', '*R,3-hydroxybutan-2-one*', or '*(+)-aplysinillin*'). These ambiguous characters can lead to erroneous sentence splitting and word tokenization.

- **Complexity of drug named entities**

Drug NER seems to be a relatively simple task, since the number of possible drug names is smaller compared to those of other biomedical entities, such as genes or proteins. Moreover, there are different controlled vocabularies and lists of drugs collecting them, (e.g., RxNorm (Nelson et al., 2011) or the ATC classification system (WHO, n.d.)). However, current automatic information extraction systems are not able to properly extract drug names from biomedical texts without human intervention (Jagannathan et al., 2009). This finding highlights that drug names are complex named entities. There are different characteristics of drug names contributing to this complexity.

First, the same drug can have different generic and several trade names in different countries. For example, the drug '*paracetamol*' is named '*acetaminophen*' in the USA. Some of its branded names include *Acephen*® (in the USA), *Efferalgan*® (in Spain) or *Ultralief*® (in the UK). Moreover, drug names can have different abbreviations or synonyms. In addition, some drugs are approved in some countries, while they are not in others. Therefore, there is not a comprehensive list of drugs collecting all drug names approved in the world as well as all their synonyms. For example, RxNorm provides normalized names for clinical drugs approved in the USA, linking them to different synonyms, such as branded names. On the other hand, the ATC classification system refers to each pharmacological substance by one official name only, excluding possible synonyms. In addition to this, another important barrier to the maintenance of an updated controlled list is that new discovered drugs are continuously approved for sale.

Secondly, in pharmacological texts the mention of terms describing a group of drugs is frequent. These are complex terms, since they are usually represented by nested multi-word terms including protein names, numbers, abbreviations, or adjectives, as well as punctuation marks such as parentheses or hyphens (Kolárik, Hofmann-Apitius, Zimmermann, & Fluck, 2007). Moreover, group of drugs names are terms with multiple possible variants that usually are not collected in a comprehensive way in a controlled vocabulary or database. For example, the group of drugs '*Beta Blocking Agents*' (term used in the ATC classification system) can be described as well as '*Adrenergic beta-Antagonists*', '*beta-Adrenergic Receptor Blockaders*' (terms collected in the MeSH thesaurus) or, simply as '*β-blockers*', among many others (Paolillo et al., 2013).

- **Complexity of biomedical texts**

Manual annotation of biomedical texts should be carried out by domain experts capable of understanding the information described in the corpus. However, not all corpora are equally complex. Complexity varies depending on several facts, such as the source of the documents (e.g., manually curated databases or primary scientific literature), the type of study described in the text (e.g., clinical study or *in vitro* study) as well as content-related and linguistic aspects, such as the use of technical vocabulary, complex sentences, and so forth. The level of complexity of text will determine which annotators should be selected for the annotation task (e.g., a pharmacovigilance expert or a life science bachelor student), the length of training for annotators, and the explicit rules that should be described in the annotation guidelines.

- **Lack of standard or reference works in the specific domain**

A set of standard rules for manual annotation of pharmacological substances or drugs has not been established. This is a difficult task because different corpora are annotated with different final objectives. For example, the DDI corpus has been annotated for the extraction of DDIs, while the aim of the CLEF corpus is to extract clinically significant information from clinical texts (Roberts et al., 2007). Therefore, different corpora require different annotation schema and annotation guidelines.

However, detailed reference works could improve the objectives achieved by future research groups in the specific domain. Therefore, when research groups create a new

manually annotated corpus, comprehensive annotation guidelines should be written. These documents should reflect how the annotation task should be carried out, as well as how annotators should deal with specific or complex linguistic phenomena. For example, in the annotation of a pharmacological corpus, it should be important to specify how annotators should annotate stereoisomers of drugs. These chemical entities are usually described by adding a letter S or R before the drug name (e.g., '*S-warfarin*'). To ensure consistency between different annotators, a simple rule describing if the annotator should include in the annotation span only the drug name ('*warfarin*') or the stereoisomer specification ('*S-warfarin*') should be described in the annotation guidelines.

Thus, when a new research group creates a related annotated corpus with similar entities, this group could base its decisions on those adopted in the reference work, leading to closest corpora that might be re-used or exchanged in future works.

3.4.2 Linguistic aspects of drug names

The manual annotation process of a pharmacological corpus can be a difficult task if the terms that should be annotated have not been established previously. As mentioned before, drug names are complex entities and they have several nomenclatures, synonyms, and term variants. In this section, we describe some of the main linguistic phenomena regarding drug nomenclature.

- **Different nomenclatures**

Each drug has a unique and globally recognized name called International Non-proprietary Name (INN) that facilitates the identification of pharmaceutical substances (Ladas, 1975). However, different countries can assign specific non-proprietary names, such us United States Adopted Name (USAN) in the USA or the *Denominación Oficial Española* (DOE) in Spain. Examples below show some of these different nomenclatures:

« The effects of **paracetamol** are possibly reduced in patients taking **anticonvulsants**. » (i)

« The absorption of **acetaminophen** may possibly be reduced if **colestyramine** is given at the same time. » (ii)

Sentences (i) and (ii) describe two different interactions of the same drug. *Paracetamol* is an INN, while *acetaminophen* is the USAN for the same drug. Therefore, there are two different names referring to the same substance and both of them should be annotated.

On the other hand, every drug can have several brand names – that is, a drug marketed under a proprietary, trademark-protected name. There can be several drug brand names for every drug in different countries.

« **Atromid-S** may displace acidic drugs such as **phenytoin** or **tolbutamide** from their binding sites. » (iii)

Sentence (iii) describes an interaction of the drug *clofibrate*. However, in this sentence it is named using a brand name: *Atromid-S*. A search in the DrugBank database will show that this drug holds more than 90 different brand names. Therefore, these brand names are different terms referring to the same substance, and any mention of them in the text should be annotated.

Drug names usually have different synonyms and abbreviations. Two examples are shown in the following sentences:

« HUMIRA has been studied in rheumatoid arthritis patients taking concomitant **MTX**. » (iv)

« This is typical of the interaction of meperidine and **MAOIs**. » (v)

Sentence (iv) refers to the drug named *methotrexate* using an abbreviation: *MTX*. Sentence (v) describes the group of drugs *Monoamine Oxidase Inhibitors* by its abbreviation *MAOIs*. All these terms are synonyms for a drug name or a group of drugs name. Therefore, they should be annotated in the corpus.

- **Multi-word terms**

Multi-word terms are frequently used to describe drug names and, more often, groups of drugs names. Usually, common nouns such as drugs, agents, or products, among others, are preceded by an adjective describing the therapeutic effect, the mechanism, or other characteristics of the group of drugs. For example:

« The treatment of depression in diabetic patients must take into account variations of glycemic levels at different times and a comparison of the available **antidepressant agents** is important. » (vi)

However, the term can be shortened and the adjective can be used as a noun.

« In the present study we evaluated the interference of **antidepressants** with blood glucose levels of diabetic and non-diabetic rats. » (vii)

Two different annotations are possible in sentence (vi): just the shorter term <antidepressant> or the larger term <antidepressant agents>. The first option, the annotation of the shorter term, agrees with the annotation in sentence (vii), where there is only one possibility: the annotation of the term <antidepressants>. However, IE systems or techniques would benefit from the addition of common nouns that could help in the identification of group of drugs names. Therefore, we decided to annotate the longer term, whenever possible.

- **Nested terms**

A frequent linguistic phenomenon in the pharmacological domain is nested named entities. They are frequently used referring to a specific subgroup of drugs within a group of drugs. Next, some examples of nested named entities are presented:

> « The concomitant use of allopurinol and **thiazide diuretics** may contribute to the enhancement of allopurinol toxicity. » (viii)

These nested terms could be annotated in three different ways: as two independent entities <thiazide> and <diuretics>; as a unique entity <thiazide diuretics>; as three different entities <thiazide>, <diuretics>, and <thiazide diuretics>. When the author refers to *thiazide diuretics,* he or she is alluding to one group of drugs (diuretics with a concrete structure defined by the adjective thiazide). The annotation of two (option one) or three (option three) different entities would lead to the annotation in the text of more entities than those intended by the author and would complicate the annotation of DDIs between them. Thus, if we annotate more than one entity, we would express that there are two different groups of drugs: one of them *thiazide* and the other, *diuretics*. Therefore, we decided to annotate a unique entity (option two) <thiazide diuretics>.

- **Discontinuous names**

Another related linguistic phenomenon is discontinuous entities. It is especially common when drug names occur in coordinate structures. For example:

> « In some patients, the administration of a non-steroidal anti-inflammatory agent can reduce the effects of **loop**, **potassium-sparing** and **thiazide diuretics**. » (ix)

In sentence (ix) bold terms describe three different groups of drugs. The first one refers to a group of diuretics acting in the loop of Hendle, which is a specific portion of nephrons in the kidney: the <loop diuretics>. The second one is a group of diuretics that do not promote the loss of potassium, the <potassium-sparing diuretics>. The third one is the abovementioned group of diuretics sharing a common structure, <thiazide diuretics>.

The term *thiazide* is used commonly without the term diuretics, preserving its meaning. However, the terms <loop> or <potassium-sparing> always act as modifiers and do not keep the meaning by themselves. This is the reason why we decided to annotate three different entities <loop diuretics>, <potassium-sparing diuretics> and <thiazide diuretics>.

- **Ambiguity**

Term ambiguity occurs when the same term refers to many concepts (Ananiadou, Kell, & Tsujii, 2006). As other biomedical entities, drug names can be ambiguous. Below, we describe two different examples of ambiguous terms:

« Therefore, in patients taking **insulin** or oral hypoglycemics, regular monitoring of blood glucose is recommended. » (x)

« There is no evidence that EPA supplements have detrimental effects on glucose tolerance, **insulin** secretion or **insulin** resistance in non-diabetic subjects. » (xi)

In sentence (x) the context implies that the word <insulin> refers to a drug, since it is stated that it is administered by or to a patient. Therefore, it should be annotated as a drug in the corpus. However, in sentence (xi) the same word <insulin> names a substance produced by the own body. In this case, we decided to do not annotate it as an entity, since it does not conform to the previously established definition of drug, which is defined in the annotation guidelines as "a substance that is used in the treatment, cure, prevention or diagnosis of diseases".

« The **CNS depressant** effect of *oxycodone hydrochloride* may be additive with that of other **CNS depressants**. » (xii)

Another source of ambiguity is group of drugs names. In sentence (xii) the first term <CNS depressant> refers to the depressant effect on the central nervous system by the drug *oxycodone hydrochloride*. Therefore, it is not a group of drugs, but an effect of a drug. However, the second bold term <CNS depressants> refers to the group of drugs sharing the common characteristic of having a depressant effect on the central nervous system. Therefore, this second term should be annotated as a group of drugs.

Ambiguity remains an important issue in the development of accurate named entity recognition systems. Therefore, the manual annotation rules established for the annotation of ambiguity terms is a relevant decision in the development of any annotated corpora.

3.4.3 Syntactic phenomena in pharmacological texts

As mentioned before, relationships can be expressed in different ways through different syntactic phenomena, such as alternation or coordination. In the DDI corpus, a DDI relationship is a binary relationship annotated at the sentence level with an attribute type (`effect`, `advice`, `mechanism`, `int`). In this section, we describe some of the main annotation problems identified during the annotation of DDI relationships in the DDI corpus.

- **Hypernymic propositions**

A hypernymic proposition represents a taxonomic relation between a hyponym and a hypernym. Hypernymic propositions, in particular appositive structures consisting of several entities, are very common in our texts.

« The effects of adenosine are antagonized by **methylxanthines** such as **caffeine** and **theophylline**. » (xiii)

In sentence (xiii) there is an interaction involving the entities described in the appositive structure. In this example, <methylxanthines> is the hypernym, while <caffeine> and <theophylline> are the hyponyms. *Methylxanthines* is a group of drugs, and *caffeine* and *theophylline* are two drugs belonging to this group. Therefore, the sentence states that an interaction can occur between the drug *adenosine* and the members of the group *methylxanthines*, for example, *caffeine* and *theophylline*. Thereby, a DDI relationship for each one of them should be annotated.

However, in some appositive structures, the scope of the interaction only remains the hyponym and not the hypernym. See the example below:

« In addition to this pharmacological interaction, this report describes a novel chemical reaction between **temazepam** (a **benzodiazepine**) and **ethanol**. » (xiv)

Sentence (xiv) contains a group, <benzodiazepine> and a drug belonging to this group, <temazepam>. In this example, the term benzodiazepine is describing a characteristic of the drug *temazepam*, and we cannot infer from the sentence that the interaction between *ethanol* and *temazepam* can occurs between *ethanol* and other members of the group *benzodiazepine*, too. Therefore, we decided to annotate one DDI relationship only between <ethanol> and <temazepam>.

- **Coordinate structures**

The same drug can be mentioned several times in the same sentence. In these cases, it could be unclear which one should be included in a DDI relationship.

« The concomitant use of **nitrofurantoin** is not recommended since **nitrofurantoin** may antagonize the effect of **norfloxacin**. » (xv)

Sentence (xv) contains two coordinate clauses with the conjunction 'since' joining them together. In the first clause, there is just one mention of a drug: <nitrofurantoin>. In the second clause, however, there are two mentions of two different interacting drugs: <nitrofurantoin> and <norfloxacin>. Therefore, we decided to annotate only those drugs mentioned in the second clause as interacting drugs in the DDI relationships.

« The concomitant use of **nitrofurantoin** and **norfloxacin** is not recommended since **nitrofurantoin** may antagonize the effect of **norfloxacin**. » (xvi)

In sentence (xvi), however, the first clause contains two different interacting drugs, <nitrofurantoin> and <norfloxacin>, in the same fashion that the second one. Therefore, two different DDIs should be annotated: one in the first clause and one in the second clause.

« The concomitant use of **nitrofurantoin** is not recommended since it may antagonize the effect of **norfloxacin**. » (xvii)

Finally, in sentence (xvii) the interaction in the second clause is described with an anaphora of the term <nitrofurantoin>. Since we did not include anaphora annotation in the DDI corpus, a unique DDI relationship should be annotated between the two mentioned drugs <nitrofurantoin> and <norfloxacin>.

3.5 Quantitative features of the DDI corpus

This section describes the main characteristics of the DDI corpus in terms of the frequency of named entities and relationships. Based on the sentence splitting during processing, the DDI-DrugBank dataset contains 6,795 sentences, and the DDI-MEDLINE dataset is made up of 2,147 sentences. The different nature of the texts determines that the types of entities and relationships have different ratios in the two subcorpora. **Table 3.3** and **Table 3.4** show the number of the named entity types and relationships annotated in each one of them, respectively.

Regarding entity annotation, the most common type is `drug` (63%) in both subcorpora. Substances not approved for human use (`drug_n`) are the second most common type of entity in DDI-MEDLINE, while these substances account for only about 1 per cent of the entities in the DDI-DrugBank dataset. Similarly, `brand` drugs are about 12% of the entities in the DDI-DrugBank dataset; however, this type had the lowest frequency in the DDI-MEDLINE dataset. These observations are to be expected because MEDLINE abstracts usually describe results from laboratory experiments, where experimental substances not approved to be used in humans are commonly employed, while DrugBank texts are mainly compiled from repositories of drug interactions. In fact, the highest frequency of brand named entities in the DDI-DrugBank dataset might suggest that the database could have used PIs or SPCs – documents created by the manufacturer for every branded drug – as an information source.

	DDI-DrugBank	DDI-MEDLINE	Total
drug	9901 (63%)	1745 (63%)	11,646 (63%)
brand	1824 (12%)	42 (1.5 %)	1866 (10%)
group	3901 (25%)	324 (12%)	4225 (23%)
drug_n	130 (1%)	635 (23%)	765 (4%)
TOTAL	15,756	2746	18,502

Table 3.3. Numbers of annotated entities in the DDI corpus

Table 3.4 shows the numbers of annotated relationships in each subcorpus. The main difference between them is that the `advice` relationship is far more frequent in DDI-DrugBank than in DDI-MEDLINE dataset. This is also consistent with the fact that the texts from DrugBank seem to be aimed at health-care professionals because these texts usually contain recommendations for avoiding any drug interactions and their side effects. The most common type of relationship in the corpus is `effect`. Thus, this corpus is annotated with a large amount of information describing PD mechanisms and interaction effects. At the same time, the corpus contains a lot of information on PK DDIs (`mechanism`). However, both DrugBank and MEDLINE documents in the corpus present a low frequency of management recommendations. These results agree with the descriptions of the type of DDI information included in the main DDI information sources (Aronson, 2004; Bergk et al., 2005).

	DDI-DrugBank	DDI-MEDLINE	Total
effect	1855 (39.4%)	214 (65.4%)	2069 (41.1%)
mechanism	1539 (32.7%)	86 (26.3%)	1625 (32.3%)
advice	1035 (22%)	15 (4.6%)	1050 (20.9%)
int	272 (5.8%)	12 (3.7%)	284 (5.6%)
TOTAL	4701	327	5028

Table 3.4. Numbers of annotated relationships in each corpus

3.6 Discussion

In this chapter, we have described the construction and annotation of the DDI corpus. Our aim has been to create a benchmark corpus for IE of DDIs that could overcome the limitations of the existing ones. As mentioned before (**Chapter 2**), a gold-standard corpus is a manually annotated corpus, whose quality is ensured by the creation of available annotation guidelines, and the measurement of the IAA. To be useful, it must provide rich information, annotating a diversity of documents with all different entities and relationships relevant to the intended task, and with a size and number of annotations large enough to ensure an appropriate training and evaluation of IE systems. Here, we compare the DDI corpus with those corpora relevant to the DDI domain based on these required characteristics.

Firstly, we focus on the annotation of drug entities to assess the value of the different reviewed corpora for the drug NER task. We have observed that most of them have been manually annotated with pharmacological entities. However, the quality of these annotations must be ensured by available annotation guidelines and the measurement of the IAA. These three characteristics are accomplished only by four corpora: the **CLEF corpus** (Roberts et al., 2009), the **PK DDI corpus** (Boyce et al., 2012), the **CHEMDNER** corpus (Krallinger et al., 2013) and the **DDI corpus**. Furthermore, other important aspects determining the usefulness of a corpus are the

diversity of annotated entities, which should represent the relevant entities for the intended task, and the number of these annotations. As we can see in **Table 3.5**, the **CLEF corpus** has been annotated only with one type of entity drug, which includes not only pharmacological substances but medical devices − such as the term "jeringa" −, too. Moreover, the number of annotated entities is just 197. In contrast, the **PK DDI corpus** provides a more detailed classification of entities and distinguishes between active ingredients, branded drugs and metabolites. The number of annotations in this corpus is larger than in the previous one, being 3,896 annotated entities. The **CHEMDNER** corpus is the largest one, with more than 80K annotations. However, this corpus focuses on chemical entities mentions, and does not focus on pharmacological substances. Regarding their annotation guidelines, drugs mentions must have been mainly annotated as types TRIVIAL (25,610 annotations) and FAMILY (11,935). The TRIVIAL type includes mentions of chemical entities, but not necessarily substances with pharmacological properties, by a generic common name or a brand name, while the FAMILY type includes chemical families that can be associated to some chemical structure, independently of whether they refer to pharmacological substances or not.

However, the annotation schema defined in the **DDI corpus** provides the most detailed classification of these entities, including active ingredients, branded drugs, groups of drugs, and active substances not approved for human use. Moreover, the annotation guidelines in the DDI corpus include a detailed description of these different types of entities, a crucial aspect for the quality of the annotations and the agreement achieved between annotators. Therefore, our annotation guidelines could serve as a standard for annotating drug names. The number of annotated entities (18,502) differs considerably from that of the **CLEF** or the **PK DDI** corpora. Besides **CHEMDNER**, this number is only exceeded by the **DrugDDI corpus** (Segura-Bedmar, 2010). However, these annotations have been made automatically, without a later manual review. The **ITI TXM** includes a similar number of annotated entities to those in the **DDI** corpus (18,000 annotated entities), although they correspond to a unique type DRUGCOMPOUND, which includes any chemical used to affect the function of an organism, cell or biological process (Alex et al., 2008).

Corpus	Annotation	Diversity of entities	Annotated entities
BioText	M	1	*np*
ITI TXM	M	1	18000
CLEF	M	1	197
BioCaster	M	2	1022
DrugDDI	A	1	23190
ADE	M	1	5063
PK DDI	SA	3	3896
Rubrichi	M	3	*np*
EU-ADR	A	1	1753
PK	A	1	*np*
CHEMDNER	M	8	84355
DDI	SA	4	18502

Table 3.5. Comparison of corpora annotated with pharmacological substances (M for manual; A for automatic; SA for semi-automatic; *np* for not provided)

Corpus	Diversity of documents	Size		IAA	AAG
		Documents	Sentences		
BioText	1	*np*	*np*	×	✓
ITI TXM	1	400	*np*	✓	*
CLEF	1	150	*np*	✓	✓
BioCaster	1	1000	*np*	×	*
DrugDDI	1	579	5806	×	×
ADE	1	*np*	4272	✓	*
PK-DDI	1	68	*np*	✓	✓
Rubrichi	1	100	*np*	×	×
EU-ADR	1	300	*np*	✓	*
PK	1	428	5026	✓	*
CHEMDNER	1	10,000	*np*	✓	✓
DDI	2	1025	8942	✓	✓

Table 3.6. Comparison of corpora annotated with pharmacological substances (IAA for inter-annotator agreement; AAG for available annotation guidelines; * for annotation guidelines mentioned to have been used in the annotation process, but not available; *np* for not provided)

Secondly, we analyse and compare the four corpora annotated with DDIs. These include the **PK-DDI**, the **PK** (Wu et al., 2013), the **DrugDDI** and the **DDI corpus** (**Table 3.7**). All the relationships have been manually annotated. However, an important difference between them is the scope of the annotation. While the **PK DDI** and **PK corpus** annotate only PK DDIs, the **DrugDDI** and the **DDI corpus** have been annotated with all mentions of DDIs, including both PK and PD DDIs. According to the authors of the **PK DDI corpus**, the vocabulary used to describe PK DDIs is significantly different from that used to describe PD DDIs because they are discovered in distinct ways. Therefore, the annotation of these two types in the same corpus with different labels could be useful to test the performance of IE systems dealing with the recognition of each one of them. The **DrugDDI corpus**, however, have established only one type of annotation for all DDI relationships. In contrast, the other corpora have established different types of DDIs. The **PK DDI corpus** provides a double classification, distinguishing between positives and negatives – based on their assertion or negation of the occurrence of an interaction – and quantitative and qualitative DDI statements – being the first one assigned when the statement contains quantitative data, and the latter one otherwise. Similarly, the **PK corpus** classifies DDIs based on the quality of the evidence provided by the statement – or the level of certainty – as DDIs, ambiguous DDIs or non-DDIs. Although these classifications distinguish between different types of DDI statements, they focus on linguistic aspects such as negation, and do not differentiate between the pharmacological information provided by each sentence. In contrast, the **DDI corpus** proposes a classification of DDIs based on the information requirements for the effective management of DDIs: how the interaction occurs (`mechanism`), what consequences can be expected (`effect`) and how it can be managed to avoid or reduce the associated risk (`advice`). Cases where a sentence only asserts the existence of a DDI without providing any additional information are annotated, too (`int`).

Corpus	Annotation	Scope	Diversity of DDIs	Annotated DDIs
PK DDI	M	PK	4	592
PK	M	PK	3	1333
DrugDDI	M	PK and PD	1	3160
DDI	M	PK and PD	4	5028

Table 3.7. Comparison of corpora annotated with DDIs (M for manual)

Besides the annotation-related aspects discussed above, there are other important characteristics determining the usefulness of a corpus: its size and the inclusion of diversity types of documents (**Table 3.6**). The **DDI corpus** is the only available corpora made up of two different types of text: MEDLINE abstracts and documents describing DDIs from the DrugBank database. Thus, the corpus covers two different styles of biomedical text: while the texts taken from the DrugBank database are completely focused on the description of DDIs, the main topic of the scientific texts would not necessarily be on DDIs. Moreover, while abstracts are usually written in a very scientific language, the language used in the texts from DrugBank is similar to the language used in PIs. We illustrate this with three different sentences.

« The second response is due to a release of norepinephrine from nerves and was potentiated by ouabain through the increase in the norepinephrine release, whereas the first response was not due to the norepinephrine release but to a direct action on smooth muscle cell and was inhibited by ouabain. » (xviii)

«ZEBETA should not be combined with other beta-blocking agents. Patients receiving catecholamine-depleting drugs, such as reserpine or guanethidine, should be closely monitored, because the added beta-adrenergic blocking action of ZEBETA may produce excessive reduction of sympathetic activity. » (xix)

« ZEBETA should not be combined with other beta-blocking agents. Patients receiving catecholamine-depleting drugs, such as reserpine or guanethidine, should be closely monitored, because the added beta-adrenergic blocking action of ZEBETA may produce excessive reduction of sympathetic activity […] » (xx)

Example (xviii) shows a sentence from the DDI-MEDLINE dataset, while example (xix) corresponds to two sentences from the DDI-DrugBank dataset. In example (xx) we show a fragment extracted from the "Drug Interactions" section in the PI for the same drug[9]. As we can see, while the text in the MEDLINE sentence is more complex, the texts in DrugBank and the PI are coincident.

This variety of texts provides a training corpus for IE system with different levels of complexity, and, at the same time, their performance can be evaluated against simple

[9] http://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?setid=a11548a0-9c0f-4729-907c-75d8f99a6c85. Accessed on 12/02/2015

versus complex documents. Lastly, the size of the DDI corpus is significantly larger than that of other corpora, being the largest manually annotated corpora for drugs and DDIs.

Finally, the quality of the annotations must be measured by the IAA. This aspect is described and discussed in detail in the next **Chapter 4**.

3.7 Conclusions

The DDI corpus has proven to fulfil all those characteristics required to be a gold-standard for both drug NER and DDI extraction. The size of the DDI corpus is significantly larger than that of other corpora annotated with drugs and DDIs, both in the number of documents and in the total number of annotated entities and relationships. Moreover, it provides a fine-grained description of pharmacological entities and DDIs, established after a thorough study of different types of DDI texts. In next chapter, we describe its evaluation by means of the measurement of the IAA. Moreover, we evaluate the usefulness of the DDI corpus as gold-standard in the *SemEval-2013 DDIExtraction shared task* (Segura-Bedmar, Martínez, & Herrero-Zazo, 2014).

The resources described in this work, including both the annotated corpus and the annotation guidelines, are available from http://labda.inf.uc3m.es/ddicorpus, and are described in (Herrero-Zazo, Segura-Bedmar, Martínez, & Declerck, 2013).

Chapter 4

Evaluation of the DDI corpus

In order to be useful for the intended task, the quality of a corpus must be evaluated. Manual annotation can introduce errors – mainly due to fatigue during the annotation process – and bias – or the individual preferences of different annotators. These bias can be reduced by methods such as the creation of detailed annotation guidelines, a training period prior to the annotation, and the pre-annotation of the documents by text mining techniques (Artstein & Poesio, 2005; Fort & Sagot, 2010). All these strategies are directed to the same goal: to achieve high inter-annotator reliability, or the agreement between different annotators.

On the other hand, quality and usefulness of annotated corpora should be tested for the application tasks they were created for. To this purpose, the corpus has been used in the *SemEval-2013 DDIExtraction task* (Segura-Bedmar et al., 2013), where different teams from different countries used the corpus as a gold-standard for the evaluation of IE systems applied to the recognition of pharmacological substances and the detection of DDIs from biomedical texts.

In this chapter, we describe the results of the IAA for the DDI corpus (**Section 4.1**) and provide a detailed description of the *SemEval-2013 DDIExtraction task* (**Section 4.2**). Finally, we provide the main conclusions and our future work in **Section 4.3**.

4.1 Inter-Annotator Agreement (IAA)

Inter-annotator agreement (IAA) is a quantitative index for the degree of agreement between the annotations produced by two or more annotators on the same set of texts, using the same annotation guidelines (Müller & Strube, 2006). The annotations are compared using some statistical measure, being the Kappa statistic the most commonly used standard (Cohen, 1960), which takes into account that a certain degree of agreement between annotators can also be ascribed to chance. IAA enables the assessment of the quality and consistency of the corpus and the annotation guidelines, as well as the complexity of the annotation task.

The measurement of the IAA requires that more than one annotator annotates the same texts in the corpus. This is another important factor in the quality of the annotation process, since strategies of multi-annotation, where documents are annotated by more than one annotator, have been proposed as the best way to avoid problems related to manual annotation (Jagannathan et al., 2009; Wilbur et al., 2006).

During the annotation of the DDI corpus, a set of documents from each dataset, DDI-DrugBank and DDI-MEDLINE, is randomly selected and annotated by two different expert annotators (**Section 3.3.2**). We should note that the IAA scores are measured after a rigorous process to define strict, comprehensive, and clear guidelines. For this reason, IAA scores are calculated using exact matching that requires the annotations to overlap completely. For the entities, their annotations should overlap completely and annotators should agree on the assigned types, too. Regarding the interactions, the annotators should agree on the annotation of the interacting drugs as well as on the type assigned to the interaction. IAA results are shown in the following section.

4.1.1 IAA results

Both the quality and consistency of the corpus are evaluated by measuring the IAA scores, which allows for determining the complexity of the annotation task and provides insights into the quality of the guidelines developed. Moreover, IAA also provides an upper bound on the performance of the automatic systems for the detection of pharmacological substances and the interactions between them. In our corpus, we use the Kappa statistic to study IAA. A detailed description of the kappa coefficient of agreement can be found in (Cohen, 1960). **Table 4.1** and **Table 4.2** summarize the obtained results.

	DDI-DrugBank	DDI-MEDLINE
$\mathbf{K_{drug}}$	0.9534	0.8467
$\mathbf{K_{brand}}$	0.9569	0.8853
$\mathbf{K_{group}}$	0.9563	0.8299
$\mathbf{K_{drug\_n}}$	0.4422	0.8122
K	0.9104	0.7962

Table 4.1. IAA results of the annotated entities in the DDI corpus

In terms of document source type, IAA is higher for the DDI-DrugBank dataset than for the DDI-MEDLINE dataset in both entities and relationships. One explanation for this is that MEDLINE abstracts have far more complexity than texts from the DrugBank database, which are usually expressed in simple sentences. Similar to other annotated corpora (Rosario & Hearst, 2004; van Mulligen et al., 2012), IAA scores are higher for entities than for relationships.

In terms of the type of entity, the highest IAA score is obtained for the `brand` type in both DDI-DrugBank and DDI-MEDLINE subcorpora (see **Table 4.1**). This may be because branded drug names are carefully selected by the manufacturer to be short, unique, and easy to remember (Boring, 1997). Similarly, a high level of agreement is observed for `drug` and `group` type entities. These high IAA scores may indicate that these types are more clearly defined than others in the annotation guidelines (Pustejovsky & Stubbs, 2012). For example, annotators have found the identification of experimental drugs (e.g., *pempidine*), which should be annotated as `drug_n` type, more difficult than the identification of the names referring to approved drugs (`drug` and `brand`) or groups of drugs (`group`).

On the other hand, IAA results show moderate agreement for `drug_n` entities. These results can be due to the large variety of substances included in this type. Additionally, since some of these substances can be both endogenous − produced inside an organism − and exogenous − produced outside the body −, such as the terms *calcium* or *dopamine*, their recognition depends substantially on the context in which they appear. As we have explained in **Section 3.4.2**, the mentions of endogenous substances should not be annotated as pharmacological substances. In particular, the agreement was lower for `drug_n` in the DrugBank dataset than in the DDI-MEDLINE.

One of the main sources of discrepancy are metabolite names (e.g., *descarboethoxyloratadine*), which are very similar to drug names (e.g., *loratadine*) and are very frequent in DrugBank texts. We have observed that annotators often have difficulty distinguishing between both types. Similarly, another main reason for disagreement between annotators is the classifications of substances such as vitamins, since some of them can be considered as a group of drugs (e.g., *vitamin A*) while others are `drug` entities (e.g., *betacarotene*). All the mentioned differences are discussed and resolved in the harmonization process, and subsequently more accurate explanations are included in the annotation guidelines (**Section 3.3.1**).

In conclusion, the IAA scores show that annotation guidelines have been successfully developed and validated for the annotation of complex drug names such as stereoisomer (e.g., *S-warfarin*), drug salts (e.g., *oxycodone hydrochloride*), or nested named terms (e.g., *thiazide diuretics*). Therefore, the DDI corpus may be a valuable resource for developing systems for drug NER.

In general, fairly high IAA results have been obtained per type of interaction (see **Table 4.2**). The `int` type presents the highest IAA scores in both the DDI-DrugBank and the DDI-MEDLINE datasets. However, this is the least common type of relationship (less than 6%) annotated in the corpus (see **Section 3.5**). The second DDI relationship with higher IAA results is the `advice` type. This type of DDI information is very clear and can be easily identified by manual annotators in both types of document. On the other hand, two main reasons for disagreement in DDI type `advice` have been observed.

Firstly, annotators are frequently confused with sentences containing a recommendation for a specific DDI effect, as in the following sentence:

> « Consider additive sedative effects and confusional states to emerge if **chlorprothixene** is given with **benzodiazepines** or **barbiturates**. » (i)

Similarly, annotators also have problems with sentences describing a PK mechanism and suggesting a change in the dosage schedule to avoid undesired consequences. For example, the sentence below is considered as `advice` type by one annotator, while the other one classifies it as `mechanism` type.

> « **Fenofibrate** should be taken at least 1 h before or 4–6 h after a **bile acid binding resin** to avoid impeding its absorption. » (ii)

As shown in **Table 4.2**, the `mechanism` type shows the lowest IAA scores in the DDI-DrugBank dataset. One reason for this result is that annotators find it difficult to distinguish between sentences describing a PD mechanism or an effect. This observation has led to the final annotation of PD interactions with the `effect` type.

We have observed that most disagreements may be because many sentences provide various textual evidence of the same interaction and each piece of textual evidence may correspond to a different type of drug interaction. This is very common in complex sentences because subordinate clauses often describe different properties of the same interaction. In these cases, the guidelines have proposed a priority rule to assign the type of interaction. However, sometimes the annotators incorrectly applied this rule, and they often tended to assign the first type described in the sentence instead of the type according to the priority rule. On the other hand, the guidelines state that clauses in compound sentences should be considered as independent sentences, and thereby annotators should annotate each of drug interactions described in their clauses.

	DDI-DrugBank	DDI-MEDLINE
K_{effect}	0.7525	0.5548
$K_{mechanism}$	0.4214	0.5577
K_{advice}	0.9428	0.5587
K_{int}	0.9558	0.7252
K	0.8385	0.6213

Table 4.2. IAA results of the annotated relationships in the DDI corpus

4.2 DDIExtraction shared task series

The *SemEval-2013 DDIExtraction shared task* is the second edition of the *DDIExtraction Shared Task series*, a community-wide effort to promote the

implementation and comparative assessment of NLP techniques in the field of the pharmacovigilance domain and, in particular, to address the extraction of DDIs from biomedical texts.

The first edition, the *DDIExtraction 2011 task*, attracted the attention of 10 teams that submitted a total of 40 runs (Segura-Bedmar, Martínez, & Sánchez-Cisneros, 2011). In this edition, the DrugDDI corpus developed by Segura-Bedmar (2010) was used as a gold-standard for the training and evaluation of the different participating systems for the task of extraction of DDIs.

		Training	Test for NER task	Test for RE task
DDI-DrugBank	documents	572	54	158
	sentences	5675	145	973
	drug	8197	180	1518
	group	3206	65	626
	brand	1423	53	347
	drug_n	103	5	21
	mechanism	1260	0	279
	effect	1548	0	301
	advice	819	0	215
	int	178	0	94
DDI-MEDLINE	documents	142	58	33
	sentences	1301	520	326
	drug	1228	171	346
	group	193	90	41
	brand	14	6	22
	drug_n	401	115	119
	mechanism	62	0	24
	effect	152	0	62
	advice	8	0	7
	int	10	0	2

Table 4.3. Frequencies in the DDI corpus

In the second edition, several improvements in terms of organization, participation and results have been achieved. The *6th International Workshop on Semantic Evaluations (SemEval)*, held in summer 2013, scheduled the *"Extraction of drug-drug interactions from biomedical Texts task"*[10] (Segura-Bedmar, Martínez, & Sánchez-Cisneros, 2011). For this edition, two different subtasks were proposed. The first one consisted in the recognition and classification of drug names (NER task), while the second one focused on the extraction and classification of DDIs (RE task). The DDI corpus, manually annotated with both pharmacological substances and DDIs, was used as a gold-standard

[10] http://www.cs.york.ac.uk/SemEVAL-2013/task9/

for the training and testing of the systems developed by a total of 14 participating teams, which submitted 38 runs. Although participants were allowed to use other training resources, only one of them (**LASIGE**) used additional training data collection to develop its system.

The DDI corpus was provided as a training dataset and as two different test dataset for the NER and RE tasks. **Table 4.3** shows the frequencies of annotations in the DDI corpus for both the DDI-DrugBank and the DDI-MEDLINE train and test datasets.

In this section, we describe the two different tasks and the results obtained by the participating systems. In addition to this, for the RE task, we describe the major sources of errors in these systems, and present a study as to whether the results are significant statistically. Furthermore, for the top three methods in RE, we propose an ensemble system using majority and union voting strategies. Finally, we close this section with a discussion of possible future steps of the *DDIExtraction Shared Task series*.

4.2.1 NER task: recognition and classification of pharmacological substances

This task concerned the named entity extraction of pharmacological substances in text. This named entity task is a crucial first step for IE of DDIs. In this task, four types of pharmacological substances were defined: `drug` (generic drugs), `brand` (brand drugs), `group` (group of drugs) and `drug_n` (active substances not approved for human use). A detailed description of these entity types is provided in **Section 3.3.1**. For evaluation, a part of the DDI corpus consisting of 52 documents from DrugBank and 58 MEDLINE abstracts, was provided with the golden annotations hidden. The goal for participating systems was to recreate the standard annotations. Each participant system must output an ASCII list of reported entities, one per line, and formatted as:

IdSentence|startOffset-endOffset|text|type

Thus, for each recognized entity, each line must contain the id of the sentence where this entity appears, the position of the first character, and the one of the last character of the entity in the sentence, the text of the entity, and its type. When the entity was a discontinuous name (e.g., '*aluminum and magnesium hydroxide*'), this second field must contain the starting and end positions of all parts of the entity, separated by semicolon. Multiple mentions from the same sentence should appear on separate lines.

Participants could send a maximum of three system runs. After downloading the test datasets, they had a maximum of two weeks to upload the results. A total of six teams participated in this task, submitting 16 system runs. **Table 4.4** lists the teams, their affiliations, their countries, and a brief description of their systems. **Table 4.5**, **Table 4.6** and **Table 4.7** show the F1 scores for each run in alphabetic order. The full ranking information can be found on the *SemEval-2013 Task 9* website[11], and a detailed description of the teams and their systems is available in Segura-Bedmar et al. (2013).

[11] http://www.cs.york.ac.uk/SemEVAL-2013/task9/index.php?id=evaluation

	Team	Institution	Country	Description
NER Task	LASIGE	University of Lisbon	Portugal	CRF
	NLM_LHC	National Library of Medicine, National Institute of Health	USA	Dictionary-based approach
	UMCC_DLSI	Matanzas University Alicante University	Cuba Spain	j48 classifier
	UEM_UC3M	European University of Madrid, Carlos III University of Madrid	Spain	NCBO tool
	UTurku	University of Turku	Finland	TEES system
	WBI	Humboldt University of Berlin	Germany	Token sequence labelling approach

Table 4.4. Summary of the *SemEval-2013 DDIExtraction* NER task participating teams

The best results were achieved with a token sequence labelling approach proposed by the **WBI team**. Their model was trained on the training dataset as well as on entities of the test dataset for the RE task (**Section 4.2.2**). We should mention that very similar results are obtained by the **NLM_LHC** with a dictionary-based approach combining biomedical resources such as DrugBank (Law et al., 2014), the ATC classification system (WHO, n.d.), MeSH (Lipscomb, 2000), RxNorm (Nelson et al., 2011) and the UMLS Methatesaurus (Bodenreider, 2004), among others. In fact, **Table 4.7** shows that the dictionary-based approach outperformed the sequence labelling approach in both exact and partial evaluations on the DDI-DrugBank dataset.

Regarding the classification of each entity type, we have observed that `brand` drugs were easier to recognize than the other types. This could be due to the fact when a drug is marketed by a pharmaceutical company, its brand name is carefully selected to be short, unique, and easy to remember (Boring, 1997). Moreover, there are different pharmacological databases and terminologies collecting brand names that can be used as dictionaries by the participating systems. In contrast, substances not approved for human use (`drug_n`) are more difficult, due to the greater variation and complexity in their naming. In fact, only the **UEM_UC3M** team was able to recognize this type of substances on the DDI-DrugBank dataset. In addition, this may indicate that this type is less clearly defined than the others in the annotation guidelines. Another possible reason is that the presence of such substances in this dataset is very scarce (less than 1%).

Finally, the results on the DDI-DrugBank dataset were much better than those obtained on the DDI-MEDLINE dataset. While DDI-DrugBank texts focus on description of drugs and their interactions, the main topic of DDI-MEDLINE texts would not necessarily be on DDIs. Coupled with this, it is not always trivial to distinguish between substances that should be classified as pharmacological substances and those not. For example, *insulin* is a hormone produced by the pancreas, but can also be synthesized in the laboratory and used as drug to control insulin-dependent diabetes mellitus. The participating systems should be able to determine if the text was describing a substance originated within the organism, or in contrast, it described a process in which the substance is used for a specific purpose and thus should be identified as pharmacological substance.

Team	Run	STRICT	EXACT	PARTIAL	TYPE	drug	brand	group	drug_n	MAVG
LASIGE	1	0,656	0,781	0,808	0,69	0,741	0,581	0,712	0,171	0,577
	2	0,639	0,775	0,801	0,672	0,716	0,541	0,696	0,182	0,571
	3	0,612	0,715	0,741	0,647	0,728	0,354	0,647	016	0,498
NLM_LHC	1	0,698	0,784	0,801	0,722	0,803	0,809	0,646	0	0,57
	2	0,704	0,792	0,807	0,726	**0,81**	0,846	0,643	0	0,581
UMCC_DLSI	1	0,275	0,3049	0,367	0,334	0,297	0,313	0,257	0,124	0,311
	2	0,275	0,3049	0,367	0,334	0,297	0,313	0,257	0,124	0,311
	3	0,275	0,3049	0,367	0,334	0,297	0,313	0,257	0,124	0,311
UEM_UC3M	1	0,458	0,528	0,585	0,51	0,718	0,075	0,291	0,185	0,351
	2	0,529	0,609	0,669	0,589	0,752	0,094	0,291	**0,264**	0,38
UTurku	1	0,579	0,639	0,719	0,701	0,721	0,603	0,478	0,016	0,468
	2	0,641	0,659	0,731	0,766	0,784	0,901	0,495	0,015	0,557
	3	0648	0,666	0,743	**0,777**	0,783	**0,912**	0,485	0,076	0,604
WBI	1	0,692	0,772	0,807	0,729	0,768	0,787	0,761	0,071	0,615
	2	0,708	0,831	0,855	0,741	0,786	0,803	0,757	0,134	**0,643**
	3	**0,715**	**0,833**	**0,856**	0,748	0,79	0,836	**0,776**	0,141	0,652

Table 4.5. F1 scores for NER task on the whole dataset (DDI-DrugBank + DDI-MEDLINE datasets). (MAVG for macro-average)

Team	Run	STRICT	EXACT	PARTIAL	TYPE	drug	brand	group	drug_n	MAVG
LASIGE	1	0,771	0,834	0,855	0,799	0,817	0,571	0,833	0	0,563
	2	0,771	0,831	0,852	0,799	0,823	0,553	0,824	0	0,568
	3	0,682	0,744	0,764	0,713	0,757	0,314	0,756	0	0,47
NLM_LHC	1	0,869	0,902	**0,922**	0,902	0,909	0,907	0,766	0	0,646
	2	0,869	**0,903**	0,919	0,896	0,911	0,907	0,754	0	0,644
UMCC_DLSI	1	0,424	0,4447	0,504	0,487	0,456	0,429	0,371	0	0,351
	2	0,424	0,4447	0,504	0,487	0,456	0,429	0,371	0	0,351
	3	0,424	0,4447	0,504	0,487	0,456	0,429	0,371	0	0,351
UEM_UC3M	1	0,561	0,632	0,69	0,632	0,827	0,056	0,362	0,022	0,354
	2	0,595	0,667	0,721	0,667	0,842	0,063	0,366	**0,028**	0,37
UTurku	1	0,739	0,753	0,827	0,864	0,829	0,735	0,553	0	0,531
	2	0,785	0,795	0,863	**0,908**	0,858	0,898	0,559	0	0,581
	3	0,781	0,787	0,858	0,905	0,847	**0,911**	0,551	0	0,578
WBI	1	0,86	0,877	0,9	0,89	0,905	0,857	0,782	0	0,636
	2	0,686	0,894	0,914	0,897	0,909	0,865	0,794	0	0,642
	3	**0,878**	0,901	0,917	**0,908**	**0,912**	0,904	0,806	0	**0,656**

Table 4.6. F1 scores for NER task on the DDI-DrugBank dataset. (MAVG for macro-average)

Team	Run	STRICT	EXACT	PARTIAL	TYPE	drug	brand	group	drug_n	MAVG
LASIGE	1	0,567	0,74	0,772	0,605	0,678	0,667	0,612	0,183	**0,577**
	2	0,54	0,733	0,763	0,576	0,631	0,444	0,595	0,196	0,512
	3	0,557	0,693	0,723	0,596	0,702	0,667	0,56	0,171	0,554
NLM_LHC	1	0,559	0,688	0,702	0,575	0,717	0,429	0,548	0	0,462
	2	0,569	0,702	0,715	0,586	**0,726**	0,545	0,555	0	0,486
UMCC_DLSI	1	0,187	0,2228	0,287	0,245	0,2	0,091	0,191	0,13	0,23
	2	0,187	0,2228	0,287	0,245	0,2	0,091	0,191	0,13	0,23
	3	0,187	0,2228	0,287	0,245	0,2	0,091	0,191	0,13	0,23
UEM_UC3M	1	0,39	0,461	0,516	0,431	0,618	0,111	0,238	0,222	0,341
	2	0,479	0,564	0,628	0,529	0,665	0,182	0,233	**0,329**	0,387
UTurku	1	0,435	0,538	0,623	0,556	0,614	0,143	0,413	0,016	0,328
	2	0,502	0,528	0,604	0,628	0,703	**0,923**	0,436	0,016	0,533
	3	0,522	0,551	0,634	**0,656**	0,716	**0,923**	0,426	0,08	0,582
WBI	1	0,545	0,681	0,726	0,589	0,634	0,353	0,744	0,074	0,479
	2	0,576	**0,779**	**0,807**	0,612	0,673	0,444	0,729	0,14	0,534
	3	**0,581**	0,778	0,805	0,617	0,678	0,444	**0,753**	0,147	0,537

Table 4.7. F1 score for NER task on the DDI-MEDLINE dataset. (MAVG for macro-average)

Team	NLP tools	Knowledge resources
LASIGE	MALLET Adapted tokenizer from Corbett et al., 2007 FiGO	Patent corpus (ChEBI team) ChEBI ontology DrugBank
NLM_LHC		DrugBank ATC classification system MeSH RxNorm UMLS Methatesurus
UMCC_DLSI	Freeling tool	WordNet
UEM_UC3M	Mgrep analyzer	MDDB NDF NDDF Ontology for Drug Discovery Investigations MeSH DrugBank PubChem ATC classification system KEGG
UTurku	Charniak-Johnson reranking parser McClosky's biomodel Stanford parser MMTx	DrugBank
WBI	ChemSpot Jochem ABBREV OPSIN	PHARE ontology ChEBI ontology

Table 4.8. NLP tools and other resources used by the NER participating teams (References for the NLP tools and knowledge resources can be found in each individual paper in the *SemEval-2013* proceedings).

Table 4.8 shows the NLP components and external resources used by the participating systems. The NLP tools often integrated both syntactic (e.g., Standford parser) and semantic information (e.g., MetaMap that uses information from UMLS, or ChemSpot from the ChEBI ontology). All the teams used external terminological resources and/or databases to either create term dictionaries for entity recognition, such as **UEM_UC3M** and **NLM_LHC** teams, or to post-process and improve the results. All the systems used domain-specific resources, such as pharmacological terminologies and databases, while only the **UMCC_DLSI** used the general English lexical database WordNet. Finally, **LASIGE** is the only team that used an additional training data collection to develop its system. In addition to the DDI corpus, this team used a patent document corpus created by the ChEBI team.

4.2.2 RE task: extraction of drug-drug interactions

The goal of this subtask was the extraction of DDIs from biomedical texts. While the previous *DDIExtraction 2011 task* focused on the identification of all possible pair of interacting drugs, the latest edition *SemEval-2013 DDIExtaction task* also pursued the classification of each DDI according to one of the four following types: `advice`, `effect`, `mechanism`, and `int`. A detailed description of these types can be found in **Section 3.3.1**. Gold-standard annotations (correct, human-created annotations) of pharmacological substances were provided to participants for both training and test data. The test data for this subtask consisted of 158 DrugBank documents and 33 MEDLINE abstracts. Each participant system must output an ASCII list including all pairs of drugs in each sentence, one per line (multiple DDIs from the same sentence should appear on separate lines), its prediction (1 if the pair is a DDI and 0 otherwise), and its type (label null when the prediction value is 0) and formatted as:

IdSentence|IdDrug1|IdDrug2|prediction|type

The task of extracting DDIs from biomedical text attracted the participation of eight teams (see **Table 4.9**). A detailed description of them and their systems can be found in Segura-Bedmar et al. (2014).

For the evaluation, the test dataset with the golden annotations only for pharmacological substances was released to participants. Then, the evaluation was conducted by comparing the annotation predicted by each system to the golden annotations. The evaluation results were reported using the standard recall/precision/f-score metrics.

Table 4.10 shows the results of the DDI detection task. These results are not directly comparable with those reported in the *DDIExtraction 2011task* due to the use of different training and test datasets in each edition. However, it should be noted that there has been a significant improvement in the detection of DDIs: almost all participants (except for the two worst teams) achieved an F-score above 65.4% (the best F1 in the *DDIExtraction 2011 task*). The increase in the size of the corpus made for the *SemEval-2013 DDIExtraction task*, the inclusion of different types of documents, and the quality of their annotations might have been a significant contribution to this improvement.

The best system (*Run1* submitted by the **FBK-irst** team) had precision of 83.8% and recall of 83.8% (F1 82.7%) on the DDI-DrugBank dataset, while it had precision of 55.8% and recall of 55.5% (F1 53%) on the DDI-MEDLINE dataset. It should be noted that there was almost a 30 point F-score difference between the DDI-DrugBank and the DDI-MEDLINE datasets. Indeed a common characteristic observed in all systems was the strong decrease in their results on the DDI-MEDLINE dataset compared to the DDI-DrugBank dataset. This may be justified by the different styles of the two sources.

In the one hand, the texts taken from DrugBank are manually curated to provide brief descriptions of DDIs. Therefore, DrugBank contains short and concise sentences. On the other hand, the main topic of the scientific texts from MEDLINE would not necessarily be on DDIs. Moreover these texts are characterized by a very scientific language and it is common the use of long and subordinated sentences. The error analysis (see **Section 4.2.3**) shows that the systems failed drastically for long and complex sentences.

Another possible reason might be the different size between the two subcorpora. In addition, while the best system obtained balanced results in both precision and recall, the rest of the participants showed biased scores towards one or the other metric. We have observed that the use of biomedical parsers seemed to provide better performance than parsers trained for a general domain, and that the kernel-based systems in general overcame the feature-based ones.

The DDI classification task did not only consist of the identification of all possible pairs of interacting drugs, but also their classification. The results did not exceed an F1 of 65.1% (**FBK-irst** team) on the DDI-DrugBank dataset and 42% (**SCAI** team) on the DDI-MEDLINE dataset (see **Table 4.11**). These results clearly demonstrate that the identification of what type of information (such as an advice, an effect or information about the way the interaction occurs) is being used to describe a DDI may be a very complex task.

As in the DDI detection task, all systems (except the runs submitted by the **FBK-irst** team) showed a marked disparity between precision and recall. **Figure 4.1** and **Figure 4.2** show the results for each type of DDI on the DDI-Drug-Bank and DDI-MEDLINE test datasets, respectively. From each participant, we only select its best run. **Figure 4.1** suggests that some types of DDI were more difficult to classify than others on the DDI-DrugBank dataset, being the `advice` type relationship the easiest one. One possible explanation for this could be that recommendations or advice regarding a drug interaction are typically described by very similar text patterns such as *'DRUG should not be used in combination with DRUG'* or *'Caution should be observed when DRUG is administered with DRUG'*. The participating systems achieved very similar performance for the `mechanism` and `effect` relationships, while the `int` relationships seemed to be the most difficult to extract. This may be because the proportion of instances of `int` relationship (5.6%) in the DDI corpus is much smaller than those of the rest of the relations (41.1% for effect, 32.3% for `mechanism`, and 20.9% for `advice`).

	Team	Institution	Country	Description
RE Task	FBK-irst	Fondazione Bruno Kessler	Italy	hybrid kernel + scope of negations and semantic roles
	NIL_UCM	Complutense University of Madrid	Spain	SVM classifier (Weka SMO)
	SCAI	Fraunhofer SCAI	Germany	SVM classifier (LibLINEAR)
	UC3M	Carlos III University of Madrid	Spain	SL Kernel
	UCOLORADO_SOM	University of Colorado School of Medicine	USA	SVM classifier (LIBSVM)
	UTurku	University of Turku	Finland	TEES system
	UWM-TRIADS	University of Wisconsin-Milwaukee	USA	SVM classifier
	WBI	Humboldt University of Berlin	Germany	APG, SL, ST, SST, SpT kernels + TEES system + Moara system

Table 4.9. Summary of the *SemEval-2013 DDIExtraction* RE task participant teams

Team	Run	DrugBank				MEDLINE				Overall			
		Rank	P	R	F1	Rank	P	R	F1	Rank	P	R	F1
FBK-irst	1	1	0.816	0.838	0.827	1	0.558	0.505	0.53	1	0.794	0.806	0.8
	2	2	0.186	0.838	0.827	2	0.558	0.505	0.53	2	0.794	0.806	0.8
	3	3	0.186	0.838	0.827	3	0.558	0.505	0.53	3	0.794	0.806	0.8
WBI-DDI	1	6	0.857	0.686	0.762	8	0.63	0.358	0.456	6	0.841	0.841	0.736
	2	5	0.874	0.696	0.775	12	0.651	0.295	0.406	5	0.861	0.861	0.745
	3	4	0.814	0.755	0.783	4	0.625	0.421	0.503	4	0.801	0.801	0.759
UTurku	1	10	0.846	0.614	0.712	20	0.724	0.221	0.339	11	0.841	0.841	0.684
	2	9	0.861	0.624	0.724	19	0.778	0.221	0.344	8	0.858	0.858	0.696
	3	8	0.843	0.638	0.726	15	0.658	0.263	0.376	9	0.833	0.833	0.699
SCAI	1	11	0.836	0.619	0.711	7	0.688	0.347	0.462	10	0.826	0.826	0.69
	2	12	0.837	0.617	0.71	17	0.686	0.253	0.369	12	0.829	0.829	0.683
	3	7	0.796	0.681	0.734	6	0.431	0.526	0.474	7	0.748	0.748	0.704
UC3M	1	12	0.656	0.758	0.703	13	0.392	0.421	0.406	13	0.632	0.632	0.676
	2	19	0.415	0.814	0.549	10	0.313	0.642	0.421	19	0.404	0.404	0.537
NIL_UCM	1	16	0.615	0.615	0.615	22	0.419	0.137	0.206	16	0.608	0.608	0.588
	2	14	0.673	0.688	0.68	21	0.548	0.242	0.336	14	0.667	0.667	0.656
UWM_TRIADS	1	17	0.525	0.689	0.596	11	0.415	0.424	0.419	17	0.517	0.517	0.581
	2	15	0.573	0.665	0.616	9	0.427	0.446	0.436	15	0.561	0.561	0.599
	3	18	0.465	0.746	0.573	5	0.387	0.63	0.479	18	0.458	0.458	0.564
UColorado_SOM	1	22	0.387	0.739	0.508	16	0.256	0.663	0.37	21	0.37	0.37	0.492
	2	20	0.391	0.765	0.518	14	0.28	0.663	0.394	22	0.378	0.378	0.504
	3	21	0.422	0.646	0.511	18	0.253	0.6	0.356	23	0.398	0.398	0.491

Table 4.10. Results for the DDI detection task on test dataset (P for precision; R for recall; F1 for F-score)

Team	Run	DrugBank				MEDLINE				Overall			
		Rank	P	R	F1	Rank	P	R	F1	Rank	P	R	F1
FBK-irst	1	3	0.654	0.672	0.663	4	0.407	0.368	0.387	3	0.633	0.642	0.638
	2	1	0.667	0.686	0.676	2	0.419	0.379	0.398	1	0.646	0.656	0.651
	3	2	0.664	0.682	0.673	3	0.419	0.379	0.398	2	0.643	0.653	0.648
WBI-DDI	1	6	0.702	0.561	0.624	7	0.463	0.263	0.336	6	0.685	0.532	0.599
	2	5	0.707	0.563	0.627	12	0.488	0.221	0.304	5	0.695	0.53	0.601
	3	4	0.657	0.609	0.632	5	0.453	0.305	0.365	4	0.642	0.579	0.609
UTurku	1	9	0.723	0.525	0.608	18	0.517	0.158	0.242	9	0.714	0.489	0.581
	2	7	0.738	0.535	0.62	16	0.593	0.168	0.262	7	0.732	0.499	0.594
	3	8	0.706	0.534	0.608	13	0.5	0.2	0.286	8	0.694	0.502	0.582
NIL_UCM	1	12	0.54	0.541	0.54	22	0.387	0.126	0.19	12	0.535	0.501	0.517
	2	10	0.566	0.579	0.573	19	0.357	0.158	0.219	10	0.557	0.538	0.548
UC3M	1	11	0.518	0.598	0.555	15	0.265	0.284	0.274	11	0.495	0.568	0.529
	2	21	0.231	0.454	0.306	21	0.138	0.284	0.186	21	0.222	0.437	0.294
SCAI	1	15	0.546	0.404	0.464	1	0.625	0.316	0.42	14	0.551	0.395	0.46
	2	16	0.545	0.402	0.463	8	0.6	0.221	0.323	16	0.548	0.384	0.452
	3	14	0.513	0.439	0.473	6	0.31	0.379	0.341	15	0.486	0.433	0.458
UWM_TRIADS	1	17	0.407	0.534	0.462	10	0.309	0.315	0.312	17	0.4	0.513	0.449
	2	13	0.452	0.524	0.485	9	0.312	0.326	0.319	13	0.439	0.505	0.47
	3	18	0.361	0.579	0.445	11	0.247	0.402	0.306	18	0.35	0.562	0.432
UColorado_SOM	1	22	0.166	0.317	0.218	20	0.13	0.337	0.188	22	0.161	0.319	0.214
	2	20	0.258	0.503	0.341	14	0.196	0.463	0.275	20	0.25	0.499	0.334
	3	19	0.288	0.441	0.349	17	0.173	0.411	0.244	19	0.272	0.438	0.336

Table 4.11. Results for the DDI detection and classification task on test dataset (P for precision; R for recall; F1 for F-score)

A common characteristic of all participating systems is the use of support vector machines (SVMs). While most systems used feature-based methods, only three teams (**FBK-irst**, **WBI-DDI** and **UC3M**) applied kernel-based methods, which in general achieved better performance than the feature-based ones. Unlike feature-based methods, kernel-based methods do not require the explicit definition of feature vectors. A kernel-based method contains a kernel function and a kernel learner. A kernel function is a function that computes the similarity between two instances (for example, drug pairs). A kernel learner (such as SVM) is a learning algorithm that performs a learning task in a feature space.

Most participating systems separated the learning problem into two stages: first, the DDIs were detected and then, they were classified into one of the types proposed in the guidelines. The only exceptions are the **UTurku** and **NIL_UCM** teams. The TEES system, developed by the **UTurku** team, uses a multiclass SVM on a rich graph-based feature set. The **NIL_UC3M** team trained a multi-class SVM classifier with five classes (`mechanism`, `effect`, `advice`, `int` and null for negative instances). The **NIL_UC3M** also developed an approach in which the DDI detection and classification stages were separated. The evaluation on test dataset showed that the two-stage approach yielded better results than those achieved by the multi-class classifier. As regards the two-stage approaches, the first stage, the detection of DDIs, was always performed by a binary classifier responsible for distinguishing between negative and positive DDIs instances. Most teams treated each DDI type as a single classification sub-problem (one-vs-all). The **SCAI** team was the only one that did not use any machine learning techniques in the classification task. DDI instances detected in the previous step were classified according to a set of trigger words related to each type of DDIs.

Regarding the NLP tools often integrated into the participating systems, stemming, POS tagging, and syntactic parsing were the most common ones. Stanford parser tools (Klein & Manning, 2003) were widely used by most systems. Around half of the participant systems used the Charniak–Johnson parser (Charniak & Johnson, 2005) with David McClosky's biomodel (McClosky, 2010) trained on the GENIA corpus and unlabeled PubMed articles. From the results of the **FBK-irst**, **WBI,** and **UTurku** teams, we can confirm that the parsers for the biomedical domain provided better performance than parsers trained for a general domain. Some systems also used additional elements, such as lemmatization (**WBI** and **UWM_TRIADS** teams), semantic parsing provided by MMTx (**UTurku** and **NIL_UCM** teams), or disease named entity recognition (team **FBK-irst**). Negation detection was only used by one team (**FBK-irst**). Surprisingly, only half of the participating systems used external lexical resources such as dictionaries or ontologies. **Table 4.12** shows the NLP components and external resources used by the participating systems. None of them made use of any additional training data collections to develop their systems, which implies that all systems relied only on the training dataset provided by the DDI corpus.

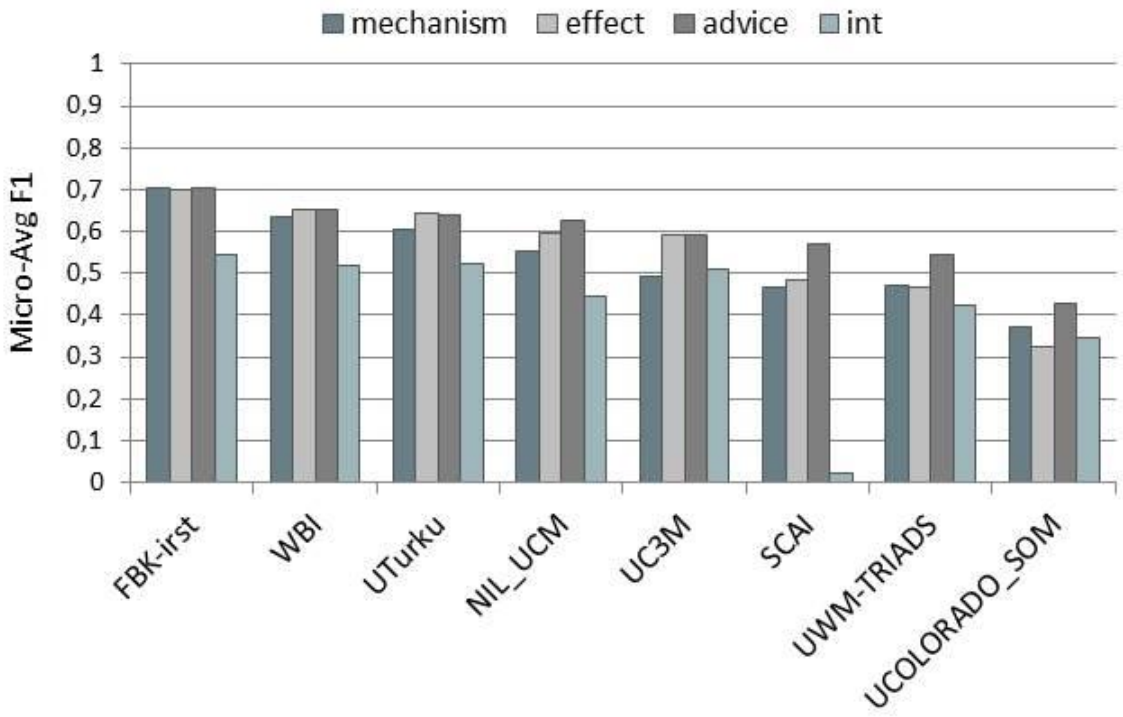

Figure 4.1. Micro-Avg F1 scores by DDI type on the DDI-DrugBank test dataset

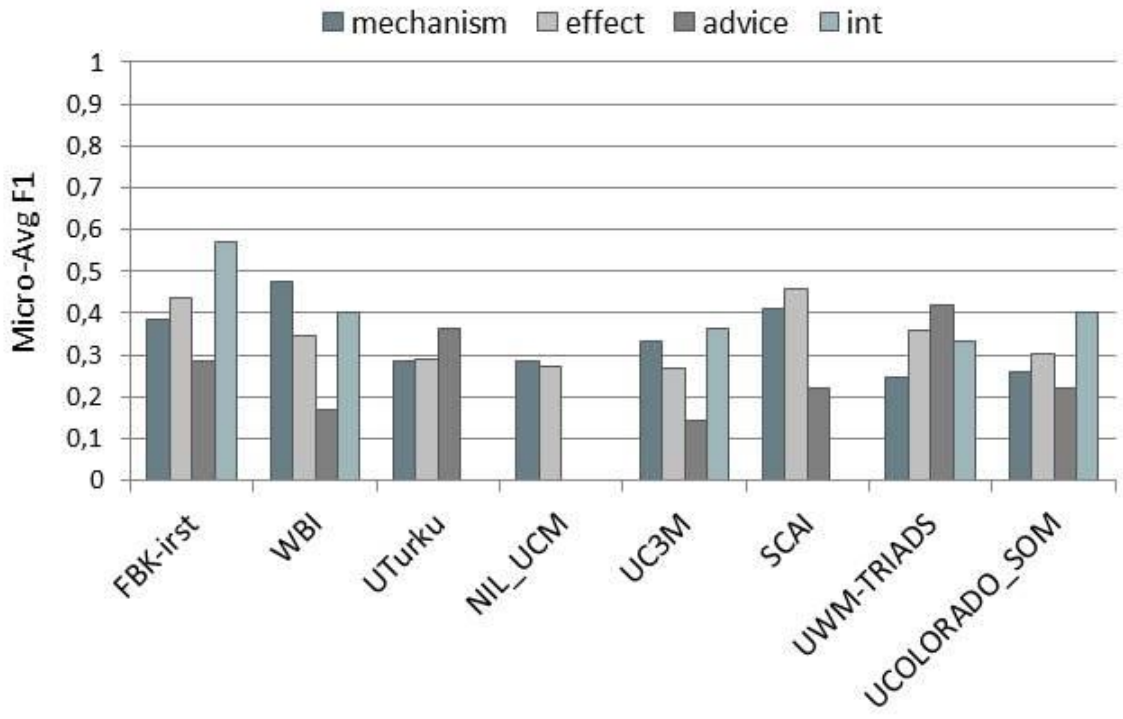

Figure 4.2. Micro-Avg F1 scores by DDI type on the DDI-MEDLINE test dataset

Team	NLP tools	Knowledge resources
FBK-irst	Charniak-Johnson reranking parser McClosky's biomodel Standford parser BioEnEx (NER tool for diseases)	
WBI-DDI	Charniak-Johnson reranking parser Mc-Closky's biomodel Standford converter BioLemmatizer	DrugBank PHARE ontology
UTurku	Charniak-Johnson reranking parser Mc-Closky's biomodel Standford parser tools MetaMap	WordNet DrugBank
NIL_UCM	TreeTagger PaiceHusk Stemmer Standford parser tools MMTx	
UC3M	GATE Standford parser plug-in	ATC system
SCAI	Porter Stemming algorithm Charniak-Lease parser	
UCOLORADO_SOM	Genia Dependency Parser Porter Stemmer	OpenDMAP WordNet
UWM_TRIADS	Standford NLP tools Dragon tool (a lemmatizer)	FDA Drug classification

Table 4.12. NLP tools and other resources used by the RE participating teams (References for the NLP tools and knowledge resources can be found in each individual paper in the *SemEval-2013* proceedings).

4.2.3 Error analysis of RE systems

The aim of this section is to perform a detailed error analysis of the results of the RE systems with the objective of providing a road map for future work in the extraction of DDIs from text, as well as to identify further improvements in the DDI corpus. To this end, we focus on the study of the main source of errors produced by the systems developed by the following teams: **FBK-irst**, **WBI-DDI,** and **Uturku**. For each team, we only analyze their best runs. The reason for this choice is that these systems were the top-performing in *SemEval-2013 DDIExtraction task*.

- **Analysis of false negatives:**

Table 4.13 and **Table 4.16** present the main causes for the false negatives in the DDI-DrugBank and the DDI-MEDLINE datasets, respectively. From **Table 4.13**, we can see that one of the most important factors contributing to false-negatives in DrugBank texts was the lack of cataphora resolution in the three systems. The resolution of the appositions in sentences, prior to the detection of DDIs, could allow furthering improving

the performance, particularly the **FBK-irst** and **UTurku** systems. Similarly, the resolution of anaphora and the detection of coordinate structures may also help to reduce false negatives, though fewer than the resolution of cataphoras and appositions. Another major cause of false negatives is that many DDIs are described with very unusual text patterns. The high variability of natural language expression allows DDIs to be able to be composed using many different lexical and syntactic realizations. Classifiers have problems in detecting these cases since they are probably unrepresented in the training data.

Error cause	FBK-irst	WBI	UTurku	Example
Detection of coordinate structures required	14	23	17	E1, E2
Detection of appositions required	28	18	96	E3, E4
Unusual patterns for coadministration	11	18	27	E5, E6
Unusual patterns for DDI	28	51	20	E7, E8
Long DDI descriptions	6	27	20	E9
Unobvious DDIs	10	6	27	E11, E12
Resolutions of percentages, dosages and temporal expressions required	8	4	26	E13, E14
Resolution of anaphora required	7	17	17	E15
Resolution of cataphora required	27	40	46	E16
Resolution of complex and compound sentences required	6	13	36	E17
Total	143	217	332	

Table 4.13. Analysis of false negatives in the DDI-DrugBank dataset

Table 4.14 and **Table 4.15** show some examples of false negatives in the DrugBank dataset. Long, complex, and compound sentences are other sources of false negatives. Many DDIs are described in long and complex sentences, which usually have a complex syntactic and lexical structure. Sentences with several embedded subordinated clauses are often encountered in both DrugBank and MEDLINE texts. Moreover, these sentences also pose a challenge to syntactic parsers due to their high levels of ambiguity. This might be one of the reasons explaining why the methods using syntactic features from parsers (e.g., Standford parser) were not capable of dealing with these type of sentences. The **FBK-irst** system showed a lower rate of false negatives (only 2% were classified as long DDI descriptions, and only 4% as complex and compound sentences) compared to the other two systems. In this case, the use of semantic roles, which are used to rule out possible negative instances, could be helping to overcome a wrong syntactic analysis.

Some sentences describe DDIs without giving an absolute certainty of their existence or using uncommon patterns. For example, in the sentence '*Lapatinib may have the potential to convert Herceptin-refractory to Herceptin-sensitive tumors in HER2-positive breast cancer by up-regulation of the cell surface expression of HER2'*, it is even difficult for a human being to determine whether the sentence describes a DDI or not. The detection of dosages and numeric and temporal expressions can also help to improve the performance of the systems, since many sentences describe DDIs including additional information such as dosages, dosage regimen, or per cents of change of parameters, among others.

ID	Example	DDIs not detected
E1	Several studies demonstrate a decrease in the bioavailability of **methyldopa**$_{e_1}$ when it is ingested with **ferrous sulfate**$_{e_2}$ or **ferrous gluconate**$_{e_3}$	(e_1, e_3)
E2	**Sulfoxone**$_{e_1}$ may increase the effects of **barbiturates**$_{e_2}$, **tolbutamide**$_{e_3}$, and **uricosurics**$_{e_4}$	(e_1, e_2); (e_1, e_4)
E3	Concurrent administration of bacteriostatic **antibiotics**$_{e_1}$ (e.g., **erythromycin**$_{e_2}$, **tetracycline**$_{e_3}$) may diminish the bactericidal effects of **penicillins**$_{e_4}$ by slowing the rate of bacterial growth	(e_3, e_4)
E4	Other inhibitors of the cytochrome P450 3A4 enzyme system, such as **anitmycotic agents**$_{e_1}$ (e.g., **itraconazole**$_{e_2}$ and **miconazole**$_{e_3}$) or **macrolide antibiotics**$_{e_4}$ (e.g., **erythromycin**$_{e5}$ and **clarithromycin**$_{e_6}$), may alter **oxibutynin**$_{e_7}$ mean pharmacokinetic parameters (i.e., Cmax and AUC)	(e_1, e_7); (e_2, e_7); (e_3,e_7); (e_4, e_7); (e_5, e_7); (e_6, e_7)
E5	The occurrence of stupor, muscular rigidity, severe agitation, and elevated temperature has been reported in some patients receiving the combination of **selegiline**$_{e_1}$ and **meperidine**$_{e_2}$	(e_1, e_2)
E6	The addition of **aspirin**$_{e_1}$ to **Streptokinase**$_{e_2}$ in the risk of minor bleeding	(e_1, e_2)
E7	There is usually complete cross-resistance between **PURINETHOL**$_{e_1}$ and **TABLOID**$_{e_2}$	(e_1, e_2)
E8	Concomitant treatment with **NEXAVAR**$_{e_1}$ resulted in a 21% increase in the AUC of **doxorubicin**$_{e_2}$	(e_1, e_2)
E9	Other drugs such as **cisapride**$_{e_1}$ or **pimozide**$_{e_2}$, which are metabolised by hepatic CYP3A isozymes have been associated with QT interval prolongation and/or cardiac arrhythmias (typically torsades de pointe) as a result of increase in their serum levels subsequent to the interaction with significant inhibitors of the isozyme, including some macrolide **antibacterials**$_{e_3}$	(e_1, e_3); (e_2, e_3)

Table 4.14. Examples of false negatives in the DDI-DrugBank dataset

ID	Example	DDIs not detected
E10	Certain **macrolides**$_{e1}$ interact with **terfenadine**$_{e2}$ and **astemizole**$_{e3}$ leading to increased serum concentrations of the latter	(e_1, e_2); (e_1, e_3)
E11	**Furosemide**$_{e1}$ and probably other **loop-diuretics**$_{e2}$ given concomitantly with **metolazone**$_{e3}$ can cause unusually large or prolonged losses of fluid and electrolytes	(e_2, e_3)
E12	Concomitant administration of **alcohol**$_{e1}$ had a minimal effect on plasma levels of **mirtazapine**$_{e2}$	(e_1, e_2)
E13	Concomitant administration of **aspirin**$_{e1}$ (1000 mg TID) to healthy volunteers tended to increase the AUC (10%) and Cmax (24%) of **meloxicam**$_{e2}$	(e_1, e_2)
E14	All patients taking **NSAIDs**$_{e1}$ should interrupt dosing for at least 5 days before, the day of, and 2 days following **ALMINTA**$_{e2}$ administration	(e_1, e_2)
E15	Although **minoxidil**$_{e1}$ does not itself cause orthostatic hypotension, its administration to patients already receiving **guanethidine**$_{e2}$ can result in profound orthostatic effects	(e_1, e_2)
E16	Drugs which may potentiate the myeloproliferative effects of **Leukine**$_{e1}$, such as **minoxidil**$_{e2}$, **lithium**$_{e3}$ and **corticosteroids**$_{e4}$, should be used with caution	(e_1, e_2); (e_1, e_3); (e_1, e_4)
E17	**Mexitil**$_{e1}$ does not alter serum **digoxin**$_{e2}$ levels but **magnesium-aluminium hydroxide**$_{e3}$, when used to treat gastrointestinal symptoms due to **Mexitil**$_{e4}$, has been reported to lower serum **digoxin**$_{e5}$ levels	(e_3, e_5)

Table 4.15. Examples of false negatives in the DDI-DrugBank dataset (cont. 2)

As regards the DDI-MEDLINE dataset (see **Table 4.16**), false negatives have similar error sources to those in the DDI-DrugBank dataset. The major cause of false negatives for all three systems was their inability to detect those DDIs described by patterns that are very scarce, even unrepresented, in the training data. This may be due mainly to the small size of the training dataset from MEDLINE. The detection of doses and numerical and temporal expressions also seem to be another significant problem that these systems would have to face in order to improve their performance in the detection of DDIs. Anaphoras, cataphoras, coordinated structures, and appositions have a much less significant effect on the false negatives in the DDI-MEDLINE dataset than in the DDI-DrugBank dataset. A possible reason for this could be that many texts in DrugBank provide descriptions of DDIs involving a drug and a list of drugs. The use of these linguistic structures is very common and useful in providing these kinds of description. **Table 4.17** shows some examples of false negatives in the DDI-MEDLINE dataset.

Error cause	FBK-irst	WBI	UTurku	Example
Detection of coordinate structures required	3	0	7	E18
Detection of appositions required	5	6	3	
Unusual patterns for coadministration	17	17	15	E20
Unusual patterns for DDI	5	10	30	E21
Long DDI descriptions	3	4	2	E22
Unobvious DDIs	3	3	2	E23
Resolutions of percentages, dosages and temporal expressions required	4	7	9	E24
Resolution of anaphora required	2	2	2	E25
Resolution of cataphora required	0	0	2	E26
Resolution of complex and compound sentences required	5	6	2	E27
Total	47	55	74	

Table 4.16. Analysis of false negatives in the DDI-MEDLINE dataset

ID	Example	DDIs not detected
E18	**AAV2**$_{e_1}$–mediated retinal transduction is improved by co-injection of **heparinise III**$_{e_2}$ or chondroitin **ABC lyase**$_{e_3}$	(e_1, e_3)
E19	It is better to avoid prescribing isoenzyme CYP 2D6 inhibitors to women treated with **tamoxifen**$_{e_1}$ for breast cancer, especially **SSRI antidepressants**$_{e_2}$ such as **paroxetine**$_{e_3}$ and **fluoxetin**$_{e_4}$	(e_1, e_2); (e_1, e_3); (e_1, e_4)
E20	**Warfarin**$_{e_1}$ users who initiated **citalopram**$_{e_2}$, **fluoxetine**$_{e_3}$, **paroxetine**$_{e_4}$, **amitriptyline**$_{e_5}$, or **mirtazapine**$_{e_6}$ had an increased risk of hospitalization for gastrointestinal bleeding	(e_1, e_2); (e_1, e_3); (e_1, e_4); (e_1, e_5); (e_1, e_6)
E21	Reduction of PTH by **cinacalcet**$_{e_1}$ is associated with a decrease in **darbepoetin**$_{e_2}$ requirement	(e_1, e_2)
E22	In an in vitro assay, **lapatinib**$_{e_1}$ induced HER2 expression at the cell surface of HER2-positive breast cancer cell lines, leading to the enhancement of **Herceptin**$_{e_2}$–mediated ADCC	(e_1, e_2)

Table 4.17. Examples of false negatives in the DDI-MEDLINE dataset

ID	Example	DDIs not detected
E23	However, the evidence for a **calcium**$_{e_1}$ effect on **iron**$_{e_2}$ absorption mainly comes from studies that did no isolate the effect of calcium from that of other dietary components, because it was detected in single-meals studies	(e_1, e_2)
E24	Systemic and apparent oral **midazolam**$_{e_1}$ clearance were 24% (269,73 vs. 354,102 ml/min, P=0.022) and 31% respectively, lower in **cyclosporine**$_{e_2}$-treated patients (n=20) than in matched tracrolimus-treated patients (n=20)	(e_1, e_2)
E25	Acute administration of **hemantane**$_{e_1}$ or **doxycycline**$_{e_2}$ failed to influence locomotion in mice, while their combination normalized motor activity	(e_1, e_2)
E26	Regulatory agencies state that the combination of **clopidogrel**$_{e_1}$ and the CYP2C19 inhibitors **omeprazole**$_{e_2}$ and **esomeprazole**$_{e_3}$ should be avoided	(e_1, e_2); (e_1, e_3)
E27	Exposure to oral **S-ketamine**$_{e_1}$ is unaffected by **itraconazole**$_{e_2}$ but greatly increased by **ticlopidine**$_{e_3}$	(e_1, e_3)

Table 4.17 (cont). Examples of false negatives in the DDI-MEDLINE dataset

- **Analysis of false positives:**

Table 4.18 and **Table 4.20** show the main causes of false positives in the DDI-DrugBank and DDI-MEDLINE datasets, respectively. The major cause of false positives in DrugBank refers to sentences in which interacting drugs have more than one mention. The systems were able to detect that there was an interaction between two drugs, but failed to identify the mentions that were actually involved in this DDI.

Error cause	FBK-irst	WBI	UTurku	Example
Incorrect pair	57	60	36	E28
Annotation error	27	19	19	E29
Resolution of coordinated structures required	31	21	12	E30
Same drug	8	28	10	E31
Lack of evidence	41	21	9	E32
Resolution of apposition structures required	3	3	3	
Total	167	152	89	

Table 4.18. Analysis of false positives in the DDI-DrugBank dataset

The first example in **Table 4.19** (see E28) shows a sentence describing a DDI between *ibuprofen* and *ALIMTA®*. We can see that both drugs appear twice in the sentence, but only their last two mentions are involved in the description of a DDI.

However, the three systems failed to detect this DDI, because they proposed the pair formed by the first two mentions of the drugs. Annotation errors (see E29) are the second source of false positives in DrugBank. The candidate pairs are correctly detected by the systems, but are not annotated in the DDI corpus. Another cause of false positives is the systems' incapability to distinguish between drugs constituting a coordinate structure, and therefore, to recognize that they are not describing a DDI (see E30). Notably, one of the main sources of false positives would be resolved with a simple rule preventing mentions of drugs referring to the same drug, which could be considered as a candidate DDI (see E31). The lack of evidence to confirm the existence of a DDI is another source of false positives (see E32). In fact, it is the main cause of false positives in the DDI-MEDLINE dataset (see **Table 4.20** and **Table 4.21**).

ID	Example	FP	Gold DDIs
E28	Although **ibuprofen**$_{e_1}$ (400 mg qid) can be administered with **ALIMTA**$_{e_2}$ in patients with normal renal function (creatinine clearance 80 mL/min), caution should be used when administering **ibuprofen**$_{e_3}$ concurrently with **ALIMTA**$_{e_4}$ to patients with mild to moderate renal insufficiency (creatinine clearance from 45 to 79 mL/min)	(e_1, e_2)	(e_3, e_4)
E29	Careful monitoring of **phenytoin**$_{e_1}$ concentrations in patients receiving **DIFLUCAN**$_{e_2}$ and phenytoin is recommended	(e_1, e_2)	
E30	It may also interact with **thiazides**$_{e_1}$(increased thrombocytopenia), **cyclosporine**$_{e_2}$(increased nephrotoxicity), **sulfonylurea agents**$_{e_3}$ (increased hypoglycemic response), **warfarin**$_{e_4}$ (increased anticoagulants effect), **methotrexate**$_{e_5}$(decreased renal excretion of **methotrexate**$_{e_6}$), **phenytoin**$_{e_7}$(decreased hepatic clearance of **phenytoin**$_{e_8}$)	(e_1, e_7); (e_2, e_7); (e_3, e_7); (e_5, e_7); (e_5, e_7)	
E31	Severe toxicity has also been reported in patients receiving the combination of **tricyclic antidepressants**$_{e_1}$ and **ELDEPRYL**$_{e_2}$ and **selective serotonin reuptake inhbitors**$_{e_3}$ and **ELDEPRYL**$_{e_4}$	(e_2, e_4)	(e_1, e_2); (e_3, e_4)
E32	There are no clinical data on the use of **MIVACRON**$_{e_1}$ with other **non-depolarizing neuromuscular blocking agents**$_{e_2}$	(e_1, e_2)	
E33	**Tetracycline**$_{e_1}$, a **bacteriostatic antibiotic**$_{e_2}$, may antagonize the bactericidal effect of **penicillin**$_{e_3}$and concurrent use of these drugs should be avoided	(e_2, e_3)	(e_1, e_3)

Table 4.19. Examples of false positives in the DDI-DrugBank dataset

Error cause	FBK-irst	WBI	UTurku	Example
Incorrect pair	11	10	2	E34
Annotation error	2	1	1	E35
Lack of evidence	35	13	3	E36
Total	48	24	6	

Table 4.20. Analysis of false positives in the DDI-MEDLINE dataset

ID	Example	FP	Gold DDIs
E34	**Moxifloxacin**$_{e_1}$ and **lomefloxacin**$_{e_2}$ reacts faster with **sucralfate**$_{e_3}$ and **gelusil**$_{e_4}$ in acidic media whereas with **erythromycin**$_{e_5}$ in basic media and multi-minerals in neutral media	(e_3, e_4)	(e_1, e_2); (e_1, e_4); (e_1, e_5); (e_2, e_3); (e_2, e_4); (e_2, e_5)
E35	Improved parathyroid hormone control by **cinacalcet**$_{e_1}$ is associated with reduction in **darbepoetin**$_{e_2}$requirements in patients with end-stage renal disease	(e_1, e_2)	
E36	On day 8, a single **panobinostat**$_{e_1}$ dose was co-administered with **ketoconazole**$_{e_2}$	(e_1, e_2)	

Table 4.21. Examples of false positives in the DDI-MEDLINE dataset

4.2.4 Conclusions and future directions

The goal of the *SemEval-2013 DDIExtraction task* was to promote the development of IE techniques applied to the detection of drug names and DDIs from biomedical texts. There were a total of 38 runs submitted by 14 different teams from 7 different countries (6 of the teams participated in the drug name recognition task, while 8 participated in the DDI extraction task). The highest F1 scores obtained were 71.5% for drug name recognition and classification and 65.1% for extraction and classification of DDIs.

In the NER task, the participant systems performed well in recognizing generic drugs, brand drugs, and groups of drugs, but they failed in recognizing active substances not approved for human use. The **WBI** team achieved the highest F-score on DrugBank texts (65.6%), while **LASIGE** was the best team on the DDI-MEDLINE dataset (57.7%). Therefore, although the results were positive, there is still ample room to improve in drug NER.

Concerning the task of detection of DDIs, the participating systems demonstrated substantial progress over the previous *DDIExtraction 2011 task*. The best team, **FBK-irst**, achieved a competitive F-score of 82.7% on DrugBank texts. However, performance on the DDI-MEDLINE dataset was lower mainly due to the limited size of its training

dataset. Another possible reason may be that MEDLINE texts have a greater complexity than DrugBank texts. All teams have used machine-learning methods, specifically SVM. In general, non-linear kernel-based methods overcame linear SVMs.

We conclude that research into DDI extraction must continue. The error analysis points out the main limitations of the participating systems. Current approaches have focused on syntactic aspects, drawing their attention to the sentence structure. The resolution of linguistic phenomena such as cataphora, anaphora, appositive and coordinate structures, and complex sentences, among others, might lead to better performance.

However, few participating systems took into account the sentence meaning. Approaches using domain knowledge have been recently applied with success to the pharmacological domain (Garten, Coulet, & Altman, 2010; Kang et al., 2014). The use of knowledge resources could reduce the number of false positives generated by the current DDI extraction systems, because these resources can help to distinguish between those pairs of drugs that are DDIs from those that are not. The information required for a semantic-based IE system could be taken, for example, from pharmacological databases such as DrugBank, PharmGKB (Hewett et al., 2002), SIDER (Kuhn, Campillos, Letunic, Jensen, & Bork, 2010) and KEGG (Kanehisa, Goto, Sato, Furumichi, & Tanabe, 2012), among others. Some of them describe specific pairs of interacting drugs. For example, in DrugBank 39 different drugs that interact with *ciprofloxacin* are described. On the other hand, a larger number of DDIs can be deduced indirectly by exploiting, for example, the drug-protein relationships. Thus, the relationships of two different drugs with the same protein can be used to infer the mechanism leading to a DDI (Hage & Tweed, 1997). For example, *ciprofloxacin* is described to inhibit the activity of the metabolic enzyme *CYP1A2*, and *duloxetine* is described to be metabolized by *CYP1A2*. Therefore, there could be an interaction between *ciprofloxacin* and *duloxetine*. Similarly, the relationships of two different drugs with the same ADR can be used to infer possible DDIs (Campillos, Kuhn, Gavin, Jensen, & Bork, 2008). For example, *morphine* is related to the side effect central nervous system depression. Therefore, other drugs related to the same ADR, such as *oxycodone*, could interact with *morphine*.

Up to now, the main limitation for the development of semantic-based approaches has been the availability of appropriate knowledge bases in a machine-readable format. However, the creation of these knowledge bases is becoming more feasible and common in the pharmacological domain (Khelashvili et al., 2010; Whirl-Carrillo et al., 2012). This is due to the increasing number of databases and web servers providing structured and semi-structured pharmacological information, such as DrugBank or KEGG. Moreover, there are different community projects such as Bio2RDF (Belleau, Nolin, Tourigny, Rigault, & Morissette, 2008) or LODD (Samwald et al., 2011), which work to link the various sources of biological and pharmacological data together, enabling the integration of several pharmacological aspects described in different databases (Pathak, Kiefer, & Chute, 2013). Another important factor is the proliferation of biomedical ontologies to store and formally represent domain knowledge. Ontologies enable the integration of the information disperse through different and heterogeneous databases, and provide resources that can be exploited by IE systems (Wimalasuriya, 2010).

Therefore, future directions for DDI extraction might entail the combination of syntactic and semantic information. In addition, increasing the size of training dataset, in

particular for the DDI-MEDLINE dataset, would also have a very positive impact on the results.

4.3 Conclusions

In this chapter, we have corroborated the quality of the DDI corpus as a gold-standard for drug NER and DDI extraction from text.

On the one hand, the quality of the annotations has been ensured by the IAA. However, based on the analysis of these results, we propose further activities that will enhance the quality of the corpus.

- Multiple-annotation process involving at least three annotators. In this strategy, all documents are annotated by more than one annotator (Jagannathan et al., 2009; Wilbur et al., 2006).

- More detailed description of `drug_n` entities, which can be divided into different groups, in order to reduce the bias in the annotation of this entity type.

- Inclusion of a larger number of MEDLINE abstracts, in order to obtain a more balanced corpus respecting to the number of DrugBank texts.

On the other hand, we have evaluated the usefulness of the DDI corpus in the final application task. In this way, we have used it as a gold-standard in the *SemEval-2013 DDIExtraction task*. The DDI corpus has provided a common framework for the evaluation of different drug NER and DDI RE systems, which have shown a significant improvement with respect to the previous 2011 edition. The increase in size of the corpus, the inclusion of different types of documents, and the quality of their annotations might have contributed significantly to this improvement. The analysis of the results of the participating systems has provided future directions in order to enhance the utility of the corpus:

- Increasing the size of the DDI-MEDLINE dataset, to provide a larger coverage of patterns with a scarce representation in the current version.

- Annotation of linguistic phenomena required for a better understanding of the text, such as negation, modality, cataphora, or anaphora.

- Annotation of relevant pharmacological information, including quantitative information (drug dosage, time interval between administration of the drugs, alterations in PK parameters, drug concentration, etc.), and qualitative information (ADRs, indications, pharmaceutical forms, administration routes, etc.). This information can be useful for the development of new IE systems for both drug NER and DDI extraction.

Chapter 5

Semantic resources in the pharmacological domain: State of the art

The second main contribution of this thesis is the creation of a comprehensive ontology for the representation of all DDI-related knowledge, a resource designed for the computational community working on applications within the DDI domain. Prior to the development of this new ontology, we have studied the current state of the art in the field. Here, we provide a review of existing semantic resources in the pharmacological domain from a pharmacovigilance perspective, and specially focusing on their scope to represent DDI-related information.

The great amount of biomedical and pharmacological knowledge is organized in different terminological resources. Machine-processable artifacts have become important resources for NLP techniques. Controlled vocabularies, taxonomies, and thesaurus can be used as input for the creation of term lists for NER tasks (Grego & Couto, 2013; Lamurias et al., 2013). In addition to this, ontologies provide a contextual framework and semantic knowledge base that can be exploited for RE tasks (Yulan He, Road, & Ex, 2008; Huang, Zhu, Ding, Yu, & Li, 2006; Müller et al., 2004). During the last years, the number of biomedical ontologies has increased leading to a great availability of related resources. For example, the open repository BioPortal (Noy et al., 2009) provides access to more than 360 biomedical ontologies. Integration and mapping among them is a key point for the optimal sharing and interchange of information (Bodenreider, 2008; Smith et al., 2007).

There are several reviews of biomedical ontologies (Bodenreider & Stevens, 2006) and their applications (Bodenreider, 2008; Spasic, Ananiadou, McNaught, & Kumar, 2005; Stevens, Goble, & Bechhofer, 2000). From a pharmacological perspective, ontologies have been studied from the point of view of drug discovery (Vázquez-Naya et al., 2010), drug repurposing (Andronis, Sharma, Virvilis, Deftereos, & Persidis, 2011) and medicinal chemistry (Gómez-Pérez, Martínez-Romero, Rodríguez-González, Vázquez, & Vázquez-Naya, 2013). Since one of the main goals of this project is to create an ontology that represents all DDI-related information, the aim of this chapter is to provide an overview of those ontologies relevant in pharmacovigilance, that is, the science and activities related to the collection, analysis, and prevention of ADRs or any other drug-related problem, with a special interest in the representation of DDI-related information. To this purpose, we focus on the following characteristics:

1. representation of pharmacological substances: scope and level of granularity.
2. representation of DDIs
3. representation of other concepts related to pharmacovigilance: ADRs.
4. intra-relationships: relations between different concepts in every ontology.
5. inter-relationships: relations between concepts among different ontologies.
6. use of unique identifiers, synonyms, and definitions in natural language.
7. format and availability.

This section is organized as follows. Firstly, we review in **Section 5.1** those terminological resources collecting chemical substances, including drugs. **Section 5.2** focuses on the terminologies specifically created as repositories for pharmacological substances. Then, in **Section 5.3**, an overview of ontologies for ADRs is provided. **Section 5.4** deals with ontologies related, in some way, to the DDI domain. The strengths and weaknesses of the reviewed resources are discussed in **Section 5.5**, and a review of the main limitation that should be addressed in this thesis is provided in **Section 5.6**. Finally, the main conclusions of this chapter are highlighted in **Section 5.7**.

5.1 Terminological resources for chemical substances

The first aspect in the study of DDI-related terminologies is how those substances relating to the domain are represented. There are different types of chemical substances relevant to the biomedical domain. Pharmacological substances have a key role due to their capacity to cure, prevent, or diagnose diseases. However, other entities such as proteins or toxins are important, too. All of them are chemical substances that can be classified based on their structural characteristics (e.g., small molecular weight chemicals such as the vast majority of drugs, or macromolecules such as proteins), their applications

(e.g., chemicals used as drugs or chemicals used as pesticides), or their effects in the body (e.g., pharmacological substances or toxic substances).

Comprehensive ontologies in the biomedical domain contain, therefore, different types of chemical substances, including drugs. These drugs can be described at different levels of granularity or detail. **Figure 5.1** shows a simplified view of these levels of granularity.

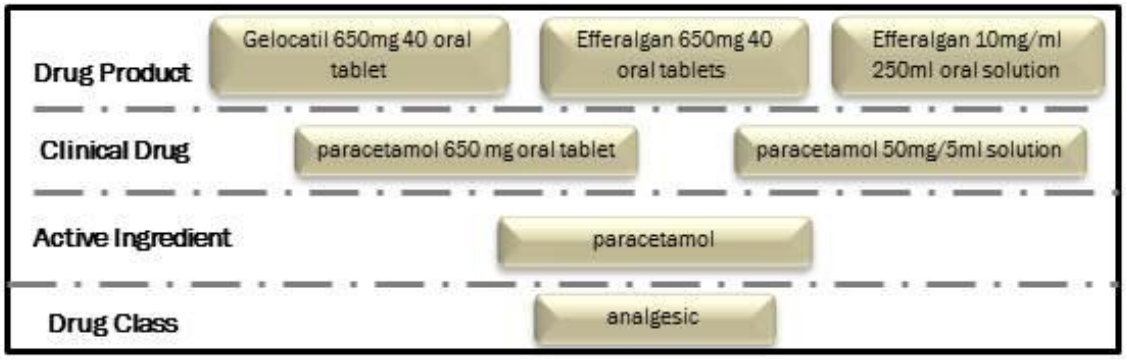

Figure 5.1. Representation of different levels of granularity for drugs

The word "drug" can be used with different meanings in different sources. An "active ingredient" is the specific molecule that bears some pharmacological activity, such as *paracetamol* or *omeprazole*. They are usually classified according to some relevant characteristic, such as chemical structure or pharmacological activity, in different "drug classes" (e.g., *benzodiazepine*, *analgesic*, etc.) Active ingredients are administered as tablets, capsules, solutions, and so forth, with a specific strength or dose. An active ingredient, its strength, and the pharmaceutical form are called "clinical drug". Therefore, '*paracetamol 650 mg oral tablet'*, '*paracetamol 50mg/5ml oral solution',* and '*paracetamol 1000 mg oral capsules'* are three different clinical drugs. Every one of these clinical drugs can be commercialized with different brand names (e.g., *Efferalgan*® or *Gelocatil*®), in packages with a specific number of units. They are called "drug products". Therefore, '*Efferalgan 10mg/ml 250 ml oral solution'*, '*Efferalgan 650mg 40 oral tablets',* and '*Gelocatil 650mg 40 oral tablets'* are three different drug products. Finally, for safety and commercial reasons, every drug product is unambiguously identified in a specific country by a national code. Therefore, in order to compare and relate the information collected among different ontologies, it is important to establish the level of granularity of drugs represented in each one of them.

One of the most important controlled vocabularies in biomedicine is the **Medical Subject Headings (MeSH) thesaurus** (Lipscomb, 2000), developed and maintained by the U.S. National Library of Medicine (NLM). The broad scope of MeSH represents the knowledge from different areas related to biomedicine, including diseases, proteins, or therapeutic techniques, organized by taxonomic relationships. Pharmacological substances are arranged under the top-level category '*Chemicals and Drugs Category*', and are hierarchically represented at different levels of granularity: as active ingredients organized under different drug classes. However, this main category does not provide a

unique subclass for pharmacological substances, but along with other chemicals such as proteins, lipids, or toxins.

Another important top-level category in MeSH is the '*Pharmacological Actions Category*', which represents those pharmacological actions that can be exhibited by drugs or chemicals. Each individual drug is manually linked to one or more pharmacological actions. In this way, MeSH provides an indirect classification of drugs based on their pharmacological activity. Synonyms are provided as '*Entry terms*', along with other related terms (**Figure 5.2**). Each individual chemical entity is unequivocally identify in MeSH through a Registry Number (RN), a reused identifier such as the CAS number[12], the EC code (Tipton & Boyce, 2000) or the FDA Unique Ingredient Identifier (UNII)[13]. MeSH trees can be navigated using the MeSH browsers and the complete thesaurus can be downloaded in different formats, including XML or ASCII.

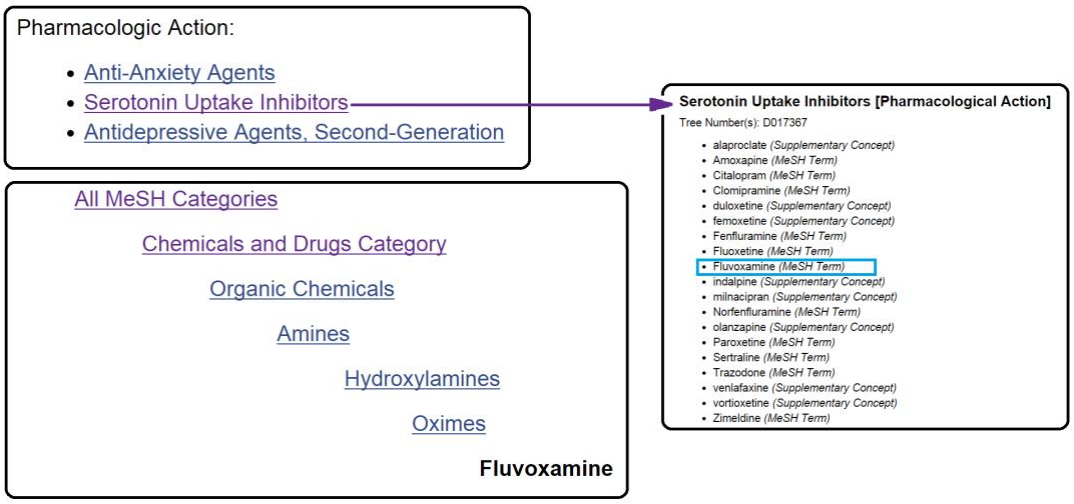

Figure 5.2. Different hierarchies and classifications for the drug *fluvoxamine* in MeSH

Another important terminological resource in medicine is **SNOMED Clinical Terms (SNOMED CT)** (Stearns et al., 2001), a comprehensive clinical terminology that provides clinical content and expressivity for clinical documentation and reporting. Its main purpose is to encode the meanings that are used in health information. The International Health Terminology Standards Development Organisation (IHTSDO) is the owner and is responsible for this terminology. The core of SNOMED CT is the international release, available as English and Spanish versions. Every IHTSDO Member Country creates their own national extension, which is adapted to their special conditions and necessities, including the different drug products commercialized in every country.

Several important areas of the pharmacological domain, including diseases, are organized in SNOMED CT into hierarchies or top-level concepts. Pharmacological substances are described in two of them: '*Substances'* and '*Pharmacological/biological product*'. The main category '*Substances'* organizes active ingredients, such as *diazepam*

[12] http://www.cas.org/content/chemical-substances
[13] http://www.fda.gov/forindustry/datastandards/substanceregistrationsystem-uniqueingredientidentifierunii/default.htm

or *penicillin*, as well as other non-pharmacological chemicals and related terms (e.g., '*silk'*, '*red wine*' or '*oncogene protein P53*'). The top-level '*Pharmacological/biological product*' organizes clinical drugs, distinguishing them from their chemical constituents or active ingredients. Explicit relationships between both hierarchies are established in this ontology. For example, '*furosemide (substance)*', '*furosemide (product)*' and '*furosemide 80mg tablet (product)*' are three different concepts in SNOMED CT. The latter two are hierarchically related through an '*is_a*' relationship. Both of them are related to the concept '*furosemide (substance)*' by a '*has_ingredient*' relation. Commercialized drug products are not included in the international release of SNOMED CT. However, they can be represented in the different national versions and related to the corresponding clinical drugs and active ingredients. Therefore, SNOMED CT includes drugs at the drug class, active ingredients, clinical products, and, for the national extensions, drug product levels. SNOMED CT provides unique identifiers for all concepts. It has been converted to the common UMLS format, its concept names have been connected to those already in the UMLS Metathesaurus, and its content has been assigned UMLS identifiers (CUI), semantic types. Content of SNOMED CT is available under licence conditions, which allow consultation but limits its use for open source final applications.

The **NCI Thesaurus** (NCIT) (Sioutos et al., 2007) is the National Cancer Institute (NCI) reference terminology and biomedical ontology. It covers a broad scope of cancer related terms, including drugs. The NCI drug model organizes active ingredients based on functional, structural, and therapeutic intent hierarchies. Relationships with other concepts, such as mechanism of action, physiologic effects, or molecular targets are formally represented. Clinical drugs and drug products are not included in this terminology, limiting its scope to the active ingredient and drug class levels. However, it provides different synonyms, including brand names, and natural language definitions. Every class has its own NCI URI, and it is mapped to other ontologies, such as the ChEBI ontology. There is information regarding other code systems, including the FDA UNIIs. The NCIT is publicly available in the Web Ontology Language (OWL)[14], a World Wide Web Consortium (W3C) standard language for representing ontological information on the semantic web.

Finally, the European Bioinformatics Institute (EMBL-EBI)[15] created and maintains the ontology for **Chemical Entities of Biological Interest (ChEBI)** (Degtyarenko et al., 2008). This ontology organizes "small" chemical compounds with a relevant role in the biomedical domain, including pharmacological substances. Each one of these pharmacological substances is related to at least one drug application, such as *analgesic* or *antiemetic*. To date, there are 3,540 classes in ChEBI that have some drug-related role, that is, bearing some pharmacological activity[16]. Pharmacological substances are classified, as well, regarding their molecular structure or other characteristics, such as biological or chemical role. Therefore, ChEBI represents drugs at the active ingredient and drug class levels. Every chemical entity has a unique ChEBI identifier and ChEBI URI. Information of other IDs, such as DrugBank, ChEBML, KEGG, or UniProt IDs, is provided, too. The ontology includes synonyms, brand names, and natural language definitions for each class. The ontology can be queried and downloaded in different formats, including OWL.

[14] http://www.w3.org/2001/sw/wiki/OWL
[15] http://www.ebi.ac.uk/
[16] http://www.ebi.ac.uk/chebi/tools/ontoquery/response.jsp?submit=Submit&hiddenQuery=has_role+some+drug+&page=1 [Accessed 14/01/2013]

5.2 Terminological resources for pharmacological substances

In the previous section, we have reviewed those terminological resources for chemical substances that include drugs. Specific terminologies focusing on pharmacological substances are available, too. However, although the number of possible drug names is limited compared to those of other biomedical entities, such as genes or proteins, there is not a single drug vocabulary providing complete and interoperable drug names, codes, and relevant information (Cimino et al., 1998; Solomon et al., 1999). Different countries use different nomenclatures and synonyms for drugs, which are commercialized under different branded or generic names. There are different ontologies created to collect all pharmacological substances and their nomenclature variations at different levels of granularity.

The **Anatomical Therapeutic Chemical (ATC) Classification System** (WHO, n.d.) was developed by the World Health Organization (WHO) to provide a standard and international taxonomy used to classify and compare therapeutic compounds. Drugs are named using their International Non-proprietary Name (INN) and it is one of the most important code systems for the identification of pharmacological substances. Each of them has an ATC code representing its level in the hierarchy. Drugs are hierarchically organized within five different levels of granularity: from the main system where the drug realizes its pharmacological activity to the individual active ingredient. Clinical drugs, brand names, or synonyms are not included. The ATC classification system has been represented in OWL (Croset, Hoehndorf, & Rebholz-Schuhmann, 2012). However, most of the ontological design principles have not been addressed, limiting its reuse in other ontologies.

While the aim of the ATC classification system is to provide an international resource for the identification of drugs, national drug terminologies are important to support the development of national clinical information systems. To this purpose, the United States Veterans Health Administration (VHA) created the **National Drug File (NDF)** (Carter et al., 2002), a nationally maintained medication terminology. The aim of the NDF is to provide a hierarchical classification of pharmacological substances with different levels of granularity. It includes drug classes, active ingredients, clinical drugs, and U.S. National Drug Codes. This terminology evolved into a formalized reference terminology, the **NDF-RT**, created as a Description Logic-based reference model. Unlike NDF, where drugs are classified into a single-inheritance hierarchy of drug classes, NDF-RT provides multiple hierarchies for drug classification through the use of specific relationships such as '*may_treat*' or '*has_mechanism_of_action'*, which relate drugs with their mechanisms of action, physiological effects, clinical kinetics, and therapeutic diseases. Drug classes are maintained, as well, to provide users with a familiar and clinically relevant drug terminology. In this way, different levels of pharmacological substances are related through formal relationships, such as '*has_ingredient'* or '*product_component_of*'. Every concept has a unique identifier NUI and there is a relationship with other codes such as FDA UNIIs. NDF-RT is included in the UMLS Metathesaurus and in the RxNorm vocabulary. There is an OWL version of this ontology.

An important aspect of NDF-RT is that this terminology includes DDIs for specific pairs of drugs. They are subclasses of the main class '*VA Drug Interactions [VA Drug Interaction]*' category, with 7,305 subclasses to date. Every *[VA Drug Interaction]* subclass represents an interaction between two drugs organized in the '*Chemical Ingredients [Chemical/Ingredient]*' top-level category. They are related to the DDI through two attributes: '*Ingredient_1*' and '*Ingredient_2'*. Finally, every DDI has related information regarding its degree of severity: '*Critical'* or '*Significant'*. The information about specific DDIs and their level of severity is developed by a committee at the Department of Veterans Affairs National Drug File Support Group and used in the computerized patient record system (CPRS) throughout the VA health care system to generate alerts when a interacting drug combination is prescribed by a clinician (Ko et al., 2007; Olvey, Clauschee, & Malone, 2010a). However, DDI information has been recently removed from NDF-RT (Peters, Bodenreider, & Bahr, 2014).

Another important U.S. drug vocabulary is **RxNorm** (Nelson et al., 2011), the U.S. NLM standardized drug vocabulary. It was built upon other drug vocabularies, bridging the gap between different drug information sources, such as SNOMED CT, the FDA National Drug Code Dictionary[17], or the previously described NDF-RT. The main characteristic of RxNorm is that this resource allows interoperability between different concepts describing drug products at different levels of granularity. The core concepts in RxNorm are clinical drugs, which are related to active ingredients and drug products through specific relationships. Therefore, RxNorm provides the most comprehensive description of commercialized drug products and their active ingredients and other important information, such us strength or pharmaceutical form. The content of different resources is mapped through the use of the RxNorm Concept Unique Identifier (RxCUI). Other drug codes systems, such as the FDA UNII, are represented as attributes at different levels of abstraction.

These terminologies can be reused in new ontologies created for specific purposes. For example, the **Drug Ontology (DrOn)** (Hogan, Hanna, Joseph, & Brochhausen, 2011) has been developed for the specific purpose of retrieving National Drug Codes (NDC) through different queries, such as active ingredient, mechanism of action, physiological effect, or therapeutic class. Since no existing resource was appropriate for this purpose, authors developed this new ontology, which was populated with RxNorm. It provided the NDCs that were later related to the corresponding branded and generic drug products, clinical drugs, and active ingredients. DrOn was mapped, as well, to the ChEBI ontology, through mapping of their URIs when possible. It is available in OWL format.

5.3 Terminological resources for adverse drug reactions

The concept disease, in its broad sense, can be related to drugs in three main ways. A disease can be treated by a drug; therefore, the disease is the *indication* of the drug. A drug can be contraindicated in patients with a specific disease; therefore, it is a *contraindication*. Finally, a disease can be an undesirable and harmful consequence of the

[17] http://www.accessdata.fda.gov/scripts/cder/ndc/

use of a drug; in this case, it is an *adverse drug reaction* (ADR). In the DDI domain, the observed consequence will be the altered effect of one or both interacting drugs. Therefore, knowledge about associated ADRs for each drug is important in the development of a DDI ontology.

The **Medical Dictionary for Regulatory Activities (MedDRA)** (Brown, Wood, & Wood, 1999) is a standard multilingual medical terminology developed under the auspices of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). It is a fundamental tool in pharmacovigilance and is used in the pre- and postmarketing phases of the medicines regulatory process by the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA). Terms in MedDRA were derived from several sources including the formers WHO's Adverse Reaction Terminology (WHO-ART) (WHO, 1992) or the Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART)[18], among others. MedDRA provides a classification of ADRs at different levels of granularity and with a deep degree of detail. The lowest or most specific one collects more than 70,000 terms, which correspond to how information about ADRs should be communicated. Therefore, it is useful for recording adverse events and medical history in electronic health records. The English version of MedDRA is distributed as sets of extended ASCII delimited files.

The **Adverse Event Reporting Ontology (AERO)** (Courtot, Brinkman, & Ruttenberg, 2011) was developed to support clinicians at the time of data entry, increasing quality and accuracy of reported adverse events. It was created to provide an adverse event terminology that address some of the current drawbacks observed with broad ontologies such as MedDRA. One of the most important is the lack of definitions in natural language, which can lead to heterogeneity in the coding of these events. With 398 classes and 62 object properties, the scope of this ontology is much lower than that of MedDRA. AERO is available in OWL and OBO[19] formats.

The **Ontology of Adverse Events (OAE)** (He, Xiang, Sarntivijai, Toldo, & Ceusters, 2011) is a standardization and integration for data on biomedical adverse events. This ontology provides a broad scope of adverse events, with natural language definitions and mapping to other resources such as SIDER (Kuhn et al., 2010), a database for drugs and their side effects, or the aforementioned MedDRA. The ontology is available in OWL format and it has been developed following the OBO Foundry design principles (Smith et al., 2007), a collaborative effort for the development and maintenance of biomedical ontologies to ensure their integration and interoperability.

These ontologies have been created to provide formal representations of ADR or AE related concepts. However, pharmacological substances are not represented, leading to a lack of a formal ontology that relates drugs with theirs ADRs. To address this problem, the **Adverse Reactions and Mechanism Ontology (ARM)** (Zhichkin, Athey, Avigan, & Abernethy, 2012) has been proposed. This is an ongoing project for a systematic organization of all ADR-related data and information. The aim of the authors is to create a knowledge base that will enable linking ADR with biological mechanisms and functions. However, for the time being, the ontology has not been published or described in any source.

[18] http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CST/
[19] Ontologies of Biological and Biomedical Interest format

5.4 Ontologies related to the DDI domain

Up to now, we have described terminological resources for chemical entities – focusing on pharmacological substances –, drug terminologies, and ontologies for ADRs. However, the pharmacological domain involves other relevant areas, such as pharmacodynamics, pharmacokinetics or pharmacogenomics. All of them are related to DDIs, and have been addressed in different ontologies. We have identified those ontologies that could be useful for our purpose of development of a comprehensive ontology for DDIs. Here, we describe them in a chronological order.

The **Drug Interaction Ontology (DIO)** (Yoshikawa, Satou, & Konagaya, 2004) is a formal representation of drug pharmacological actions, depicted by drug-biomolecule interactions that are the underlying mechanism in some types of DDIs. The conceptual model of this ontology enables the representation of complex pharmacological processes. However, relationships are established only at a high level of reality since individual drugs and biomolecules are not comprehensively represented. The number of classes is 179 and the number of relationships is 14. Natural language definitions are provided only for the half of the classes. Anatomical-related classes are linked to the Foundational Model of Anatomy (FMA) (Rosse & Mejino, 2003) and UMLS CUIs are provided for a total of 34 classes. The ontology is available in OWL format.

The **drug-mechanism evidence taxonomy** was created in the framework of the **Drug Interaction Knowledge Base** (**DIKB**) (Boyce, Collins, Horn, & Kalet, 2009) project with the aim to provide a way to assign the level of confidence of experts with a particular drug-related fact. This evidence taxonomy covers all the kinds of evidence that, when combined with a set of inclusion criteria, enables drug experts to specify their confidence in a specific drug mechanism assertion. Therefore, the taxonomy does not represent specific drugs or interactions between individual pairs of them. It is implemented in OWL, and it consists of 125 classes and 70 object properties. Ninety-five per cent of the classes are defined in natural language, and one of each four is formally defined in the ontology by an OWL definitional axiom (or '*equivalent to*' relationships). Links or mappings to other resources are not provided.

Rubrichi & Quaglini created an ontology by acquiring the domain knowledge from the drug interactions and contraindications sections from Italian Summaries of Product Characteristics (SPCs) (Rubrichi & Quaglini, 2012). The aim of this project was the development of a system able to extract automatically drug-related information from texts. Therefore, the ontology should represent all DDI-related information that could be found in that type of texts. However, the represented concepts are not described at a deep level of detail and specific drugs or DDIs are not included. The final ontology, which can be obtained by requesting the authors, has been implemented in OWL and has 23 classes, 17 object properties or relationships, and 7 individuals. Natural language definition or OWL definitional axioms are not provided.

The **Pharmacokinetics ontology (PKO)** (Wu et al., 2013) is a recent work developed for the representation of drug pharmacokinetics information. This ontology does not represent specific DDIs or DDI mechanisms, but collects concepts related to the DDI domain. Specifically, authors focused on the representation of PK DDI studies and their components. PK DDI studies are experiments developed *in vitro* or *in vivo* to study

the existence of drug interactions affecting some of the PK parameters of the interacting drugs. The main contribution of this ontology is a manually created representation of PK parameters and their definitions in natural language. The utility of this ontology has been evaluated in the annotation of texts describing a PK study of DDIs (Wu et al., 2013). Pharmacological substances are imported from the ChEBI ontology, although related information such as synonyms or cross-references has not been included. The PKO is available in OWL format.

In contrast to the PKO, the **Pharmacodynamics ontology (PDO)** (Imai, Hayakawa, & Ohe, 2013) focuses on the description of the pharmacodynamics domain. Researchers at the University of Tokyo have developed a description framework of PDO with the aim to support machine-reasoning systems for the detection of possible DDIs based on PD mechanisms. However, as the PKO, it does not represent specific DDIs or DDI mechanisms. This ontology is only part of a national project for a Medical ontology in Japan, and currently it is available in Japanese only. The ontology has not been implemented in OWL format.

In **Table 5.1**, we provide a summary of the figures comparing the OWL versions of these five ontologies. We include here the NDF-RT, too, since, as mentioned before, it represents specific DDIs between drugs.

		NDF-RT	DIO	DIKB	Rubrichi	PKO	PDO
Classes	Own	-	169	91	23	1,381	-
	Imp	-	10	34	0	44,955	-
	Tot	48,978	179	127	0	46,336	-
Relationships	Own	0	9	37	17	12	-
	Imp	0	5	30	0	216	-
	Tot	0	14	70	0	228	-
Classes with NLD		0	89 (49%)	119 (95%)	0	434 (31.4%)	-
Related IDs		CUI NUI UNII	FMAID CUI	0	0	0	-
OWL definitional axioms		0	0	32 (25,6%)	0	2 (0.15%)	-

Table 5.1. Metrics[20] and comparison of DDI-related ontologies in OWL format (Own for entities created in the ontology; Imp for imported entities; Tot for the total number of entities; NLD for natural language definitions.)

[20] Metrics from http://bioportal.bioontology.org/ontologies/, http://www.ontobee.org/ (accessed 24/01/2014) or the manually study of each terminology when appropriate.

5.5 Discussion

In this chapter, we have reviewed existing semantic resources from a pharmacovigilance perspective, specially focusing on their scope to represent DDI-related information. **NDF-RT** represents specific DDIs between specific pairs of drugs with one of two degrees of severity: critical or significant. However, other related information such as the mechanism leading to the DDI, the effect or consequence for the patient, recommendations for avoiding the DDI, patient-related (age, diseases, genetics, etc.) or drug-related (dose, administration route, etc.) affecting factors, are not represented. The ontology created by **Rubrichi** aims to include all these aspects. However, the level of detail in the description of concepts is too shallow to provide an appropriate description of the DDI domain. **DIO** provides a formal representation of interactions between drugs and biomolecules, which has been used to infer DDIs between one particular pair of interacting drugs. This ontology provides a framework for the formal representation of DDI mechanisms. However, the scope of the ontology, with 180 classes, 14 object properties, and 1194 axioms, is too small to represent all the different DDI-mechanisms. In a similar way, the description framework of the **PDO** provides a first step in the formal representation of PD processes. However, there is still a lack of explicit representation of PD DDI mechanisms. Finally, the **PKO** and the **evidence taxonomy** (or **DIKB**) focus on limited areas related to the DDI domain, but do not provide a comprehensive representation of the domain.

With the exception of the **PDO**, the other four ontologies have been implemented in OWL, the World Wide Web Consortium (W3C) standard ontology language for the semantic web[21]. Using standard formats allows interoperability among ontologies and increases its usefulness. This recommendation is one of the OBO Foundry principles, a collaborative effort for the development and maintenance of biomedical ontologies (Smith et al., 2007). These recommendations for building ontologies cover aspects such as format, content, naming conventions, collaboration between similar projects, among many others. However, the ontologies related to the DDI domain reviewed here have not followed clear ontological design principles, such as those proposed by the OBO Foundry, limiting their reuse in other ontologies. Therefore, from this study we conclude that the knowledge about DDIs has not been comprehensively represented in any existing ontology. **Table 5.2** summarizes the strengths and limitations of these resources.

Regarding other pharmacological terminologies, most of them could provide lists of drugs (**MeSH**, **ChEBI**, **ATC**, **NDF**, **RxNorm,** or **DrOn**) to be imported in our ontology. **MedDRA**, **AERO,** or **OAE** could be imported as a source of ADR terminology. However, relationships between drugs and ADRs are not established, limiting the representation of DDI-related information. The effect of a DDI is an altered effect of one or both interacting drugs. Therefore, drug-ADR relationships are necessary for the formal representation of DDI effects. In contrast, the drug model in the **NCIT** links drugs to other concepts, such as mechanism of actions or proteins. Nevertheless, only cancer-related drugs are included, limiting its use for DDIs representation between other classes of drugs. **SNOMED CT** could be a useful resource for our purposes, since representation of pharmacological drugs is comprehensive and there are relationships between them and

[21] http://www.w3.org/2001/sw/wiki/OWL

other concepts, such as drug-disease relationships. However, licence conditions limit its reuse for open source final applications.

Semantic resource	Strengths	Limitations
NDF-RT	◆ Representation of specific DDIs between drugs. ◆ Inclusion of degree of severity.	◆ No other DDI-related information included. ◆ DDIs have been removed from the NDF-RT.
DIO	◆ Represents drug-biomolecule interactions. ◆ Useful to infer DDIs on the basis of a PK mechanism.	◆ No specific or general DDIs represented. ◆ Low scope and coverage. ◆ Not ontological design principles followed.
Evidence taxonomy (DIKB)	◆ Represents the evidence for specifying the confidence in a DDI assertion.	◆ Focuses on a limited area of DDI knowledge (evidence). ◆ Not ontological design principles followed.
Rubrichi	◆ Represents DDI-related knowledge (mechanism, effect, affecting factors…).	◆ No specific DDIs represented. ◆ Shallow representation of DDI-related concepts. ◆ Not ontological design principles followed.
PKO	◆ Focuses on PK DDI studies and their components.	◆ Focuses on a limited area of DDI knowledge (PK DDI studies). ◆ Not ontological design principles followed.
PDO	◆ Focuses on PD processes. ◆ Useful to manually infer DDIs on the basis of a PD mechanism.	◆ Lack of explicit representation of PD DDI mechanisms. ◆ Not ontological design principles followed.

Table 5.2. Summary of strengths and limitations of current DDI-related ontologies

Therefore, no existing resource provides an appropriate representation of all DDI-related information. For this reason, a new ontology is developed in the framework of this thesis. With the ontological representation of the DDI knowledge in this new resource, we provide a formalized model for the general pharmacological principles common to all DDIs. These include the formal representation of DDI mechanisms, which are comprehensively represented in the ontology, including all possible PK and PD mechanisms. The conceptualization of these mechanisms requires the representation of different drug-protein relationships, which are included in our ontology. Alongside the general representation of DDI knowledge, the ontology is populated with specific DDIs between specific pairs of drugs. All this information is integrated in a conceptual model

implemented in Description Logics (DL), enabling the use of a reasoning engine to check the consistency of the ontology and the inference of new classification hierarchies and relationships. Finally, the interoperability of this new ontology is ensured by the adoption of ontological design principles proposed by the OBO Foundry. In this framework, the ontology is developed as an open source artefact, which is available for the research community.

5.6 Unresolved issues

Automatic systems for extraction of DDIs can be built upon ontologies providing the knowledge of the domain, as a source of the relevant entities and the relationships between them. Moreover, combination of semantic annotations among concepts and DL can provide a framework for the inference of DDIs based on their mechanisms and the intrinsic characteristics of interacting drugs, enabling the development of new systems for the early prediction of unknown DDIs in pharmacovigilance. To achieve these goals, however, some limitations should be addressed:

1. There is not a comprehensive ontology for DDI-related information.

2. DDIs between individual drugs have been represented in the NDF-RT, but without providing any additional information such as mechanism, effect, or recommendations for avoiding the DDI.

3. Most DDI mechanisms occur because two different drugs bind the same protein in the body. These mechanisms can be formally represented in an ontology through the representation of drug-protein relationships, enabling the inference of DDIs. However, the different mechanism types leading to DDIs have not been represented comprehensively in any ontology.

4. Ontological representation of DDI mechanisms has been tackled only from a PK point of view, but not PD. Moreover, these works have not studied all different types of PK mechanisms that can be involved in DDIs and which are important for their appropriate understanding and manage.

5. The consequence of a DDI is frequently related to the exacerbation of the ADR of one or two interacting drugs. For example, some *antihistaminic drugs* used to reduce allergic symptoms cause somnolence. With the concomitant use of a tranquilizer, such as *diazepam*, the degree of somnolence in the patient is higher. Therefore, the formal representation of the relationships between drugs and their related ADRs is important to characterize the consequences of DDIs. However, there is not available ontology for drug-ADR representation.

6. The study of those characteristics influencing DDIs, such as chemical structure, ADRs profile, or mechanism of action, could provide interesting approaches for the inference of DDIs or the selection of non-interacting

alternatives. However, there is not existing ontologies representing comprehensively all this information and its relationship with DDIs.

7. Current DDI-related ontologies have not followed ontological design principles. This fact reduces their integration and interoperability with other ontologies.

Thus, one of the main contributions of this thesis is to create a DDI ontology that covers each of the unsolved issues described in the above list.

5.7 Conclusions

This chapter has reviewed the current semantic resources in the pharmacological domain covering some aspect related to DDIs. Firstly, we have analysed the terminologies for chemical entities that include pharmacological substances. The **MeSH thesaurus**, **SNOMED CT**, the **NCIT**, and the **ChEBI** ontology are the outstanding resources in this group. Secondly, we have reviewed the terminologies focusing on pharmacological substances, such as the **ATC classification system**, the **NDF-RT,** or **RxNorm**. In addition to this, we have reviewed terminologies and ontologies for ADRs, which include **MedDRA**, **AERO**, and **OAE**. Finally, we have paid special attention to those ontologies specifically created to represent DDIs or their mechanisms: **DIO**, the drug-mechanism **evidence taxonomy** or DIKB, **Rubrichi** et al. SPCs ontology, the **PKO** and the **PDO**.

The existence of these resources proves the interest of the research community in the formal representation of DDI knowledge. However, the analysis described in this chapter has shown that there are still limitations that should be addressed in a comprehensive ontology for DDI knowledge. The development of this new resource requires a detailed analysis of how DDI-related knowledge has been represented by other researches and the different concepts and relationships covered in current ontologies, in order to reuse useful information and to identify deficiencies. Thus, in next **Chapter 6**, we review in detail current modeling efforts in the DDI domain.

Chapter 6

DDI-knowledge modeling: State of the art

In the previous chapter, we have identified and reviewed the semantic resources in the pharmacological domain covering some DDI-related aspect, focusing on the analysis of resources for DDIs or their mechanisms. In this chapter, we complete this review by analysing the main efforts of several research groups to model this DDI-knowledge. In other words, we focus here not in the analysis of the final resources, but in the modeling approaches followed to create them.

To perform useful actions for people working in a given domain, a system requires knowing something about that domain. Concrete knowledge about a domain (e.g., *Maria has one sister and one brother, Ana and Diego*) requires prior general knowledge or how concrete objects are related in the world (e.g., *A person can have none or several siblings, that can be sisters or brothers. A sister is always a woman and a brother is always a man*). Conceptual modeling is the activity that elicits and describes the general knowledge of a particular domain. The set of concepts used in a particular domain constitutes a conceptualization of that domain, and its graphical description is the conceptual model (CM) (Olivié, 2007). Usually, the design of a CM relies on the perspective that experts have of a specific domain. Since it is the result of an intellectual activity performed by humans that serves different objectives, there is not a unique representation. Consequently, different CMs representing the same domain can exist. These CMs are abstract models that can be translated into different implementations and interpretable schemata such as OWL ontologies, relational databases, XML schema, and so forth.

Therefore, as any other formal representation, the creation of a new ontology for DDIs requires a previous conceptualization of the domain knowledge describing how relevant concepts relate to each other in the world. Before performing this task, we have identified and analysed current research projects that have focused or have required the conceptualization of the DDI domain. With this review, we aim to identify which aspects of DDIs have been currently conceptualised, how this information has been modelled by different research groups, how the different CMs have been translated, and the applications given to the final models.

This chapter is organized as follows. **Section 6.1** describes the method used to create a common framework for the comparison of different conceptualizations of the DDI domain. These different projects are summarized in **Section 6.2** and compared in **Section 6.3**. Results of our analysis are discussed in **Section 6.4**, and the unresolved issues that will be addressed in the framework of this thesis are presented in **Section 6.5**. Finally, the main conclusions are provided in **Section 6.6**.

6.1 Creation of a common framework

To the best of our knowledge, there are seven different projects that have required a total or partial representation of the DDI domain knowledge, which have been described in scientific journals or conference proceedings between 2004 to date (July 2014). These projects have in common that, for different purposes, they developed and described their conceptualizations of the DDI domain, which were subsequently implemented in different models. However, every work describes its conceptualization in different ways. To compare the different CMs created by each research group, we study: 1) the original CM created by the authors; 2) their natural language description of the domain; 3) the final implemented model.

However, a comparative analysis of the seven conceptualizations is difficult. In most cases, CMs are not provided and those explicitly included in the publications differ considerably. In addition to this, natural language descriptions can lead to a subjective interpretation of the described domain, which can be different from that intended by the authors. Finally, the final implemented models are complex artifacts (sets of rules in first order logic (FOL) or OWL ontologies) and therefore, a straight comparison between them is difficult.

Consequently, to compare the similarities and differences between the seven models, we have represented all of them as Unified Modeling Language (UML) class diagrams, a standard modeling language that can be applied to diverse independent domains.[22] In brief, classes are represented as boxes with three parts. The top part contains the name of the class, the middle part contains the attributes of the class, and the bottom part contains the methods the class can execute. The taxonomical relation '*is_a'* is represented as a line ending in a hollow triangle shape on the superclass. Other relationships between classes are represented as arrows, while aggregation or composition relationships are represented by a shallow or filled diamond shape, respectively.

[22] http://www.uml-diagrams.org/

For those projects that already provided a diagram showing their CMs, we only adapt their models to the common one. In cases when we study the final ontologies, they are analysed and transformed into the corresponding CM. We do not represent any attribute that is not explicitly mentioned in the original project. Moreover, to avoid subjective interpretation of ontologies, we include in the model only the relationships that explicitly establish their range and domain, and we rule out those where this information is missed. In the case of conceptualizations implemented as set of rules in FOL, we consider variables as classes and predicates as relationships. Those predicates defining objects are represented as attributes or classes as appropriate. Finally, due to the lack of space, we only include in the diagrams the top-level classes and their most relevant subclasses. Those classes having subclasses not represented in the CM are shown as shaded boxes.

With this approach, we are able to provide a common representation framework, which enables the correct comparison of different conceptualizations. The resulting CMs for each conceptualization are shown in the next section with a brief description of the projects where they were created.

6.2 Modeling approaches in the DDI domain

In this section, we describe the seven different approaches that have dealt, in some way, with the task of conceptualization of DDI knowledge. We refer along this chapter to each project by the name provided to the final implemented model. When there is not a name assigned, we refer to the project by the first author's name in the corresponding publication.

- **Modeling DDI knowledge acquired from DDI monographs (Mille et al.):**

One of the earliest efforts in modeling the DDI domain was carried out in 2007 by Mille et al. (Mille, Degoulet, & Jaulent, 2007). They created a simple CM aimed to represent the whole DDI domain consisting in only six main classes (**Figure 6.1**). Knowledge represented in this CM was acquired from the study of natural language descriptions of DDIs obtained from French drug-drug interaction monographs (Agence française de sécurité sanitaire des produits de santé, 2006) with the final aim to create a DDI structured knowledge base that could be used by CDSS. The information needed to populate the knowledge base was acquired from descriptions of DDIs obtained from the previously mentioned monographs. To do this, the CM was used to create an XML schema of DDI knowledge for the encoding or markup of textual documents.

This CM represents the DDI as the central concept, which is related to the other ones. The model covers most of the important questions related to the DDI domain: *how does the interaction occur?* ('Mechanism'); *which ones are the interacting drugs?* ('Partner'); *what is the consequence of the DDI?* ('Consequence'); *which factors can increase the risk of the DDI?* ('Risk Factor'; 'Risk Association'); *which factors or measures can decrease the risk of the DDI?* ('Precaution of Use'; 'Limitation'). Although the model provides a wide coverage of the DDI domain, a deeper description of each one of these aspects is not represented. Even so, this model proved to be useful to encode a total of

one thousand and six DDI monographs and to create a knowledge base with the extracted information.

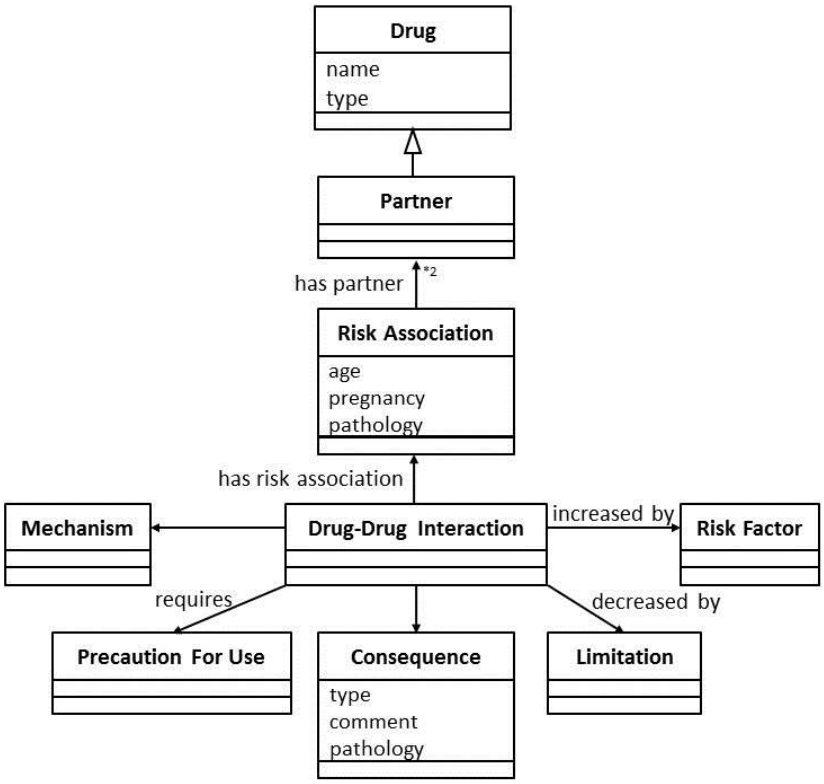

Figure 6.1. Conceptual model in Mille et al.

- **An ontology for SPCs representation (Rubrichi & Quaglini):**

Rubrichi & Quaglini required the conceptualization of the DDI domain for the creation of an ontology that would be used to create a system able to automatically extract drug-related information from text (Rubrichi, Quaglini, Spengler, Russo, & Gallinari, 2013).

This CM explicitly represents the concept of DDI as a class (**Figure 6.2**). As well as in Mille's CM, this representation includes some of the most important aspects of DDIs: *what is the consequence of the DDI?* ('Interaction Effect'); *which factors can increase the risk of the DDI?* ('Intake Route'; 'Posology'; 'Personal Conditions') and *which actions or measures can decrease the risk of the DDI?* ('Recovering Action'). An important characteristic is that this is the only work that includes the concept 'Side Effect' in the conceptualization. However, an important aspect is not represented: the 'Mechanism', or how the interaction occurs, which is a crucial aspect in the study, understanding, and management of DDIs (see **Section 7.1.3**).

As mentioned before, the CM was used to create an ontology, which was used to annotate a set of SPCs texts for the training and testing of an IE system (Rubrichi & Quaglini, 2012). The extracted information was recently used by the authors in a new

experiment aimed to populate automatically the ontology through the representation of the extracted entities as instances of concepts. The final number of entities in the populated ontology is, however, not provided in the publication (Rubrichi et al., 2013).

- **The National Drug File Reference Terminology (NDF-RT):**

The National Drug File Reference Terminology (NDF-RT) is the formalized reference terminology of the U.S. Veterans Health Administration (VHA) nationally maintained medication terminology, the National Drug File (NDF) (Carter et al., 2002). NDF-RT is used for modeling drug characteristics including ingredients, chemical structure, dose form, physiological effect, mechanism of action, pharmacokinetics, and related diseases. It also includes DDIs between pairs of drugs (U. S. Department of Veterans Affairs, 2012).

NDF-RT is used by the VHA computerized systems. Information about specific DDIs is established by an expert committee at the Department of Veterans Affairs National Drug File Support Group and is used in the computerized patient record system (CPRS) throughout the VA health care system to generate alerts when a interacting drug combination is prescribed by a clinician (Ko et al., 2007; Olvey et al., 2010a).

In spite of the detailed description of drug characteristics shown in **Figure 6.3**, DDI-related information is represented in a very simple way in the model. In NDF-RT, a DDI is related to exactly two active ingredients and has an attribute 'Severity' that takes one of two values: '*Significant*' or '*Critical*'. No other information related to the domain is included in this model. However, it is important to note that all DDI-related information has been recently removed from NDF-RT (Peters et al., 2014).

- **The Drug Interaction Ontology (DIO):**

The Drug Interaction Ontology (DIO) is an OWL-DL ontology developed for the formal representation of pharmacological actions depicted by drug-biomolecule interactions. The interference of different drugs in the same drug-biomolecule interaction is the underlying mechanism of most DDIs. If the interaction between a drug and the biomolecule is related to the pharmacodynamics of the drug, the resulting DDI is a PD DDI. In contrast, if the drug-biomolecule interaction regulates some of the pharmacokinetic processes of the drug in the body, it will lead to a PK DDI (see **Section 7.1.3** for a comprehensive description of these concepts). DIO focuses only on the latter one and therefore PD knowledge is not included. It is important to emphasize that DIO represents interactions between a drug and a biomolecule, and not interactions between two drugs. However, as we explain below, drug-biomolecule interaction information can be exploited by a system to predict DDIs.

Yoshikawa et al. (2004) described their CM as a triadic relationship between three concepts: 1) the effector that triggers the interaction; 2) the counterpart object; and 3) the output or consequence of that interaction. Since the model represents drug-biomolecule interactions and not drug-drug interactions, the concept of DDI was not represented. However, a relevant aspect in this model was the location of the drug-biomolecule interaction, or where in the body the interaction occurs.

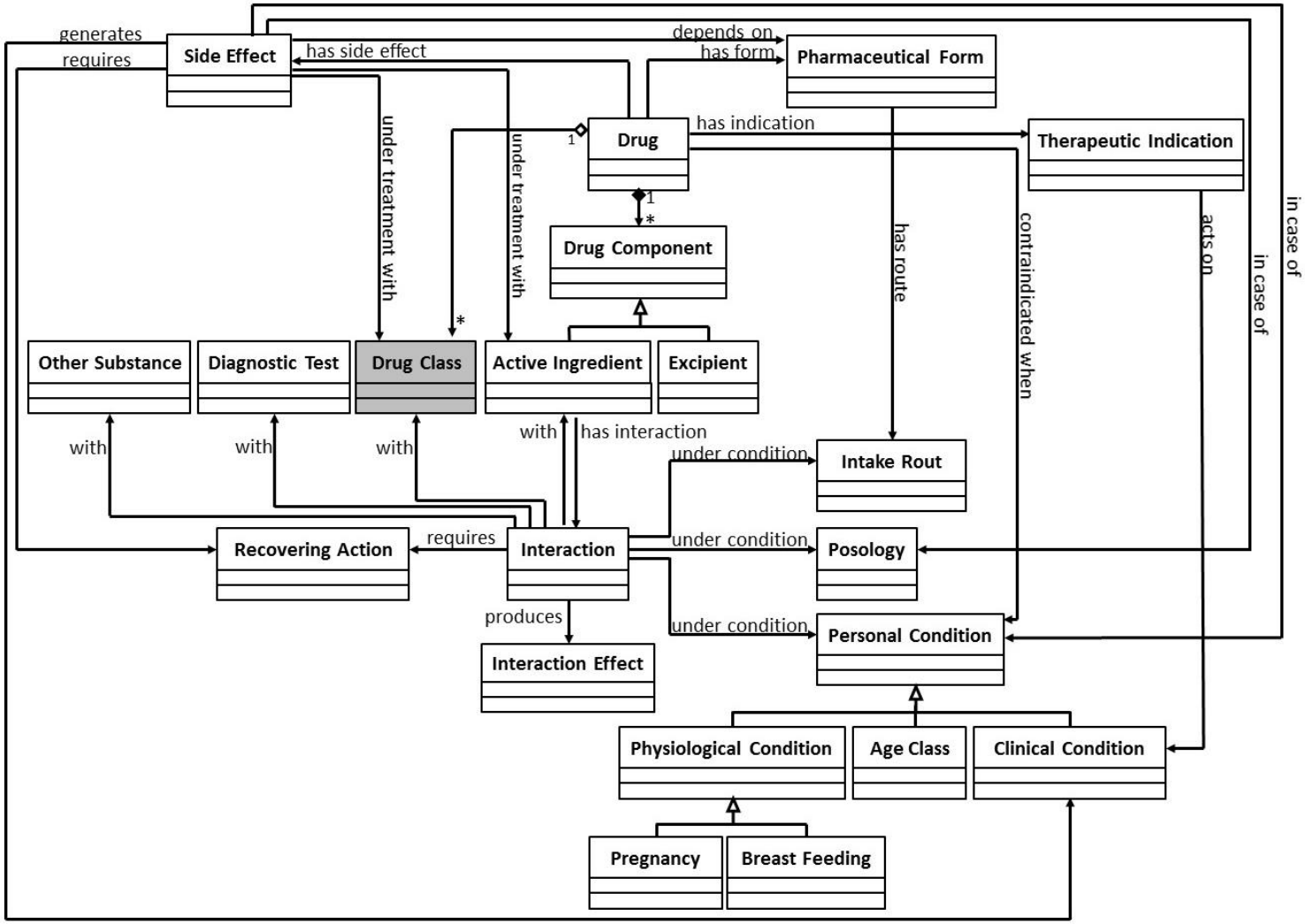

Figure 6.2. Conceptual model in Rubrichi et al. Shaded boxes represent classes with additional subclasses (Adapted from Rubrichi et al., 2012)

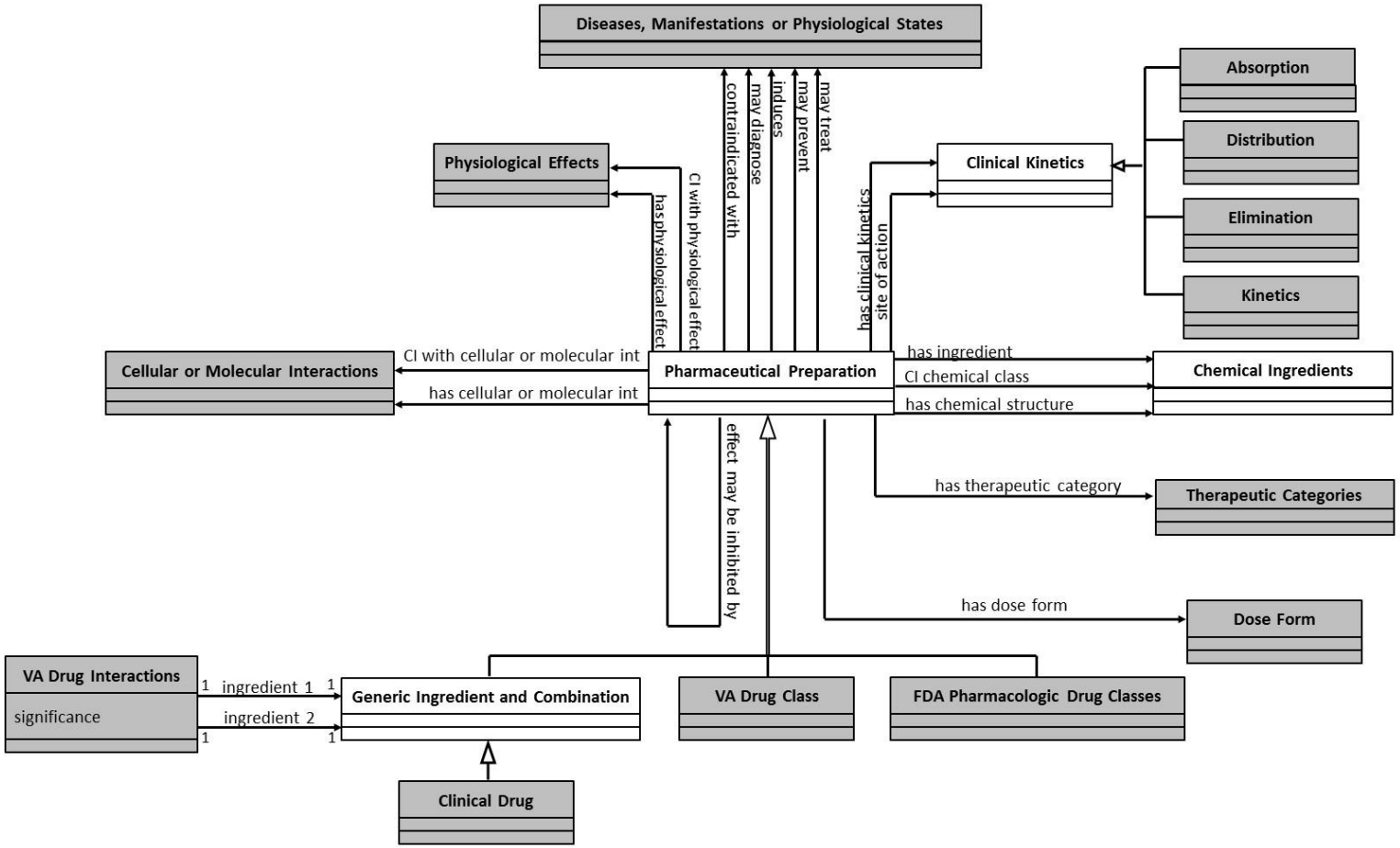

Figure 6.3. Conceptual model in NDF-RT. Shaded boxes represent classes with additional subclasses

The aim of this project was the development of a data model that could represent drug-biomolecule interactions for knowledge sharing and functional usage. DIO's developers applied their ontology for prediction of PK DDIs and created a system able to identify correctly several possible DDI mechanisms between a specific pair of drugs (*irinotecan* and *ketoconazole*). However, interactions between other drugs have not been studied in the framework of this project (Konagaya, 2012).

The adaptation of the final implemented DIO to the common CM format used in this review (**Figure 6.4** and **Figure 6.5**) leads to a more complex CM than that shown in their original work (Yoshikawa et al., 2004). In this extended CM, we can see that anatomical concepts and cellular components are included in the model to represent the anatomical and cellular locations of interactions. 'Drug', 'Metabolites', and 'Proteins' are hierarchically represented as chemicals. The intermediate products formed in a drug-biomolecule interaction, called 'Drug-Protein Complex', are represented, too. These chemicals participate in different processes that are classified at the organismal level (including the PK processes 'Drug Absorption', 'Drug Distribution' and 'Drug Excretion'), the cellular level (such us the PK process 'Drug Metabolism'), or the molecular level (e.g., different types of enzymatic reactions). Therefore, the conceptualization implemented as DIO provides a detailed description of PK processes. Moreover, this CM is unique among those studied in this review in terms of including the localization of these reactions in the organism.

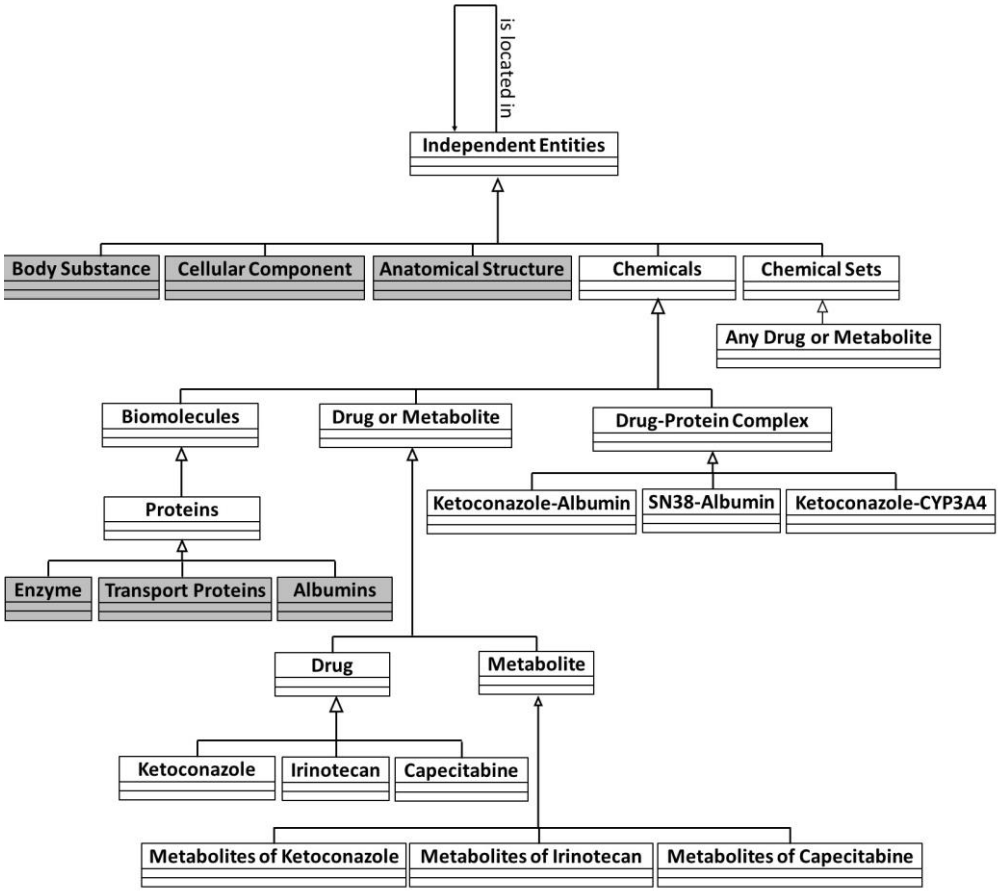

Figure 6.4. Conceptual model in DIO for the 'Independent Entities' top-level class. Shaded boxes represent classes with additional subclasses.

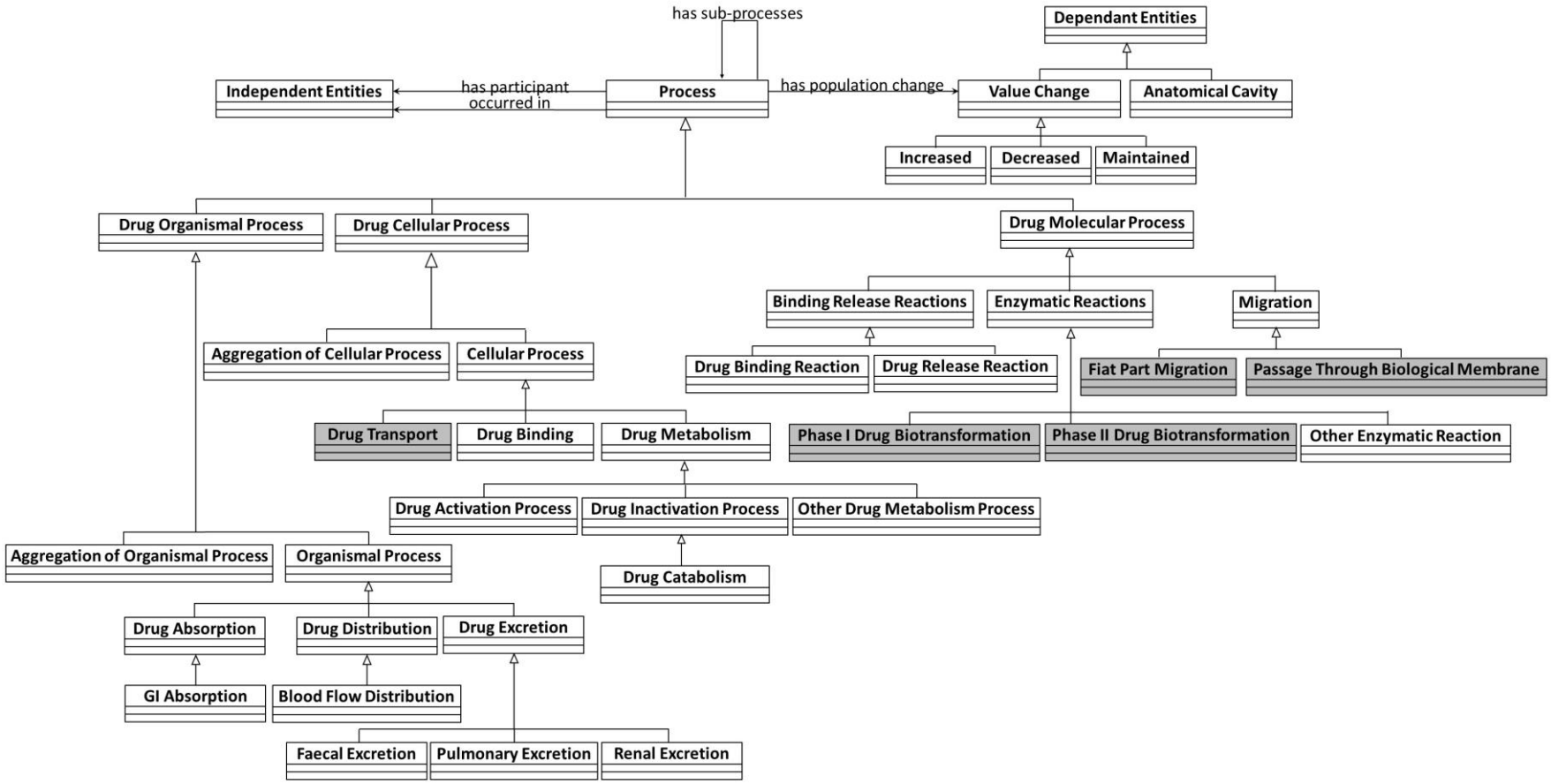

Figure 6.5. Conceptual model for DIO showing the 'Process' top-level class and its relationships to the 'Independent Entities' and 'Dependant Entities' classes. Shaded boxes represent classes with additional subclasses.

- **The Drug Interaction Knowledge Base (DIKB):**

Boyce et al. have worked since 2004 in the domain of DDI knowledge representation (Boyce, Collins, Horn, & Kalet, 2004). One of their main contributions to this field is the Drug Interaction Knowledge Base (DIKB), a knowledge representation system designed to predict DDIs on the basis of their underlying mechanisms and the evidence supporting the drug-related facts (Boyce, Collins, Horn, & Kalet, 2010a). These predictions are possible through the formal representation of the mechanisms that lead to DDIs, which are modelled as a set of rules in FOL (Boyce, Collins, Horn, & Kalet, 2007).

In their earliest efforts, Boyce et al. studied the formal representation of different types of PK mechanisms: induction and inhibition of metabolizing enzymes, change in gastro-intestinal pH, and change in gastro-intestinal motility (Boyce et al., 2004). However, in further work they focused on the extension of rules that modeled the conditions for the specific process of metabolic inhibition, resulting in a closer description of this specific DDI mechanism (Boyce et al., 2007).

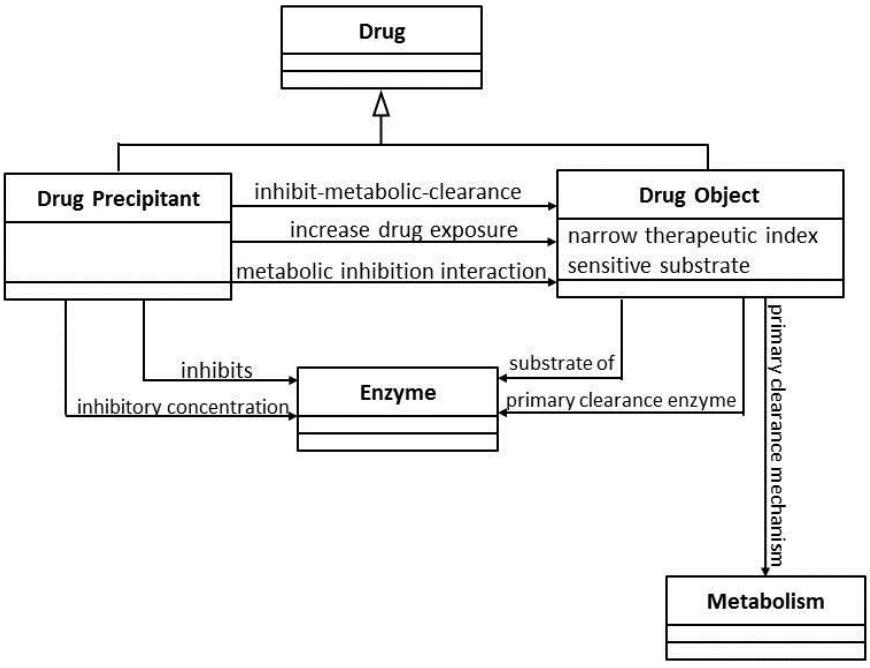

Figure 6.6. Conceptual Model in the DIKB

This CM (**Figure 6.6**) focuses on the relationships that exist between the principal actors that take part in a DDI occurring by the inhibition of the enzymatic metabolism of a drug: the precipitant drug, the object drug, and the metabolic enzyme. In this model, the PK process 'Metabolism' is represented as a concept and related to the object drug. For a deeper description of DDIs, two drug characteristics that determine the incidence and significance of DDIs are included, too. They are the characterization of the object as a drug with 'Narrow therapeutic index' and/or as a 'Sensitive substrate'. With the inclusion of these characteristics, the model focuses on DDIs that potentially could be more relevant in the clinical domain.

These classes, relationships, and attributes are combined in a rule-based theory of how drugs interact based on this mechanism. Using information of a manually curated database of drug-related facts – structured information of specific drugs, metabolites, and metabolic enzymes and relationships between them – and the rule-based theory, they developed a machine-reasoning system able to predict interactions between individual pairs of drugs (Boyce, Collins, Horn, & Kalet, 2010b). The final DIKB contains quantitative and qualitative assertions about drug mechanisms and PK DDIs for over 60 drugs, primarily *psychotropics* and *HMG-CoA reductase inhibitors*[23].

In the framework of the same project, uncertainty behind drug information was identified as one of the main challenges to represent and use drug-mechanism knowledge. To overcome this issue, authors created an evidence-base for the assignment of drug experts' confidence level for every drug-mechanism fact. For this purpose, they combined a set of inclusion criteria with a new evidence taxonomy containing 36 evidence types (Boyce et al., 2010a). This ontology does not model the DDI domain. However, it could be reused to categorize the level of evidence of DDI-related information.

- **The Pharmacodynamics ontology (PDO)**

Imai et al. (2013), from the University of Tokyo, have focused on the field of PD DDIs. With the aim to progress in the development of machine reasoning systems to detect DDIs occurring by this type of mechanism, they created the description framework of the PDO, which has been created as part of a national project for the creation of a Medical Ontology in Japan[24]. As in DIO, specific information regarding DDIs is not included in the model. However, their descriptions of pharmacological processes can be used to predict interactions between specific pairs of drugs.

Five fundamental classes and several relationships among them model the domain of drug pharmacodynamics – or how a drug produces a pharmacological effect by interacting with a target in the body (**Figure 6.7**). Therefore, in this CM processes are predominantly represented. Specifically, different signal transduction sub-processes that can occur when a drug interacts with a target are represented. In these cellular processes a signal is conveyed to trigger a change in a cell, leading to a chain of physiological responses that will finally produce the pharmacological response. As mentioned before, the concept of DDI is not represented in the model. However, the domain was modeled in order to allow inferring a DDI when two different single drug molecules have in common some of the represented signal transduction processes, or when they have the same physiological response. The suitability of the model to predict PD DDIs has been tested manually on a limited number of drugs related to the *noradrenaline*-signal transduction process.

[23] http://dbmi-icode-01.dbmi.pitt.edu/dikb-evidence/front-page.html
[24] http://www.m.u-tokyo.ac.jp/medinfo/medont2010-2012proj/

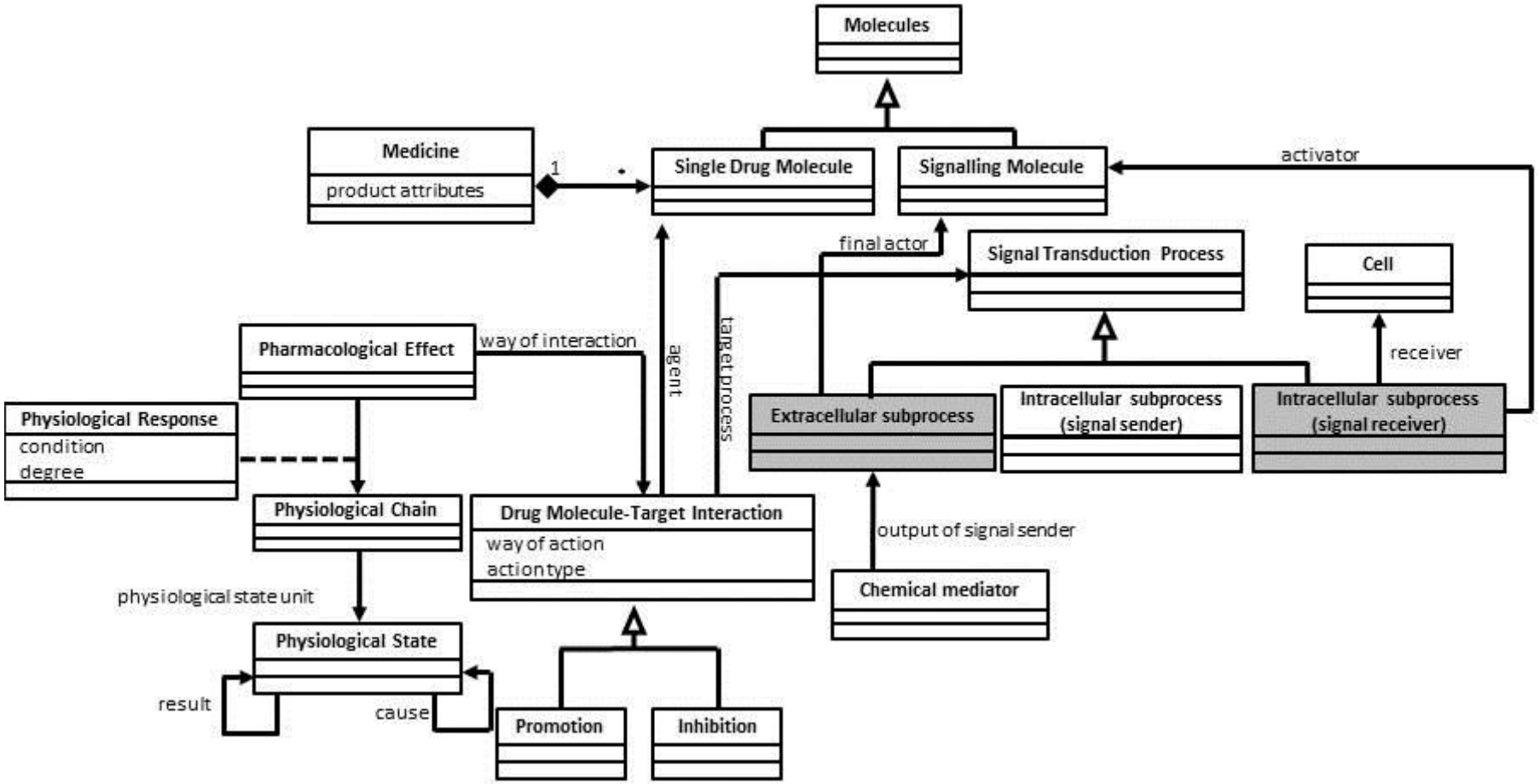

Figure 6.7. Conceptual model in the PDO. Shaded boxes represent classes with additional subclasses

- **The Pharmacokinetics Ontology (PKO):**

The aim of the PKO, developed at Indiana University, was to represent PK-related information (Wu et al., 2013). Although the final ontology integrates information from different resources, modeling efforts in this project focused on the representation of different types of PK DDI studies. PK DDI studies are experiments developed *in vitro* or *in vivo* to study the existence of drug interactions affecting some of the PK parameters of the interacting drugs.

As shown in the CM in **Figure 6.8**, the PKO does not include the concept DDI. There are five main classes representing the different types of PK studies ('Pharmacokinetic Experiments') and the entities relevant in that studies ('Drug', 'Metabolizing enzymes', 'Transporters', and 'Subjects').

The CM was implemented as an OWL ontology that imports other ontological resources (such as the ChEBI ontology or SOPHARM (Coulet, Smaïl-Tabbone, Napoli, & Devignes, 2006)), and used for annotation of documents describing PK DDI experiments.

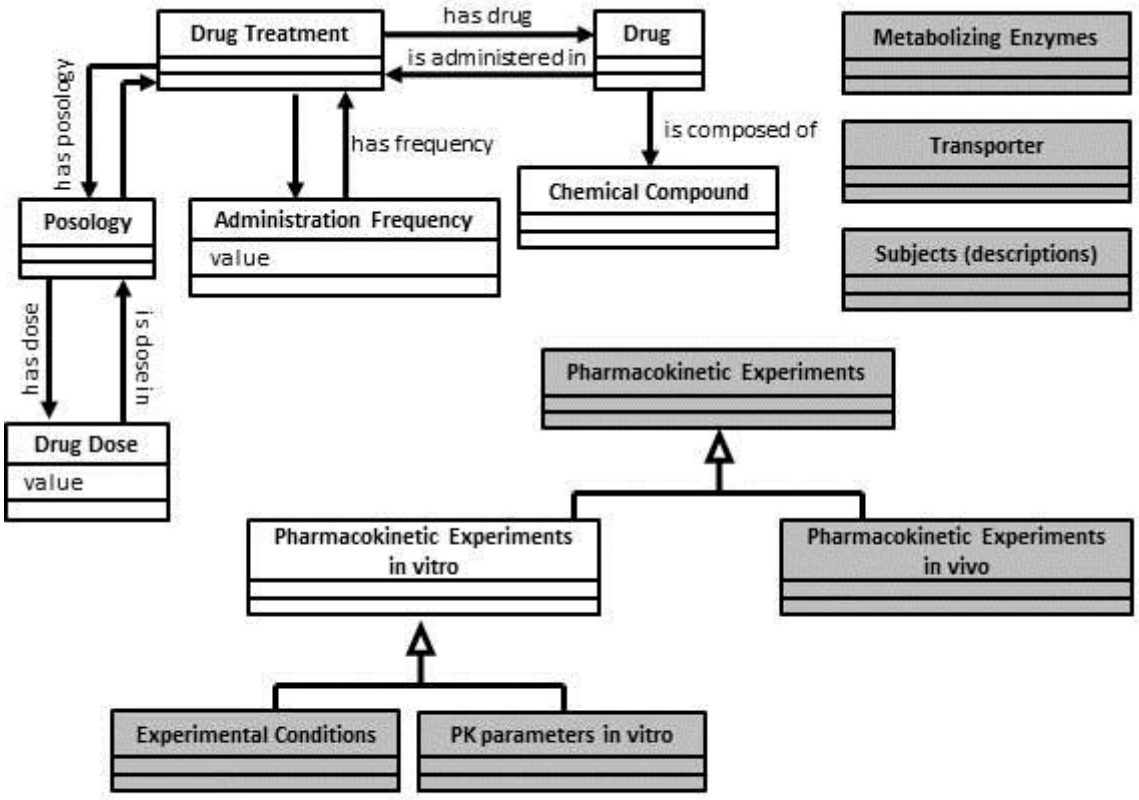

Figure 6.8. Conceptual model in the PKO. Shaded boxes represent classes with additional subclasses.

6.3 Comparison of DDI knowledge modeling approaches

In this review, we have identified seven research works describing their conceptualizations of the DDI domain in the last ten years. The main difference between their CMs is their scope – that is, the part of the DDI domain studied and covered by their representations. The scope determines the representation of the concept DDI in the model, as well as which other related concepts are included. Here, we compare the different conceptualizations analyzing the representation of the concepts 'Drug' and 'Drug Class', other relevant objects in the domain such as metabolites or proteins, the concept 'DDI', and other qualities and processes required to model the domain.

As we explained in **Section 5.1**, the word "drug" can be used with different meanings in different sources: as an "active ingredient" or the specific molecule that bears some pharmacological activity (e.g., *paracetamol*), as a "drug class" that gathers active ingredients on the basis of some relevant characteristic in the same group of drugs (e.g., *benzodiazepine*, *analgesic*, etc.) or as a "clinical drug" or "drug product", which refer to the unitary dose of a medicine (e.g., a specific tablet of *paracetamol*) or a commercial unit of a medicine (e.g., a pack containing 20 tablets of *paracetamol*), respectively. In the case of the CMs reviewed in this work, we have observed that all of them represent the concept 'Drug' at the 'Active Ingredient' level. Therefore, all of them agree to consider that a DDI should be described to occur between two active ingredients. Hence, four of them (**Rubrichi**'s CM, **NDF-RT**, **PDO** and **PKO**) include in their models the fact that an active ingredient is a component of a medicine – that is, an active ingredient is the component of a clinical drug or a drug product.

Regarding the concept of 'Drug Class', it is included in the CMs created by **Mille** and **Rubrichi**, **NDF-RT** and the **PKO**. However, all of them represent this concept in very different ways. **Mille**'s represents drug classes as an attribute 'type' of an 'Active Ingredient'. In this way, the active ingredient *paracetamol* would have type *analgesic* if we represent it in this model. In contrast, **Rubrichi**'s CM establishes that a 'Clinical Drug' or 'Drug Product' belongs to a 'Drug Class', instead of establishing the relationship for the 'Active Ingredient'. In a similar way, **NDF-RT**, which includes different classifications of drugs, relates drug classes and clinical drugs hierarchically. In this way, the clinical drug *acetylcysteine 20% inhalation solution* is a subclass of the drug class *mucolytics*. This clinical drug is related to an active ingredient class in **NDF-RT** through the relationship '*has ingredient*' (e.g., *acetylcysteine 20% inhalation solution* – '*has ingredient*' – *acetylcysteine*). Therefore, in this model active ingredients and drug classes are indirectly related through their respective relationships with clinical drugs. A special case is the representation of drug classes in the **PKO**. As mentioned before, this ontology has imported information from other ontologies. In this way, the model corresponding to the representation of the concept 'Drug' shown in **Figure 6.8** was imported from SOPHARM, which in turn integrates information from the ChEBI ontology. As a result, in this CM an 'Active Ingredient' is a subclass of one or more 'Drug Class', which are subclasses of the top-level class 'Drug'.

Besides drugs, other object entities are relevant in the DDI domain. Drug metabolites (or the molecules produced as the consequence of the metabolism of a drug in the body)

are represented only in **DIO** and **DIKB**. The reason is that these two works focus on the representation of the processes that alter the metabolism of drugs. Therefore, metabolites are key concepts in their models. Another important object in the DDI domain is 'Proteins', since they are involved in the mechanisms – both PD and PK – of most DDIs. In spite of this, they are represented only in three of the reviewed models. **DIO** represents 'Enzymes', 'Transporters' and 'Albumins' as three different types of proteins. Meanwhile, 'Enzymes' is the unique type represented in the **DIKB** CM. Finally, the **PKO** represents 'Metabolizing Enzymes', 'Transporters', and 'Targets'.

Regarding the representation of DDIs, we have observed that the concept of a DDI is only explicitly represented in four CMs. **NDF-RT** and **Mille** and **Rubrichi**'s CMs use the concept DDI as the central class in their models, while the **DIKB** represents it as the relationship '*metabolic inhibition interaction*', which explicitly represents an interaction between two different drugs. The CMs created by **Mille** and **Rubrichi** represent the relationships between a DDI and other related aspects (risk factors, mechanism, effect, etc.). In contrast to them, **NDF-RT** includes only the relationship between a DDI and the two interacting drugs. In spite of their differences, the four of them share a binary representation of the DDI. In other words, all the models that explicitly describe a DDI consider that it involves exactly two entities. In this way, an interaction is described to occur in **DIKB**'s model between a precipitant and an object drugs. **NDF-RT** and **Mille**'s CMs specify that an interaction occur between two drugs. In contrast, the CM developed by **Rubrichi** asserts that an interaction occurs between two drugs or between a drug and a group of drugs, a drug and a diagnostic test, or a drug and a different substance.

Another important difference between the seven projects is that only two of them (**Mille** and **Rubrichi**'s CMs) have attempted a global representation of the DDI domain, while the other five focus on some concrete aspect, providing a partial representation of the domain. Specifically, the **PDO** has focused on the representation of drug-molecule interactions that lead to different PD processes in the body, and how they are related to drugs, pharmacological effects, and the consequent physiological effect. This information is useful to represent the underlying processes altered in a PD DDI. In a similar way, but focusing on PK instead of PD processes, **DIO** represents drug PK processes and their relationships with drugs, metabolites, and proteins, providing a representation of those processes that can be altered in a PK DDI. Meanwhile, the **DIKB** model shows how one of these specific processes (i.e., metabolism) is altered in a PK DDI caused by the inhibition of a metabolic enzyme. This model includes the different drug characteristics that influence or determine the significance of a DDI. Meanwhile, the **PKO**'s CM does not focus on the representation of DDI mechanisms as the ones described before, but on the representation of PK DDI studies and their components. Finally, as mentioned before, **NDF-RT** focuses only on the representation of the relationship between a DDI and the two interacting drugs.

Up to now, we have focused on the study and comparison of relevant objects in the DDI domain that are represented in some of the seven reviewed CMs. However, processes and qualities are important concepts that have been represented, too. We have identified four main types of processes and/or qualities related to the DDI domain and included in some of these models: the effect of the DDI, its mechanism, factors that can increase the risk or severity of the DDI, and measures to avoid or manage the DDI.

The consequence or effect occurring as a result of a DDI is only represented in the CMs created for the global representation of the DDI domain (**Mille** and **Rubrichi**'s

CMs). However, they represent the concept at a very high level of granularity, and specializations of the different types of consequences of a DDI are not detailed. In contrast, **DIKB**'s CM does not represent this effect of a DDI as a class, but as relationships between different drugs. These relationships represent that a drug can alter the clearance of another drug – there is a change in the metabolism of the drug and therefore an alteration in its global clearance or elimination from the body ('*inhibit metabolic clearance*') – or that a drug can alter the exposure or concentration of the other drug in the body ('*increase drug exposure*').

Mechanism, or the process that leads to the occurrence of a DDI, is represented as a class only in the CM created by **Mille**, while it is not included in the one created by **Rubrichi**. In the same fashion that it represents effects, **DIKB**'s CM represents DDI mechanisms as relationships between two different drugs and an enzyme (e.g., the precipitant drug '*inhibits*' an enzyme while the object drug '*is substrate of*' the same enzyme). PK processes and PK parameters can be altered in a PK DDI mechanism. They are represented in **DIO** and the **PKO**, respectively. In contrast, **DIKB**'s CM represents only the PK process metabolism. In the same way that PK processes are altered in PK DDIs, PD processes and pharmacological effects are altered by a PD DDI mechanism, and are represented in the **PDO**.

There are different ways to avoid or manage a DDI (Lea, Rognan, Koristovic, Wyller, & Molden, 2013). However, only **Mille** and **Rubrichi**'s CMs introduce this concept, although without a deep level of detail ('Precaution of Use' and 'Recovering Action', respectively). Factors that can increase the severity of a DDI are represented only in three models. **Mille**'s CM includes patient-related factors, while the **DIKB** represents some of the most important drug-related features that can influence the severity of a DDI. In contrast, **Rubrichi**'s model considers both patient and drug related factors.

Finally, we have compared how the authors implemented their different CMs. **Rubrichi**'s CM, **DIO** and the **PKO** were implemented as OWL ontologies. The **DIKB** represented the model as a set of rules in FOL, while **Mille**'s CM was used to build an XML schema. The public release version of **NDF-RT** is available in several forms that include XML and OWL formats.[25] Finally, the **PDO** has not been implemented yet.

To conclude, we provide a summary of the results in **Table 6.1**. In addition to this, in **Annex 1** we compare the concepts included in each one of the reviewed CMs, and contrast them with the CM in our ontology DINTO (**Section 7.1.3**).

6.4 Discussion

In the last ten years, there have been different projects that have required the conceptualization of DDI-related knowledge. The CMs created by these research groups have been developed independently and from scratch and, since each one of them was created with different purposes, there is not a high degree of overlapping among them.

[25] http://www.cancer.gov/cancertopics/cancerlibrary/terminologyresources/fmt

The result is a group of different isolated CMs dealing with the representation of different aspects of the same domain.

The two closest works are **Mille** and **Rubrichi**'s CMs, which were developed to be used in NLP and have been applied to the annotation of texts. These CMs share a wide coverage of the DDI domain, but with a very low level of detail. Although their structure is similar, there are some differences between them. In this way, **Rubrichi**'s CM does not represent the concept of mechanism while side effects or pharmaceutical characteristics of drugs are not included in **Mille**'s one. Therefore, both are complementary to each other and could be combined to provide a more comprehensive view of general aspects related to the DDI domain. Another work created for NLP tasks is the PKO. The main difference between this model and the two previous ones in that it was designed to describe only PK DDI studies and therefore it does not overlap with them.

The other main application of DDI-related CMs has been the prediction of DDIs on the basis of their mechanisms. Three different CMs have demonstrated that the formal representation of DDI knowledge can be used successfully to infer or predict interactions between specific pairs of drugs. The larger prediction experiment was developed in the framework of the **DIKB** project that predicted PK DDIs occurring by enzymatic metabolism inhibition for over 60 drugs. In contrast, the model proposed in **DIO** was used to detect PK DDIs between a specific pair of drugs (*irinotecan* and *ketoconazole*). Although the evaluation was limited to only one pair of drugs, they identified through this model four different processes or DDI mechanisms that could explain the DDI between these two drugs. This detailed description of how a PK DDI may occur is important and consistent with the pharmacological fact that most DDIs are caused not by a single mechanism, but often by two or more mechanisms acting in concert (Baxter, 2013). In a similar way, the CM implemented as the **PDO** was used to identify the different sub-processes involved in the interactions between drugs participating in the same signal transduction process, leading to the identification of different types of PD DDI mechanisms.

Contrary to NLP intended models, CMs used to predict DDIs have in common that they focus on a specific aspect of the DDI-related knowledge and model it in detail. Consequently, the number of classes in these CMs is larger than in the former ones. However, these three CMs have important differences in their scopes. As mentioned before, the **DIKB** model is more specific, since it focuses only on one particular PK DDI mechanism. In contrast, the **PKO** and the **PDO** cover PK and PD processes, respectively, and any DDI mechanism is explicitly represented. The level of detail in which PD and PK processes are represented is similar in both models. Therefore, they could be considered as complementary to each other.

However, although these three models proved that they could be used to correctly predict DDIs between pairs of drugs, their use in a final system, such as a CDSS for DDI alerts, is limited because they rely on the availability of explicit information for every individual drug or drug-related facts according to the CM. For example, to be able to predict the interaction between two drugs (e.g., *irinotecan* and *ketoconazole*) following the CM represented in the **DIKB** project, it is necessary to know that *irinotecan* is a substrate of the enzyme *CYP3A4* and that *ketoconazole* inhibits this enzyme. Also, it is required in this model as well to establish if this is the main clearance route for *irinotecan* or not, and if the drug has narrow therapeutic index or if either is a sensitive substrate. In the case of **DIO,** it is necessary to know, as well, where every process occurs (e.g., in the

liver, a vein, an artery, etc.). Meanwhile, the **PDO** requires explicit knowledge about how a drug acts in a specific signal transduction process (and its sub-processes), and how it is related to causal chains of physiological states that lead to the physiological response. Therefore, as the degree of detail of the CM increases, the number of necessary drug-related facts increases, too. The discussed projects identified and gathered these required drug-related facts through manual search of the literature. However, manual curation is a highly time-consuming and expensive task and, as a consequence, the models could be applied only for a reduce set of drugs. Moreover, new drug-related facts are discovered every day and published in the scientific literature (Hunter & Cohen, 2006). Therefore, keeping up to date a system for DDI identification that relies on these CMs would be a challenging task.

An alternative to manual curation of drug-related facts could be to exploit the increasing pharmacological knowledge stored in structured and machine-readable formats in public pharmacological databases and knowledge bases (Khelashvili et al., 2010; Whirl-Carrillo et al., 2012) A new approach for the efficient and sustainable prediction of DDIs could be the design of CMs that support DDI prediction on the basis of their mechanisms, taking into account the availability of structured pharmacological information that can be exploited automatically. This method would lead to less detailed CMs that, nevertheless, would be able to predict DDIs for larger sets of drugs.

The aim of this thesis is to create an ontology that provides a complete representation of the DDI domain and that, at the same time, describes in a comprehensive way all the different aspects of DDIs mentioned in pharmacological text. During the annotation and analysis of the DDI corpus (see **Section 3.1**), we observed that DDIs are mentioned in multiple ways: describing how the interaction occurs, through the relations of two different drugs with the same protein, as a contraindication, as the effect of the DDI, and so forth. Therefore, the CM of an ontology useful for extracting DDIs from text should have a global scope of the DDI domain with a deep degree of detail, at the same time. In addition to this, inference of DDIs requires a detailed representation of all different pharmacological processes leading to different types of PK and PD DDIs. These characteristics, however, are not met by models described in this review. On the one hand, CMs created by **Mille** and **Rubrichi** are designed at a too high level of abstraction. Therefore, different types of DDI mechanism, effects, or risk factors are not represented. On the other hand, **DIKB**'s model focuses only on one of the many PK DDI mechanisms that can lead to DDIs, while the **PDO** and **DIO** describe pharmacological processes, but not explicitly DDI mechanisms. Finally, the **PKO** focuses on the representation of PK studies, and do not meet the requirements described before.

Consequently, the creation of our ontology requires the design of a new CM representing, in an appropriate degree of detail, all DDI-related knowledge. Nevertheless, current efforts described in this review will be considered and adopted when appropriate, as a way of gathering consensual knowledge regarding the formal representation of DDI knowledge.

	DIO	DIKB	Mille	NDF-RT	Rubrichi	PDO	PKO
Year	2004	2005-2007	2007		2012	2013	2013
Intention	Modeling of drug-biomolecule interactions	Prediction of DDIs	Development of Clinical Decision Support Systems	Modeling of drug related information	Automatic information extraction from texts (NLP)	Prediction of DDIs	Annotation and extraction of DDIs from text (NLP)
Representation of DDI in the CM	No	Yes	Yes	Yes	Yes	No	No
Main application	Prediction of PK DDIs	Prediction of PK DDIs	Encoding of textual documents	Support of computerized systems	Annotation of text	Prediction of PD DDIs	Annotation of text
Intentional scope	Partial	Partial	Global	Partial	Global	Partial	Partial
Degree of deepness/detail	Detailed	Detailed	Shallow	Shallow	Shallow	Detailed	Detailed
Knowledge source	-	Lectures and class notes of a graduate class on drug interactions	DDI monographs	-	SPCs	-	Textbooks and literature sources
Implementation Language	OWL	FOL	XML	DL (and OWL)	OWL	OWL	OWL

Table 6.1. Results of the comparison of the seven different conceptual models

6.5 Unresolved issues

Current efforts in modeling DDI-knowledge represent a significant contribution to the state of the art in this research area. However, some limitations should be still addressed to create a comprehensive CM for the DDI domain.

1. Current CMs do not provide, at the same time, a wide and detailed enough representation of the DDI domain, which would entail their use in different applications, such as NLP tasks and inference of DDIs.

2. The CMs created for NLP applications provide a global representation of the domain, but with a low level of detail. Therefore, their application to IE tasks has been limited to encoding and manual annotation of text. However, their performance in extraction of DDIs from text, such as scientific abstracts, has not been evaluated.

3. CMs for prediction of DDIs provide a detailed representation of pharmacological processes leading to DDIs. However, they rely on manual curation of drug-related facts to be used in inference of DDIs, which restricts the number of drugs included in the evaluations and makes it difficult to update new information.

4. There are two main types of mechanisms leading to DDIs: PK and PD DDI mechanisms. However, current CMs have not represented both of them in the same framework. Therefore, applications relying on these CMs must focus only on one type of DDIs, PK or PD DDIs, but not on both of them at the same time.

5. There are different subtypes of PK and PD DDIs. However, current CMs have focused only on some of these subtypes, and there is not a comprehensive representation of all of them in any CM.

6. The possible effects of a DDI (e.g., beneficial or harmful, clinically relevant or non-clinically relevant effect) have not been represented in the reviewed CMs.

7. Proteins are relevant substances in DDIs, since they are involved in most DDI mechanisms. However, they have been poorly represented in current CMs, and their relationships with different drugs have not been represented in detail.

8. The different strategies to avoid or manage DDIs have not been represented in detail in these CMs.

9. Drug and patient-related factors affecting DDIs are not comprehensively included in the reviewed CMs.

The new CM in our ontology for DDI knowledge will attempt to cover each of the unresolved issues described in the above list.

6.6 Conclusions

In this review, we have provided an integrated view of current efforts for the representation of DDI knowledge. These CMs have been used for two main applications: NLP of pharmacological texts and inference of DDIs. Models created for NLP require a global representation of the DDI domain, while models applied to DDI inference need a more detailed representation. These projects have shown that formal representation of DDI knowledge can be used successfully in both NLP and prediction of DDIs on the basis of their mechanisms. However, they have not been used in DDI extraction from text, and the number of drugs included in the inference experiments has been determined by the required manual cuaration of drug-related facts. For better performance in IE tasks, it is necessary that future CMs combine a global scope of the DDI domain and deeper level of detail of the concepts and relationships included in the representation. At the same time, we believe that inference of DDIs on a large scale requires the creation of models that do not strongly rely on manually curated drug-facts, but which could exploit existing computerized drug information sources. Based on these conclusions, we develop the conceptualization of our ontology, which is described in **Section 7.1.3**.

Chapter 7

The Drug-Drug Interactions Ontology: DINTO

As we have seen in previous chapters, there is any ontology representing, in a comprehensive way, all of the information related to the DDI domain. Therefore, in this thesis we propose the construction of a new ontology, the Drug-Drug Interactions Ontology (DINTO). This chapter details in **Section 7.1** the process of building the ontology, and the resulting ontology is described in **Section 7.2**.

7.1 Building the ontology

There are different approaches for building ontologies that can be divided into two main categories: constructing the ontology from scratch and reusing other ontologies. On the one hand, the creation of an ontology from scratch implies the design of the ontological structure without any general or reference model (Cristani & Cuel, 2004). It is generally carried out manually and the ontologists need to identify the concepts in the domain and the relationships between them, and implement the ontology in a machine-readable format. On the other hand, reusing ontologies is becoming a relevant process in ontology engineering, mainly due to the increasing amount of available resources (Simperl, 2009). This approach enables the interoperability among existing ontologies and avoids the duplicity of efforts and content overlap between them.

The creation of an ontology is a complex and time consuming task. As a consequence, semi-automatic approaches have been used for the partial automation of the ontology development process (Gómez-Pérez, Fernández-López, & Corcho, 2004). The main contribution of these approaches is the reduction of the efforts required during knowledge acquisition (KA), one of the most time and resource-consuming activities in the development of an ontology (Payne, Mendonça, Johnson, & Starren, 2007). This automated ontology construction, named ontology learning, is based on natural language analysis and machine learning techniques (Gómez-Pérez & Manzano-Macho, 2003). Therefore, the main limitation of this approach is that it is strongly dependent on the employed technique and the selected data source. Consequently, an extensive verification of the final ontology is required to ensure its quality.

These different approaches can be used in combination. For example, the general structure of an ontology can be manually developed from scratch in the first steps, while information gathered in other ontologies is reused subsequently (Suárez-Figueroa, Gómez-Pérez, & Fernández-López, 2012). In the same way, a manually created ontology can be populated later through NLP techniques (Rubrichi et al., 2013).

Once the most appropriate approaches have been identified, following a method defining what specific activities and the order in which they should be carried out can be useful in performing the task of ontology creation (Uschold, 1996). Different researchers have proposed methods for building ontologies, with the aim to provide guides that help ontologists in the development of a new ontology, and different authors have studied and reviewed them. A new analysis of all of them is out of the scope of this chapter. However, interested readers can find in Gómez-Pérez et al. (2004) a comprehensive review and comparison of classical methodologies and methods for building ontologies from scratch or by reusing other ontologies – including a review of ontology learning methods.

For the creation of DINTO, we require starting the development process from scratch, since, as we have seen in **Chapter 5** and **Chapter 6**, no existing ontologies have dealt with a comprehensive representation of the DDI domain. However, some relevant information might be found in other ontologies and databases. Therefore, we need to integrate information from different resources, too. To the best of our knowledge, the methodology that best fits these requirements is the Neon Methodology (Suárez-Figueroa et al., 2012), developed by the *Ontology Engineering Group* at *"Universidad Politécnica de Madrid"*.[26] This methodology focuses on ontology engineering by reuse and supports different aspects of the ontology development process. It encompasses, as well, a previous work by the same research group: the methodology for building ontologies from scratch METHONTOLOGY (Fernández-López, Gómez-Pérez, & Juristo, 1997).

[26] http://www.oeg-upm.net/

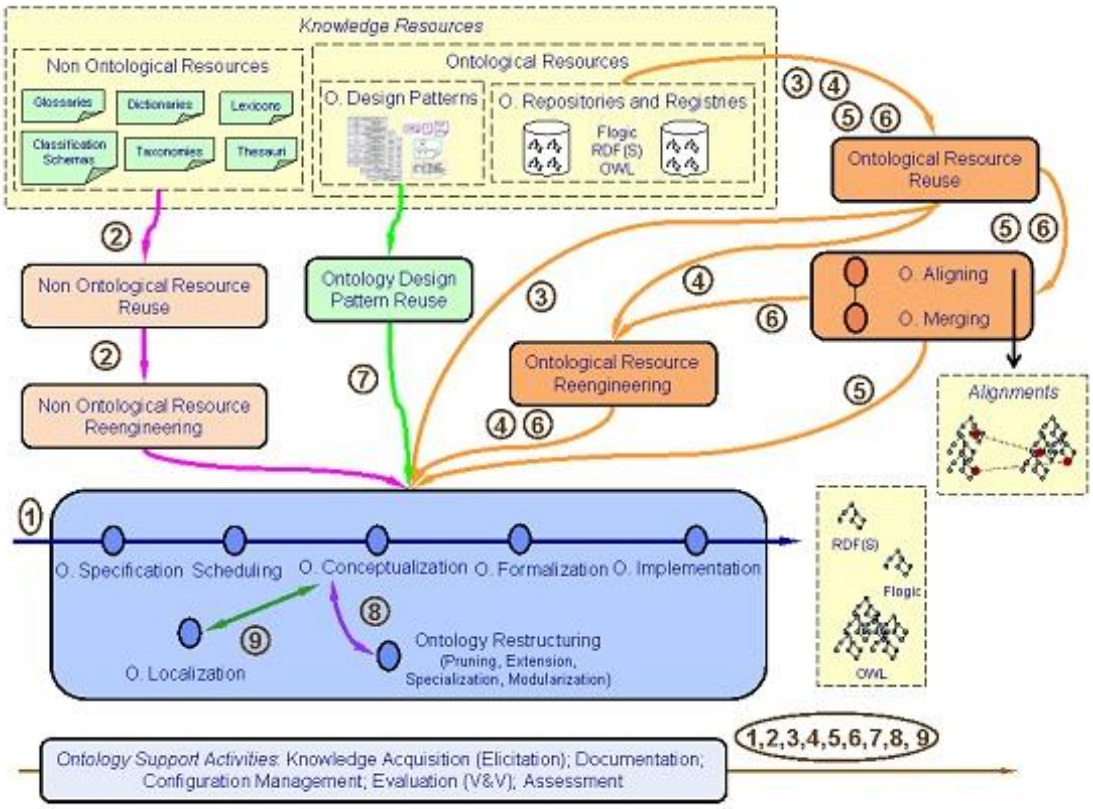

Figure 7.1. Neon Methodology (from Suárez-Figueroa et al., 2012)

The Neon Methodology proposes a variety of pathways for developing ontologies or scenarios which cover commonly occurring situations. **Figure 7.1**, originally published in (Suárez-Figueroa et al., 2012), represents them. Scenario 1, which describes the different activities required for the creation of an ontology from scratch, and the Ontology Support Activities (shown in the light blue box at the bottom of the figure) have been previously described in the framework of METHONTOLOGY. It proposes an iterative process where the different activities are not intended to be carried out in a sequential way.

Below, we provide a description of the most important activities developed during the construction of DINTO: 1) specification, 2) knowledge acquisition, 3) conceptualization, 4) implementation, 5) ontological resources reuse, 6) non-ontological resources reuse, 7) creation of rules for inference of DDIs, and 8) maintenance. The support activity "evaluation" is not included in this section, since it is described in detail in **Chapter 8**, **Chapter 9**, and **Chapter 10**.

7.1.1 Ontology specification

The aim of the specification activity is *"to state why the ontology is being built, what its intended uses are, who the end users are, and which requirements the ontology should fulfil"* (Suárez-Figueroa, Gómez-Pérez, & Villazón-Terrazas, 2009). The output of this activity is the ontology requirements specification document (ORSD), which describes

the purpose, scope, and intended users and scenarios of the ontology. It also includes a series of competency questions (CQs), or natural language questions representing the aspects that the final ontology should address, such as "*Is the effect of DrugA modified by DrugB?*" or "*Is there a PK interaction between DrugA and DrugB?*"

We have defined the ontological requirements for a comprehensive ontology that should represent all the information related to the DDI domain, and that should be mapped to other ontologies. It is primarily intended to be used by text miners and software developers, although we anticipate that it may find additional use amongst domain experts once it is further developed. For a successful application in NLP, concepts in the ontology should include alternative terms representing every different terminological variation (lemmas, synonyms, abbreviations) and should include semantic relationships characteristics of the domain, which is the input to the RE task, for example, based on patterns (Lönneker, 2003). On the other hand, the ontology might be used by the computational community working on applications within the DDI domain, such as the development of CDSS or signal detection systems in pharmacovigilance. For this purpose, the ontology must include known DDIs, and allow the inference of new ones on the basis of their mechanisms.

In order to ensure reusability of our ontology, we establish as a requirement that the ontology must follow the OBO Foundry recommendations for building ontologies (Smith et al., 2007). The OBO Foundry is a collaborative effort for the development and maintenance of biomedical ontologies, which aims to ensure their integration and interoperability through the establishment of a set of principles that are voluntarily accepted by its participants (see **Section 8.1.3**).

The CQs identified in this activity are in turn used as a type of requirement specification and evaluation for the finished ontology (see **Section 8.1.2**). The ORSD resulting from this activity is shown in **Annex 2**.

7.1.2 Knowledge Acquisition

The knowledge acquisition (KA) activity consists in capturing the knowledge of the domain that should be represented in the ontology. This activity is considered in Neon Methodology as a support activity that is coincident with other ones. Specifically, the effort on KA is greater during the specification and conceptualization activities, and decreases during formalization and implementation (Gómez-Pérez et al., 2004).

There are many different knowledge sources from which the domain knowledge can be extracted − domain experts, scientific literature, books, or even other ontologies − and different techniques that can be used to elucidate the knowledge from them − interviews with experts, brainstorming, informal and formal analysis of texts, or use of knowledge acquisition tools (Fernández-López et al., 1997). During our review of current approaches for modeling the DDI domain (see **Chapter 6**), we have observed that most projects acquired the domain knowledge from different textual sources: lectures and class notes of a graduate class on drug interactions (Boyce et al., 2007), textbooks and literature sources (Wu et al., 2013), DDI monographs (Mille et al., 2007), and Summaries of Products Characteristics (SPCs) (Rubrichi & Quaglini, 2012). Therefore, it seems that the study of

DDI-related texts has proven to be a useful source for the elicitation of this domain knowledge. This fact agrees with the theory of sublanguages proposed by Zelling Harris (Harris, 1982, 1991), who proposed that the languages of technical domains, such as the biomedical, have a limited number of words and possible relationships between them, and a structure and regularity that can be observed by examining the corpora of the domain and which can be represented in a form suitable for computation (Friedman, Kra, & Rzhetsky, 2002).

According to this, we have selected the previously created DDI corpus as the most appropriate resource for KA in the creation of DINTO (Herrero-Zazo, Segura-Bedmar, Martínez, et al., 2013). This corpus collects documents from different sources (the database DrugBank and scientific abstracts from MEDLINE), which provide a global coverage of the DDI domain and descriptions at different levels of detail. Hence, the corpus has been previously annotated with different types of DDIs, which enables the study of different aspects of the domain (e.g., the manner in which the interaction occurs, the effects of the DDI or the different recommendations for avoiding a DDI). An additional advantage of using the DDI corpus is that it is available in XML format and can be used by different tools that support text analysis, facilitating the examination of this large set of documents.

The final goal of the analysis of the corpus is to identify relevant concepts in the DDI domain and the different relationships that can be established between them. Then, these concepts and relationships are represented in a CM and implemented in the final ontology. However, the corpus includes a large number of documents and sentences, where usually different terms refer to the same concept, or different patterns describe the same relationship.

Corpus linguistic techniques and tools can be useful to analyze such a large textual information source, since they allow for the reliable and exhaustive analysis of text. In this project, we have employed two different strategies. The first one is linguistic pattern analysis, which allows identifying the relevant concepts in the domain and the different terms used to describe them. The second one is word frequency and concordance analyses, which are performed to identify the terms most commonly used in the texts, and the different relationships that are established among them. These two approaches are described below, along with other general information sources and ontologies also used in this KA activity.

- **Linguistic pattern analysis**

The first step in KA is to identify the relevant concepts in the domain, which are organized subsequently into taxonomies. The complexity of the DDI domain leads to a large number of different concepts, and a huge body of different terms referring to them. Their identification can be difficult due to linguistic aspects such as ambiguity, nested terms, or coordinate structures (see **Section 3.4**). However, corpus linguistic tools can be useful, since they support the annotation of sentences and the subsequent analysis of the annotated segments.

In this project, we use the UAM Corpus tool[27], a free environment enabling the annotation of selected segments in the corpus with various features or labels. First, we create an initial label-schema list representing the main general concepts in the domain, such as 'PK Mechanism' or 'Study Subject'. Secondly, the documents in the DDI corpus are read, and terms or groups of terms are annotated. The list is iteratively refined during the progress of the analysis, when new concepts are identified. The complete annotation label-schema list is shown in **Annex 3**.

When the desired texts have been annotated, the tool retrieves the segments annotated with the same label, which can be studied for the identification of linguistic patterns or refinement of the annotation. The most important conclusions of our analysis refer to the description of DDI mechanisms and DDI effects in text, which are described in detail below.

- *Description of DDI mechanisms in text:*

The analysis of sentences annotated in the DDI corpus as type 'Mechanism' (i.e., sentences that describe how an interaction occurs) leads to the identification of important concepts and the relationships between them expressed in text. Here, we summarize the most relevant conclusions.

1. The mechanism of a PK DDI – or how a PK DDI occurs – is expressed as the alteration in a PK process: e.g., "*increases the absorption of*", "*altered the metabolism of*".

2. The consequence of a PK DDI mechanism is expressed as the alteration in a PK parameter or in the levels or concentrations of the drug: e.g., "*an increase in the AUC*", "*has been described to increase the mean half-life*", "*decreases the levels of*".

3. There are consequences of a PK DDI mechanism that can be caused by the alteration of several PK processes. Therefore, it is not possible to establish a causal relation between that consequence and a mechanism: e.g., the consequence "*increases the levels of*" can be caused by an increase in the absorption, an alteration in the distribution, or a decrease in the metabolism or elimination of the drug.

4. Mentions of metabolites are frequently related to DDIs occurring via a PK mechanism: e.g., "*MetaboliteA can be formed with concurrent ingestion of DrugA and DrugB*".

5. There are some mechanisms that can lead to unpredictable DDI effects: e.g., "*Since DrugA may reduce the gastrointestinal absorption of both DrugB and vitamin K, the net effects are unpredictable*".

6. Some terms used in the pharmacological domain to describe PD mechanisms (antagonism, synergism, potentiation, etc.) can be used in texts to describe a PK

[27] http://www.wagsoft.com/CorpusTool/

mechanism: e.g., "*DrugA may potentiate the action of DrugB by inhibiting its metabolism*".

7. Numerical values are usually used when describing DDIs occurring by a PK DDI mechanism: e.g., "*resulted in an approximately 60% increase in the AUC*".

- *Description of DDI effects in texts:*

During the analysis of annotated segments in the corpus labelled as 'Pharmacodynamic Effect of a Drug', we identified five main ways a clinical consequence of a DDI can be described:

1. The effect of a DDI is the effect of a drug: e.g., "*increase the adrenergic effect of*", "*reduced the action of*".

2. The effect of a DDI refers to a negative effect of a drug: e.g., "*the adverse effect of*", "*the ototoxic potential of*".

3. The effect of a DDI refers to the clinical manifestations, generally related to an ADR without explicitly relating them with a drug: e.g., "*serious reactions such as rigidity, myoclonus, or autonomic instability*".

4. The effect of the DDI provides information related to some aspect of the DDI: e.g., "*additive CNS depressant effect*", "*observed an excessive reduction of blood pressure*".

5. The effect is expressed through a modification in some analytical test result: e.g. "*increase in prothrombin time*".

All this knowledge is used during the next activity, conceptualization of the domain. Besides the identification of relevant concepts, it enables the identification of relationships between them in an accurate way, such as that the consequence of a DDI (the 'DDI Effect') is a special type of effect caused by a drug, and therefore, should be represented as a subclass of the concept 'Drug Effect'; or that a 'PK DDI Mechanism' is always related to one or more 'PK Process', and therefore, a specific relationship between them should be represented in the ontology.

It is worthy to note that, in contrast to the annotation of the DDI corpus, this exercise does not have the final aim to provide an annotated document. As we have explained in this section, the objective is to support the analysis and labelling of sentences to provide insights in how information is described in texts, and to identify relevant concepts and relationships in the domain. The gained knowledge is later represented in intermediate conceptual models and implemented in an ontology. Therefore, annotation of the complete corpus or the creation of annotation guidelines is not necessary.

- **Frequency and concordance analysis**

The second approach for the analysis of the DDI corpus is word frequency and concordance analyses. The aim is to identify the terms most commonly used in the DDI

corpus, and then analyze their use in context to identify their relationships with other terms and concepts. To this purpose, we use the freeware AntConC[28] corpus analysis toolkit, which has been previously used in KA for the construction of biomedical ontologies (Mendonça, Coelho, Andrade, & Almeida, 2012).

We perform a word frequency analysis to identify the occurrences of words in the 1,017 documents forming the whole DDI corpus. We use a stoplist to filter out words that are not relevant to this study, such as pronouns or prepositions. The final list includes 90,382 words distributed over 6,842 word types. Since we are not interested in identifying different drug names (e.g., *paracetamol*, *ciprofloxacin*, *Gelocatil*®, etc.), we substitute them by the words *druglabel*, *grouplabel,* and *brandlabel*, following the manual annotation of the DDI corpus as `drug`, `group`, or `brand` entity types, respectively. Using these labels, we can generalize the linguistic patterns and focus on the study of other concepts. Indeed, the three most frequent words correspond to these three different groups of drug names. We review the 2,436 most frequent types of words, excluding those terms that are found in the corpus only once, twice, or three times. This list is used then to identify and represent relevant concepts in the conceptualization activity.

In addition to the word frequency analysis, and in order to study how different concepts relate to each other in the domain, we conduct a concordance analysis of relevant terms. During the linguistic pattern analyses, we have observed that most DDIs are expressed through the modification of an effect, PK process, or PK parameter. Therefore, we identify the terms used to describe a modification or alteration (increase, decrease, potentiation, etc.) and analyze the concepts in the domain that are usually related to them. For example, the concept enhanc* can be used with terms labelled as 'PK Parameter', 'PK Process', 'Effect', or 'Toxicity'; however, the term elevat* is only used in our corpus with terms labelled as 'Concentration' and 'PK Parameter'. This knowledge is used to establish relationships between concepts in the conceptualization activity. The complete analysis is shown in **Annex 4**.

- **General pharmacological information sources and ontologies in KA**

Another important source of information is provided by other ontologies that can potentially be reused and imported into DINTO. As a first step, we have reviewed in **Chapter 5** those different ontologies relevant to the DDI domain. The final selected ontologies and the way in which they are reused in DINTO is described in **Section 7.1.5**.

Finally, different pharmacological sources have been used as a source of information and support during the creation of this ontology, including drug interaction compendia (Baxter, 2013; Hansten, 2003; Tatro, 2010), pharmacology books (Levine, Walsh, & Schwartz-Bloom, 2005; Sweetman & Martindale, 2006), and specialized publications (DuBuske, 2005; Huang et al., 2008; Olvey, Clauschee, & Malone, 2010b; Zhang, Reynolds, Zhao, & Huang, 2010; Zhou et al., 2005), among many others.

[28] http://www.antlab.sci.waseda.ac.jp/

7.1.3 Conceptualization

This activity consists in capturing the different concepts relevant to the domain and their relationships in a graphical representation called conceptual model (CM) (Olivié, 2007). We have designed a CM for DINTO that reuses and integrates information currently available in public information resources, such as chemical entities and roles from the ChEBI ontology (Degtyarenko et al., 2008) and drug-protein relationships from the database DrugBank (Knox et al., 2011), among others. Therefore, with this approach the development of our CM is driven by the combination of both the requirements of the final application and the information available to be imported into the ontology.

In this section, we describe the conceptualization of DINTO depicted in relevant aspects within the DDI domain. We describe why that specific information should be included in an ontology for DDIs and how we have conceptualised the knowledge. The CMs are represented following the common representation framework in UML class diagrams explained in **Section 6.1**. A global view of the final CM is shown in **Figure 7.2**. Due to the lack of space, this figure does not include all classes, relationships, and attributes in the ontology, and the model has been simplified. However, this representation and the partial CMs shown in this section, provide a detailed description of the conceptualization of DINTO.

- **Chemical entities, pharmacological entities, and protein entities:**

A comprehensive representation of the DDI domain requires the inclusion in the CM of drugs and proteins, all of them represented as classes. Drugs are imported into DINTO from the ChEBI ontology (Degtyarenko et al., 2008), which represents chemical entities with small molecular weight as subclasses of the top-class level '*chemical entity*'. Proteins are also chemical entities, but with high molecular weight and, therefore, they are not included in ChEBI.

In our CM, we represent this knowledge by creating a new class '*pharmacological entity*', which is the parent class of all imported drugs, independently of their nature. The class '*pharmacological entity*' is defined as "*a chemical entity which possesses some pharmacological activity. Therefore, it has the capacity to interact within a biological system and to produce a physiological effect*".

We have mentioned several times in this thesis that the term "drug" is used with different meanings, which are important from an ontological point of view. Following the conceptualization adopted in ChEBI, we consider that the class '*drug*' is a role, that is, a particular behaviour that some entity can exhibit. Therefore, in our ontology there are not "drugs" but chemical entities that can exhibit some drug-related role. When we refer to these classes in this text, we use the term '*pharmacological entity*', which is the specific class that we have created to list all of them, and to distinguish them from other chemical entities such as proteins. In contrast, when referring to the concept of drugs independently of any ontological representation, we continue using simply the term "drug".

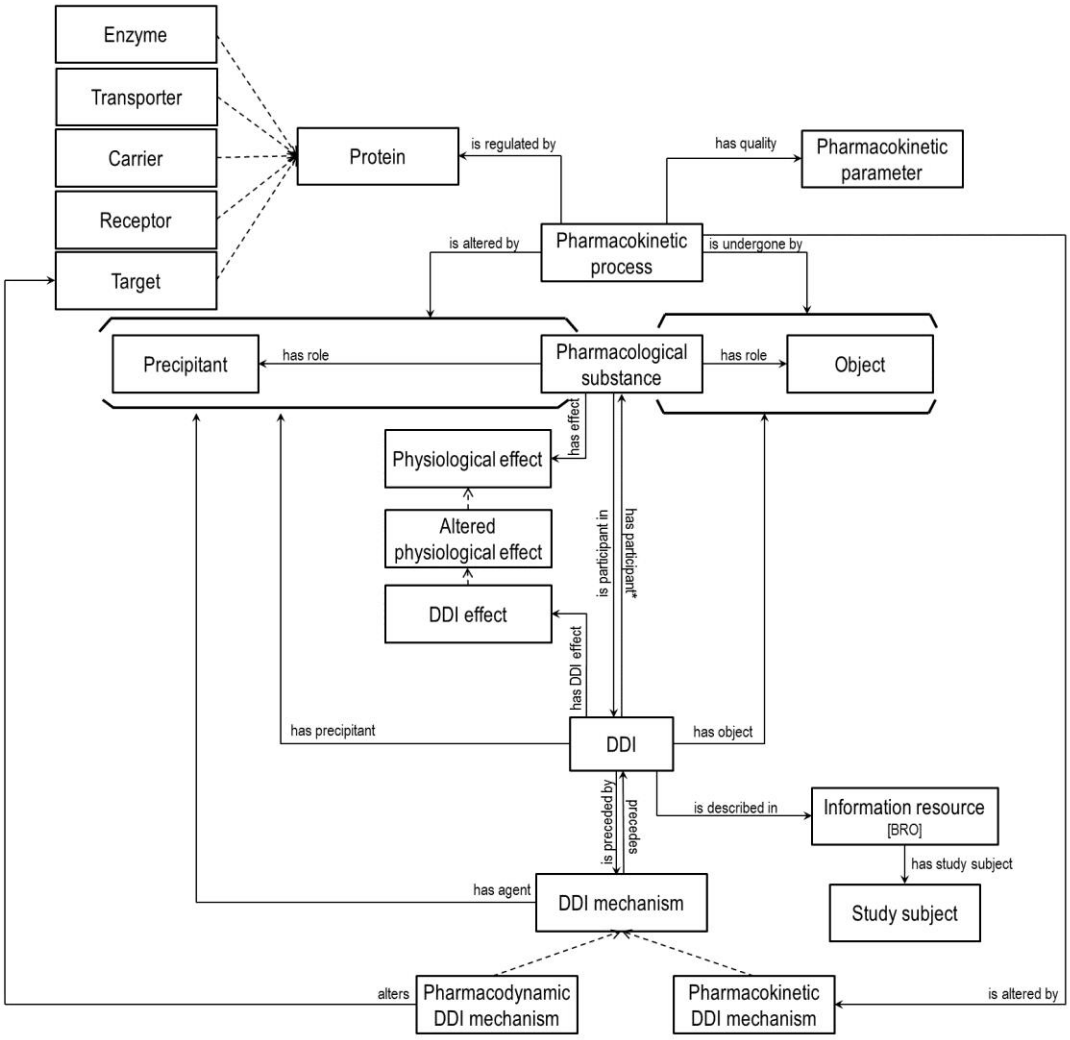

Figure 7.2. Simplified conceptual model representing the main classes and relationships in DINTO

The CM showing the hierarchy for chemical entities is shown in **Figure 7.3**. It includes an important aspect that is frequently mentioned in DDI texts: the concentration of a drug in the body. As shown in the figure, it is represented in the CM as an attribute of the class '*pharmacological entity*'.

- **Role**

We have said that a role is the different behaviour that a material entity can exhibit, such as '*drug*', '*anti-allergic agent*', or '*hepatotoxic agent*'. These roles are important to describe the activity of pharmacological entities and are imported from ChEBI (**Figure 7.4**). There are three roles that a '*pharmacological entity*' can exhibit in a DDI and that have been manually added to this hierarchy: '*participant*', '*precipitant*', and '*object*'. A DDI involves two participants; the one that leads to the interaction is the precipitant or perpetrator drug, while that one which effect or levels are altered because of the DDI is the object or victim.

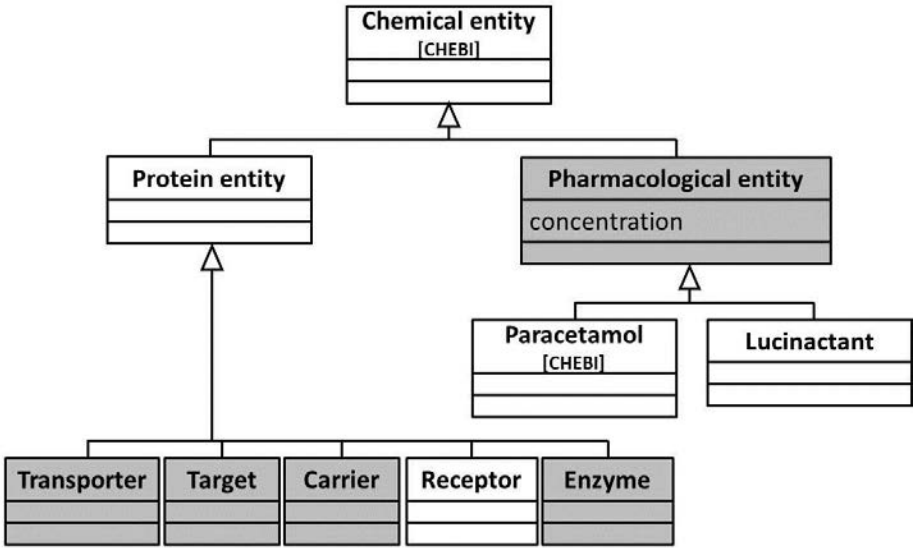

Figure 7.3. Conceptual model representing the hierarchy for chemical entities in DINTO. Classes imported from the ChEBI ontology are labelled as [CHEBI]. Shaded boxes represent classes with additional subclasses.

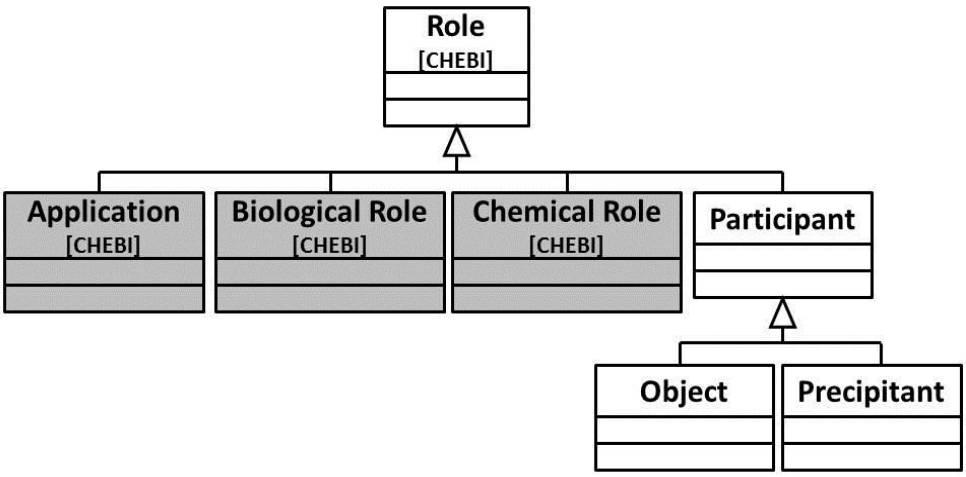

Figure 7.4. Conceptual Model representing the hierarchy for roles in DINTO. Classes imported from the ChEBI ontology are labelled as [CHEBI]. Shaded boxes represent classes with additional subclasses.

From an ontological point of view, a drug such as *paracetamol*, cannot be said "to be" a *precipitant,* for example. *Paracetamol* is a chemical entity that, under certain circumstances, can exhibit a role *precipitant* when it participates in a DDI. Therefore, *paracetamol* is a subclass of the anonymous class shown below:

'pharmacological entity' and *'has role'* some *'precipitant'*

That is, '*paracetamol*' is a '*pharmacological entity'* that fulfils the condition of having a '*has role'* relationship with the class *'precipitant'*. Therefore, a DDI does not have a participant and an object drugs, but a '*pharmacological entity*' with a *'participant'* role and a *'pharmacological entity'* with an *'object'* role, as shown in **Figure 7.5**.

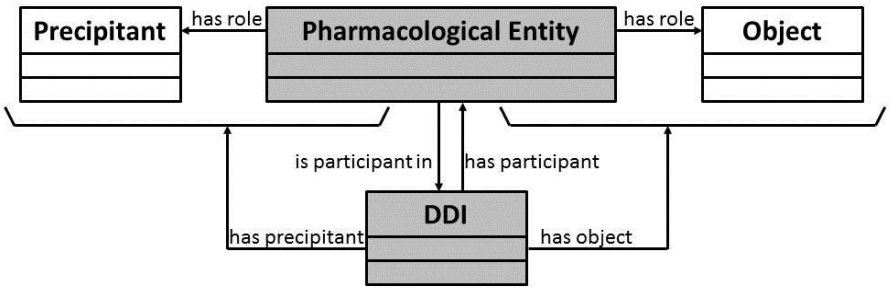

Figure 7.5. Conceptual model representing the roles '*participant*', '*precipitant*', and '*object*' in DINTO. Shaded boxes represent classes with additional subclasses.

- **DDI Mechanism:**

A very important aspect of DDI-related information is how the interaction occurs. This information is crucial in the study, understanding, and management of DDIs. For example, knowing how an interaction occurs enables the prediction of the possible consequences of the DDI (e.g., the increase in the toxicity of a drug due to a decrease in its enzymatic metabolism) or to identify possible therapeutic alternatives (e.g., the selection of a related drug that does not inhibit that specific enzyme).

The process leading to a DDI is the DDI mechanism. DDIs are broadly classified in two main groups by their mechanisms: pharmacokinetic (PK) and pharmacodynamic (PD) DDIs. A PK mechanism occurs when one drug affects some PK process of the other drug leading to an alteration of the levels of the drug in the body. On the other hand, a PD mechanism occurs when one drug alters the effects of another drug without altering its levels in the body, but affecting its final effect or its activity in the target. On a molecular basis, most DDI mechanisms – including both PK and PD mechanisms – are caused by the interactions between the interacting drugs and proteins in the body. For example, two drugs interact when one of them inhibits the activity of the target responsible of the activity of the other drug (PD DDI), or when one drug alters the activity of a protein mediating the absorption of the second one (PK DDI). Therefore, PK or PD mechanisms can be sub-classified by the type of protein that is altered. It is important to note, as well, that most DDIs occur not by a single mechanism, but often by two or more mechanisms

acting in concert (Baxter, 2013). Therefore, an interaction between a unique pair of interacting drugs can be preceded by more than one DDI mechanisms.

This knowledge is represented in the CM shown in **Figure 7.6**. Two classes represent the two types of mechanisms, PK and PD mechanisms, which have different subclasses according to the type of protein involved in the mechanism. At the same time, each one of them is subdivided on the basis of how the precipitant drug – aka perpetrator or the drug that produces the DDI – acts on the protein (*agonism* and *antagonism* for targets and *inhibition*, *induction,* and *saturation* for other proteins).

Although most DDI mechanisms involve a protein, some DDIs do not. A common example is DDIs caused by the physicochemical properties of the drugs and not by the interaction with a protein. This is a PK mechanism occurring by the formation of non-absorbable complexes.

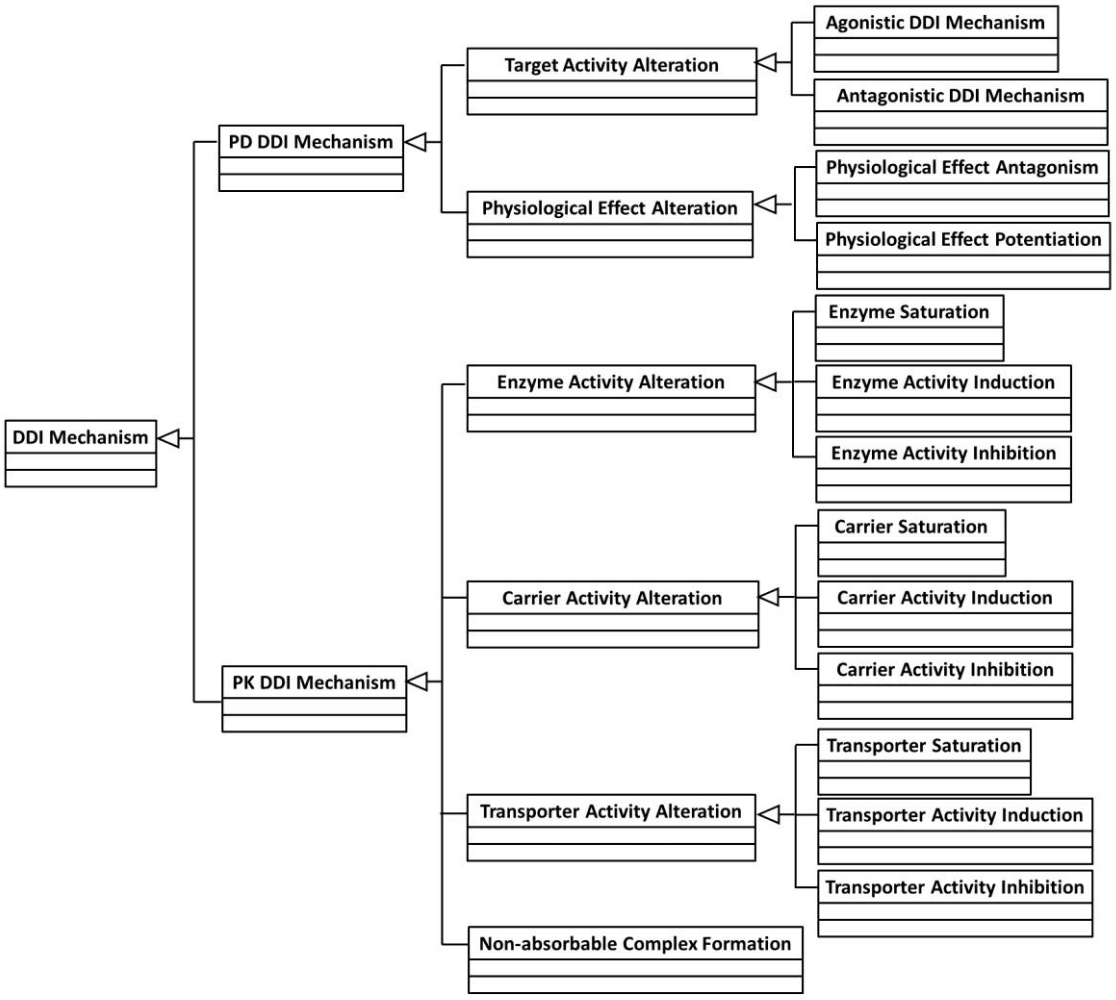

Figure 7.6. Conceptual model representing the hierarchy for DDI mechanisms in DINTO

The DDI mechanism is a process that always precedes the DDI, and the occurrence of the DDI is determined by the existence of at least one DDI mechanism process. Therefore, these two classes are related in DINTO through the inverse relationships '*precedes*' and '*is preceded by*', and formally defined through the '*equivalent to*' (≡) axiom as shown below:

'DDI' ≡ *'is preceded by'* some *'DDI mechanism'*

'DDI mechanism' ≡ *'precedes'* some *'DDI'*

These relationships are used in DINTO to link classes at lower levels of granularity, too. They relate a specific DDI with its specific mechanism – or mechanisms – when this information is available (**Figure 7.7**). This information is translated from the database DrugBank, which provides three possible types of mechanisms for some of the DDIs included in this database: "Possible target-based interaction", "Possible enzyme-based interaction", and "Possible transporter-based interaction". We added a fourth type of mechanism – the previously mentioned '*non-absorbable complex formation*' – extracted from the brief natural language description of the DDI in the database.

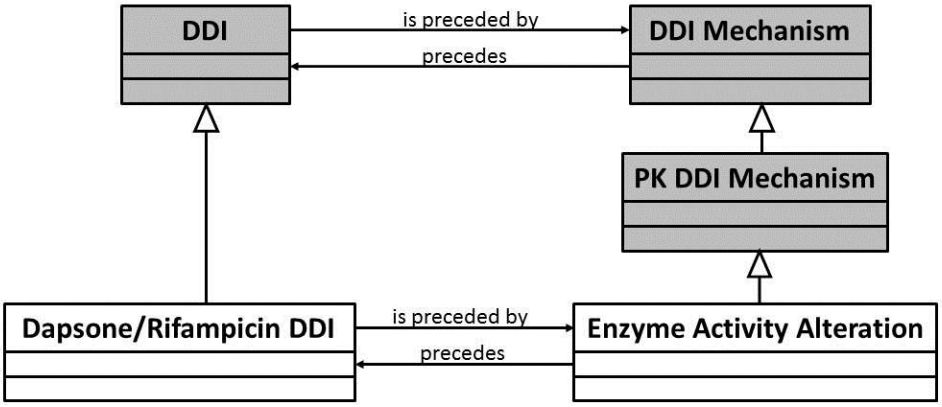

Figure 7.7. Conceptual model representing the relationships between a DDI and a DDI mechanism in DINTO. Shaded boxes represent classes with additional subclasses.

Finally, the last relationship described in this section is the one between the DDI mechanism and the interacting drugs. As mentioned before, a DDI has two different participating drugs. The one leading to the DDI mechanism (precipitant) is the one which triggers the DDI mechanism. These two classes are related through the inverse relationships '*is agent in*' and '*has agent*' (**Figure 7.8**).

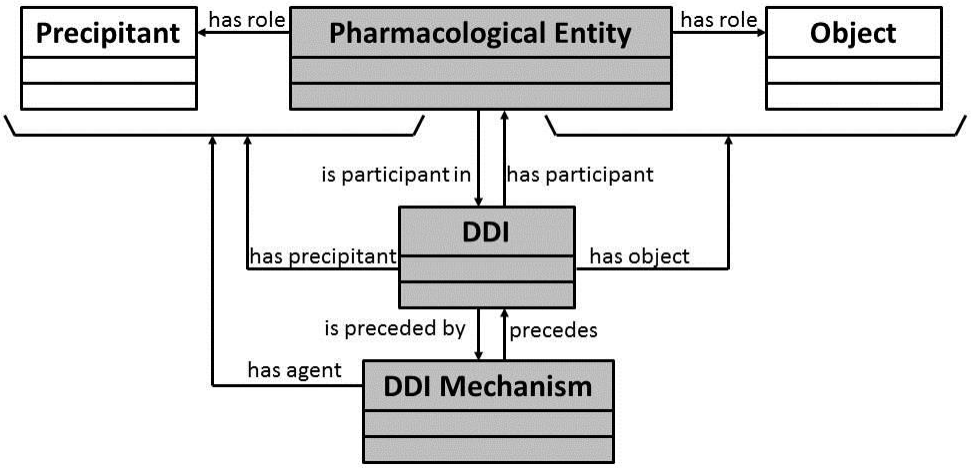

Figure 7.8. Conceptual model representing the relationships between a DDI mechanism and the precipitant drug. Shaded boxes represent classes with additional subclasses.

- **Pharmacokinetic processes:**

The term "pharmacokinetics" refers to what the body does to the drug, while "pharmacodynamics" describes what the drug does to the body (Benet, 1984). There are different PK processes that a drug undergoes in the organism from its administration to its elimination from the body. These processes are absorption, distribution, metabolism, and excretion.

Most drugs – with the exception of those acting locally or those administered intravenously – have to be absorbed from the site of administration in order to reach the bloodstream and its site of action. Absorption requires the passage of drugs across biological membranes. Sometimes this pass is facilitated or impeded by certain proteins, which activity can be altered by other substances, including other drugs, leading to DDIs caused by the alteration of the absorption of a drug.

Distribution is another important PK process by which a drug is transported through the bloodstream – frequently binding plasmatic proteins that act as transporters – and reaches the different tissues. To do this, the drug has to pass across different membranes. For example, only those drugs with certain physicochemical characteristics can cross the blood-brain barrier and reach the central nervous system. Alterations in drug distribution can be a caused of DDIs, too.

This is one of the several mechanisms that the body possesses to protect itself from strange (and possibly dangerous) substances such as drugs. The most important is metabolism. During metabolic processes, different chemical reactions take place mainly in the liver in order to transform drugs in inactive and more soluble substances. Important types of proteins that catalyse these chemical reactions are enzymes, including the *cytochrome P450 family (CYP450)*, which are especially relevant to DDIs. The products derived from metabolic reactions are metabolites. Most times metabolites are inactive or less active than the drug itself. However, other metabolites are toxic and responsible of the harmful effects of the drug. In contrast, some drugs, such as *fosamprenavir*, are administered as inactive compounds. When these substances – known

as prodrugs – are metabolized in the body, the changes in their chemical structures lead to the formation of an active product, which is responsible of the pharmacological effect.

Finally, the drug is eliminated from the body through the excretion process. The most important excretion route is by the kidneys in the urine, although drugs can be also excreted in the faeces, breast milk, sweat, and so forth.

All these PK processes determine the levels of a drug in the body. Therefore, the alteration of one of them can lead to unexpected changes in drug concentration. This is what happens during a PK DDI, which occurs when one drug alters some of the PK processes of the other drug. All the different PK processes are represented in our CM under the homonymous class, which is shown in **Figure 7.9**.

PK processes are related to other concepts in our CM. In this way, the object drug *'undergoes'* a PK process, while the precipitant drug *'alters'* (which can be an increase or a decrease) a PK process. On the other hand, a PK DDI mechanism, as we will see below, *'alters'* a PK process, while a protein *'regulates'* (*'facilitates'* or *'impairs'*) a PK process (**Figure 7.10**).

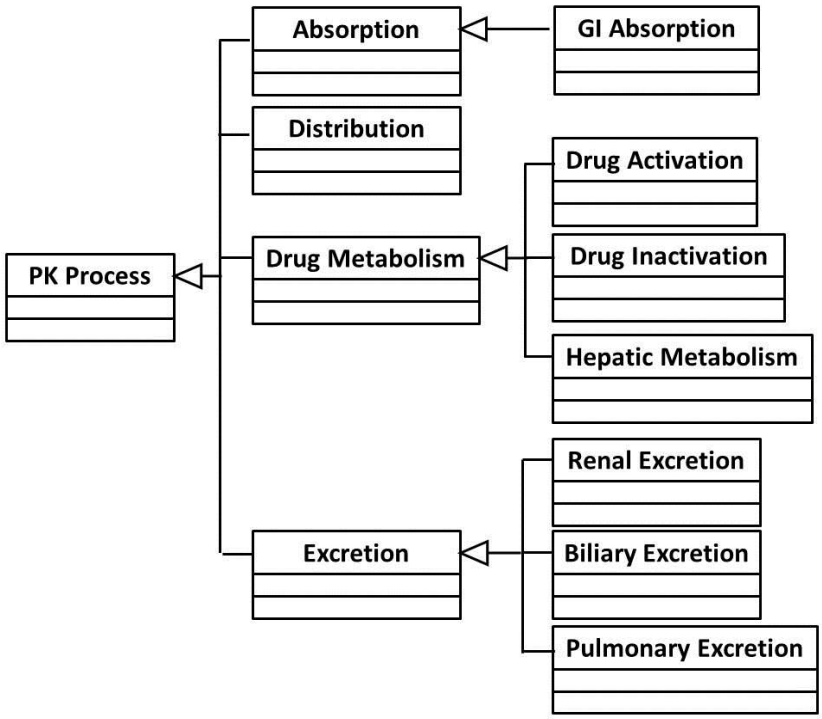

Figure 7.9. Conceptual model representing the hierarchy for PK processes in DINTO

- **Pharmacokinetic parameters:**

PK processes, as well as the concentration of the drug in the body, are quantitatively defined by PK parameters. A fundamental understanding of these parameters is required to design appropriate drug regimens for patients. In fact, the effectiveness or toxicity of a dosage regimen is determined by the concentration of the drug in the body.

PK DDIs produce an alteration of a PK process and, as a consequence, the related PK parameters are altered, too. During the linguistic analysis of the corpus we identified that sentences reporting a PK DDI usually describe an alteration in a PK parameter (e.g., "*sertraline increased the AUC and the Cmax of a single dose of pimozide by about 40%*"). Therefore, the representation of PK parameters is important in DINTO.

The Pharmacokinetics ontology (PKO) (Wu et al., 2013) represents drug PK information, including a manually created representation of PK parameters and their definitions in natural language. We create a subset of PKO for these concepts, which is imported in DINTO under the top-level class *'pharmacokinetic parameter'*.[29]

As mentioned before, PK parameters quantify PK processes. Therefore, the classes *'pharmacokinetic process'* and *'pharmacokinetic parameter'* are related by the inverse relationships '*has quality*' and '*is quality of*' (**Figure 7.11**).

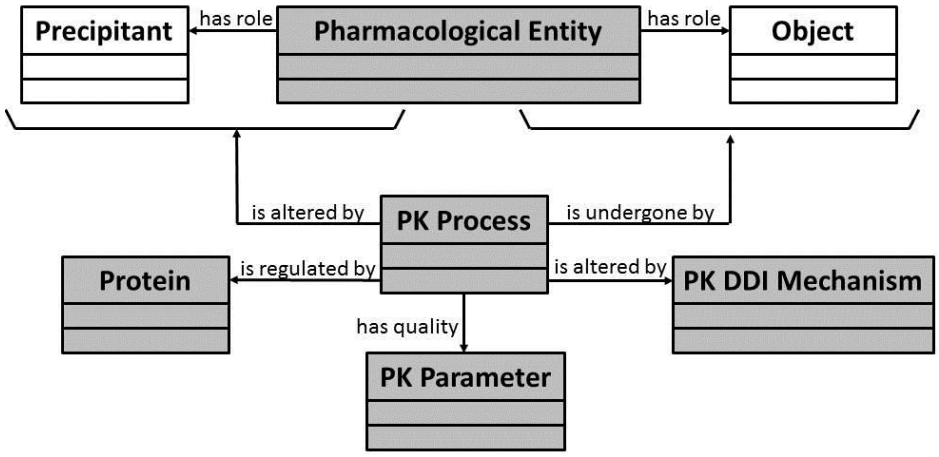

Figure 7.10. Conceptual model representing the relationships between a PK process and other concepts in DINTO. Shaded boxes represent classes with additional subclasses.

- **Physiological effect and DDI effect:**

The effect produced by a drug is always an alteration in a physiological function or process that maintains the existence of the living organism. Drugs may increase or decrease the normal function of tissues or organs, but they do not confer any new functions on them (Levine et al., 2005). For example, the drug *glibenclamide*, which is used to control blood sugar levels in diabetic patients, produces its effect by increasing the normal production of *insulin* (the hormone responsible of regulating the amount of glucose in the blood) by the organism. Therefore, the effect produced by *glibenclamide* is an alteration of the physiological production of *insulin*.

[29] PKO_DINTO_subset.owl downloadable from https://code.google.com/p/dinto/

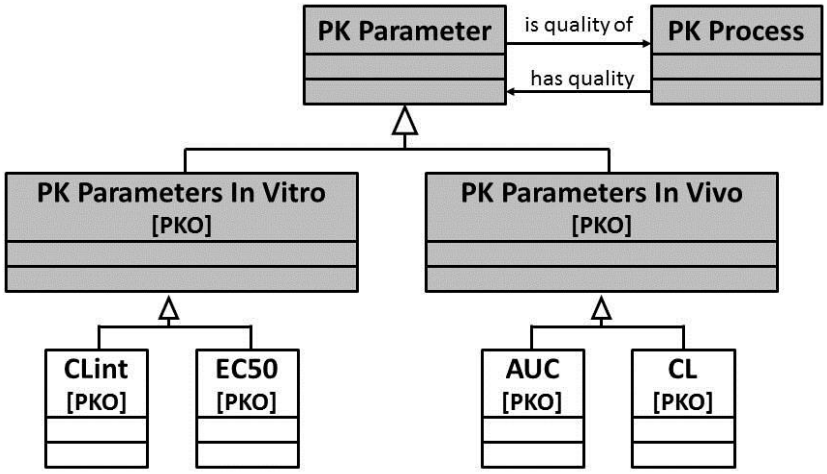

Figure 7.11. Conceptual model representing the hierarchy for PK parameters and their relationship with PK processes in DINTO. Classes imported from the PKO are labelled as [PKO]. Shaded boxes represent classes with additional subclasses.

This concept is represented in DINTO through the class *'physiological effect'*. Drugs can produce different effects that are classified as therapeutic (or the intended effect of the drug), adverse (undesirable effects of the drug), and toxic effects (an exaggeration of the desired therapeutic effect which appears at high doses). These three possible drug effects are represented in DINTO as subclasses of *'physiological effect'*.

In the same way that drugs do not produce any "new" function in the body, the consequence of a DDI is never a "new" effect. The effect of a DDI is always the alteration of some of the effects of one or both interacting drugs. Therefore, a *'DDI effect'* is a physiological effect that has been altered (increased or decreased) because of a DDI.

During a DDI, the adverse effect of one or both interacting drugs may be potentiated or increased. For example, concomitant administration of several *central nervous system depressors* can increase their adverse effects and produce an excessive depressive response. In other cases, the increase in the concentration of a drug or its metabolites can produce an exacerbation of its toxic effects. This occurs, for example, when the concentration of a toxic metabolite of *paracetamol* is increased as a consequence of its interaction with *rifampicin,* leading to an exacerbation of *paracetamol's* hepatotoxic effects (Baxter, 2013). In contrast, a decrease in the concentration of a drug can lead to therapeutic failure. For example, the administration of *rifampicin* with the inmunosupressor *cyclosporin,* used to avoid rejection in transplanted patients, produces a decrease in the concentration of *cyclosporin* and a decrease in the therapeutic effect, and the treatment will not be effective (Naqvi, 2000).

These situations (increase in adverse and toxic effects or decrease in therapeutic effect) lead to DDIs that are potentially harmful for patients. However, other DDIs can have a beneficial effect. For example, *ritonavir* is administered in conjunction with protease inhibitors used in the treatment of patients with HIV since *ritonavir* increases the absorption of these drugs in the body and allows to reduce the required dose and the dosing schedule (Edwards, 2012). As well, it will be a beneficial DDI those situations in

which one drug, such as *naloxone,* decreases the toxic effect of the other interacting drug, such as *morphine*, being an important treatment in cases of overdose (Sweetman & Martindale, 2006).

Both harmful and beneficial DDI effects can have an important impact on patients and therefore are relevant from a clinical point of view (*'clinically relevant DDI effect'*). However, some DDI effects are unnoticeable or unobservable in most patients, and experienced physicians accommodate the effects (such as rises or falls in serum drug levels) without consciously recognizing that it was the result of an interaction (Baxter, 2013). These DDI effects are said to have no relevance from a clinical point of view (*'non-clinically relevant DDI effect'*).

The CMs representing physiological and DDI effects are shown in **Figure 7.12** and **Figure 7.13**.

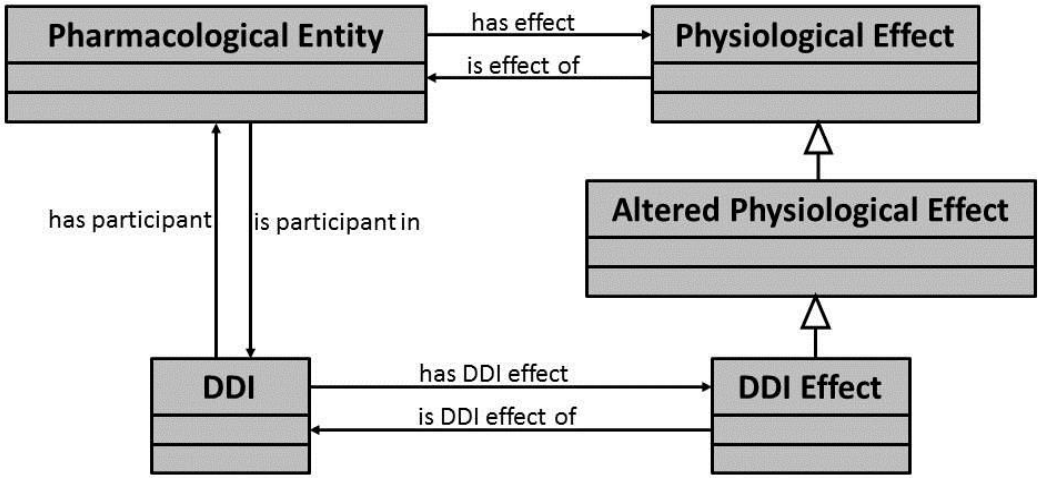

Figure 7.12. Conceptual model representing the relationships between a physiological effect, a DDI effect and other concepts in DINTO. Shaded boxes represent classes with additional subclasses.

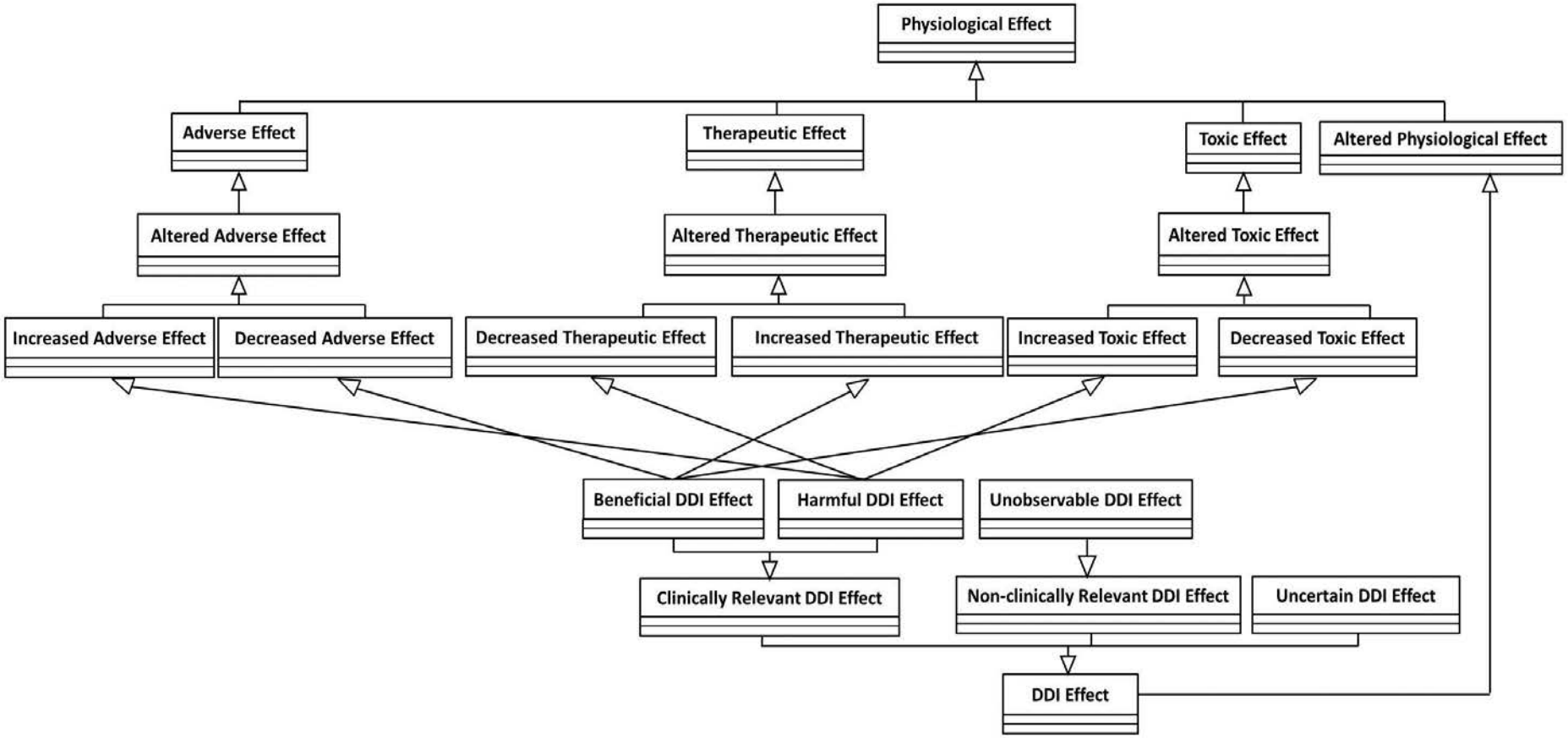

Figure 7.13. Conceptual model representing the hierarchies for physiological and DDI effects in DINTO

- **Drug-drug interactions:**

A drug interaction occurs when the levels or effects of one drug are altered by another substance, such as environmental substances, food and drinks, herbal medicines or other drugs. Interactions between drugs are called drug-drug interactions (DDIs). Therefore, a DDI is a process that occurs between two interacting drugs or participants. The one producing the interaction is called the precipitant or perpetrator, while the other one is called the object or victim.

On the one hand, a DDI can be classified by its mechanism as *'pharmacokinetic DDI'* or *'pharmacodynamic DDI'*. Both of them are classified in our CM, in the same way that the preceding mechanisms, based on the involved protein and the action performed by the precipitant drug (**Figure 7.14**). On the other hand, DDIs can be classified, as well, on the basis of their effects. Thus, we represent the classes '*clinically relevant DDI*', '*non-clinically relevant DDI*' or '*uncertain DDI*'.

Specific interactions between two individual drugs have been imported automatically from the database DrugBank and represented in two different ways in our ontology. On the one hand, there are classes representing the processes that occur when two specific drugs interact. For example, the interaction between the pharmacological entities '*desvenlafaxine'* and *'amitryptyline'* is represented as the class *'desvenlafaxine/amitriptyline DDI',* which is formally defined by having them as its participating drugs:

'desvenlafaxine-amitryptyline DDI' ≡ *('has participant'* some *'amitriptyline')* and *('has participant'* some *'desvenlafaxine')*

In other words, any individual of the '*desvenlafaxine-amitryptyline DDI*' class is equivalent to (≡) any individual that has a '*has participant*' relationship with at least one (*some*) individual of the class '*amitriptyline*' and, at the same time, has a '*has participant*' relationship with at least one (*some*) individual of the class '*desvenlafaxine*'.

On the other hand, both classes *'desvenlafaxine'* and *'amitryptyline'* are related by the symmetric relationships *'may interact with'*, which represents the possible interaction of *'amitryptyline'* with *'desvenlafaxine'* and vice versa.

Since the DDI mechanisms for the imported DDIs have been also translated from DrugBank, the use of a reasoner engine (Cuenca, 2011) enables their classification on the basis of the underlying mechanisms (see **Section 9.2.1**).

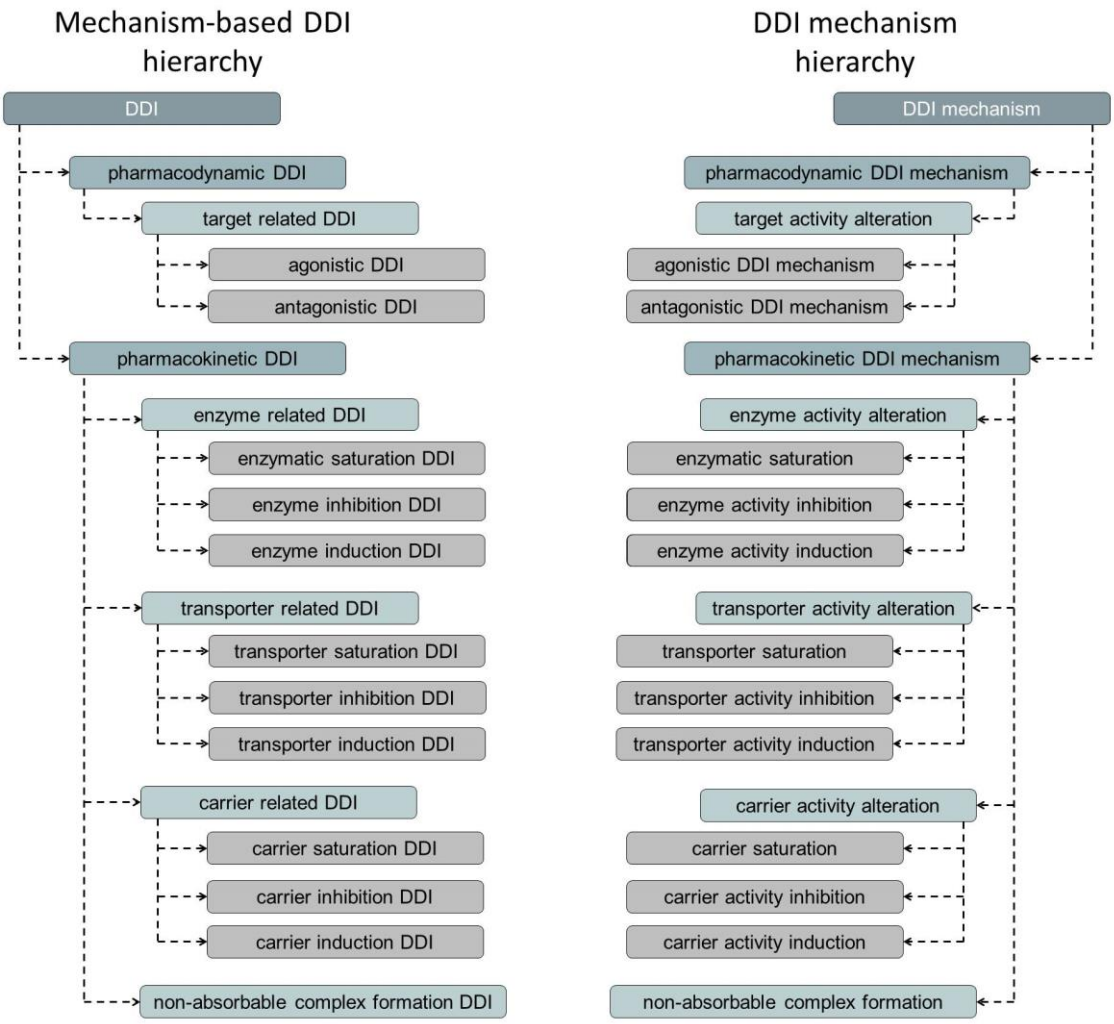

Figure 7.14. Hierarchies for DDIs and their related mechanisms in DINTO

- **<u>Drug-protein relationships:</u>**

The role of proteins in our bodies is essential. They are constituents of our cells, trigger physiological responses, and catalyse chemical reactions that maintain our normal physiological functions. Proteins are, furthermore, determinants for the activity of drugs. They influence the different PK processes of a drug in the body, and trigger the effects (including therapeutic, toxic, and adverse effects) of most drugs.

The type of interaction between a drug and a protein is determined by the type of protein (receptor, enzyme, transporter, etc.) and the activity that the drug produces on that protein. For example, a drug might bind a receptor without leading to any response. However, this union would prevent other substances – i.e., normal substances originated in our body and externally administered substances such as other drugs – from binding to that protein. In this case, the drug '*blocks*' or '*antagonizes*' the protein. In contrast, when a drug decreases the activity of an enzyme (a protein which catalyses chemical reactions in the organism), the drug '*inhibits*' the activity of the enzyme.

Therefore, in our CM we have included different possible relationships between proteins and pharmacological entities. They have been adapted from the database DrugBank, where the relations between drugs and proteins are described as semi-structured text. With our approach, we can translate in further steps the drug-protein information from DrugBank into DINTO (see **Section 7.2.2** and **Annex 6** for a detailed description of these and other relationships).

- **Description of a DDI**

There are different aspects characterizing a DDI that appear frequently in texts (Tatro, 2010). They are represented as attributes in our CM, as shown in **Figure 7.16**.

Firstly, '*documentation*' is an important factor to determine the significance of a DDI (Tatro, 2010), since it is related to the degree of confidence in the causal association between a DDI and an altered clinical response. That is, the more the DDI is documented, the more confidence exists in its causal relation with the effect. The DDI compendium '*Drug Interaction Facts*' (Tatro, 2010) establishes five types of documentation levels for a given DDI (*established, probable, suspected, possible, and unlikely*) in a way that a well-known and documented DDI described in different sources is considered to be *established,* while a *possible* or *unlikely* DDI would be that one that has insufficient evidence supporting the existence of a clinically relevant interaction. The level of documentation of a DDI is represented in DINTO as the data property *'documentation level',* which can adopt one of the five mentioned levels.

Another two important aspects are the '*incidence*' of a DDI - or the relative frequency of occurring of that interaction −, and its '*onset*' − or how rapidly the clinical effects of a DDI can occur, which determines the urgency with which preventive measures should be instituted to avoid the consequences of the interaction. There are two levels of onset in our CM: '*rapid'* and '*delayed'*.

There are another two concepts frequently used to describe a DDI. On the one hand, the potential '*severity*' of the interaction is particularly important in assessing the risk versus benefit of therapeutic alternatives. There are three degrees of severity: '*major'*, '*moderate',* and '*minor'*. On the other hand, the concept '*relevance'* is used, from a clinical point of view, to describe the real importance that the DDI has in the clinical practice. It can adopt two values: '*clinical relevance'* or '*non-clinical relevance'*.

Figure 7.15 shows these five attributes, along with other four attributes describing not a DDI, but the strategy to manage or avoid it. They are different types of recommendation related to a change in one of the interacting drugs, monitoring of the patients, change in the dose of one or both drugs, or performance of some clinical test. This model has been adapted from Samwald et al. (2013), where they show a descriptive RDF model for representing pharmacogenomics statements in SPCs. Specifically, the model represents the different recommendations for those situations when pharmacogenomics might influence the levels and effects of some drug treatments. We have studied the model and decided that it could be extrapolated to the domain of DDIs. Therefore, it has been included in our CM.

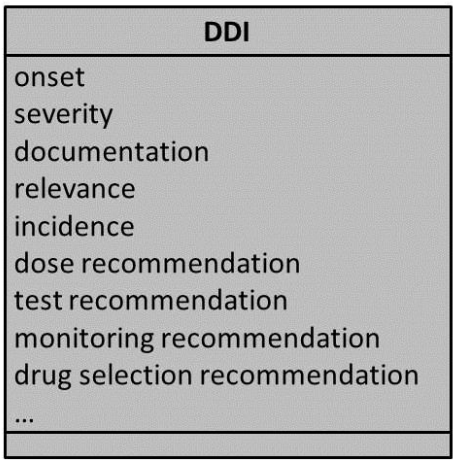

Figure 7.15. Conceptual model representing the attributes of a DDI and the possible recommendations to avoid it in DINTO. Shaded colour represents that this class has additional subclasses.

- **Study subject**

An important aspect used when describing a DDI and the study or situation where it has been observed is the type of subject. We include this information in our CM, as it is shown in **Figure 7.16**. It is related to the information resource through the relationship *'has study subject'* and has three attributes: '*age', 'gender',* and '*race or ethnic'*.

- **DDI Information Resources:**

A DDI can be described in different sources, such as scientific publications, databases, DDI compendia, individual case reports, and so forth. Therefore, they are important concepts that should be included in our ontology. The Biomedical Resource Ontology (BRO) (Tenenbaum et al., 2011) is a controlled terminology of resources used by the biomedical research community created to improve the sensitivity and specificity of web searches. This ontology represents different information resources in a single hierarchy, providing natural language definitions for each one of them. We created a subset of BRO to be reused in DINTO consisting in 52 classes.[30] Although BRO does not conform to the naming conventions followed in the construction of DINTO, we decide not to modify the original labels provided by the authors (see **Annex 10**).

During the linguistic analysis of the DDI corpus, however, we have identified other terms used to describe DDI information resources and not included in BRO. We observed that the most frequent type of information resource cited in the texts is different types of studies (e.g., *'animal study'*, *'in vitro'*, *'in vivo'*, *'controlled study'*, etc.). They have been included in the CM as subclasses of the BRO's class *'Clinical_Research_Data'*.

[30] DINTO_BRO_subset.owl downloadable from https://code.google.com/p/dinto/

An information resource can be characterized by the type of subjects participating in the study and the number of them. This information is represented through the relationship between a *'study subject'* and the *'information resource'* and by the attribute '*subject number'*, as it is shown in **Figure 7.16**. Here, we can see as well that a DDI and an information resource are related by the inverse relationships '*is described in*' and '*describes'*.

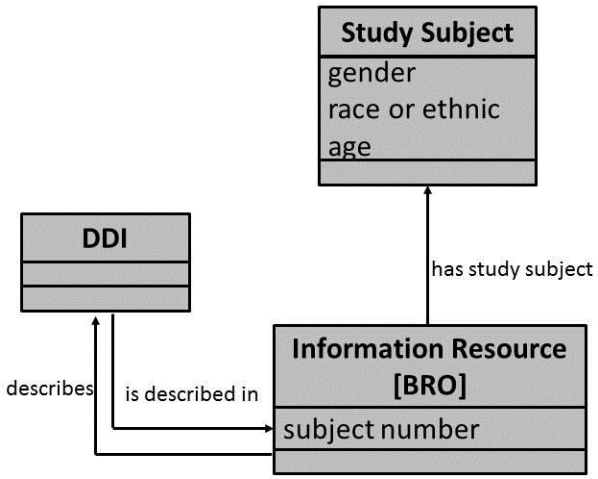

Figure 7.16. Conceptual model representing the hierarchy for information resources and study subjects, and their relationships with a DDI in DINTO. Classes imported from the BRO are labelled as [BRO]. Shaded boxes represent classes with additional subclasses.

- **Description of the interacting drugs**

We finish the description of the conceptualization of DINTO with a very important aspect: how the interacting drugs should be represented in an ontology. In the state of the art in DDI modeling efforts (**Chapter 6**), we have observed that all reviewed CMs represented the interacting drugs at the "active ingredient" level. Therefore, all previous efforts agreed to consider that a DDI should be described to occur between two active ingredients.

However, a recent approach aimed to develop a foundational representation of evidence for DDIs (called potential DDIs or PDDIs), postulated that this representation of DDI knowledge is not correct (Brochhausen et al., 2014). These authors stated that active ingredients should not be assigned the status of drugs, because the excipients, route of administration, and dose are determinants for their actions, and that therefore DDIs should be described at the "clinical drug" level. As we have explained in **Section 5.1**, active ingredients are administered as clinical drugs, that is, as the combination of the active ingredient with pharmacological inactive substances (excipients) in a specific pharmaceutical form (e.g., tablet, solution, etc.) and with a specific strength or dose. Such a representation avoids the extrapolation of DDIs described at the active ingredient level to pharmaceutical forms and administration routes, such as interactions with topical drugs, which are commonly non-clinically relevant due to the scarce quantity of active ingredient absorbed. This would be especially important for clinical oriented applications,

such as the creation of CDSS, where a high frequency of irrelevant alerts leads to alert fatigue and clinicians getting overloaded with alerts (Bryant, Fletcher, & Payne, 2014).

However, this new approach possesses some limitations, too. Firstly, it is not an accurate representation of the pharmacological domain, since an interaction between two drugs occurs at the molecular level, triggered by the molecular structure and/or mechanism of action of the active substance. Although the characteristics of the clinical drug, such as strength or administration route, can influence this process, it is ultimately related to the active ingredient level.

The representation at the clinical drug level relies on the fact that a single molecule of a drug cannot bear by itself a drug role, neither to interact with another drug. This premise is correct, since the effect leading to the interaction requires the action of many molecules acting in concert in the body. However, the contrary point of view, attributing the interaction to the clinical drug, could be considered incorrect in a similar way. Some DDIs do not occur subsequently to the single administration of one clinical drug, but during a treatment, which implies a posology with repeated administrations of the clinical drug. Therefore, the effect leading to an interaction might require, in these cases, the administration of several units of a clinical drug. Thus, both approaches fail in the representation of this pharmacological knowledge in a similar fashion.

Figure 7.17 shows an example of how some DDIs rely on the dosage of the interacting drugs, being therefore related to a repetitive administration of the clinical drugs. In this simplified situation, a patient is in treatment with DrugA, which is metabolized by enzyme E. A new drug DrugB is prescribed, which should be administered daily. Since DrugB is an inhibitor of enzyme E, the administration of both drugs will lead to a DDI. However, when the patient takes the first tablet of DrugB, the doses reached in the body are not high enough to produce a potent inhibition of the enzyme. Therefore, the levels of DrugA are not altered in the body. The administration of a second dose of DrugB leads to a more prolonged inhibition of E, and to an alteration in the metabolism of DrugA, leading to an increase of its levels in the body. However, the concentrations are not raised high enough to produce a clinical manifestation. With a third administration of DrugB, the concentrations of DrugA rises even more. At this moment, a laboratory test might identify the abnormal concentration. However, signs or symptoms do not appear yet. Finally, with a fourth administration, the clinical manifestations of an adverse DDI arise. Therefore, although it is possible that the clinical manifestations of a DDI appear after a single administration, sometimes they are not apparent until the repeated administration of the clinical drug (Hojo, Echizenya, Ohkubo, & Shimizu, 2011; Zaccara et al., 1993).

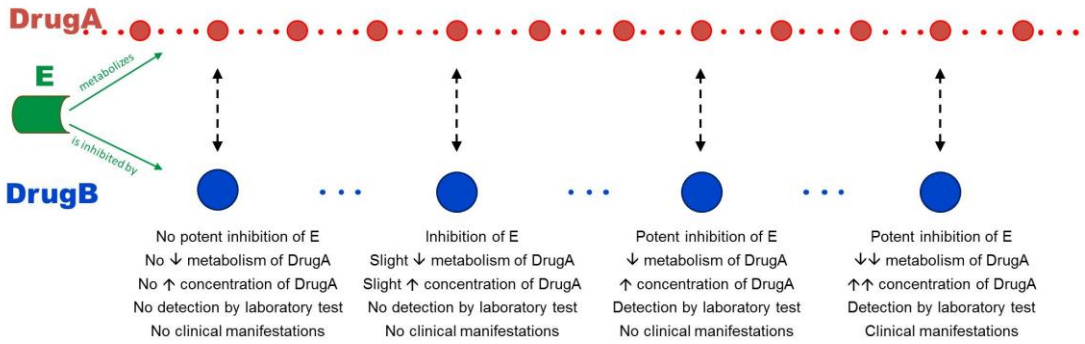

Figure 7.17. Example of influence of repetitive administrations of a clinical drug in a DDI

In addition to this, other aspects must be taken into account to determine the correct representation of the interacting drugs. Firstly, the representation of DDIs in our ontology must allow for the representation of their mechanisms on a molecular basis, as well. This requires the representation of drug-protein relationships, which occur between proteins and drug molecules. As we have observed based on our review on the state of the art, this type of relationship is always represented at the active ingredient level.

Secondly, during the analysis of the DDI corpus, we have observed that descriptions of DDIs in text can be made at different levels of granularity. General texts usually describe the interaction at the active ingredient (or brand) level, without providing information about the drug strength or dose (sentence (i)). In contrast, clinical studies or case reports usually refer to the specific clinical drugs suspected to interact (sentence (ii)).

> « The occurrence of stupor, muscular rigidity, severe agitation, and elevated temperature has been reported in some patients receiving the combination of **selegiline** and **meperidine**. »[31] (i)
>
> « A patient taking **iproniazid 150 mg daily** developed severe orthostatic hypotension on two occasions within an hour of taking **selegiline 5 mg**. »[32] (ii)

From all these facts, we conclude that the most appropriate representation of interacting drugs in our ontology is at the active ingredient level. Nevertheless, information regarding the strength, pharmaceutical form, and dose of the interacting drugs is included, too. It is represented as attributes in our CM. However, as we have explained, these attributes cannot be linked with any particular drug, but with a drug that has been described to interact with another one in a specific study. Sentences (iii) and (iv) show how two different studies describe an interaction between the same interacting drugs, *carbamazepine* and *simvastatine*, but with different doses.

[31] Sentence extracted from the package insert o ELDEPRYL®, retrieved from the database DailyMed (http://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?setid=106429ad-859a-4b29-babf-42cb85f7236e)
[32] Sentence extracted from the DDI information compedia Stockley's Drug Interactions (Baxter, 2013)

« Carbamazepine reduced the AUC of simvastatin in twelve patients taking a **single oral dose** of **simvastatin 80 mg** and **a daily tablet** of **carbamazepine 600 mg**. » (iii)

« Carbamazepine reduced the AUC of simvastatin in twelve patients taking **a single oral dose of simvastatin 40 mg** and **a daily tablet** of **carbamazepine 300 mg**. » (iv)

From the above examples, we could assert that an interaction occurs with *simvastatin 40 mg* and *simvastatin 80mg*, but we could not confirm or deny that the interaction might occur with *simvastatin 10mg* or *simvastatin 90 mg*, too. Indeed, these sentences are describing only an observation made in a clinical study or case report. Therefore, information regarding dose, administration route, and pharmaceutical form is related to a pharmacological entity that has been observed to participate in a DDI and is described in an information resource. Therefore, in this ontology, characteristics such as '*pharmaceutical form*', '*administration route*', or '*dose*' are attributes of the anonymous class '*pharmacological entity*' that is participant in a '*DDI*' that is described in some '*information resource*'.

'*pharmacological entity*' and ('*is participant in*'
some ('*DDI*' and ('*is described in*'
some '*information resource*')))

7.1.4 Implementation

During the implementation activity, the intermediate CMs need to be implemented in a formal language. The OBO Foundry (Smith et al., 2007), a collaborative effort for the development and maintenance of biomedical ontologies, recommends the use of a common shared syntax in order to ensure the integration and interoperability among different ontologies. Specifically, it is recommended that ontology members are expressed in either the Open Biomedical Ontology (OBO) format or the Web Ontology Language (OWL). The OBO format was created for the Gene Ontology and supports most of the biomedical ontology content (Tirmizi et al., 2011), while OWL is the World Wide Web Consortium (W3C) standard ontology language for the semantic web.[33] The OBO flat format is human-friendly and is supported by OBO-Edit (Day-Richter, Harris, Haendel, & Lewis, 2007), an ontology editing tool. However, OBO-Edit does not support importing fragments of other ontologies, which would require the use of different strategies in order to reuse ontologies in our project. In contrast, OWL syntax, although understandable by humans, can be difficult for non-computer scientist experts. However, it provides a higher level of expressive power and better inference capabilities than OBO. Moreover, the widely used OWL ontology editor Protégé[34] supports importing fragments or whole ontologies, and provides a large number of new plugins trying to make easier the building of ontologies in OWL.

[33] http://www.w3.org/2001/sw/wiki/OWL
[34] http://protege.stanford.edu/

For all these reasons, we have selected as implementation language the Web Ontology Language OWL 2[35] and Protégé 4.3 as the ontology editing software (**Figure 7.18**)

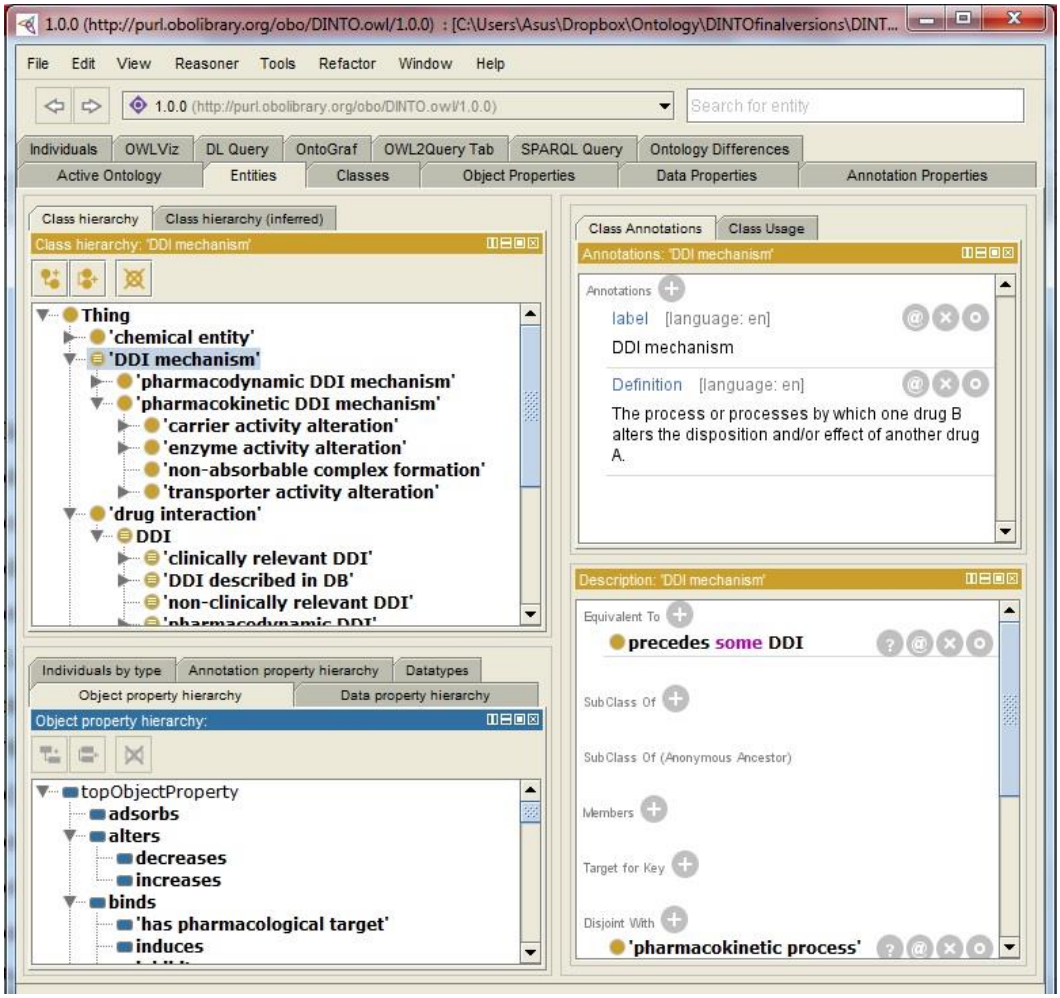

Figure 7.18. Screenshot of DINTO in the ontology development environment Protégé

7.1.5 Ontological resources reuse

We have identified and reused those ontologies that have currently represented aspects of the DDI domain required by our ontology, and imported them keeping their original Unique Resource Identifiers (URIs), identifiers, and labels. The imported fragments are described below.

1. The Chemical Entities of Biological Interest ontology (Degtyarenko et al., 2008) (Original URI: http://purl.obolibrary.org/obo/CHEBI_11111)

[35] http://www.w3.org/TR/owl2-overview/

The ChEBI ontology organizes "small" chemical compounds with a relevant role in the biomedical domain, including pharmacological substances. Each one of these pharmacological substances is related to at least one drug application, such as '*analgesic'* or '*antiemetic'*, which are subclasses of the top level class '*role'*. Specifically, we have imported the following information:

- *'Pharmacological entity' classes*: From the total number of chemical entities included in ChEBI[36], we have selected the 3,026 having some drug-related role − that is, all those chemical entities bearing some pharmacological activity. We import the classes and their metadata, which is represented in OWL as annotation properties: InChI, InChIKey, synonyms, definition, and cross references (3,027 imported classes).
- *'Role' classes:* We import the top-level class '*role'* and all its subclasses. Therefore, the ontology includes not only the '*drug'* role, but also other possible '*application'* roles, such as '*biological'* and '*chemical'* roles (903 imported classes).
- *'has role' relationships:* We import all the relationships between the imported *'pharmacological entity'* classes and the corresponding '*role'* classes (7.845 '*has role'* relationships).

2. The Pharmacokinetics Ontology (PKO) (Wu et al., 2013) (Original URI: http://www.owl-ontologies.com/2009/11/5/PKO.owl#label)

The PKO has been used as a source for terminology, definitions, and units in this area. Imported information is described below:

- *'Pharmacokinetic parameter' classes*: We identify and select those classes in the ontology referring to PK parameters. We also import the annotation properties included by the original authors, which provide units for the PK parameters, definitions and, in some cases, references. This process is carried out manually and we generate a fragment of the ontology that should be imported and merged in DINTO later (77 imported classes)[37].

3. The Biomedical Resource Ontology (BRO) (Tenenbaum et al., 2011) (Original URI: http://biontology.org/ontologies/BiomedicalResourceOntology.owl#label)

Although BRO is not specifically related to the DDI domain, it is a useful representation of different resources used by the biomedical research community to conduct research. We have imported into DINTO the following information:

- *'Information Resources' classes*: Instead of creating new classes to represent those possible information sources where DDIs can be described, we manually create a fragment of BRO including that information relevant to our

[36] The latest version available at the moment of building DINTO is ChEBI realease 112.
[37] PKO_DINTO_subset.owl downloadable from https://code.google.com/p/dinto/

domain, which has been imported and merged in DINTO later (50 imported classes)[38].

4. *OBO Ontology Metadata*[39] (Original URI: http://www.w3.org/2002/07/owl#label and http://purl.obolibrary.org/obo/IAO_1111111).

The OBO ontology metadata project standardises annotation properties for common annotation types such as definition, synonym, and so on. We import this metadata ontology (a subset of the IAO),[40] which provides 38 annotation properties.

5. *Basic Formal Ontology (BFO)*[41] (Original URI: http://ifomis.org/bfo/1.1#label).

We follow BFO's definitions for upper level classes such as *'process'* and *'continuant'* in our work. The complete ontology is imported in the corresponding version of DINTO[42] (39 imported classes).

To increase the interoperability of our ontology, we have mapped classes and attributes in DINTO with other relevant ontologies. This mapping is represented in the ontology through the annotation property '*maps to*'. The most important mapped ontologies are:

- *Relation Ontology (RO)* (Smith et al., 2005): RO is a collection of relations intended primarily for standardization across ontologies in the OBO Foundry and wider OBO library. We map our relations to RO where appropriate.

- *Semanticscience Integrated Ontology (SIO)* (Dumontier et al., 2014): SIO is an ontology to facilitate biomedical knowledge discovery that features a simple upper level comprised of essential types and relations for the rich description of objects, processes, and their attributes. We map our relations to SIO where no appropriate RO mapping is available.

- *Evidence taxonomy for DIKB* (Boyce et al., 2009): We map our *'has object'* and *'has precipitant'* relationships to their equivalents in this taxonomy.

7.1.6 Non-ontological resources reuse

Up to now, we have described the conceptualization and implementation of the basic skeleton of our ontology and how it has been enriched by importing information from other ontologies. However, information included in this version is still limited in terms of its applicability to IE and inference of DDIs.

[38] BRO_DINTO_subset.owl downloadable from https://code.google.com/p/dinto/
[39] https://code.google.com/p/information-artifact-ontology/wiki/OntologyMetadata
[40] http://code.google.com/p/information-artifact-ontology/
[41] http://www.ifomis.org/bfo
[42] DINTO_1BFO.owl downloadable from https://code.google.com/p/dinto/

Firstly, the number of pharmacological entities included in the ontology is limited in comparison with all those possible entities manually annotated in the DDI corpus (see **Section 10.1**). Secondly, although general mechanisms of DDIs have been represented, specific DDIs between pairs of drugs have not been included. Moreover, relationships between specific drugs and proteins for the indirect representation and inference of DDIs are not represented, either. Finally, the number of database and code identifiers for drugs − which are necessary for the correct mapping and merging of other information sources and our ontology − does not show enough coverage.

To overcome all these limitations, we have integrated information from the pharmacological database DrugBank (Wishart et al., 2006), a rich resource combining chemical and pharmaceutical information of approximately 4,900 pharmacological substances. It covers identification, taxonomical, and interactions aspects, among many others. We use the version available at the time of developing the ontology DrugBank 3.0 as a source of information about drugs, their different code and database IDs, and synonyms. We also import specific DDIs between pairs of drugs and interactions between drugs and proteins. The activities performed in this process are described below.

1. Exact mapping between ChEBI and DrugBank

As mentioned before, 3,026 *'pharmacological entity'* classes have been imported in DINTO from the ChEBI ontology. Integration of additional information for these classes requires the unambiguous identification of the same concept in the two sources. However, this mapping is not always easy because, among other reasons, two resources can represent concepts at different levels of granularity (Brenninkmeijer et al., 2013).

The most characteristic example is the representation of drug salts. On the one hand, ChEBI considers that a drug and its different salts are not the same concept. For example, in ChEBI the drug '*lidocaine'* and its two salts '*lidocaine hydrochloride*' and '*lidocaine hydrochloride monohydrate'* are considered three different chemical entities and each one of them has its own ChEBI unique identifier. On the other hand, in DrugBank there is only one entry corresponding to the drug '*lidocaine*', which describes the drug '*lidocaine hydrochloride*' as its salt, while its relationship with '*lidocaine hydrochloride monohydrate*' is not specified. However, cross-references in ChEBI to DrugBank establish that the three of them ('*lidocaine*', '*lidocaine hydrochloride*', and '*lidocaine hydrochloride monohydrate*') link with the same DrugBank ID entry, which corresponds to the entry '*lidocaine*'. We consider that the extrapolation of DDIs described for '*lidocaine*' in DrugBank to its different salts '*lidocaine hydrochloride*' and '*lidocaine hydrochloride monohydrate*' might lead to the inclusion of unknown or unproven interactions and, in consequence, to a source of error in the ontology.

To avoid this possible source of errors, we have adopted a strategy of exact mapping between the two sources, which establishes that a drug will be considered to be the same in both sources if the cross-referenced IDs are coincident in both ChEBI and DrugBank (**Figure 7.19**). According to this, in the previous example there is only one mapping between '*lidocaine*' in ChEBI and '*lidocaine*' in DrugBank. Therefore, information from DrugBank is imported only for the class '*lidocaine*'. Meanwhile, the classes '*lidocaine hydrochloride*' and '*lidocaine hydrochloride monohydrate*' are still present in the final ontology, but without incorporating any additional information.

2. **Creation of new drug classes:**

The coverage for drugs in ChEBI is limited in comparison to those drug entities annotated in the DDI corpus (see **Section 10.1**). Therefore, it is necessary to include an additional source of drugs. Some drug entries in DrugBank are not represented in ChEBI. Therefore, in addition to the 3,026 drug classes imported from ChEBI, we create 5,760 new classes for those drugs present in the database but not included in the ontology.

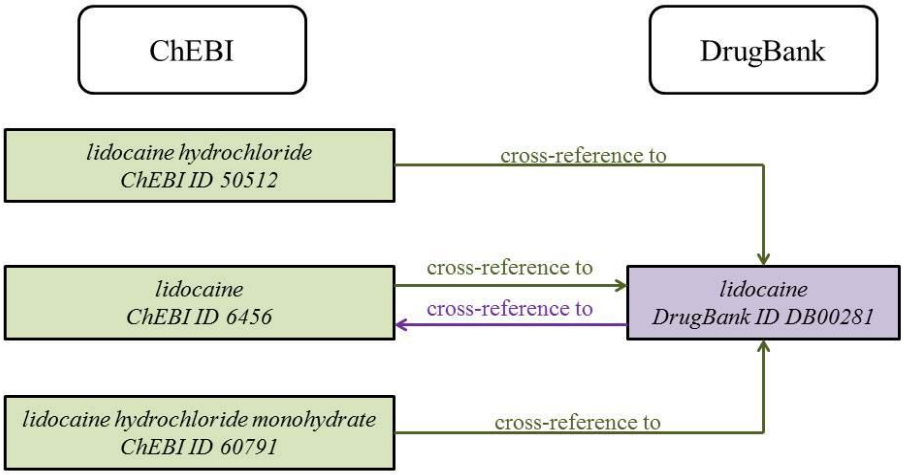

Figure 7.19. Cross-references between ChEBI and DrugBank for the drug *lidocaine* and its salts

In contrast to classes imported from ChEBI, which maintain their original URIs and IDs in DINTO, new classes translated from DrugBank are provided a new DINTO URI as follows: http://purl.obolibrary.org/obo/DINTO_DBXXX, where DBXXX is the unique identifier provided for that drug in DrugBank. In this way, it is easy to know whether a drug class in DINTO is originally from ChEBI only based on the URI.

3. **Manual refinement of the automatic mapping:**

Once new drug classes have been added to those previously imported from ChEBI, we identify cases where two different classes (one from ChEBI and one from DrugBank) have the same label. The reason is that these classes share the same preferred name in both DrugBank and ChEBI, but have not been identified as the same concept through our exact matching strategy.

We conduct a manual review to identify and correct all duplicated cases. There are 66 classes imported from both DrugBank and ChEBI. Most cases are due to the lack of the cross-reference to the DrugBank ID in ChEBI or vice versa. To ensure that the two classes refer to the same concept, we check the CAS Registry Number (CASRN) in both the database and the ontology. CASRNs are unique numerical identifiers assigned by the *Chemical Abstracts Service* to every chemical substance included in their CAS

Registry.[43] If the CASRN matches between the two sources, we consider them a unique class in DINTO, which conserves the original ChEBI URI and integrates additional information from DrugBank. The list of duplicate classes and the error sources is provided in **Annex 5**.

4. Drug identification information:

The traditional identification of drugs with natural language names is essential for human communication, although associated problems such as ambiguity might arise (see **Section 3.4.2**). To avoid these issues different drug codes systems have been created, such as the well-known ATC classification system (WHO, n.d.) or the FDA Substance Registration System and its unique ingredient identifiers (UNIIs) for ingredients in drugs.[44] Similarly, in the field of pharmacological databases and other computerized information sources, the problem of unique identification of a drug has been tackled by the assignment of unique numerical or alpha-numerical identifiers in a way that each dataset has created their own IDs (Brenninkmeijer et al., 2013).

This framework has led to a proliferation of different identifiers for the same concept among different information resources, which can play a crucial role in data integration (Hassanzadeh, Zhu, Freimuth, & Boyce, 2013). As we have seen, information from DrugBank and ChEBI can be integrated by using the correspondent cross-references between ChEBI and DrugBank IDs (**Figure 7.19**), while another identifier such as the CASRN can be used to identify unequivocally those cases where correspondence between concepts is not clear. However, not all data sets have cross-references to other identifiers. In these cases, this strategy for data integration cannot be used. For example, NDF-RT and MeSH do not have cross-references to ChEBI or DrugBank IDs. In the same way, the latter ones do not have cross-references to NUIs (NDF-RT unique identifiers) or MeSH IDs. Therefore, it would not be possible to integrate data from the four sources. However, NDF-RT provides MeSH IDs, while MeSH provides CASRN for drugs. Therefore, it would be possible to integrate data from ChEBI or DrugBank and MeSH by using the CASRNs and then, between these and NDF-RT by means of the MeSH IDs.

Therefore, inclusion of cross-references to different data identifiers and drug systems codes could increase the interoperability and further development by data integration of DINTO. Some cross-references (such as the KEGG database[45]) are currently imported from ChEBI, but some of them are missing. Therefore, we integrate identification information from DrugBank in order to increase the number of cross-references for each drug.

Following the model used in ChEBI, this information is represented in the ontology as annotation properties (see **Section 7.2.4**). During the process, we have identified that the annotation property for CASRN is provided in terms of different annotation properties in ChEBI (e.g., it is usually described as a cross-reference to ChemIDPlus, but sometimes it can be found as a cross-reference to KEGG Compound or KEGG Drug). As

[43] http://www.cas.org/content/chemical-substances
[44] http://www.fda.gov/forindustry/datastandards/substanceregistrationsystem-uniqueingredientidentifierunii/default.htm
[45] http://www.genome.jp/kegg/compound/

an additional step, we identify and provide a unique format for all values, through the creation of the annotation property CASRN.

5. Proteins and drug-protein relationships:

DrugBank provides information about interactions between drugs and proteins, which is translated into our ontology. Firstly, we create proteins as classes, providing them with a URI, an identifier, and identification information such as the UniProt IDs (Consortium, 2008). Then, we represent relationships between them and individual drugs at the class level (see **Section 7.1.3**).

6. DDI relationships:

Finally, we have translated information about DDIs drugs from DrugBank into DINTO. DDIs are represented in two different ways: *i)* through the *'may interact with'* relationship between two pharmacological entities; and *ii)* as classes representing an interaction between two pharmacological entities (**Figure 7.20**). Information related to the mechanism leading to the DDI is also imported from DrugBank, when available.

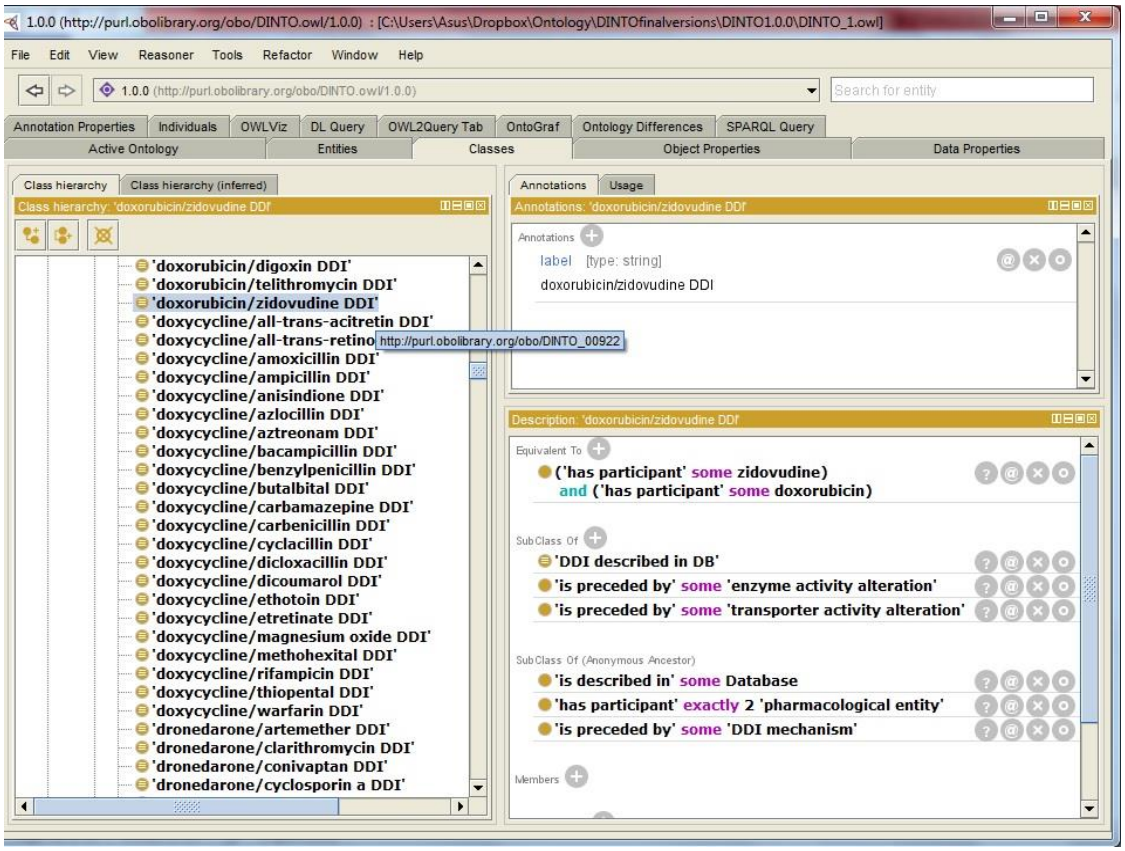

Figure 7.20. Representation of the DDI between '*doxorubicin*' and '*zidovudine*' in DINTO

7.1.7 Rules for DDI mechanisms representation

One of the objectives of this thesis is to create an ontology supporting different application and, in particular, NLP tasks and inference of DDI-knowledge. The latter one might be useful for the development of pharmacovigilance tools, such as CDSS or signal detection systems, among others. Specifically, our aim is to infer new DDIs exploiting structured knowledge reused from other resources and automatically imported in DINTO, in a way that would avoid the high costs of manual curation of pharmacological information.

To do this, we propose to infer new relationships *'may interact with'* between classes in our ontology on the basis of existing drug-protein relationships. Despite the richness of OWL's set of relational properties, its expressivity does not allow us to express all possibilities for object relationships (Holford, Khurana, Cheung, & Gerstein, 2010; Horrocks, Patelschneider, Bechhofer, & Tsarkov, 2005). This limitation can be overcome by the combination of an OWL ontology with inference rules (Golbreich, 2005). The Semantic Web Rule Language (SWRL)[46] is an expressive OWL-based rule language that provides more powerful deductive reasoning capabilities than OWL alone and has become the most widely used rule language in the semantic web community (Holford et al., 2010). SWRL rules are expressed in terms of OWL concepts (classes, properties, individuals), and it is supported by the ontology development environment Protégé and the reasoner engines HermiT (Glimm, Horrocks, Giorgos, & Zhe, 2014) and Pellet (Sirin, Parsia, Grau, Kalyanpur, & Katz, 2007).

In the recent years, SWRL rules have been used to model biomedical knowledge aimed to support the development of clinical decision support systems (CDSS) applied to clinical pathways compliance checking (Alexandrou, Pardalis, Bouras, Karakitsos, & Mentzas, 2012; Z. Huang, Bao, Dong, Lu, & Duan, 2014; Ye, Jiang, Diao, Yang, & Du, 2009), supervision and treatment of critical patients (Martínez-Romero et al., 2013), or remote patient monitoring systems (Shojanoori & Juric, 2013), for example. The combination of OWL and SWRL has been used to represent EHRs information and support the interoperability between heterogeneous systems (Lezcano, Sicilia, & Rodríguez-Solano, 2011), and the reasoning capabilities provided by SWRL rules have been exploited to query patient datasets at different levels of abstraction (Taboada et al., 2012) or to infer new relationships in the pseudogenes domain (Holford et al., 2010). Furthermore, these rules have been used to evaluate the performance of novel semantic-based methods for IR (Hassanpour et al., 2011).

SWRL rules are written as antecedent-consequent pairs, where the antecedent is referred to as the rule "body" and the consequent is referred to as the "head". The head and body consist of a conjunction of one or more atoms. Atoms can be of the form *C(x)*, *P(x, y)*, *sameAs(x, y)* or *differentFrom(x, y)*, where *C* is an OWL DL description, *P* is an OWL property, and *x*, *y* are either variables, OWL individuals, or OWL data values. The example below shows a SWRL rule describing that two drugs interact when one inhibits the metabolizing enzyme of the other one:

[46] http://www.w3.org/Submission/SWRL/

```
inhibits(?othery, ?z), metabolizes(?z, ?y),
DifferentFrom (?othery, ?y) -> 'may interact
with'(?othery, ?y)
```

This rule establishes that if one pharmacological entity (?othery) inhibits the activity of an enzyme (?z), which metabolizes another pharmacological entity (?y), therefore there might be an interaction between the two pharmacological entities ?othery and ?y. Since SWRL adopts OWL's open world assumption (OWA), it is not straightforward to assume that two individuals are automatically distinct if they have different names. Therefore, it is necessary to include the differentFrom atom, which determine that the variables ?othery and ?y do not refer to the same underlying individual. SWRL rules reason about OWL individuals. Therefore, to perform the inference, the ontology should include at least three individuals (two pharmacological entities and one enzyme) and the relationships '*inhibits'* and '*metabolizes'* among them.

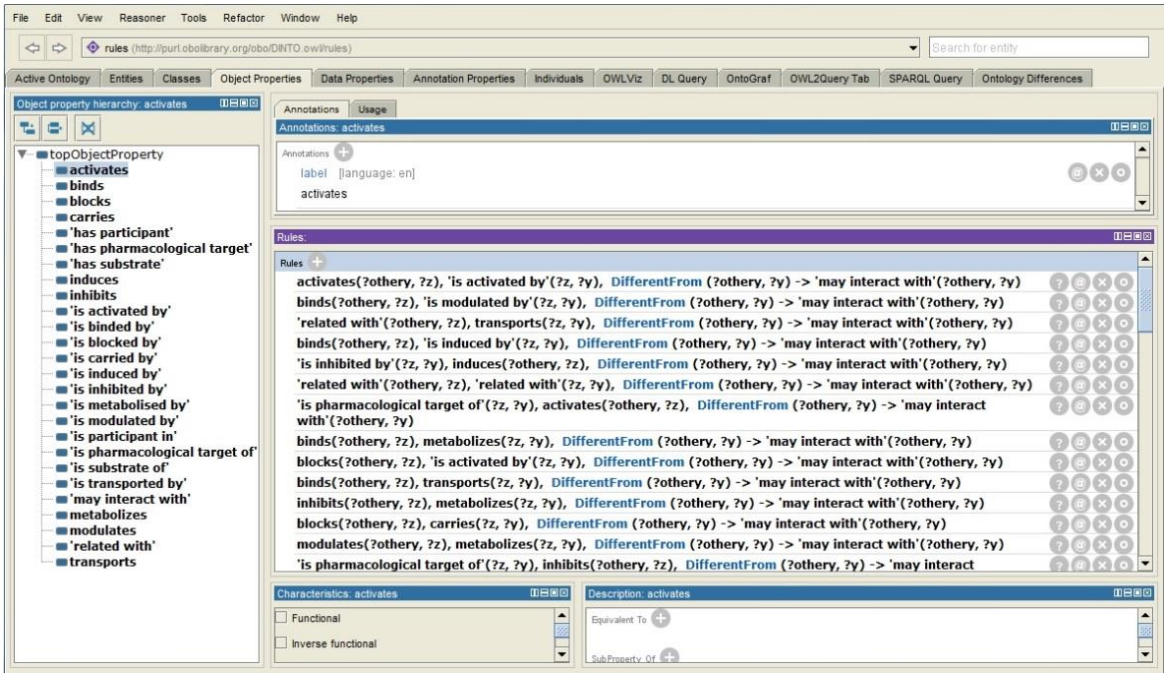

Figure 7.21. SWRL rules in Protégé

In this project, we have created two main groups of rules in SWRL, each one of them consisting of 59 rules. The first one formalizes different types of DDI mechanisms, including both PD and PK mechanisms, in a set of rules that allows for inferring new *'may interact with'* relationships between drugs (**Figure 7.21**). The second one assigns a specific type to a DDI on the basis of its inferred mechanism, which is inferred from drug-protein relations information. A detailed description of these rules is provided in **Section 7.2.5**, and the results obtained when combining our OWL ontology and the SWRL rules are described in **Section 9.2.2** and **Section 9.2.3**.

7.1.8 Maintenance

An important aspect of any ontology is its evolution and maintenance. Ontologies might include errors, lack of information, or may need to evolve in order to adapt to changes or new discoveries in the domain (Stojanovic, 2004). There is not a standard for ontology maintenance. Nevertheless, we have established a strategy for the maintenance of DINTO, which will ensure consistency between different releases, documentation of any change made, and tracking of users' suggestions or requests (Shaban-Nejad & Haarslev, 2009). The proposed activities are described below.

1. Assignment of one person responsible for maintaining DINTO.

2. Creation of an online system enabling users to make requests for modifications and addition of new terms, which tracks the suggestions and manages the change requests (https://code.google.com/p/dinto/issues/list).

3. Enhancement of the ontology through activities aimed to:

    - identify missing or misplaced relationships and terms,

    - add new information not covered in the current version of the ontology,

    - provide definitions for remaining undefined terms,

    - identify redundancy.

4. Notification of changes via reports at DINTO site https://code.google.com/p/dinto/

5. Management of identifiers and tracking of all those that are deleted or merged in order to provide lists of this changes with every release of DINTO.

6. Technical evaluation (as described in **Chapter 8**) of each new version prior to its public release.

7.2 Description of DINTO

In the previous section, we have provided a detailed description of the process of building the ontology. Here we describe the final resource and the different constituents of its CM – that is, the elements that, in combination, formally represent the pharmacological domain of drug interactions knowledge.

On the one hand, concepts in the domain are represented as classes in DINTO and organized into nine different hierarchical taxonomies. Top-level classes represent concepts that are more general, while subclasses represent specializations of them. On the other hand, object properties represent relationships between classes, while data properties represent characteristics of concepts. Additional information or metadata is

represented in the ontology as annotation properties, including mapping to other ontological resources. All these entities are described in the following sections.

7.2.1 Classes

Each class in DINTO represents a unique concept. They are organized in nine non-overlapping (disjoint) taxonomies or top-level classes that represent all the DDI-related knowledge. A description of them is provided below.

- '*chemical entity*': The top class '*chemical entity*' is imported from the ChEBI ontology and maintains its original URI. However, only a fragment of ChEBI is imported in DINTO, and the structure is therefore different between both sources. This top class has two subclasses in DINTO: '*pharmacological entity*' and '*protein entity*'.

  - '*pharmacological entity*': this class collects all chemical entities that can exert a pharmacological activity. In other words, this class represents all substances that can produce any pharmacological effect in the organism, which are described as active ingredients. Most of these subclasses have been imported from the ChEBI ontology corresponding to those classes having some *'drug'* role. However, to increase the coverage of DINTO, other pharmacological entities not present in ChEBI have been imported from the DrugBank database. All of them are subclasses of *'pharmacological entity'* class and siblings between them. However, it is possible to distinguish their original source since classes imported from the ChEBI ontology maintain their original URI (http://purl.obolibrary.org/obo/CHEBI) while the new ones created from DrugBank have a DINTO URI (http://purl.obolibrary.org/obo/DINTO).

  - '*protein entity*': the class *'protein entity'* collects proteins related to the previously describe *'pharmacological entities'*. They are organized into five subclasses ('*carrier'*, '*enzyme'*, '*target'*, '*transporter',* and '*receptor'*) although the same protein (e.g., '*5'-nucleotidase*') can be a member of more than one of them (e.g., '*target'* and '*enzyme'*).

- '*DDI mechanism'*: this class represents the different processes that can lead to a DDI. It is subdivided into two subclasses: '*pharmacodynamic DDI mechanism*' and '*pharmacokinetic DDI mechanism*'. Each one of them is subdivided into more specific mechanisms (**Figure 7.14**)

- '*drug interaction*': This class represents all possible drug interactions. The subclass '*DDI*' represents interactions between two drugs (or drug-drug interactions). This structure allows for the representation, in future versions of the ontology, of interactions other than DDIs, such as food-drug interactions or environmental substance-drug interactions.

  - The class '*DDI*' represents DDIs by different classification criteria. Subclasses *'pharmacodynamic DDI'* and *'pharmacokinetic DDI'* represent DDIs on the basis of their preceding mechanisms, while subclasses

'clinically relevant DDI', *'non-clinically relevant DDI'* and *'uncertain DDI'* aim to classify DDIs by their type of effects or consequences.

- The subclass '*DDI described in DB*' includes specific interactions between pairs of drugs with a common source of information. Each one of the subclasses represents the interaction between two subclasses of the '*pharmacological entity*' class.[47]

- '*Information Resource*': This top class represents the different information resources that can describe some aspect of a DDI. This class has been imported from the BRO (Tenenbaum et al., 2011), although some additional relevant concepts have been included as subclasses of the class '*Clinical_Research_Data*'.

- '*pharmacokinetic parameter*': This top class represents different PK parameters, which are concepts frequently used in the DDI literature. Subclasses have been imported from the PKO (Wu et al., 2013) and maintain their original URIs.

- '*pharmacokinetic process*': It represents the different processes that a pharmacological substance undergoes in the body ('*absorption'*, '*distribution'*, '*metabolism',* and '*excretion'*) and that are altered in a PK DDI.

- '*physiological effect*': The class '*physiological effect*' represents the different effects that a pharmacological substance can produce in the body. Based on their consequences for patients, drug effects are classified as adverse, toxic, or therapeutic effects. In the domain of drug interactions, the effect of a DDI is the consequence of the alteration of the effects of one or both interacting drugs. Therefore, the class '*altered physiological effect*' has a subclass named '*DDI effect*', which can be an altered adverse, toxic, or therapeutic effect. Regarding the consequences for patients, the effect of the DDI can be classified by its clinical relevance, and/or as a potentially beneficial or harmful DDI effect.

- '*role*': A role represents a particular behaviour which a material entity may exhibit and describes their activities. This class and its subclasses have been imported from the ChEBI ontology and maintained their original URIs. Only one new class '*participant*' and its two subclasses '*object'* and '*precipitant*' have been created to represent the different roles that an interacting drug can have in a DDI.

- '*study subject*': This class represents the different types of individuals that can take part in a DDI study.

7.2.2 Object properties

Object properties represent relationships between two classes. They are used to describe and define concepts in the ontology. As definitional relationships, they support

[47] In order to follow ontological design principles, in the next release of DINTO this information will be substituted by an annotation property called '*Provenance'* and the string value '*DrugBank database'*.

automated classification by reasoning engines, and are inherited into descendant concepts in the resulting inferred taxonomy. There are 72 object properties in DINTO. Some of them are organized into hierarchies showing specialization, while others are organized in property chains to allow the inference of new relationships between concepts.

Domain – the class whose definition may use the object property – and range – the class that the object property can refer to – are established when appropriate to provide accurate definitions of both object properties and classes, and all of them have a natural language definition.

Figure 7.22 shows part of the object property hierarchy in Protégé. A complete list and description of these relationships can be found in **Annex 6**.

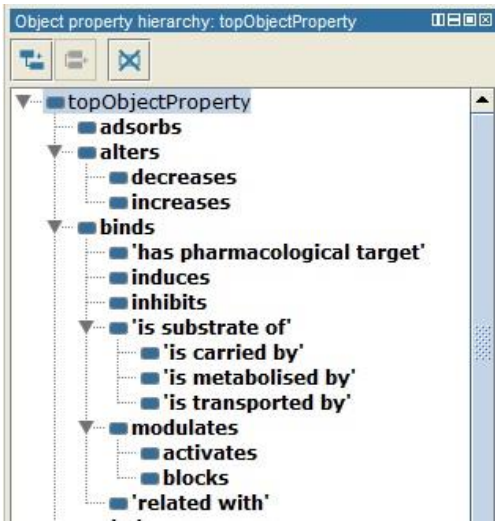

Figure 7.22. Partial view of the object property hierarchy in DINTO

7.2.3 Data properties

Some important characteristics of concepts need to be represented as data. Relationships between classes and data are represented in the ontology through data properties, which are informational attributes of concepts. A data property value is a text string or an integer attached to a single concept, without any inheritance to descendant concepts in the inferred taxonomy. There are 17 data properties in DINTO that relates five classes in the ontology with data values. These classes are '*pharmacological entity*', '*study subject*', '*DDI*', '*Information Resource*', and the anonymous class representing a DDI interacting drug.[48]

[48] This anonymous class is defined in the ontology as: '*pharmacological entity*' and ('*is participant in*' some ('*DDI*' and ('*is described in*' some '*Information Resource*'))). For abbreviation purposes, we refer to it in the text as *DDI interacting drug*.

- *'pharmacological entity'*: There is only one data property used to describe information about a pharmacological substance.
    - *has concentration* → `integer`
      The measure of the amount of pharmacological substance in a determine tissue and at a specific moment.

- '*DDI*': There are different data properties providing information about a specific DDI.
    - *has documentation level* → `{"established", "possible", "probable", "suspected", "unlikely"}`
      The degree of confidence that the interaction can cause a specific effect.
    - *has incidence* → `string`
      Relative frequency of occurrence of the DDI.
    - *has onset* → `{"delayed" , "rapid"}`
      How rapidly the clinical effects of an interaction can occur.
    - *has relevance* → `{"clinical relevance", "non-clinical relevance"}`
      The real importance that the DDI has in the clinical practice.
    - *has severity* → `{"major" , "minor" , "moderate"}`
      Degree of the undesirable effect of the DDI.
    - *has dose recommendation* → `{"change dose schedule", "decrease from baseline", "increase from baseline", "no change necessary", "use specific dose"}`
      A recommendation to avoid or decrease the effect of a DDI related to the dose of the interacting drugs.
    - *has drug selection recommendation* → `{"add medication", "change administration route", "no change necessary", "not-restart", "use alternative"}`
      A recommendation to avoid or decrease the effect of a DDI related to the selection of the interacting drugs.
    - *has drug monitoring recommendation* → `{"change monitoring strategy", "not-necessary", "recommended", "required"}`
      A recommendation to avoid or decrease the effect of a DDI related to monitoring some aspect or parameter of patients.
    - *has test recommendation* → `{"not-necessary", "recommended", "required", "take note it is available"}`
      A recommendation to avoid or decrease the effect of a DDI related to testing recommendations.

- '*study subject*': The main characteristics of subjects participating in a DDI study are described through data properties.
    - *has age* → `integer`
      The age of the subject participating in a study or other information source describing a DDI.
    - *has gender* → `{"female", "male"}`
      The gender of the subject participating in a study or other information source describing a DDI.
    - *has race or ethnic* → `string`
      The race or ethnic of the subject participating in a study or other information source describing a DDI.

- '*Information Resource*': In some cases, it is important to know the number of subjects participating in a DDI study. This information is described through a data property.
    - *has subject number* → `integer`
      The number of subjects participating in a study or other information source describing a DDI.

- *DDI interacting drug*[49]: Characteristics related to the pharmaceutical presentation of the interacting drugs in one specific DDI study or information resource are described as data properties in DINTO.
    - *has administration route* → `string`
      The route use to administrate the drug described in the study or other information source describing a DDI.
    - *has dose* → `integer`
      The dose of the drug described in the study or other information source describing a DDI.
    - *has pharmaceutical form* → `string`
      The dose of the drug described in the study or other information source describing a DDI.

7.2.4 Annotation properties

Annotation properties represent metadata of concepts in the ontology. This information is established at the class level. It is non-definitional and is not used in automated classification. Therefore, it is not inherited into descendant concepts in the resulting inferred taxonomy.

[49] This anonymous class is defined in the ontology as: '*pharmacological entity*' and ('*is participant in*' some ('*DDI*' and ('*is described in*' some '*Information Resource*'))). For abbreviation purposes, we refer to it in the text as *DDI interacting drug*.

DINTO imports the OBO ontology metadata file,[50] a standard representation of annotation properties to be used across OBO ontologies with 38 annotation properties. In those cases where the ontology requires additional metadata, we have created specific annotation properties.

Annotation properties can represent metadata for any class in the ontology. However, they are most frequently used to provide information about chemical entities. Here, we describe the most important annotation properties used in DINTO.

- **Annotation properties for synonyms and alternative names:**

DINTO provides different synonyms and alternative names for classes, especially for pharmacological entities, that are included in the ontology as annotation properties.

  - *Synonym* (original source: ChEBI): It is the annotation property used in ChEBI for alternative names and synonyms. This annotation property is used for synonyms entered manually in DINTO.

  - *DBSynonyms* (original source: DINTO): This annotation property represents all possible alternative generic names described in the database DrugBank for every drug entry.

  - *DBName* (original source: DINTO): Annotation property specifically created to represent the preferred name for a pharmacological substance used in the DrugBank database.

  - *DBBrand* (original source: DINTO): This annotation property represents the different brand names for a pharmacological entity translated from DrugBank.

  - *DBSalt* (original source: DINTO): Annotation property specifically created to represent salts of a pharmacological entity described in the database DrugBank.

- **Annotation properties for code systems and chemical identification:**

There are several chemical and pharmacological code systems included in the ontology as annotation properties. They are important for the unambiguous identification of chemical entities among different information sources and necessary for integration of information in the ontology or related projects.

  - *AHFSCode* (original source: DINTO): The American Hospital Formulary Service (AHFS) Pharmacologic-Therapeutic Classification code(s) for a pharmacological entity.

  - *ATCCodes* (original source: DINTO): The Anatomical Therapeutic Chemical Classification System code(s) for a pharmacological entity.

[50] https://code.google.com/p/information-artifact-ontology/wiki/OntologyMetadata

- *CASRN* (original source: DINTO): The Chemical Abstracts Service Registry Number (CASRN) for a pharmacological entity.

- *InChI* (original source: ChEBI): The IUPAC International Chemical Identifier for a pharmacological entity.

- *InChIKey* (original source: ChEBI): The InChIKey for a pharmacological entity.

- *SMILES* (original source: ChEBI): The Simplified Molecular-Input Line-Entry System for a pharmacological entity.

- *altId* (original source: ChEBI): Previously assigned ChEBI IDs for a pharmacological entity.

- *xref* (original source: ChEBI): Cross reference to other pharmacological resources and database IDs. The following are of special interest for mapping purposes:

    - KEGG DRUG
    - KEGG COMPOUND
    - DrugBank
    - National Drug Code Directory
    - Drugs Product Database (DPD)
    - Drugs.com
    - Protein Data Bank (PDB)
    - Wikipedia
    - PharmGKB
    - RxList
    - UniProt Knowledge Base (UniProtKB)
    - PubChem Substance and PubChem Compound
    - Reaxys
    - ChemSpider
    - etc.

- **<u>Annotation properties for useful information and data:</u>**

There is relevant information providing a better description of entities that, however, is not formally represented. Therefore, it is included in natural language as annotation properties.

- *Definition* (original source: ChEBI): A brief natural language description of the entity.

- *Gene* (original source: DINTO): Regarding protein entities, the gene that codifies for that protein.

- *OrganismClass* (original source: DINTO): Regarding protein entities, the corresponding organism for that protein.

- *mapsTo (original source: DINTO):* Link to an entity in a different ontology that represents the same entity in DINTO.

7.2.5 SWRL rules

SWRL rules are used to extend the expressivity of OWL 2 and to exploit their reasoning capabilities for the inference of new information in DINTO. There are a total of 118 SWRL rules that can be divided into two main groups: *i)* rules modeling DDI mechanisms that define new *'may interact with'* relationships between pharmacological entities, and *ii)* rules assigning a DDI type on the basis of the inferred mechanism for a DDI. These types include *'target related DDI'* and its subtypes *'agonistic DDI'* and *'antagonistic DDI'*; *'enzyme related DDI'* and the subtypes *'enzyme inhibition DDI'*, *'enzyme induction DDI',* and *'enzymatic saturation DDI'*; *'carrier related DDI'* and its subtypes *'carrier induction DDI'*, *'carrier inhibition DDI',* and *'carrier saturation DDI'*; *'transporter related DDI'* and the subtypes *'transporter inhibition DDI'*, *'transporter induction DDI',* and *'transporter saturation DDI'*. Other 17 rules simply assign the type *'DDI'* when the relationships are too ambiguous to know the type of interaction between two drugs (e.g., when the only known relationship between two pharmacological entities and a protein is the object property '*related to*' (**Table 7.1**).

SWRL rules in DINTO	**Number**
SWRL rules to infer DDIs	59
SWRL rules to infer DDI mechanisms	59
'target related DDI' type	17
'agonistic DDI' subtype	4
'antagonistic DDI' subtype	3
'enzyme related DDI' type	9
'enzyme inhibition DDI' subtype	1
'enzyme induction DDI' subtype	2
'enzymatic saturation DDI' subtype	2
'transporter related DDI' type	3
'transporter inhibition DDI' subtype	2
'transporter induction DDI' subtype	2
'transporter saturation DDI' subtype	1
'carrier related DDI' type	3
'carrier inhibition DDI' subtype	2
'carrier induction DDI' subtype	2
'carrier saturation DDI' subtype	1
'DDI' type	5

Table 7.1. Types of SWRL rules in DINTO

7.2.6 DINTO in numbers

To end this chapter, we provide a table describing the final number of entities in the first version of DINTO 1.0.0 (**Table 7.2**) and describe the different owl files that can be downloaded from https://code.google.com/p/dinto/ (**Table 7.3**).

Ontology Metrics	Total
Number of classes	25,809 11,555 DDIs 8,786 drugs
Number of individuals	0
Number of properties	161
Number of object properties	73
Number of data properties	17
Number of annotation properties	71
Classes with natural language definition	9,392
Classes with OWL definitional axiom	11,587

Table 7.2. Final number of entities in DINTO

Files available to download	Description
DINTO_1.owl	The first version of DINTO.
DINTO_1BFO.owl	The first version of DINTO mapped to the top-level ontology BFO.
BRO_DINTO_subset.owl	The fragment imported from BRO to DINTO.
PKO_DINTO_subset.owl	The fragment imported from PKO to DINTO.
DINTO_rules_inferenceDDI.owl	The file containing the 59 SWRL rules created to infer DDIs in DINTO.
DINTO_rules_inferenceDDI_types.owl	The file containing the 59 SWRL rules created to infer types of DDIs in DINTO.
DINTO_inf_classification.owl	The version of DINTO intended to be used with a reasoner engine to obtain an inferred classification of DDIs on the basis of their asserted mechanism (*IExp1 in* **Section 9.2.1**).
DINTO_SEMEVAL2_inf.owl	The version of DINTO created to be evaluated in the framework of the *SemEval-2013 DDI Extraction task*. It contains the DDIs inferred using a reasoner for a reduced number of pharmacological entities (*IExp2* in **Section 9.2.2**)
DINTO_inf_mech.owl	The version of DINTO intended to be used with a reasoner engine to obtain and inferred classification of DDIs on the basis of their inferred mechanisms (*IExp3* in **Section 9.2.3**).

Table 7.3. Description of the DINTO-related files available to download

Chapter 8

Evaluation of DINTO

Ontologies are engineering artifacts that, as all other engineering artifacts, need to be evaluated (Vrandečić, 2010). This evaluation is required during the development of a new ontology, not only to ensure the quality of the resulting resource but also to guide ontologists during the construction process and refinement steps. Moreover, ontology reuse requires previous evaluation of ontology candidates in order to provide new users with a way to assess their quality (Brank, Grobelnik, & Mladeni, 2005).

In spite of the relevance of ontology evaluation, this area is still an open research field (Hastings, Brass, Caine, Jay, & Stevens, 2014). There is a lack of clear evaluation methodologies that leads to the adoption of different approaches by different ontology development groups (Gómez-Pérez, 1999). The main reason for this lack of consensus is that the establishment of objective measures for ontology evaluation is limited by the intrinsic nature of these resources.

Ontology evaluation can be divided into two main groups: technical and application-based evaluation. The former one is a validation of the content and consistency of the ontology, while the latter one assesses the quality and usefulness of the ontology for an intended application.

On the one hand, there are ontological aspects that can be objectively measured from a technical perspective. These include taxonomical aspects, naming conventions, completeness, or consistency. During the recent years, there have been important efforts to identify these aspects and to provide standards to ensure the development of reusable ontologies. There are three research groups and/or projects that have strongly contribute to the development of principles and standards for ontology technical evaluation. Firstly, the most important efforts regarding domain-independent ontology evaluation have been

addressed by the *Ontology Engineering Research Group* at "*Universidad Politécnica of Madrid*".[51] Secondly, principles for the creation of reusable and high quality ontologies in the biomedical domain have been established by the OBO Foundry community, a collaborative effort for the development and maintenance of biomedical ontologies to ensure their integration and interoperability (Smith et al., 2007). Thirdly, another important work in the field of controlled medical vocabularies is Cimino et al.'s Desiderata (Cimino, 1998), which delineates desirable characteristics for controlled medical terminologies and attempts to summarize emerging consensus regarding structural issues of such terminologies. These three works provide standards, principles and recommendations for ontologies in the general and biomedical domain that contribute to the creation of high quality and reusable ontologies.

On the other hand, there are ontological aspects that cannot be assessed objectively. The reason is that, although domain ontologies are formal representations of a body of knowledge, conceptualization is a subjective procedure open to different representations. In this way, the same aspect of the real world (e.g., a pharmacological process) can be represented in different ways in different ontologies (Olivié, 2007). Consequently, there can be several ontologies representing the same knowledge. Assessing their quality and usefulness depends on users' requirements and their suitability for the application where they are used. Therefore, application-based evaluations are necessary to ensure the quality and usefulness of ontologies.

The evaluation of DINTO integrates both mentioned approaches: technical and application-based evaluations. With the aim to provide an objective and complete evaluation of this new ontology, we propose a strategy based on the combination of different evaluation methods that, in conjunction, provide a comprehensive description of the consistency, quality, and re-usability of the ontology. This strategy is summarized as follows, and described in this chapter and the following **Chapter 9** and **Chapter 10**:

1. TECHNICAL EVALUATION: The ontology is evaluated in its form and content to assess its consistency and expressivity, as well as for the detection of errors such as circularity, partition, or semantic inconsistency (Gómez-Pérez et al., 2004). This evaluation process is divided into the following subtasks:

    - **Classification scenario testing**: evaluation of the CM of the ontology is carried out by means of the representation of real examples for DDIs taken from the pharmacological literature. Protégé enables the logical inference of new entity classification and new relationships using a reasoning engine. The representation of real examples from the DDI domain allows the checking of consistency and expressivity of the CMs.

    - **Supporting or answering of previously established competency questions**: CQs are pre-established and natural language questions created during the specification activity representing the aspects that the final ontology should address. They are used in the evaluation of ontologies as a pre-established set of content requirements for the new ontology.

    - **Peer-review or human performed evaluation**: A manual review of the ontology is carried out to ensure that the ontology meets a set of predefined

[51] http://www.oeg-upm.net/

criteria, standards, or requirements (Lozano-Tello & Gómez-Pérez, 2004). Specifically, the ontology is created on the basis of the principles of the OBO Foundry community (Smith et al., 2007) to standardize and reuse ontologies. Indeed, DINTO has been reviewed by the OBO Foundry and has been accepted and listed in their website as one of the "*OBO Foundry candidate ontologies and other ontologies of interest*".[52]

2. APPLICATION-BASED EVALUATION: Biomedical ontologies should be evaluated against the task for which they were developed (Hoehndorf, Dumontier, & Gkoutos, 2012). Moreover, an ontology is said to be robust if it performs well in multiple heterogeneous tasks. Therefore, the usefulness of our ontology is evaluated in two different scenarios:

- **Inference of DDIs**: The ontology is used to infer DDIs on the basis of their mechanisms. Information represented in the ontology is combined with SWRL rules to infer DDIs between individual pairs of drugs and their mechanisms. These inferences are evaluated against asserted information imported from the database DrugBank (see **Chapter 9**).

- **DDI Information Extraction**: DINTO is used by an IE system in order to recognize pharmacological substances and extract DDIs from texts. This system is evaluated on the DDI corpus, and results are compared to those obtained by participating systems on the *SemEval-2013 DDIExtraction task* (Segura-Bedmar et al., 2013) (see **Chapter 10**).

8.1 Technical evaluation

Technical evaluation checks the quality of the ontology from an objective point of view and independently of the requirements of final users and applications. Because there is not a unique methodology for ontology technical evaluation (Gómez-Pérez, 1999), our approach consists in combining three different methods in order to evaluate the ontology in a comprehensive way. These methods are: *i)* classification scenario testing, *ii)* supporting or answering of competency questions, and *iii)* peer-review evaluation. With this strategy, it is possible to evaluate if the ontology is consistent, complete,and free of taxonomical errors

8.1.1 Classification scenario testing

Classification scenario (CS) testing is performed to check if the knowledge represented in the ontology is consistent and complete. As we explained in **Section 7.1.3**, during conceptualization of DINTO the domain knowledge is organized and structured into CMs that are then implemented using the ontology editor Protégé to construct a formal representation of the DDI domain. During these processes, different errors could

[52] http://www.obofoundry.org/

be introduced, leading to the inference of inconsistent information. A useful method to identify these conceptualization and implementation mistakes is the representation of real DDI examples taken from the pharmacological literature in the ontology. Protégé enables the logical inference of new entity classification and new relationships through the use of a reasoning engine, a program that infers logical consequences from a set of explicitly asserted facts or axioms (Cuenca, 2011). Therefore, the representation of real examples from the DDI domain as individuals in the ontology allows for validating the consistency and the expressivity of the CMs.

This evaluation is carried out iteratively during the processes of conceptualization and implementation, making possible the correction of inconsistences during the creation of the ontology. Hence, in a final evaluation, we select three CS testing examples to evaluate the consistency of the final ontology. The two first are well known interactions described in different sources representing one of the two main types of DDIs (PK and PD DDIs). The third one is an unknown DDI recently described in a case report and published in a medical journal. These examples are explained below (CS1, CS2, and CS3, respectively) and results of the CS testing are graphically represented in **Figure 8.1**, **Figure 8.2**, and **Figure 8.3**. In these figures, boxes represent individuals, black arrows are explicitly represented relationships between individuals and red dashed arrows represent inferred relationships obtained when using a reasoner engine. Green ellipses are attributes manually added for individuals. Every individual's box shows the '*is a*' relationships used to classify them in the ontology. Black lines show information manually added in the experiment, while red italic lines show the inferred classification of individuals in the ontology.

- *CS1: PK DDI between rifampicin and cyclosporin A*

  Rifampicin is an antibiotic used in the treatment of tuberculosis. This drug is important in the field of DDIs since it is a potent inducer of numerous metabolic enzymes and transporters. Therefore, it interacts with a large number of drugs. One of them is the immunosuppressant cyclosporine A. Among other applications, this drug is used in transplanted patients to avoid organ rejection. Therefore, therapeutic failure in patients treated with cyclosporin A can have serious consequences.

  Concomitant administration of rifampicin and cyclosporin A decreases the serum levels of the immunosuppressant due to the increase in its metabolism. This occurs because rifampicin induces the activity of different CYP 450 isoenzymes, including CYP 3A4, which is involved in the metabolism of cyclosporin A.

  This is a very well documented, established, and clinically important DDI. Transplant rejection may occur unless the cyclosporin A dosage is markedly increased. The interaction develops within a few days (within a single day in some cases). It is necessary to monitor the effects of concurrent use and increase the cyclosporin A dosage appropriately (Baxter, 2013; Tatro, 2010); (see **Figure 8.1**).

- *CS2: PD DDI between morphine and naloxone*

Morphine is the principal member of a group of drugs called opioid analgesics. These drugs are potent pain killers used to treat moderate to severe pain. Their effects are exerted by their activity in the opioid receptors family. Activation of these receptors leads to the analgesic effect. However, other severe adverse effects, such as depression of the central nervous system (CNS), are produced, too. Therefore, administration of excessive doses of opioid analgesics can produce severe respiratory depression, coma, and even death.

The drug naloxone is used to reverse these symptoms. Naloxone has affinity for the opioid receptors but, in contrast to morphine, which activates these receptors, naloxone blocks them. Therefore, naloxone decreases the activity and the effects of morphine, as well as other opioid agonists, and is used in recovery of opiated-induced CNS depression (Baxter, 2013; Tatro, 2010); (see **Figure 8.2**)

- *CS3: "Propafenone associated severe central nervous system and cardiovascular toxicity due to mirtazapine: a case of severe drug interaction"* (Rajpurohit, Aryal, Khan, Stys, & Stys, 2014)

«We describe a rare case of severe drug-drug interaction between propafenone and mirtazapine leading to propafenone toxicity. A 69-year-old Caucasian male taking propafenone for atrial fibrillation was prescribed mirtazapine for insomnia. Subsequent to the first dose of mirtazapine the patient experienced seizures, bradycardia and prolonged QRS as well as QTc intervals on EKG. The patient was admitted to the ICU and recovered after supportive management. Propafenone is an established class IC antiarrhythmic drug commonly used in the treatment of atrial fibrillation. It is metabolized through the CYP4502D6 pathway. Five to 10 percent of Caucasians are poor metabolizers. Mirtazapine is a commonly prescribed antidepressant drug, which is also metabolized through and may modulate the CYP4502D6 pathway leading to altered metabolism of propafenone and possible adverse effects. In this case, toxicity was reversed once the offending drugs were discontinued. An extensive review of the literature revealed this to be the first described case of drug interaction between propafenone and mirtazapine» (original abstract in (Rajpurohit et al., 2014); (see **Figure8.3**).

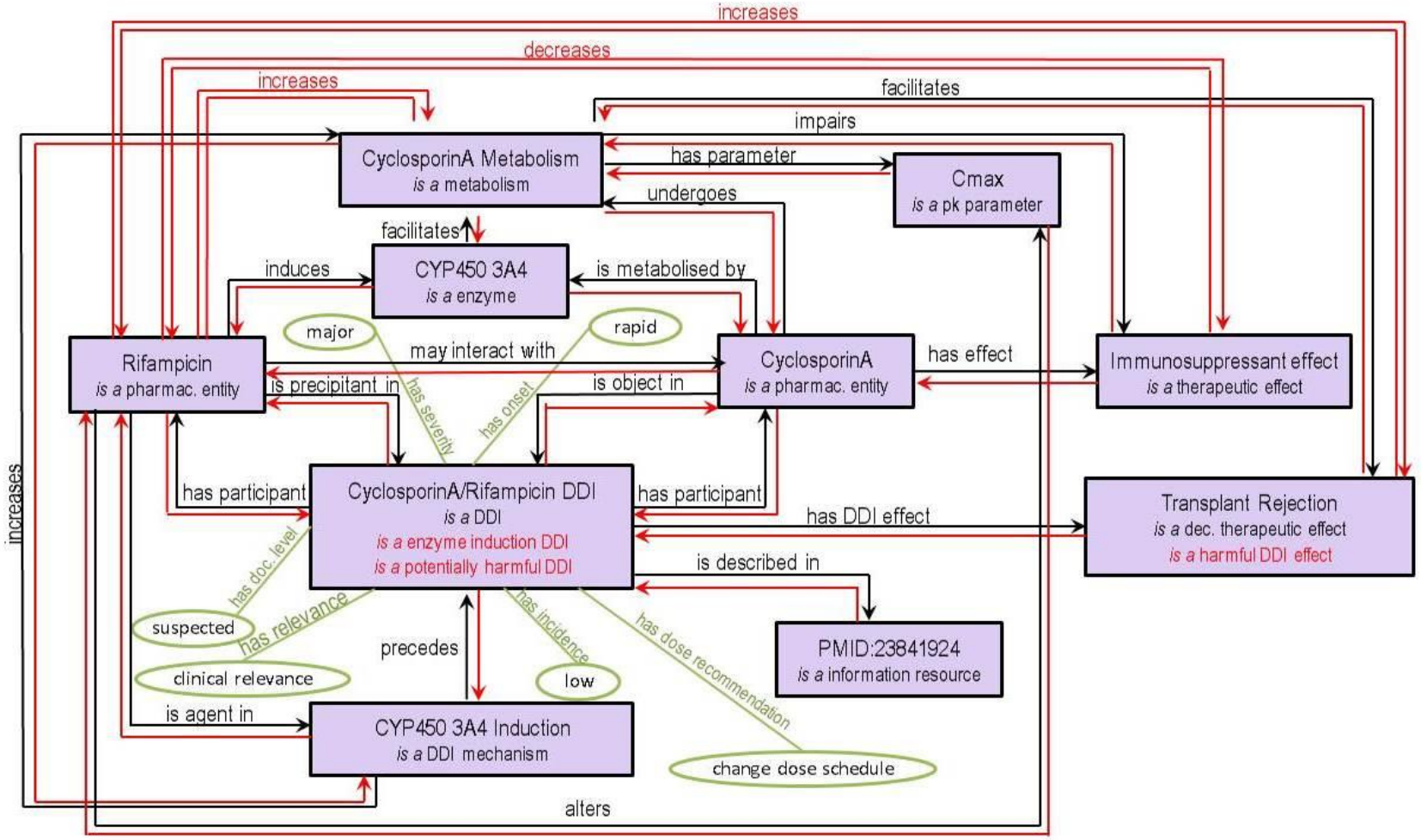

Figure 8.1. Classification Scenario 1 (CS1) representing the interaction between *rifampicin* and *cyclosporin a.*

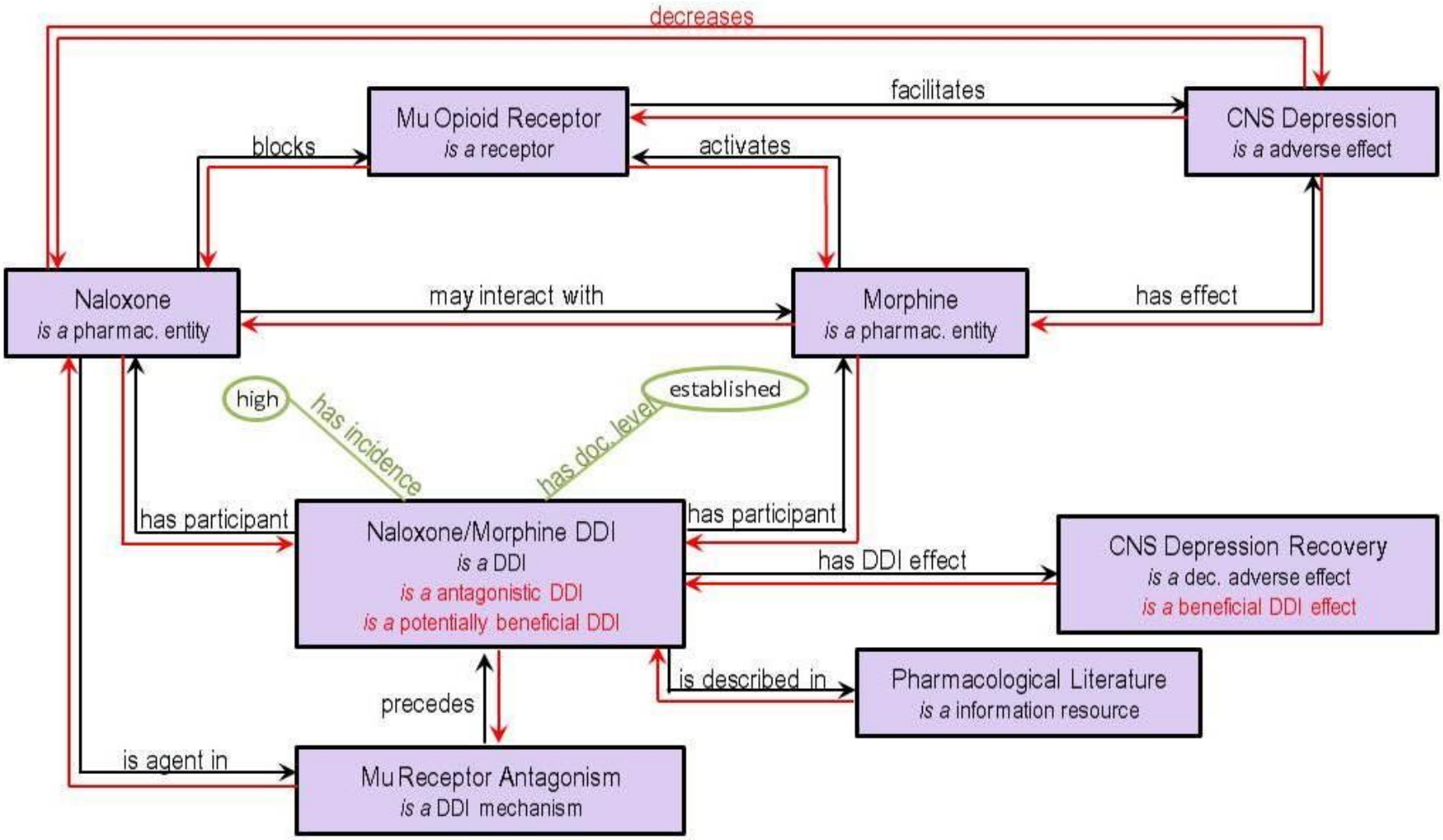

Figure 8.2. Classification Scenario 2 (CS) representing the interaction between *morphine* and *naloxone*

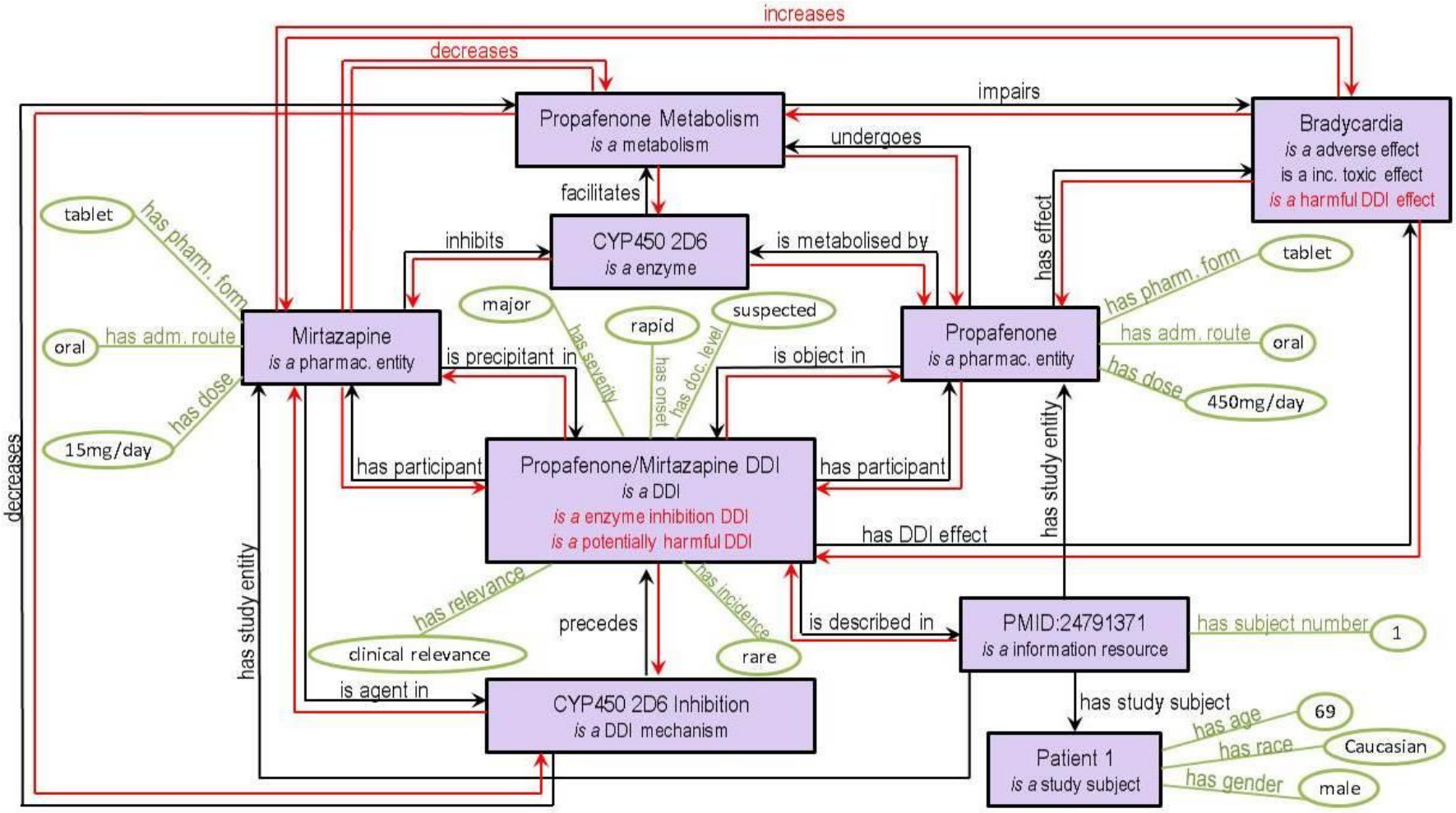

Figure 8.3. Classification Scenario 3 (CS3) representing the interaction between *propafenone* and *mirtazapine*

- **Results and conclusions**

In this evaluation approach, we have represented three different real DDIs as individuals in the ontology (represented as boxes in **Figure 8.1**, **Figure 8.2**, and **Figure 8.3**). Each one of these individuals is classified in the corresponding classes (shown as '*is a*' relationship in each box) in the ontology, and we establish different relationships between them (black arrows) following the real information provided by the pharmacological literature. Attributes, represented as data properties (green ellipses), are manually included, too. After leveraging a reasoner engine, we obtain inferred information in the form of inferred classification of individuals (red italic '*is a*' lines in boxes) and inferred relationships (represented as red dashed arrows). From this evaluation, we have drawn two main conclusions, which are described below.

Firstly, we have observed that the ontology is consistent, since two different reasoner engines − FacT++ (Tsarkov & Harrocks, 2006a) and Hermit 1.3.8 (Glimm et al., 2014) − can be leveraged without leading to inconsistences within the three versions representing the DDI examples. This means that the ontology conforms to the underlying logical theory −a variant of DL (Plessers & De Troyer, 2006) − of the OWL ontology language and that, therefore, the knowledge in the ontology has been correctly represented from a logical point of view.

Secondly, this method has demonstrated that the knowledge represented in the ontology allows for inferring important implicit information, and that these inferences are correct from a pharmacological point of view.

On the one hand, individuals are correctly classified in different classes. In this way, the interaction between '*rifampicin*' and '*cyclosporin A*' and the interaction between '*propafenone*' and '*mirtazapine*' are classified as '*pharmacokinetic DDIs*', that is, as DDIs occurring due to a PK mechanism. Moreover, a more detailed description of the DDI mechanism is inferred in a way that the first one is classified as an '*enzyme induction DDI*', and the second one as an '*enzyme inhibition DDI*'. Furthermore, these DDIs are classified in the ontology as '*potentially harmful*', since their consequences could be dangerous for patients. In contrast, the interaction between '*naloxone*' and '*morphine*' is classified as a '*pharmacodynamic DDI*', and specifically as an '*antagonistic DDI*', a type of PD DDI occurring due to the opposed effect of both drugs on the same target. Furthermore, regarding the potential consequences for the patient, the DDI is correctly classified as a '*beneficial DDI*'.

On the other hand, important information is inferred as new relationships between individuals. In addition to the inference of inverse relationships for every explicitly established relation (e.g., the relationship '*is participant in*' is inferred when the relationship '*has participant*' is represented between two individuals in the ontology, and vice versa), DINTO allows for inferring new relationships that provide a more detailed description of the DDI mechanism and the effect. In this way, it is inferred that '*rifampicin*' increases the metabolism of '*cyclosporin A*' (mechanism) and decreases its '*immunosuppressant effect*' (effect). Regarding the *mirtazapine-propafenone* example, it is inferred that '*mirtazapine*' decreases the metabolism of '*propafenone*' (mechanism) and increases its adverse effect '*bradycardia*' (effect). Finally, it is inferred that '*naloxone*' decreases the CNS depression induced by '*morphine*' (effect).

Therefore, besides proving the consistency of the ontology, this technical evaluation has shown other important aspects. In this way, the results prove that the CM in the ontology is a correct representation of different types of DDIs including those occurring by a PD mechanism and those occurring by a PK mechanism. Moreover, the CM enables the representation and inference of important aspects of the DDI domain, including how a drug alters the effect of another one, how it modifies a PK process, or the type of consequence – beneficial or harmful – for patients. These findings are further confirmed in the application-based evaluation in **Chapter 9**.

8.1.2 Supporting or answering of previously established competency questions

During the specification activity, we have created a set of CQs in natural language representing the aspects that the final ontology should address (see **Section 7.1.1**). Therefore, they constitute a useful resource to check if the ontology represents all that information required in first development stages or, in other words, if the ontology is complete.

Ontology evaluation by identifying a set of CQs was firstly introduced by (Grüninger & Fox, 1995) and later on included in the methodologies for ontology creation METHONTOLOGY (Fernández-López et al., 1997) and NeOn Methodology (Suárez-Figueroa et al., 2012). Therefore, it is a key step in the creation of DINTO. During the development process, CQs are used to identify the information that should be included in the ontology in an iterative process. Finally, the list of CQs is used to evaluate the completeness of the ontology using specific examples of DDIs extracted from the pharmacological literature.

To show if content in the ontology satisfies our pre-established requirements, we use the same three DDI examples as in previous section (CS1, CS2 and CS3), which are a good representation of how DDIs are represented in text, and that include DDIs occurring by different mechanisms. **Annex 7**, **Annex 8**, and **Annex 9** show the CQs lists, the corresponding answers, and the corresponding axioms as an example of how the ontology corresponds to the CQs.

- **Results and conclusions**

To perform this evaluation we use the three DDI examples represented at the individual level in the ontology on the basis of the textual information shown in CS1, CS2, and CS3. As it is shown in **Annex 7**, **Annex 8**, and **Annex 9**, most CQs could be answered at least in one example.

This exercise shows that important DDI-related information is explicitly represented as classes in the current version of the ontology. These include interactions between two drugs (CQs 1), the type of mechanism leading to the DDI (CQs 3, 5-7), and drug-protein relationships (CQs 24-27). In contrast, other information is not explicitly described at the class level, but it can be manually added at the individual level without leading to inconsistences. This includes information related to the effect of the DDI (CQs 2, 8-11),

specific alterations in PK parameters (CQ 4), the characteristics of the interacting drugs (CQs 22), the source of information describing the DDI (CQs 29-33), patient characteristics (CQs 18-21), and recommendations (CQs 12-17). This demonstrates that the CM is complete and flexible enough to allow for the representation of this knowledge, although specific drug and DDI-related facts have not been included yet in the ontology. As we have established in the maintenance strategy for DINTO (**Section 7.1.8**), this information will be consecutively included in further versions of the ontology.

There are two questions, however, that could not be answered in any of the examples. These questions are related to the topics of description of an alternative non-interacting drug (CQ 23) and the characterization of interacting drugs as agents with narrow therapeutic index (CQ 28). Although this information is relevant to the DDI domain, it has not been included it in the current CM. Therefore, this task should be addressed during maintenance activities in order to include these aspects in next versions of the ontology.

8.1.3 Peer-review or human performed evaluation

To evaluate the content and structure of DINTO, we perform a manual review to ensure that the ontology meets a set of predefined criteria, standards or requirements (Lozano-Tello & Gómez-Pérez, 2004). These requirements are adopted from three main research efforts that have identified those characteristics desirables for reusable and high quality ontologies (Cimino, 1998; Fernández-López et al., 1997; Smith et al., 2007). A brief description of them is provided below.

1. The *Ontology Engineering Research Group*: The *Ontology Engineering Research Group* at "*Universidad Politécnica de Madrid*" has worked during the last years on the creation of methodologies for domain-independent ontology development (Gómez-Pérez et al., 2004; Suárez-Figueroa et al., 2012). They have studied and described ontology evaluation aspects to ensure the quality of an ontology from a technical point of view (Gómez-Pérez, 1999). In this way, they identified a set of criteria that should be evaluated in a given ontology (consistency, completeness, conciseness, expandability, and sensitiveness), and a set of possible errors that can be made by ontologists when building taxonomic knowledge into an ontology (circularity, partition, redundancy, and semantic errors).

2. *The OBO Foundry:* The OBO Foundry (Smith et al., 2007) is a collaborative effort based on the voluntary acceptance by its participants of an evolving set of principles specifying best practices in ontology development. These principles are designed to foster interoperability of ontologies within the biomedical domain, and to ensure a gradual improvement of quality and formal rigor in ontologies in ways designed to meet the increasing needs of data and information integration in the biomedical domain. The OBO Foundry principles establish that new ontologies should: *i)* be developed in a collaborative effort, *ii)* use common relations that are unambiguously defined, *iii)* provide procedures for user feedback and for identifying successive versions, and *iv)* have a clearly bounded subject matter. The long-term goal of the OBO Foundry is that the data generated through

biomedical research should form a single, consistent, cumulatively expanding and algorithmically tractable whole.

3. Desiderata for controlled medical terminologies: Cimino et al. described a set of delineated desirable characteristics or desiderata for controlled medical terminologies, which attempted to summarize consensus regarding structural issues of such terminologies (Cimino, 1998). The desiderata covers aspects such as content, structure, and naming conventions and provides a detailed description of the most important characteristics needed to make controlled vocabularies sharable and reusable, that can be made extensible to biomedical ontologies (Cimino, 2006).

The review of their recommendations and their combination has led to the construction of an evaluation template establishing the requirements that our ontology should adhere to. This template (see **Annex 10**) is used as a guide during the development of DINTO and as a checklist for the evaluation of the final ontology. In this way, the final version of DINTO is evaluated and refined until complete adherence to all evaluation criteria included in the template are addressed.

- **Results and conclusions**

The resulting evaluation template is used as a guide during the development process and all requirements are considered during the creation of DINTO. Once the final version of the ontology is created, we perform a deep evaluation against the template requirements. Final version has not been considered correct until all of them have been met. The final evaluation template for DINTO is shown in **Annex 10**.

As shown in the template, all the requirements have been fulfilled. Naming conventions and use of numerical labels for URIs are only violated for those entities imported from PKO and BRO. However, we maintain them in order to preserve the original source. Analysis of inconsistences is carried out using ontology reasoner engines, and manual review of incompleteness and redundancies is performed, too. Finally, it is important to note that DINTO has been developed in collaboration with an important OBO Foundry member, the ChEBI ontology.

8.1.4 Conclusions

In this chapter, we have introduced ontology evaluation and have presented our strategy to validate the quality of DINTO from a technical and application-based perspective.

The technical evaluation described here consists of three different approaches. The first one is classification scenario testing, which has shown that the knowledge represented in the ontology is consistent and complete enough to represent three real DDI examples. The second one, supporting or answering previously established CQs, has proven that the ontology is complete and covers the aspects established in the specification activity. Finally, the third technical evaluation approach is the manual evaluation of the content and structure of the ontology following a set of predefined

requirements. Supported by a specifically created evaluation template, we have shown that all the requirements have been met by the final version of DINTO.

In the next chapters, we describe two different application-based evaluations of DINTO that, in combination with the technical evaluation described here, provide a comprehensive description of the consistency, quality, and re-usability of the ontology.

Chapter 9

Inference of DDIs and their mechanisms

In the framework of this thesis, we propose the creation of a new ontology for the NLP research community working within applications in the DDI domain. However, high quality and robust ontologies are those that perform well in multiple tasks and that can be reusable across different applications. To assess its "robustness", the ontology should be evaluated on multiple heterogeneous tasks (Hoehndorf et al., 2012). Therefore, we decide that a complete evaluation of DINTO should include the assessment of its performance in an additional application.

Two main reasons have driven our selection of computational inference of DDIs as an alternative application scenario. First of all, during the study of current efforts in DDI-knowledge modeling (**Section 6.4**) we have observed that, besides NLP, prediction of DDIs has been the other main application of these projects (Arikuma et al., 2008; Boyce et al., 2010b; Imai et al., 2013). These research groups have demonstrated that the formal representation of pharmacological information can be successfully applied to the inference of DDIs on the basis of their underlying mechanisms. Therefore, this additional evaluation approach could provide a supplementary method to those described in **Chapter 8** to assess if the CM in our ontology is correct and complete.

In addition to this, computational inference or prediction of DDIs is a research field that has attracted a great deal of attention during the recent years. Many different approaches have been proposed to achieve this goal: *i)* extrapolation of *in vitro* data to the

in vivo situation; *ii)* similarity-based methods and data mining; *iii)* text mining of scientific literature; and *iv)* knowledge representation and reasoning.

Firstly, some authors have adopted traditional approaches using *in vitro* data to predict *in vivo* DDIs (Zhang, Zhang, Zhao, & Huang, 2009) in order to automatically obtain predictions for a larger number of drugs (Bonnabry et al., 1999; Fahmi et al., 2009; Ito et al., 2004; Quinney et al., 2010). In contrast to this one, that extrapolates *in vitro* information to *in vivo* situations, similarity-based methods exploit pharmacological information, such as molecular structure or target similarity, to identify potential DDIs (Gottlieb et al., 2012; Vilar et al., 2012, 2014). Mostly, these projects are based on data mining. For example, within this group several research groups have studied the use of ADRs data from spontaneous report databases to identify new DDIs (Harpaz, Haerian, Chase, & Friedman, 2010; Thakrar et al., 2007; van Puijenbroek et al., 2000).

Besides these approaches, text mining has emerged in the recent years as a new and promising approach to identify DDIs from texts (Duke et al. 2012; Tari et al. 2010; Segura-Bedmar, Martínez & de Pablo-Sánchez 2011a; Segura-Bedmar, Martínez & de Pablo-Sánchez 2011b; Zhang et al. 2012) that hence can be used in combination with other methods, such as Tari et al. (2010), who combined text mining and automated reasoning techniques, or Percha & Altman (2012), who combined text mining and semantic network representation.

Finally, pharmacological knowledge representation and reasoning has been widely used for the prediction of DDIs. It can be divided into two main subtypes: *i)* those using semantic networks or *ii)* those based on DL representations and/or rules. Semantic networks − representing drugs as nodes and interactions between drugs as edges or arcs − have been used by different authors to infer DDIs through reasoning (Cami et al., 2013; Guimerà & Sales-Pardo, 2013; J. Huang et al., 2013; Percha & Altman, 2012; Takarabe, Shigemizu, Kotera, Goto, & Kanehisa, 2011). However, there is an increasing interest on more expressive representation formalisms (Brochhausen et al., 2014), such as DL and/or rules, which support additional inferences and have proven to be useful to predict different types of DDIs (Imai et al., 2013; Moitra, Palla, Tari, & Krishnamoorthy, 2014; Tari et al., 2010; Wu et al., 2013; Yoshikawa et al., 2004).

For all these reasons, we consider that inference of DDIs – that in the framework of this project is defined as the automatic identification of DDIs or DDI-related knowledge, such as type of mechanism, from implicit information in the ontology − is an adequate application to evaluate DINTO. In this chapter, we provide a review of the projects that have studied the inference of DDIs using DL or rule-based representation of DDI knowledge (**Section 9.1**), analysing their unresolved issues and the possible contributions of our work. In **Section 9.2**, we describe three different experiments performed to evaluate the inference capabilities of DINTO. Finally, main conclusions of this chapter are provided in **Section 9.3**.

9.1 Related work on DDI prediction using DL, rules, and reasoning

In the previous section, we have provided a broad review of the state of the art in computational inference of DDIs. Here, we focus on those projects relying on knowledge representation as DL and/or rules to infer DDIs. To the best of our knowledge, five research groups have exploited this type of formally represented information to the inference of DDIs on the basis of their underlying mechanisms.

Arikuma et al. (2008) used the formal representation of PK molecular events of drugs in the OWL ontology **Drug Interaction Ontology (DIO)** and an inference program based on drug interaction rules to infer PK DDIs. In their experiment, they predicted four different mechanisms leading to a PK DDI occurring between two drugs (*irinotecan* and *ketoconozale*), that were considered as four different DDIs. These DDIs were manually evaluated against pharmacological literature. Two of them were confirmed by published papers, while the other two could not be supported by current literature and were considered to be *negligible* by the authors – i.e., clinically unobservable or irrelevant when the drug doses are low.

In the framework of the **Drug Interaction Knowledge Base (DIKB)** project, Boyce et al. (2010) conducted an experiment to infer DDIs occurring via a specific PK mechanism. They combined *i)* the drug-related facts or *assertions* manually curated in the DIKB about a total of 35 drugs and drug metabolites and their relationships with enzymes, *ii)* a set of FOL rules, called *rule-based theory*, of how drugs interact by metabolic inhibition and *iii)* the *evidence-base*, which provides evidence for, or against, each one of the assertions and is used to distinguish between relevant and negligible DDIs. A total of 586 possible interacting pairs were assigned a value *true* or *false* regarding the prediction provided by an inference program. To evaluate the inferences, the authors created an experiments' validation set, constituted by 48 interacting pairs found to be described in the pharmacological literature as interacting or non-interacting pairs. The remaining 538 pairs were not found in the literature. Therefore, it is not possible to assert if they are non-interacting pairs or they have not been described yet. A total of 65 DDIs were predicted by the system; 34 of them were in the validation set and therefore considered as correct inferences, while 31 predicted DDIs belonged to the group of 538 pairs not known to be interacting or non-interacting drugs.

DIKB information was reused in a different project. **Moitra et al**. (2014) proposed a system that integrates the capabilities of semantic modeling and temporal reasoning to identify quantitatively PK DDIs occurring via a metabolism-related mechanism. DDI information was modelled in the Semantic Application Design Language (SADL),[53] which was then translated to OWL models. The semantic model incorporates information from the DIKB, and the logic rules for DDI inference were represented in the declarative logic programming language Answer Set Programming (ASP) (Niemelä, 1999). To evaluate the system, the authors studied the interactions between three drugs (*fluoxetine*, *clozapine,* and *olanzapine*) and concluded that *fluvoxamine* had a greater impact on the metabolism of *clozapine* than on the metabolism of *olanzapine.* Manual review of the

[53] http://sadl.sourceforge.net/sadl.html

literature confirmed this results, although evaluation in a larger set of drugs was not performed yet.

Similarly focusing on metabolism-related DDIs, **Tari et al.** (2010) combined a logical representation of the domain, text mining techniques, and automated reasoning to infer DDIs. Drug-related facts, such as drug-protein relationships, were extracted through NLP techniques from MEDLINE abstracts and stored in a parse tree database. DDI domain knowledge representing how a drug alters the metabolism of another one was represented as AnsProlog (Gelfond & Lifschitz, 1991) logic rules, a declarative language useful for reasoning, and an AnsProlog solver called *clingo* (Gebser, Ostrowski, & Schaub, 2009) was used to compute the inferences. To evaluate this approach, 496 DDIs described in DrugBank between 295 drugs were used as a gold-standard. A total of 979 DDIs were inferred from drug-facts extracted from texts, of which 123 were included in the gold-standard, and 856 were not. A sample of 345 of these remaining DDIs (40%) was manually reviewed to evaluate if there was evidence supporting them − that is, if the extracted drug-protein relationships were consistent with a possible mechanism underlying the DDI −, and it was found that this was the case for 286 (82%) of the reviewed DDIs.

In contrast to the former projects, Imai et al. (2013) addressed the inference of PD DDIs. To do this, they manually included drug-related facts in the **Pharmacodynamics Ontology (PDO)** as subordinate classes for 89 drugs related to *noradrenaline*. They identified 14 different types of PD DDIs that could occur between two drugs, and checked how many pairs could lead to a DDI by one or more of these mechanisms. In contrast to the previous projects, Imai et al. did not implement any inference system, and inferences were identified manually in accordance with the information represented in the ontology and the different types of DDIs established. To evaluate the predictions made, they selected 22 drugs leading to 231 predicted DDIs and checked if they were described in the Japanese SPCs. They found descriptions for 72 of the DDIs, while the remaining 159 were not mentioned.

These five projects demonstrate that formal representation of drug-related facts and DDI knowledge can be combined to infer interactions between drugs. However, one of the main limitations encountered by most of them is the low coverage of drugs included in their experiments. Indeed, the **PDO** was used to infer DDIs between 89 drugs and the **DIKB** predicted interactions between 35 pharmacological substances, while **Moitra et al.**'s approach and the **DIO** were used to predict the interaction between only three and two drugs, respectively. The main reason for this low coverage is that all these methods rely on manual curation to identify, gather, and structure the drug-related facts required as basic information to infer the DDIs. Although expert manual curation provides high quality information, this activity is both cost-intensive and time-consuming. Moreover, new pharmacological information is discovered and published every day in the scientific literature, which makes keeping up to date a knowledge base with this information a difficult task. As an alternative to manual curation, **Tari et al**. automatically extracted drug-related facts through text mining and could test their system on 295 drugs. However, a different solution could be brought by the increase in pharmacological information stored in structured and machine-readable formats as public databases and knowledge bases that has emerged during the last years (Khelashvili et al., 2010; Whirl-Carrillo et al., 2012). Exploiting this information and integrating it automatically in a knowledge base could overcome the limitation of manual curation.

Another limitation of these works is that none of them dealt with the inference of different types of DDIs in the same framework. On the one hand, the projects using the **DIO** or the **DIKB**, as well as **Moitra et al**.'s and **Tari et al**.'s projects, focused on PK DDIs. More specifically, the four latter addressed only the inference of metabolism-related DDIs. On the other hand, the experiment performed using the **PDO** included only PD DDIs. Therefore, it has not been demonstrated yet if the same CM and inference system can be used to predict, at the same time, DDIs occurring via both PK and PD mechanisms and their respective subtypes.

Evaluation of the inferred results was a challenging task, too. Most of the projects performed a manual review of pharmacological literature to evaluate their inferences. The only exception was **Tari et al**., who created a gold-standard from DrugBank to compare their inferences, and then performed a manual evaluation of the non-matching results. However, manual review leads to three main issues. Firstly, this is a time consuming task that cannot be applied to a great set of inferred DDIs. Indeed, Arikuma et al. could only evaluate the interactions predicted based on **DIO** for 22 representative drugs since the number of all combination of the original 89 drugs was too large. Secondly, manual review of the large pharmacological literature can introduce bias in the evaluation, since many clinically-relevant DDIs could be missed during the review process or they could be not reported in the consulted sources (Boyce et al., 2010b). Thirdly, it is not possible to know when a DDI is not described in the literature because there is not an interaction between the interacting pair or because, although this interaction exists, it has not been described yet (Tari et al., 2010). Therefore, the evaluation of *false positives* − or incorrect inferences − and *false negatives* − true information that, however, has not been inferred − can never be performed in a comprehensive way.

In short, the main unresolved issues are:

1. Low coverage of drugs included in the experiments.
2. Representation of only one interaction type: PK or PD DDIs.
3. Manual evaluation in most cases.

With our experiment, we attempt to contribute to the research in the field of computational inference of DDIs by overcoming the limitations encountered by former efforts. To do this, firstly we automatically integrate pharmacological information available in databases and ontologies instead of manually curate it in our ontology. With this strategy, we expect that the coverage of drugs will be larger than in previous projects. Secondly, we address the inference of both PK and PD DDIs in order to know if they can be predicted in the same representation framework. Thirdly, in order to obtain a more detailed description, different subtypes of DDI mechanisms, such as *enzyme inhibition*, *antagonism,* or *transporter induction*, are represented. Therefore, we expect that our representation of DDI-knowledge will be correct and complete enough to infer both DDIs and the specific mechanisms preceding them.

Another contribution of our DDI-inference experiment is that we employ exclusively resources and tools for the semantic web and the ontology engineering field. It has been

postulated that *"having the inference mechanism and the descriptive knowledge combined under the same syntactic structure provides means for interoperability of rule systems"* (Lezcano et al., 2011). Moreover, some authors consider that using ontologies in conjunction with rules is a major challenge for the Semantic Web (Golbreich, 2005). Therefore, in contrast to the previous efforts that represented differently drug-related facts and inference rules, we use only the Web Ontology Language OWL 2 and its extension the Semantic Web Rule Language (SWRL) to formally represent all the information required. Our supporting tool during the whole process is the ontology editor Protégé and, instead of implementing a new inference program, we use existing ontology reasoner engines compliant with the OWL semantics.

Therefore, this experiment provides a unique and novel framework for the assessment of the contributions that semantic web technologies and ontology engineering can provide in projects requiring *i)* the storage of large amounts of structured information, *ii)* a formal and highly expressive representation of complex processes, and *iii)* powerful inference capabilities to apply machine reasoning to large amounts of semantic data.

9.2 Inference of DDIs and DDI mechanisms using DINTO

To evaluate the inference capabilities of DINTO, we have designed three different experiments. In the first experiment (*IExp1*), we analyse how a reasoner classifies known DDIs by using only explicit information about their mechanisms. The aim is to check if the information explicitly represented in the ontology is consistent, and if the relationships established between DDIs and their mechanisms allow for their correct classification. However, inferences that are more complex rely on the formal representation of DDI mechanisms. For this purpose, we have created two sets of SWRL rules to infer DDIs and their mechanisms (see **Section 7.2.5**). Therefore, in the second experiment (*IExp2*) we validate if the combination of the first set of rules and drug-protein relationships in DINTO can be used to infer new DDIs. Finally, in the third experiment (*IExp3*) we combine the second set of rules to infer both, DDIs and their mechanisms. With this third experiment, we validate if the formal representation of DDIs mechanisms by means of drug-protein relationships can be used to identify automatically the underlying mechanism of a DDI.

In this section, we firstly introduce ontology reasoner engines and discuss their current limitations, and then we describe the methods, results, and evaluation of the three inference experiments, providing a brief discussion of each one of them.

Ontology reasoning plays a key role in ontology engineering and has become an indispensable activity to help ontologists during the development and evaluation of ontologies, as well as users to query their contents. It is performed by a reasoner engine, a program that infers logical consequences from a set of explicitly asserted facts or axioms (Cuenca, 2011). With the advancement of ontology languages, many ontology reasoners have been developed. Some of them are available as plugins for the ontology editor Protégé. This provides a user-friendly interface for the study of the resulting inferences.

Some of the most widely used ontology reasoners are FacT++ (Tsarkov & Harrocks, 2006b), Pellet (Sirin et al., 2007), or HermiT (Glimm et al., 2014). The three of them support the features of OWL 2 ontology language. However, as we explained in **Section 7.1.7**, OWL 2 cannot express all type of relations and its expressivity can be extended by adding SWRL rules, which are supported by Pellet and HermiT.

Ontology reasoner engines facilitate the following tasks:

- Checking the consistency of the ontology: Reasoners check if the ontology axioms conform to the underlying logical theory of the ontology language and detect any inconsistency within it (Plessers & De Troyer, 2006). We use the reasoner engines FacT++ and HermiT to check the consistency of our ontology, as we explain in **Chapter 8**.

- Classification: Implicit classification of classes and individuals can be obtained when leveraging a reasoner engine. Different axioms can lead to these inferred classifications: *i)* the '*is a*' relationships that shows a taxonomical classification of classes and, therefore, of their individuals; *ii)* OWL definitional axioms or '*equivalent to*' assertions; *iii)* object properties domains and ranges, which are used as axioms in reasoning; *iv)* specific SWRL rules.

- Inference of relationships between classes or individuals: New relationships can be obtained when a reasoner engine is used in an ontology. They are inferred when some of the following axioms have been included in the ontology: *i)* object properties *characteristics*, such as transitivity or symmetry; *ii)* '*inverse of*' axioms between two object properties, that infers inverse relationships; *iii)* property chains, which infer a relationship *x* between two classes *a* and *c* when they are linked by two specific relationships *y* and *z* and another class *b* (**Figure 9.2**); *iv)* specific SWRL rules.

- Answering queries over ontology classes and instances: Reasoners can be used to query and search the ontology content. They are capable of finding more general or specific classes or retrieving individuals or triples matching a given query.

Although current ontology reasoner engines perform all the activities mentioned above, their performance with very large and complex ontologies is still compromised (Cuenca, 2011; Holford et al., 2010). As we have observed during the development of DINTO, reasoners can crash with a consistent ontology when: *i)* it has a very large TBox (i.e., the terminological box which handles the axioms around classes.); *ii)* it has a large ABox (i.e., the assertional axioms or assertional box used to assert the truth about individuals); or *iii)* it has a complex level of DL expressivity. Due to the complexity of our ontology (DL $\mathcal{ALCRIQ}$(D)), the large number of classes and properties that it covers (25,809 classes and 73 object properties), and the large number of corresponding individuals required to infer DDIs, we cannot use any of the available OWL reasoners with the final version of DINTO. Therefore, in order to test its consistency or to infer new information, we adopt different strategies to simplify the ontology while maintaining that information necessary to retrieve the desired inferences (Holford et al., 2010). Some of these strategies include reducing the number of classes, relationships, or individuals, for

example. In order to ensure the reproducibility of our results, we describe in detail the modifications done for each experiment, and the corresponding files are available at https://code.google.com/p/dinto/.

9.2.1 *IExp1*: Classification of DDIs on the basis of explicitly asserted mechanisms

The objective of this experiment is to evaluate if the ontology is consistent when classifying DDIs imported from DrugBank on the basis of their asserted mechanisms as PK or PD DDIs.

- **Methods:**

DINTO 1.0.0 includes a total of 11,555 classes representing DDIs that have been imported from the database DrugBank. Information regarding their mechanisms (*'target activity alteration'*, *'enzyme activity alteration'*, *'transporter activity alteration'*, or *'non-absorbable complex formation'*) has been imported as well when it was provided by the original source (**Section 7.1.3**). In this way, the interaction between the drugs '*abciximab*' and '*tirofiban*' is described in the ontology to be preceded by a *'target activity alteration'*, which is a type of PD DDI mechanism.

We have established in DINTO that all those classes preceded by a PD DDI mechanism belong to the class PD DDI. In a similar way, we have defined that every class preceded by a PK DDI mechanism is a PK DDI. In other words, any individual of the '*pharmacodynamic DDI*' class is equivalent to (≡) any individual that has a '*is preceded by*' relationship with at least one (*some*) individual of the class '*pharmacodynamic DDI mechanism*'. Therefore, when running an ontology reasoner, DDI classes are classified regarding their mechanisms as PD or PK DDIs. In this way, the interaction between *abciximab* and *tirofiban* should be classified as a PD DDI, since it is described in the ontology to be caused by the alteration of the activity of a target.

'pharmacodynamic DDI' ≡ *'is preceded by' some 'pharmacodynamic DDI mechanism'*

'pharmacokinetic DDI' ≡ *'is preceded by' some 'pharmacokinetic DDI mechanism'*

Performing this reasoning task requires a reduction on the size of the ontology, while maintaining all the DDIs imported from DrugBank and their mechanisms. We create a reduced version of DINTO including only the top classes *'DDI mechanism'* and *'drug interaction'* and their subclasses. The resulting inferred classification of DDIs has been exported as a new version that can be imported in DINTO 1.0.0 and downloaded from https://code.google.com/p/dinto/.

- **Results:**

We use the reasoner engines HermiT 1.3.8 and FacT ++ to classify the ontology, which show that there are no inconsistencies. A total of 1,101 DDI classes are classified as PD DDIs, while 5,711 are classified as PK DDIs (**Figure 9.1**). It is important to note that some DDIs are described in DrugBank to be preceded by both a PK and a PD DDI mechanism. This is the case of the interaction between '*didanosine*' and '*zalcitabine*', which is preceded by a *'target activity alteration'* and a *'transporter activity alteration'*. Therefore, the class *'didanosine/zalcitabine DDI'* is classified as both PD and PK DDI. This information is correct from a pharmacological perspective since, although most DDIs are frequently assigned a type PK DDI or PD DDI, there could be situations where both types of mechanisms can lead to the occurrence of DDIs (Baxter, 2013).

Similarly, the same pair of interacting drugs can be classified at the same time as two different subtypes of PK DDIs (e.g., *'transporter related DDI'* and *'enzyme related DDI'*). Due to this fact, from the 5,711 PK DDIs, 5,283 are classified as *'enzyme related DDI'* and 1,673 are classified as *'transporter related DDI'*. Finally, 128 DDIs are classified as *'non-absorbable complex formation DDI'*. None of them is classified, however, as *'carrier related DDI'*, since the mechanism *'carrier activity alteration'* is not assigned in DrugBank to any DDI. Finally, those DDIs for which any DDI mechanism is established in DrugBank are not classified in the ontology. These results are summarized in **Table 9.1**.

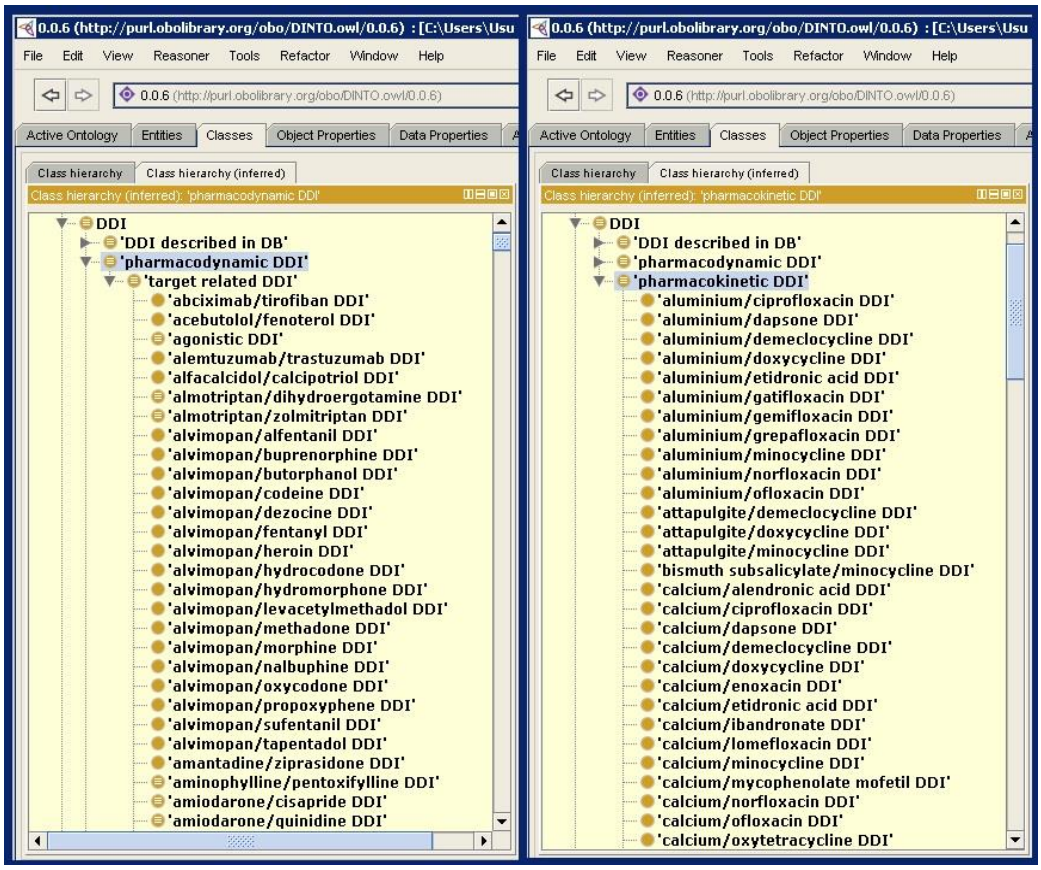

Figure 9.1. Protégé screenshot showing the inferred class hierarchies for PD and PK DDIs

Classification of DDIs	#
PK DDIs	5,711
PD DDIs	1,101
PK DDIs + PD DDIs	659
enzyme related DDIs	5,283
transporter related DDIs	1,673
non-absorbable complex formation DDIs	128
enzyme related + transporter related DDIs	1,367
enzyme related + target related DDIs	631
enzyme related + transporter related + target related DDIs	140
transporter related + target related DDIs	166
unclassified DDIs	6,061

Table 9.1. Results of the classification of DDIs on the basis of their asserted mechanisms.

- **Discussion:**

This experiment shows that the conceptualization and implementation of DINTO is a correct representation of the relations between DDIs and their mechanisms. Explicit information of the mechanism of a DDI imported into DINTO from DrugBank is used to obtain a classification of DDIs as *'PK DDI'*, *'PD DDI'*, *'enzyme related DDI'*, *'transporter related DDI'*, *'non-absorbable complex formation DDI',* or *'target related DDI'*. However, more specific descriptions, such as if the interaction is due to the *induction* of an enzyme, the *inhibition* of a transporter or the *antagonistic* effect of two drugs on the same target, cannot be obtained with that knowledge.

However, we believe that other knowledge represented in the ontology, such as drug-protein relationships, can be exploited to obtain a more detailed classification of DDIs based on their mechanisms. This premise is tested in the other two experiments *IExp2* and *IExp3*.

9.2.2 *IExp2*: Inference of DDIs

The purpose of this experiment is to determine if the knowledge represented in the ontology can be used to infer new DDIs on the basis of their mechanisms. To do this, we test two different approaches, which are described in this section: *i)* inference of DDIs using property chains and *ii)* inference of DDIs using SWRL rules.

- **Methods:**

Our first approach to infer new DDIs is the creation of different property chains describing the relationship *'may interact with'*. Property chains are a feature in OWL 2 used to assert a single property based on the existence of several properties. Through the

use of these chained properties, we represent ordered pharmacological events (Herrero-Zazo et al. 2013). An example of one of this property chains is shown in **Figure 9.2**.

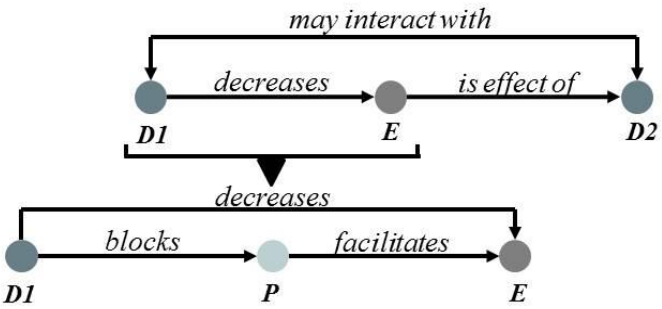

Figure 9.2. Property chain of the relationships *'may interact with'* and '*decreases'*

In this example, the object property *'may interact with'*, which is a symmetric relationship that has domain and range the class '*pharmacological entity*', is a subproperty of the property chain:

'decreases' o *'is effect of'* → *'may interact with'*

At the same time, the object property *'decreases'* is a subproperty of different chained properties, such as, for example:

'blocks' o *'facilitates'* → *'decreases'*

This means that if a drug *D1* blocks the activity of a protein *P* that, at the same time, facilitates an effect *E*, the drug *D1* decreases the effect *E*. Meanwhile, if this effect *E* is the effect produced by another drug *D2*, then the drug *D1* may interact with drug *D2* and vice versa.

These and other inferences are successfully obtained when representing different real examples of DDIs at the individual level. These examples are described in detail in a conference proceeding presented at the International *Semantic Web Applications and Tools for Life Sciences* (SWAT4LS) Workshop in Edinburgh, Scotland (Herrero-Zazo et al. 2013), where we describe our first steps and achievements in the construction of DINTO.

However, those examples are always performed representing only two drugs at the individual level. When we try to extend this approach to a version including a larger number of individuals and their relationships, we observe that the use of property chains leads to incorrect inferences, such as that every drug interacts with itself or erroneous *'may interact with'* relationships between unrelated pharmacological entities. Therefore, from this approach we conclude that the inference of DDIs by representing pharmacological mechanisms as property chains is not appropriate to infer DDIs between multiple pharmacological entities.

As an alternative to property chains, we decide to create rules representing DDI mechanisms and use them to infer new *'may interact with'* relationships between pharmacological entities. Boyce et al. (2007) demonstrated that a set of rules in FOL could represent how one drug alters the metabolism of another drug. The same PK DDI mechanism was represented by Tari et al. (2010) and by Moitra et al. (2014), who chose the logic programming language ASP (Bonatti, Calimeri, Leone, & Ricca, 2010). Although use of logic programming in combinations with ontologies can be useful, it was not conceived as an ontology language for direct interchange of knowledge, which hinders the interoperability required by the semantic web and ontologies (Hitzler, Krötzsch, & Rudolph, 2009). In contrast to them, in our approach we use the Semantic Web Rule Language (SWRL),[54] an expressive rule language built on the same description logic foundation as OWL. Moreover, instead of focussing on a specific DDI mechanism, we address the representation of different types of mechanisms leading to both PK and PD DDIs on the basis of object properties represented in the ontology. The result is a total of 59 SWRL rules representing the different pharmacological processes involved in PK and PD DDI mechanisms (**Section 7.2.5**).

The first step is to create a version of the ontology suitable for use with a reasoner supporting SWRL. As we explained before, reasoner engines cannot process the whole version of DINTO due to its large size. Moreover, when including SWRL rules, leveraging a reasoner without crashing becomes an even more difficult task. Therefore, we adopt different strategies in order to reduce the size and complexity of the ontology. Other authors, such as Holford et al. (2010), have employed this method, too. In our case, we eliminate all OWL definitional axioms (or '*equivalent to*' axioms), simplify the object properties hierarchy and delete their characteristics, and eliminate all classes but subclasses of the '*pharmacological entity*' and '*protein entity*' top classes. Even after this reduction in size, the ontology is too large to be used with a reasoner and the 59 SWRL rules. The same results are obtained when experimenting with further versions of DINTO containing the half or containing a quarter of the total 8,786 pharmacological entities in DINTO. Finally, with a version reduced to only 607 pharmacological entities and 429 proteins, and the 59 rules, the reasoner HermiT 1.3.8 can classify it and show the inferences. However, processing the whole ontology through this strategy would require the creation of 32 reduced subsets that should be combined 496 times, and would require the exclusive dedication of a person during almost 9 months.

Given this situation, we decide to create a reduced version of the ontology for the evaluation of DINTO in the two different applications proposed in this thesis: NLP and inference of DDIs. To do this, we create a version of the ontology containing only those pharmacological entities mentioned in the DDI corpus test dataset. This version is necessary to conduct the evaluation approach described in **Section 10.2** and to compare the performance of an IE system when it exploits only known DDIs versus using known and inferred DDIs. This version of the ontology includes 426 pharmacological entities and 752 protein entities having at least one relationship with any of the included drugs. It is important to note that, in order to evaluate the coverage for drugs of DINTO, those pharmacological entities mentioned in the corpus but not presented in the original version are not included in the new one. Therefore, although it is not possible to obtain an inferred version of the whole ontology, we can obtain a proof-of-concept for our hypotheses in both applications NLP and inference of DDIs.

[54] http://www.w3.org/Submission/SWRL/

Inferences from SWRL rules are made in the ABox, that is, at the individual level. Therefore, the second step to infer new DDIs is to automatically create individuals for every class in the ontology – '*pharmacological entity*' and '*protein entity*' subclasses – and the corresponding relationships among them. OWL semantics are based on the open world assumption (OWA), or the assumption that what is not known to be *true* must be *unknown* but not false (Groppe, 2011). Therefore, it is necessary to establish that all the individuals are distinct among them. This functionality is provided by Protégé and can be easily carried out using the ontology editor.

The third and final step consists in importing the file containing the SWRL rules into this version of DINTO, and leveraging the reasoner.

- **Results:**

It takes 3.86 days for the reasoner HermiT 1.3.8 to classify the ontology. A total of 59,696 new *'may interact with'* relationships are inferred. However, not all of them are established between two drugs, and some relationships are inferred to occur between proteins. We automatically rule out those relationships between proteins, although in further versions it is possible to modify the SWRL rules to avoid these nuisance inferences.

Using a functionality provided by Protégé, we can export the inferred axioms as a new ontology. Thereafter, we automatically translate the *'may interact with'* relationships between two individuals to the corresponding class level and create 10,780 new DDI classes of the type *'drugA/drugB DDI'*. Finally, all the individuals are eliminated. The result is an inferred ontology with 21,560 *'may interact with'* relationships and 10,781 inferred DDI classes. This version can be downloaded from https://code.google.com/p/dinto/.

- **Evaluation:**

To evaluate the results of this experiment, we compare the inferred DDIs (from now on named the inferred set *I*) with the DDIs in DrugBank involving some of the 426 drugs included in the ontology (the asserted set *A*). The *I* set consists of 10,780 inferred DDIs, while the *A* set includes 2,245 asserted DDIs. There is a total of 656 DDIs common to both sets. Therefore, the 29% of the DDIs in DrugBank have been inferred in DINTO. These results are summarized in **Table 9.2**.

Due to the large number of DDIs resulting from this experiment, it is difficult to study in detail the above results. However, we can study as well the coincidences and differences between the *I* and *A* sets regarding not to the DDIs, but to the drugs involved on them. In this way, we observe that from the 426 drugs included in the experiment, only 219 participate in at least one inferred DDI, while only 309 drugs are involved in at least one asserted DDI. Specifically, there are 172 that are common to both sets. Therefore, we can consider them as drugs correctly described to participate in at least one DDI (*true positives*). On the contrary, there are 70 drugs that are not included in any of the two sets. They represent, therefore, those drugs for which any interaction has been incorrectly inferred (*true negatives*). Finally, 47 drugs are involved in at least one inferred

DDI only, while 137 drugs participate in at least one asserted DDI only. The former ones can be considered as *false positives* – i.e., drugs for which at least one DDI has been incorrectly inferred – and the latter ones represent *false negatives* – i.e., drugs for which at least one DDI has not been correctly inferred. These results are summarized in **Table 9.3**.

Coincidences between *I* and *A* sets (for the 426 drugs)

	DDIs in *A*	**DDIs not in *A***	**Total**
DDIs in *I*	656	10,124	10,780
DDIs not in *I*	1,589	-	1,589
Total	2,245	10,124	

Table 9.2. Number of DDIs in the inferred (*I*) and asserted (*A*) sets for the total 426 drugs

	#	**%**
Drugs coincident in both sets ***(true positives)***	172	40,38
Drugs not present in any of the sets ***(true negatives)***	70	16,43
Drugs only in the inferred set *I* ***(false positives)***	47	11,03
Drugs only in the asserted set *A* ***(false negatives)***	137	32,16
Total	426	100

Table 9.3. Comparison of the number of drugs in the inferred (*I*) and asserted (*A*) sets.

Finally, we focus our analysis on those DDIs involving the same drugs in both sets. With this approach, we can analyse in detail the disagreements between them. In this case, the new inferred set *I2* consists of 7,039 inferred DDIs, while the new asserted set *A2* consists of 815 asserted DDIs. The number of common DDIs in both sets is the same (656 DDIs). This means that the 80% of the DDIs in *A2* have been correctly inferred. The remaining 20% represent the *false negatives,* or those DDIs in *A2* that have not been inferred by our method. In contrast, those inferred DDIs that are not included in *A2* represent the *false positives.* **Table 9.4** summarizes these results.

We randomly select 15 *false positives* and 10 *false negatives* to perform a qualitative analysis of these results. On the one hand, we find evidence supporting all the studied *false positives*. This means that there is an underlying DDI mechanism, such as, for example, that one of the drugs inhibits one or more of the metabolizing enzymes of the other drug. As a complementary evaluation source we use the tenth edition of the DDI compendia *'Stockley's Drug Interactions'* (Baxter, 2013).[55] However, none of these *false positives* is included. Regarding *false negatives,* on the other hand, only the half of the studied DDIs is included in the compedia. We identify that this DDIs are not inferred in

[55] At the time of performing this evaluation the most recent edition is the tenth (2013) edition.

DINTO because they occur through mechanisms that cannot be represented based on currently known drug-protein relationships. A more detailed description is provided in the following subsection Discussion.

Coincidences between *I2* and *A2* sets (for the 172 common drugs)

	DDIs in *A2*	**DDIs not in *A2***	**Total**
DDIs in *I2*	656	6,383	7,039
DDIs not in *I2*	159	-	159
Total	815	6,383	

Table 9.4. Number of DDIs in the inferred (*I2*) and asserted (*A2*) sets for the common 172 drugs

- **Discussion:**

In this experiment, we have inferred a large number of DDIs by means of our knowledge representation and reasoning approach based on the information represented in DINTO. Due to the limitations associated to manual evaluation discussed in this chapter, we compare our inferences with the DDIs included in the database DrugBank.

Results show a high number of *false positives*, or inferred DDIs that are not included in the pharmacological database. Similar results, with a large number of *false positives,* were obtained by Tari et al. (2010), who compared their inferences with a gold-standard created from DrugBank, too. As they did, we find evidence supporting the inferred DDIs for our studied *false positives*, which means that there is a plausible mechanism that could produce the DDIs. This reveals that the SWRL rules created in DINTO are correct representations of the pharmacological processes underlying DDIs.

However, the elevated number of inferred DDIs not included in DrugBank can be explained by the fact that our SWRL rules have been modeled to infer DDIs on the basis of different mechanisms, but independently of other related facts, such as their significance or level of documentation. Therefore, for any pair of drugs for which an underlying DDI mechanism exists, no matter if it would lead to a clinically relevant DDI or a not-clinically relevant or unobservable DDI, the interaction is inferred. However, it would be incorrect to assume that all those DDIs not included in DrugBank are non-clinically relevant. Indeed, the evaluation against DrugBank has some limitations. First of all, DrugBank − or any other DDI information source − cannot be considered a gold-standard of "true DDIs" (Imai et al., 2013). Due to the large number of possible DDIs (that can have different degrees of significance or can be influenced by drug or patient-related facts), it is not possible to manually study, identify, and collect all of them in a unique information source. The most important compendia for drug-drug interactions, such as '*Stockley's Drug Interactions*' (Baxter, 2013), '*Drug Interactions Facts*' (Tatro, 2010) or '*The Top 100 Drug Interactions*' (Hansten & Horn, 2014), have an editorial board that establishes different inclusion criteria for the described DDIs (Brochhausen et al., 2014). Even the SPCs or PIs, the documents provided by drug manufacturers and approved by the administrative licensing authorities, have been reported to be inconsistent and/or incomplete regarding DDI information (Bergk et al., 2005). For all these reasons, it is not possible to know if the *false positives* DDIs are not included in DrugBank because *i)* they are interacting pairs but have not been included yet in the

database, *ii)* they are interacting pairs but the interaction between them has not been described in the scientific literature, or *iii)* they are not interacting pairs of drugs.

In the case of *false negatives* – or DDIs included in DrugBank but not inferred in DINTO – we have observed their number is smaller in comparison to *false positives*. Most of them are DDIs whose mechanism cannot be explained by known drug-protein relationships. During our analysis, we have identified three different types of them. The first one consists of DDIs occurring by non-absorbable complex formation. These DDIs occur by the chelation or physicochemical binding of the two drugs in the gastrointestinal tract, resulting in a complex with different physicochemical characteristics that cannot be absorbed (Levy & Reuning, 1964). Therefore, they cannot be inferred by means of our inference rules. The second one is DDIs due to the additive effects of two drugs leading to the same ADR. For example, both *protriptyline* and *voriconazole* produce prolongation of the QT interval, and their concomitant administration can lead to additive effects and finally produce a ventricular tachycardia known as Torsade de Pointes (TdP). The final type is DDIs for those that, although observed in the clinical setting and described in the scientific literature, the underlying mechanism is not known or understood yet (Kulkarni, Bora, Sirisha, Saji, & Sundaran, 2013; Patel, Rana, Suthar, Malhotra, & Patel, 2014).

In spite of these limitations, it has been proven that the combination of information about drug-protein relationships and SWRL rules can be used to represent and infer different types of DDIs on the same representation framework. Moreover, the results and analysis performed in this experiment are useful to identify further information that should be included in our ontology, and to refine our inference rules to reduce the number of *false positives.* For this purpose, it could be useful to include information regarding the therapeutic index of drugs (Boyce et al., 2007) or drug bioactivity data (Gaulton et al., 2012). On the other hand, including information about physicochemical properties of drugs, ADRs (Kuhn et al., 2010), or new discoveries about drug-protein relationships could be useful to reduce the number of *false negatives*.

9.2.3 *IExp3*: Inference and classification of DDIs on the basis of implicit mechanisms

We perform this experiment to evaluate if the implicit representation of DDI knowledge as SWRL rules and drug-protein relationships information can be used to infer DDIs and provide a more accurate description of their mechanisms than asserted information only.

- **Methods:**

To do this, we randomly select 93 DDIs imported to DINTO from the database DrugBank. We create a version of the ontology containing only the drugs involved in some of the DDIs (146 '*pharmacological entities*'), their related proteins (4,141 '*protein entities*'), and the mentioned 93 DDIs. Then, we create individuals and the relationships among them for all these classes. In order to ensure that the classification is obtained only from implicit information, we delete the OWL definitional axioms (or '*equivalent to*') that relates DDIs and their mechanisms (see **Figure 7.14**).

This version is merged with the file containing the 59 SWRL rules representing the different types of DDI mechanisms (a detailed description of these rules is provided in **Section 7.1.7** and **Section 7.2.5**). The reasoner engine HermiT 1.3.8 infers a new classification of the DDIs describing the mechanisms preceding them. An example is shown in **Figure 9.3**, where the class representing the interaction between *'amiodarone'* and *'cisapride'* is classified, based on the implicit mechanisms, as a *'target related DDI'*, *'enzymatic saturation DDI',* and *'enzyme inhibition DDI'*.

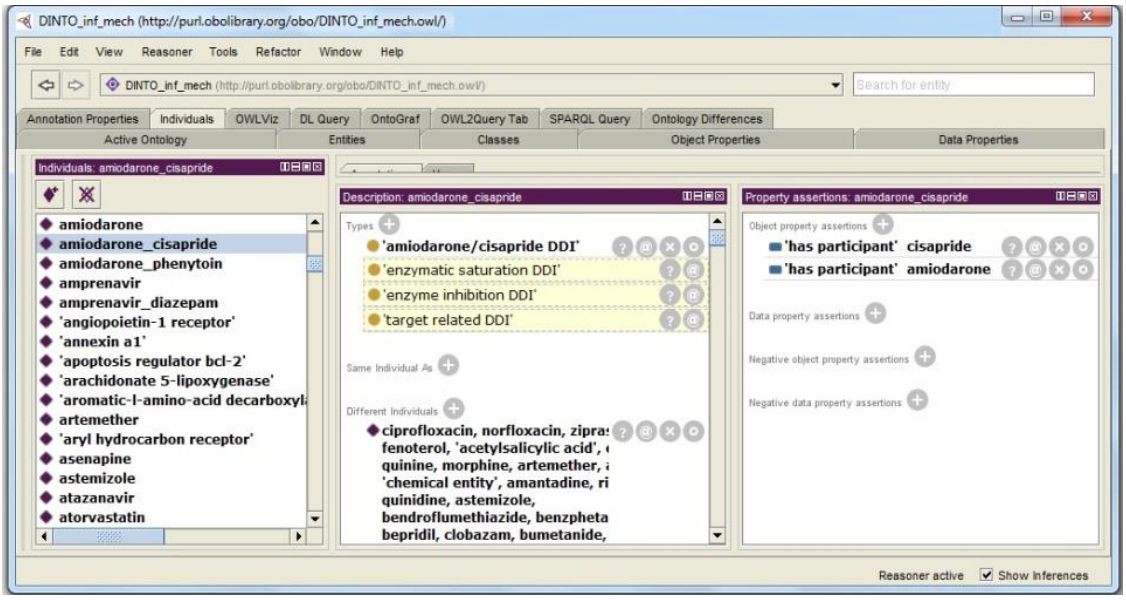

Figure 9.3. Protégé screenshot showing the inferred classification (as yellow boxes) of a DDI on the basis of its different DDI mechanisms

- **Results and Evaluation:**

Once the inferences have been obtained, we compare them with the mechanism-based classification asserted in the DrugBank dataset. As we have mentioned before, the database uses data on drug-target, drug-enzyme, and drug-transporter associations to establish the "possible based DDI mechanism" for some of the DDIs (Law et al., 2014). We translate this information into DINTO as classes and relationships and use the classification of DDIs as *'target related DDI'*, *'enzyme related DDI',* or *'transporter related DDI'* to classify DDIs on the basis of their mechanisms. However, drug-carrier associations are not included for any DDI in the original source and therefore, none of the DDIs imported from DrugBank are classified as *'carrier related DDI'*. In addition to this information, during the importing process we have added an additional mechanism *'non-absorbable mechanism formation'*, which is described in natural language for the corresponding DDIs.

In contrast to this five-level classification of DDIs in DrugBank, our SWRL rules lead to 15 different types of DDIs. These different types are the most general class *'DDI'* and its four subclasses: *i)* the target related class *'target related DDI'* and its subclasses *'agonistic DDI'* and *'antagonistic DDI'*; *ii)* the enzyme related class *'enzyme related DDI'* and its subclasses *'enzymatic saturation DDI'*, *'enzyme induction DDI',* and

'enzyme inhibition DDI'; *iii)* the transporter related class *'transporter related DDI'* and its subclasses *'transporter saturation DDI'*, *'transporter induction DDI',* and *'transporter inhibition DDI'*; *iv)* and the carrier related class *'carrier related DDI'* and its subclasses *'carrier saturation DDI'*, *'carrier induction DDI',* and *'carrier inhibition DDI'*. Since there is not information supporting it, there are no rules for the inference of the mechanism *'non-absorbable complex formation'* (**Figure 7.14**).

Classification of each DDI following the previously described hierarchy for both the inferred and asserted sets is shown in **Annex 11**. Due to the different granularity of both classifications, we consider that a DDI has compatible mechanisms in both sets if there is an exact coincidence in the type of protein(s) involved (target, enzyme, transporter, or carrier). As shown in **Table 9.5**, most of the inferences correspond to this case. Only for a single DDI the classification is not coincident between the two sources, while one of the mechanisms for seven of the DDIs asserted in DrugBank is not inferred in DINTO. In contrast, there are three DDIs with at least one additional mechanism inferred in DINTO in comparison with the asserted set. Finally, we observe there are ten DDIs for which no mechanism is inferred. They correspond in DrugBank to DDIs occurring by a *'non-absorbable complex formation'* mechanism that, as mentioned before, is not included in our SWRL rules.

	#	%
DDIs with compatible mechanisms	72	77,42
DDIs with additional mechanisms in the asserted (DrugBank) set ***(12) (38) (56) (64) (73) (74) (86)***	7	7,53
DDIs with additional mechanisms in the inferred (DINTO) set ***(2) (19) (92)***	3	3,22
DDIs with any coincident mechanism in both sets ***(82)***	1	1,08
DDIs corresponding to the *non-absorbable complex formation mechanism* type ***(4) (11) (13) (15) (16) (48) (54) (60) (81) (93)***	10	10,75
Total	**93**	**100**

Table 9.5. Results and comparison of the inferred classification of DDIs based on implicit mechanisms versus classification on the basis of asserted mechanisms (Numbers in brackets correspond to the numeration of the DDIs in **Annex 11**)

The results of this experiment show that most of our inferences are correct in comparison with DrugBank. However, there are 21 DDIs for what inferences yield to differences between both sets. Here, we analyse each one of them in order to identify the causes of theses discrepancies.

- DDIs with additional mechanisms in the asserted (DrugBank) set: There are seven DDIs for which DrugBank establishes an additional mechanism that is not inferred in DINTO. In all these cases both interacting drugs have a relationship *'induces'* or *'inhibits'* with the same protein and, as a consequence, a DDI is established between them in the asserted set. However, SWRL rules describing

these patterns have not been created in DINTO since, from a pharmacological point of view, it would not be correct to infer that this situation leads to a DDI.

To illustrate this, we discuss the example of the interaction between *midodrine* and *dexamethasone*. *Midodrine* inhibits the activity of the enzyme *Cytochrome P450 2D6 (CYP2D6).* Similarly, the drug *dexamethasone* inhibits the same enzyme. From this information DrugBank establishes that there could be a possible enzyme-related interaction between both drugs. However, this assumption is incorrect, since this interaction can occur only if this enzyme is involved in the metabolism or activity of one of the two drugs − or their metabolites. If this is not the case, both drugs alter the activity of the enzyme, but their activities or concentrations are not modified. Since any other relationship is described in DrugBank between *CYP2D6* and the two drugs, we cannot establish that this is the mechanism of interaction between them. Similar circumstances occur for the other non-coincident six DDIs, with the difference that, on these cases the protein is a transporter instead of an enzyme.

Therefore, from the discussion above, we conclude that a DDI should not be established when the information available is only that two drugs '*induce'* and/or '*inhibit'* the same protein. However, this event can be very relevant in multiple drug-drug interactions, where the adverse consequences of a DDI are observable only when a third or more concomitant drugs are added (Cone, Fant, Rohay, & Associates, 2004; Klarin, 2007).

An example could be provided by the DDI between *carvedilol, ergotamine,* and *digoxin.* The two former drugs inhibit the activity of *multidrug resistance protein 1* (commonly known as *P-glycoprotein* or *P-gp*), a protein that transports a wide range of drugs from the inside of the cell to the outside. *P-gp* located at the intestine reduces the penetration of drugs into the body. Therefore, it can be considered as a defense mechanism against xenobiotics or strange substances. When a substrate of *P-g*, such as *digoxin*, penetrates into the endothelial cells in the intestine, it binds to the transporter at the intracellular part of the membrane and is extruded to the extracellular site. As a consequence, the bioavailability − or fraction of drug absorbed that reaches the systemic circulation − is decreased (Fromm & Kim, 2011). Since neither *carvedilol* nor *ergotamine* are substrates of *P-gp*, if they are administered concomitantly, their levels in the body are not influenced by the alteration in the activity of this transporter. Therefore, there is not a transporter-based DDI between them (**Figure 9.4a**). In contrast, *digoxin* is a substrate of *P-gp* that, when administered concomitantly with *carvedilol*, can have changes on its bioavailability due to the inhibitory effects of *carvedilol.* Nevertheless, the absolute change in *digoxin* pharmacokinetics is small and not clinically significant (Wermeling et al., 1994) (**Figure 9.4b**). However, the inclusion of a third drug in the therapeutic regime, such as *ergtamine*, could produce more accentuated inhibitory effects on *P-gp*, and might lead to an increase in the bioavailability of *digoxin* and possible toxic clinical manifestations (**Figure 9.4c**).

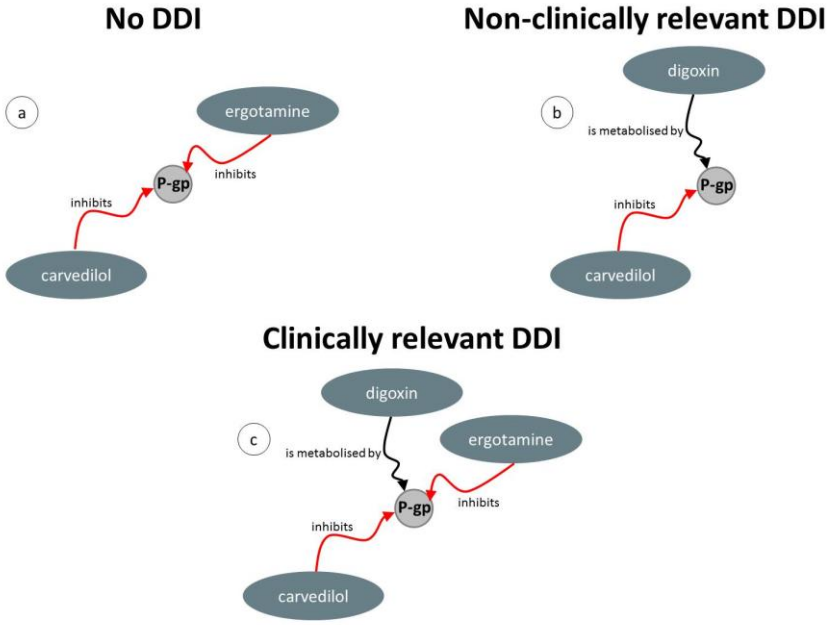

Figure 9.4. Multiple-interaction between *carvedilol*, *digoxine,* and *ergotamine*

- DDIs with additional mechanisms in the inferred (DINTO) set: There are three DDIs with an additional mechanism in the inferred set in comparison to the asserted set. Two of them have been classified as *'carrier related DDI'* and one as *'carrier saturation DDI'*. As mentioned before, any possible carrier-related mechanism has been specified in DrugBank. However, our SWRL rules enable the inference of this type of mechanism, too.

- DDIs with any coincident mechanism in both sets: We can divide the DDIs included in this section into two main groups: the ten DDIs occurring by non-absorbable complex formation, and the unique DDI for which any of the inferred and asserted mechanisms is coincident.

On the one hand, for the first ten, no DDI mechanism is inferred in DINTO because there is not any rule defining a non-absorbable complex formation DDI. The main reason is that there is not information in the ontology supporting it. Therefore, currently it is not possible to predict DDIs occurring via this mechanism. On the other hand, there is not any coincident mechanism for the interaction between *sulindac* and *bumetanide*. In DrugBank this DDI is described as a possible *'transporter related DDI'*, while the type inferred in DINTO is *'target related DDI'*. The last one is inferred because both drugs have the relationship '*has pharmacological target*' with the protein *prostaglandin G/H synthase 2*. In this case, there is a SWRL rule establishing that the coadministration of both drugs can lead to a target related DDI. On the other hand, the mechanism asserted in DrugBank is established because both drugs inhibit the protein *solute carrier family 22 member 6*. However, as we explained before, this information is not sufficient to establish a possible DDI between both drugs.

- **Discussion:**

The results of this experiment show that our representation of DDI mechanisms as SWRL rules infers correctly the mechanisms of most of the DDIs included in the study. Moreover, only three causes lead to differences for the remaining ones, and all of them have been identified.

The first one is due to the inference of PK DDIs when the two interacting pairs have a *inhibits* and/or *induces* relationship with the same protein. We have observed that the inference of a DDI just on the basis of these two relationships might be incorrect. However, this fact becomes very interesting to study the representation of multiple DDIs, where the consequence of one DDI can be exacerbated by the presence of a third or more drugs. This phenomenon has not been represented in previous approaches, and it could be relevant on the study of DDIs in polymedicated patients.

The second reason for discrepancies between the inferred and asserted sets is due to those DDIs occurring by the formation of non-absorbable complexes. There is not SWRL rules in DINTO for the inference of this type of mechanism, since the information required supporting them, such as cationic and anionic moieties of drugs, has not been included in the ontology yet.

Finally, differences between both sets arise because there are SWRL rules in DINTO to infer carrier-related DDIs and its subtypes, while in DrugBank this mechanism is not assigned to any DDI.

Therefore, our approach infers more accurate and detailed descriptions of the DDI mechanisms than those established in the asserted set, for both PK and PD mechanisms. For example, in the case of PD DDIs, for which modeling has been less deeply studied than for PK DDIs (**Section 6.4**), we have inferred that thirteen DDIs occur due to the *agonistic effect* of two drugs on the same target, five due to *antagonism*, and that in two cases both mechanisms are involved.

In fact, usually more than one mechanism is inferred for the same DDI. Sometimes they could seem incompatible: *enzyme inhibition* and *enzyme induction*, *agonism* and *antagonism*, and so forth. However, from a pharmacological point of view the information is correct, since one drug can act at the same time as both an agonist and an antoginist on the same target, or as an inducer or inhibitor on the same enzyme. The final response depends on several factors, such as the affinity, efficacy, and selectivity of a drug for the protein (Kenakin, 2012). Specifically, the attraction that a drug molecule has for a protein − or the pharmacological affinity − can be quantified, and it is described in different databases (Gaulton et al., 2012). This bioactivity data can be used to identify the mechanism that is more relevant and that determines the clinical manifestations of the DDI. Since SWRL supports the use of data values, we hypothesize that the same approach based on the combination of drug-related information represented in OWL 2 and SWRL rules could be applied to obtain descriptions of DDI mechanisms that are more accurate. Therefore, in our future work we will import into DINTO bioactivity data, and will refine our SWRL rules to provide a more detailed representation of the DDI mechanisms.

9.3 Discussion and conclusions

In this study, we have demonstrated that the use of currently available semantic web technologies, standards, and tools support the formal representation of complex pharmacological knowledge and the prediction of DDIs on a large scale by exploiting large amounts of semantic data. To the best of our knowledge, this is the first work proving that any of these two objectives can be achieved without the intervention of technologies other than those created for the semantic web.

DDIs and their underlying mechanisms are complex pharmacological processes and their conceptualization and implementation require an expressive representation language. Despite the richness of OWL's set of relational properties, its expressivity does not allow to express all possibilities for object relationships (Holford et al., 2010; Horrocks et al., 2005). Indeed, we have observed that the use of advanced OWL 2 features such as property chains is not adequate for the consistent representation of this knowledge. However, this limitation can be overcome by the combination of an OWL ontology with inference rules (Golbreich, 2005). In this work, we have demonstrated that standard language SWRL is adequate to represent different types and subtypes of DDI mechanisms in a unique framework, while, in contrast to programming languages used by former efforts (Moitra et al., 2014; Tari et al., 2010), ensuring interoperability between ontologies.

Moreover, these rules have been successfully applied to the inference of DDIs. A previous work in this domain had concluded that OWL-DL was not suitable for the detection of DDIs and that it should be used only for ontology consistency checking (Konagaya, 2012). However, in this study we have shown that DL formalisms can be successfully used for the inference of DDIs by means of combining drug-related facts as an OWL ontology, and DDI mechanisms as SWRL rules.

The bottleneck in this process has been, however, the limited performance of ontology reasoning engines with very large and complex ontologies. None of the most popular OWL reasoners, such as FacT ++, HermiT, or Pellet could process a whole version of DINTO. Therefore, following prior experiences of similar projects (Holford et al., 2010), we had to employ different strategies in order to reduce the size and complexity of our ontology. Ontology reasoning plays a key role in ontology engineering and has become an indispensable activity to help ontologists during the development and evaluation of ontologies, as well as users to query their contents. Therefore, research into ontology reasoner engines development continues, and it is boosted by efforts such as the annual OWL Reasoner Evaluation Workshop,[56] which brings together developers and users and provides an opportunity to promote their systems. Therefore, it should be expected that more robust and reliable engines would be available in the near future.

In spite of this limitation, we have achieved the largest coverage for drugs included in a DDI inference experiment. The closest project, which combined NLP with knowledge formal representation and reasoning (Tari et al., 2010), could test the inference system on a smaller number of drugs (295 drugs against the 426 included in our experiment). The main reason for the low coverage of the other projects is that they relied

[56] http://ceur-ws.org/Vol-1207/

on manual curation to identify, gather, and structure the drug-related facts required as basic information to infer the DDIs. Although expert manual curation provides high quality information, this activity is both cost-intensive and time-consuming. Moreover, new pharmacological information is discovered and published every day in the scientific literature, which makes keeping up to date a knowledge base with this information a difficult task. However, during the last years there has been a huge increase of pharmacological information stored in structured and machine-readable formats as public databases and knowledge bases (Khelashvili et al. 2010; Whirl-Carrillo et al. 2012). Exploiting this information and integrating it automatically in a knowledge base might overcome the limitation of manual curation. In this project, we have demonstrated that this premise can be applied to the ontological engineering field, too. In this way, to overcome this issue, we have designed a CM for DINTO that reuses and integrates information currently available in public information resources, such as drug-protein relationships from the database DrugBank. With this approach, the development of the ontology is driven by the combination of both the requirements of the final application and the information available to be imported into the ontology.

The inference experiments performed in this study show that DINTO is a correct and comprehensive ontology for the DDI domain. While other ontologies have focused on the separate representation of PK (Arikuma et al., 2008; Boyce et al., 2010b; Moitra et al., 2014) and PD (Imai et al., 2013) DDI mechanisms, DINTO is the first resource that represents both of them in the same framework. Moreover, previous works have addressed only the representation of one specific subtype of PK DDIs: those occurring by an alteration in drug metabolism. In contrast to this, we have represented different subtypes of both PK and PD DDI mechanisms, leading to the most complete and detailed formal description of these processes.

Evaluation of the inferences obtained in the experiments has been performed against DDIs included in the database DrugBank. However, this database, or any other DDI information source, cannot be considered a gold-standard of "true DDIs" (Imai et al., 2013). The large number of possible DDIs, with different degrees of significance and that can be influenced by drug or patient-related facts, makes it not possible to manually study, identify, and collect all of them in a unique information source. Indeed, important differences have been found in the coverage for DDIs among different information sources (Fulda, Valuck, Zanden, & Parker, 2000; Olvey et al., 2010a). As a consequence, it is not possible to ascertain if the false positive DDIs inferred in DINTO are not included in DrugBank because *i*) they are interacting pairs but have not been included yet in the database, *ii*) they are interacting pairs but the interaction between them has not been described in the scientific literature, or *iii*) they are not interacting pairs of drugs. In spite of this, the other main reason for the high number of false positives is that our SWRL rules have been modeled to infer DDIs on the basis of different mechanisms, but independently of other related facts such as their significance or level of documentation. Therefore, for any pair of drugs for which an underlying DDI mechanism exists, regardless of whether it would lead to a clinically relevant DDI or a non-clinically relevant or unobservable DDI, the interaction is inferred.

In contrast to this, other DDIs could not be predicted by our SWRL rules. Most of them are DDIs for which their mechanisms cannot be explained by known drug-protein relationships, including DDIs occurring by the chelation or physicochemical binding of the two drugs in the gastrointestinal tract (Levy & Reuning, 1964), those due to the addition of an ADR (Antonelli, Atar, Freedberg, & Rosenfeld, 2005), and DDIs for

which, although observed in the clinical setting and described in the scientific literature, the underlying mechanism is not known or understood yet (Kulkarni et al., 2013; Patel et al., 2014).

The analysis of these results is useful to identify further information that should be included in our ontology, and provides guidelines to refine our inference rules. For this purpose, in our future work we will include in the ontology the following information:

- Therapeutic index of drugs, which can be used to identify clinically significant DDIs (Boyce et al., 2007).
- Drug bioactivity data for the identification of the principal DDI mechanism (Gaulton et al., 2012).
- Physicochemical properties of drugs, which can be used to identify DDIs occurring by chelation.
- ADRs for the inference of DDIs occurring by additive effects.
- New discoveries about drug-protein relationships, which can lead to the prediction of DDIs that cannot be explained by current knowledge.

Chapter 10

DDI Information Extraction

One of the two main contributions of this thesis is the creation of a comprehensive ontology for DDI knowledge that will be made available to the NLP research community to be exploited in IE from pharmacological texts. We hypothesize that ontologies are a useful resource for IE that can contribute to a better performance of current systems (Cimiano et al., 2014; Müller et al., 2004; Wimalasuriya, 2010). The evaluation of DINTO's usefulness for this task is limited, however, since the development of a new IE system based on DINTO is out of the scope of this project. Nonetheless, to provide a proof-of-concept of its suitability for IE, we propose to use DINTO with a simple system for the two different tasks of drug NER and DDI extraction from texts. We use the previously developed DDI corpus and compare our results in the framework of the *SemEval-2013 DDIExtraction shared task* (Segura-Bedmar et al., 2014) (see **Section 4.2**). In this evaluation, we compare the results obtained by this approach based on DINTO with those obtained by the participating systems. Moreover, the best runs submitted by each team are combined with DINTO in order to study its impact on their performance.

This chapter is organized as follows. In **Section 10.1**, we describe the evaluation of DINTO for NER. After describing the system developed to perform this task, we present and discuss the achieved results. Then, we provide a brief summary of the evaluation process in the *SemeEval DDIExtraction task* and compare our results with those obtained by the participants. **Section 10.2** describes the evaluation of DINTO for DDI extraction. We provide a description of the system used in this task and the results obtained. Again, we compare our results with those achieved by the participants on the task and describe the results obtained with an ensemble for each one of them and DINTO. Finally, in

Section 10.3 we present and discuss the main conclusions of this evaluation and highlight our lines of future work.

10.1 DINTO-based named entity recognition

To evaluate DINTO for NER we develop a system combining different components from '*The General Architecture for Text Engineering*' (GATE),[57] a free open-source infrastructure for developing and deploying software components that process human language (Bontcheva, Tablan, Maynard, & Cunningham, 2004). GATE is one of the most widely used systems of its type and has many active users. It has been used in NLP tasks in different domains, including life science and medicine projects (Cunningham, Tablan, Roberts, & Bontcheva, 2013). We select GATE to create our NER system because, in contrast to other automatic annotation tools (Jonquet et al., 2009; Uren et al., 2006), it enables automatic ontology-based annotation of texts using any external ontology. Combining different components of GATE, we create a pipeline for NER based on DINTO, for which the architecture is shown in **Figure 10.1**.

First, we split the text into sentences using *Regex Sentence Splitter*, provided by GATE and based on regular expressions. GATE's default *tokeniser* is used to split the previously obtained sentences into very simple tokens (words, numbers, symbols, punctuation marks, and space-tokens). Then, the *Part Of Speech Tagger* − a modified version of the Brill tagger (Hepple, 2000) − produces POS tags as an annotation on each word or symbol using a default lexicon and ruleset. The *GATE Morphological Analyser (Morpher)* is used afterwards. This GATE plugin is based on certain regular expressions rules and it takes as input the tokenized documents. Considering one token and its part of speech tag, one at a time, it identifies its lemma and an affix. These values are then added as features of the token annotation.

GATE provides different gazetteers and gazetteer tools to perform NER. A gazetteer consists of a set of lists containing names of entities, which are used to find occurrences of these names in text. We combine two different GATE plugins to recognize drug names using the information contained in DINTO. On the one hand, *Flexible Gazetteer* provides the flexibility to choose our own customized input (in this case, the processed DDI corpus) and an external gazetteer. It performs a lookup over the document based on the values of an arbitrary feature of an arbitrary annotation type by using an externally provided gazetteer. The external gazetteer is *Ontology Resource Root (OntoRoot) Gazetteer*, a dynamically created gazetteer capable of producing ontology-based annotations over the given content according to the given ontology. The *OntoRoot Gazetteer* is created by extracting human-understandable content from DINTO (drug classes, preferred labels, synonyms, and brand names) which is then used by the *Flexible Gazetteer* to recognize drug names in the corpus.

The DDI corpus is processed with this GATE pipeline, and it is evaluated using the gold-standard annotations. In the next section we provide the results for the NER task and

[57] https://gate.ac.uk/

compare them with those obtained by participating teams in the *SemEval-2013 DDIExtraction task* (Segura-Bedmar et al., 2014).

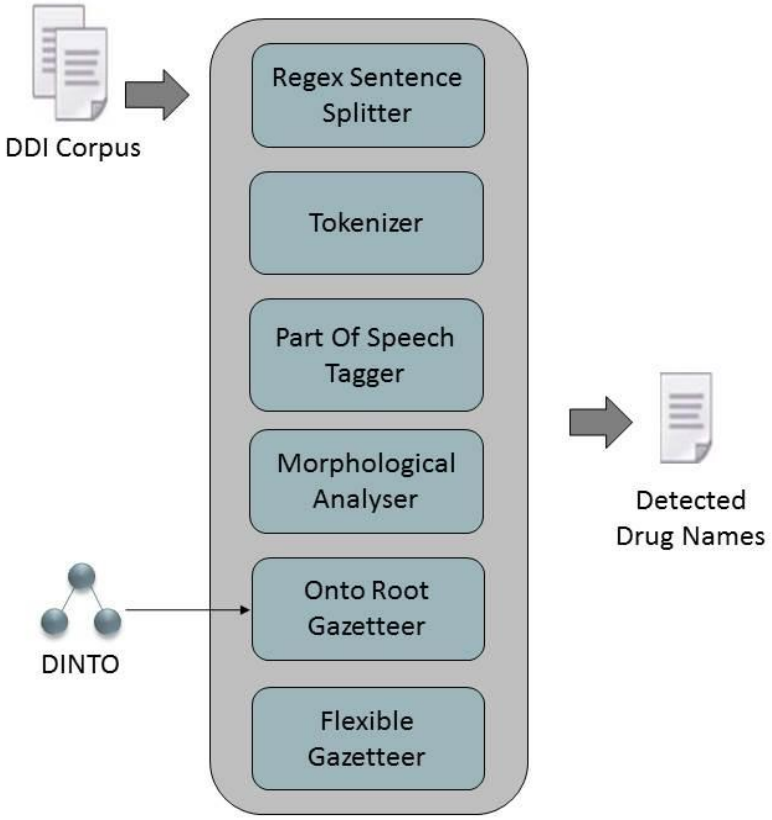

Figure 10.1. Architecture of the DINTO-based NER system

10.1.1 Results

The DDI corpus is processed with the GATE pipeline using two different versions of DINTO in two different experiments. In order to conform to the nomenclature used by the *DDIExtraction task* participating systems, we refer to them as *Run 1* and *Run 2*. *Run 1* relies on a version of DINTO containing only drugs imported from the ChEBI ontology (Hastings et al., 2013). The number of classes for pharmacological substances is 3,026. Original annotation properties are maintained. Therefore, each class has a preferred label and different synonyms – including brand names and chemical formula. In *Run 2* we use a second version of DINTO enriched with information from DrugBank and with a larger number of synonyms and brand names. Moreover, those drugs present in DrugBank and not included in the first version of DINTO are represented as new classes in the ontology, leading to a total of 8,854 classes for pharmacological substances. With these two different experiments, we aim to assess the impact of integrating different information resources in the results of the NER task.

Evaluation results are reported using the standard precision/recall/f-score metrics (Chinchor & Sundheim, 1993). **Table 10.1** and **Table 10.2** show results for *Run 1* and

Run 2, respectively. We provide results for the whole corpus and separately for the DDI-DrugBank and DDI-MEDLINE datasets.

Dataset	cor	inc	para	mis	spu	total	precision	recall	F1
DDI Corpus	198	0	6	482	40	686	0,8238	0,293	**0,4323**
DDI-DrugBank	107	0	2	195	7	304	0,931	0,3553	**0,5143**
DDI-MEDLINE	91	0	4	287	33	382	0,7266	0,2435	**0,3647**

Table 10.1. Results for NER task using a ChEBI-based version of DINTO (*Run 1*)

As shown in **Table10.1**, *Run 1* (ChEBI-based ontology) achieves moderate results for F-score. Although precision values are high, results for recall are lower. The best results are obtained on the DDI-DrugBank dataset. Therefore, a version of DINTO based only on a unique information resource provides a limited coverage for drug entities in pharmacological texts.

Dataset	cor	inc	para	mis	spu	total	precision	recall	F1
DDI Corpus	334	0	38	314	118	686	0,7204	0,5146	**0,6003**
DDI-DrugBank	187	0	15	102	28	304	0,8457	0,6398	**0,7285**
DDI-MEDLINE	147	0	23	212	90	382	0,6096	0,4149	**0,4938**

Table 10.2. Results for NER task using a ChEBI plus DrugBank-based version of DINTO (*Run 2*)

In contrast, in *Run 2* we use a version of the ontology combining two information resources, ChEBI and DrugBank. This run achieves better results for the three datasets. With this extended version of the ontology, recall increases considerably. Therefore, the inclusion of more pharmacological substances and more synonyms for the existing ones increases the coverage for drug named entities of the ontology and, therefore, the results for recall. However, a slight decrease in precision is observed. This decrease is due to an increase in false positives − or the incorrect identification of drug named entities that are not annotated in the gold-standard.

Performing error analysis of these results, we have identified two main reasons for the identification of false positives (Table 10.3). The first and dominant is the identification as named entities of non-drug names that are frequently used in

pharmacological texts. These include terms such as *'control'*, '*results'*, '*labelling'*, '*duration'* or '*mg'*. These terms are not drug names. However, they might be present in the ontology (e.g., one of the brand names for the drug '*loperamide*' imported from DrugBank is '*Pepto Diarrhea Control'*) and, therefore, extracted by the *OntoRoot Gazetteer*. Therefore, the system used in this evaluation incorrectly identifies these terms as drug named entities. A future system based on DINTO for NER might avoid this issue by using a stop list including this type of terms, for example.

The second main reason for the identification of false positives is ambiguity (see **Section 3.4.2**). During the creation of the DDI corpus, we have identified the existence of ambiguous terms as one of the most challenging problems in the manual annotation of pharmacological texts (Herrero-Zazo, Segura-Bedmar, & Martínez, 2013) and remains an important issue in the development of accurate named entity recognition systems. Therefore, our system is unable to identify them. As a consequence, each mention of the term '*insulin',* for example, is annotated as a drug named entity, independently of the term referring to the drug '*insulin'* (correct annotation) or to the substance produced by the body (incorrect annotation).

Error cause	DrugBank-DDI	MEDLINE-DDI	Example
Non-drug names	20	65	'control', 'results'
Ambiguous terms	8	25	'adrenaline', 'dopamine'
Total	28	90	

Table 10.3. Analysis of false positives for *Run2*

As mentioned before, recall results for *Run 2* increase considerably with respect to *Run 1*. This reveals an increase in the coverage for drug entities of DINTO. However, the system still fails to identify a considerable number of drug named entities. Performing the error analysis of false negatives – drug named entities annotated in the gold-standard and not identified by the system – we classified them in different types (**Table 10.4**).

The first type is terms describing groups of drugs. The DDI corpus is annotated with those terms describing a group of drugs (e.g., *'analgesic agent'*). However, these terms are not included in the GATE gazetteer and therefore, none of those terms annotated in the corpus as group type are recognized by the system. For *Run 2*, this type of error represents more than the 50% and 40% of false negatives in the DDI-DrugBank and DDI-MEDLINE dataset, respectively.

The second type of false negatives is due to the fail of the system to correctly recognize some terms included in the ontology. These include drug names such as '*warfarin'*, '*fludarabine',* or '*insulin'*. We have not been able to find a common pattern in these entities to explain this unexpected behaviour of the software. It should be expected, however, that this error would not be reproduced with a system specifically created to use DINTO for IE. This issue influences the results obtained for *Run 2* with the DDI-

DrugBank dataset, where this unrecognized terms represent more than the 30% of false negatives terms. In the DDI-MEDLINE dataset, this type of error is less frequent (minor to 5% of false negatives). However, an important reason for false negatives in the DDI-MEDLINE dataset is the non-recognition of `drug_n` type entities (representing almost the 40% of false negatives in this dataset), substances with pharmacological activity but not approved as drugs. This type of false negatives is more relevant in the DDI-MEDLINE dataset than in DDI-DrugBank dataset due to the frequent mention of this type of entities in MEDLINE texts (see **Section 3.5**). From these results, we conclude that, although ChEBI and DrugBank have information about `drug_n` substances, DINTO does not provide enough coverage for them. To overcome this issue, we plan to extend the ontology integrating resources with a broad coverage of these entities, such as the PubChem database (Bolton, Wang, Thiessen, & Bryant, 2008).

Finally, another source of false negatives are spelling variations in drug names, which the system is unable to recognize as drug named entities (e.g., the system fails to recognize the term '*temezepam'*, an incorrect variation of the drug name '*temazepam'*). A future system for NER based on DINTO could overcome this problem through the integration of lemmatization or stemming techniques.

Error cause	DDI-DrugBank	DDI-MEDLINE	Example
Missed `group` entity type	53	89	'anticoagulants', 'loop diuretic'
Missed `drug_n` entity type	5	79	'methylene blue', 'tween 80'
Missed `drug` entity type	3	26	'b-carotene', 'etizolam'
Missed `brand` entity type	4	-	'ertaczotm', 'synagis'
Entity present in the ontology but not annotated by unexpected behaviour of the software	34	17	'tofranil', 'colestipol'.
Spelling variations	3	1	'lorezepam', 'temezepam'
Total	102	212	

Table 10.4. Analysis of false negatives for *Run 2*

10.1.2 Comparison in the framework of the *SemEval-2013 DDIExtraction shared task*

In the previous section, we have presented and discussed the results achieved by a simple system exploiting DINTO for drug NER in pharmacological texts. Although the aim of this thesis is the creation of a new ontology for DDIs that could be useful for the NLP community, and the development of a new IE system is out of the scope of our

work, we want to provide better insights into the potential usefulness of DINTO for drug NER. To do this, we perform a comparative analysis with the participating systems in the *SemEval-2013 DDIExtraction task*.

In the second edition of the *DDIExtraction task*, six teams participated in the subtask of named entity extraction of pharmacological substances in text. Each participant was allowed to submit a maximum of three system runs. For evaluation, a part of the DDI corpus (the train dataset) was provided with the golden annotations hidden. The goal for participating systems was to recreate the golden annotations and to output an ASCII list of reported entities. System performance was scored automatically by how well the generated pharmacological substance list corresponded to the gold-standard annotations. In this task, several evaluation criteria were used to evaluate the results of the participant systems. These included strict matching (exact-boundary and type of entity matching), exact boundary matching (regardless of the type of entity), and partial boundary matching (regardless of the type of entity). Strict evaluation demands exact boundary match and requires that both mentions have the same entity type. This strict and exact criterion may be too restrictive for the overall goal – extraction of drug interactions – as it could miss partial matches, which can provide useful information for a DDI extraction system. Therefore, we compare the results obtained by the best run for each participant team with the result of our best run (*Run 2*, **Table 10.2**) for partial matching evaluation criteria.

A deeper description of this task, the participating systems, and their results is described in **Section 4.2.1**. To provide a brief overview we should highlight that the best results were achieved with a token sequence labelling approach based on CRF proposed by the **WBI team** (Huber, Linden, & Rockt, 2013). Substances not approved for human use (`drug_n`) were more difficult to identify by all the participant systems. In fact, only the **UEM_UC3M team** (Sanchez-Cisneros & Aparicio, 2013), using a system based on the combination of several ontologies and other information resources – ontologies from the UMLS collection (Bodenreider, 2004) combined with information extracted from MeSH (Lipscomb, 2000), DrugBank (Wishart et al., 2006), PubChem (Bolton et al., 2008), the ATC classification system (WHO, n.d.), and KEGG database (Kanehisa et al., 2012) – was able to recognize this type of substances on the DDI-DrugBank dataset. For all participant systems, results on the DDI-DrugBank dataset were much better than on the DDI-MEDLINE dataset. This fact could be explained by the higher complexity of MEDLINE texts and the current mentions of `drug_n` type entities that, as mentioned before, were challenging to identify.

Figure 10.2 shows a comparison between participant systems and the DINTO-based system for the NER task. Results are provided for the whole DDI corpus and the DDI-DrugBank and DDI-MEDLINE datasets. These results show that the system based on DINTO obtains comparable results to that obtained by a dictionary-based system developed by the participant team **UEM_UC3M** (Sanchez-Cisneros & Aparicio, 2013). Like other systems, DINTO obtains better results for the DDI-DrugBank dataset than for the DDI-MEDLINE dataset. This can be attributed to the use of the DrugBank database as a source of information for our ontology – as other systems such as **NLM_LHC** or **UEM_UC3M** – but it is also influenced by the lack of challenging `drug_n` type entities in this dataset. As mentioned above, we do not evaluate the coverage of DINTO for `group` type entities. We believe that including this type of entities would provide even better results for the NER task, and we will include them in our future work. Even so, from this evaluation in the framework of the *DDIExtraction task* we can anticipate that

future combination of DINTO with other techniques, such as CRF (Lafferty, Mccallum, & Pereira, 2001), may provide better results for the recognition of drug named entities from pharmacological texts.

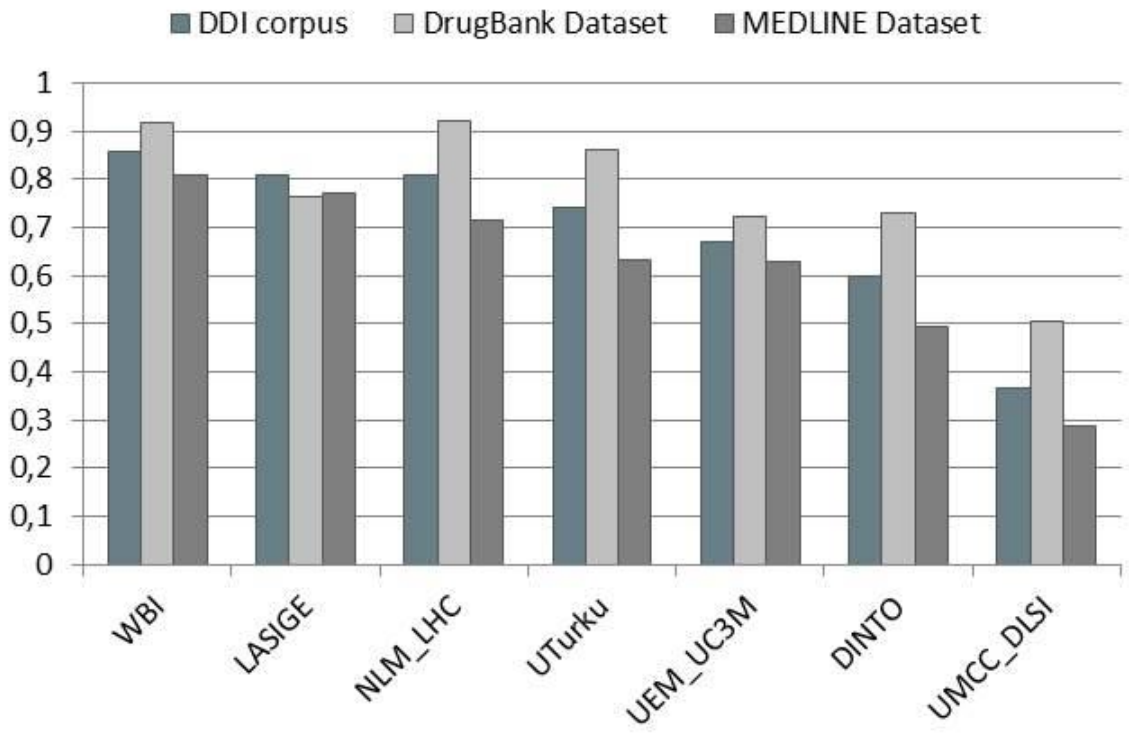

Figure 10.2. Results for the NER task for participants' best runs and for DINTO-based system (*Run 2*)

10.2 DINTO-based relation extraction

In this thesis, we have developed an ontology representing information about drugs and other pharmacological substances (8,854 classes) and the relationships between them describing a DDI. Some of these DDIs are well known and have been described in the literature and collected in the database DrugBank (11,555 DDIs). Other DDIs, however, can be inferred in DINTO on the basis of drug-protein relationships (see **Section 9.2**). These inferred DDIs might not have been described as DDIs in the literature yet, but there is a plausible mechanism underlying them. Therefore, they could potentially be described as unknown, suspected, or new DDIs in the scientific literature.

We hypothesize that a formal representation of the DDI domain can be exploited by IE systems to improve their performance in the task of DDI extraction from pharmacological texts. However, as mentioned before, the development of a new system able to use the knowledge represented in DINTO is out of the scope of this thesis. However, to provide a proof-of-concept of the potential usefulness of DINTO for DDI extraction, we develop a simple system that uses the information represented in DINTO to identify DDIs from the DDI corpus. We compare these results with those obtained by participants of the *DDIExtraction task*. Moreover, to assess the strengths and weaknesses

of our ontology, we adopt another evaluation approach. To do this, we combine by means of ensemble methods the participant systems and DINTO, in order to measure the impact of using the ontology on their results.

10.2.1 Results

To assess the contribution of DINTO to the DDI extraction task we develop a simple system able to select candidate interacting drug pairs in the DDI corpus and to identify if the interaction between them is formally represented in the ontology. We compare a baseline using a version of DINTO containing those known DDIs described in DrugBank (*Run 1*) versus a version of the ontology including both known and inferred DDIs (*Run 2*). With these two experiments, we aim to evaluate the influence of an increased number of DDIs in the ontology.

The system selects candidate drug pairs in the DDI corpus, which are labelled in the gold-standard annotations as interacting pair (*true* value in the XML) or no-interacting pair (*false* value in the XML), and looks them up in the ontology. To do this, annotation properties as preferred labels and synonyms are checked. In the case that both drugs are present in the ontology, the system looks for the *'may interact with'* relationship between them, that is used in DINTO to describe an interaction between two drugs. Those pairs of drugs in the DDI corpus that are described to interact in the ontology are considered interacting pairs (*true*), while those pairs not described to have an interaction in the ontology are considered as non-interacting pairs (*false*) (**Figure 10.3**).

Dataset	tp	fp	fn	total	precision	recall	F1
DDI corpus	164	376	815	979	0.3037	0.1675	**0.2159**
DDI-DrugBank	154	368	730	884	0.295	0.1742	**0.2191**
DDI-MEDLINE	10	8	85	95	0.5556	0.1053	**0.177**

Table 10.5. Results for the DDI extraction task using a known-DDIs version of DINTO (*Run 1*)

Table 10.5 and **Table 10.6** show results obtained using this system with two different versions of DINTO. On the one hand, *Run 1* uses a version of the ontology containing only known DDIs imported from DrugBank (**Table 10.5**). On the other hand, results for *Run 2* are achieved using a version of the ontology representing known DDIs plus DDIs inferred from drug-protein relationships (**Table 10.6**).

Dataset	tp	fp	fn	total	precision	recall	F1
DDI corpus	220	793	759	979	0.2172	0.2247	**0.2209**
DDI-DrugBank	206	759	678	884	0.2135	0.233	**0.2228**
DDI-MEDLINE	14	34	81	95	0.2917	0.1474	**0.1958**

Table 10.6. Results for the DDI extraction task using an inferred and known-DDIs version of DINTO (*Run 2*)

F1 scores achieved by both runs are low and, although there is a slight improvement in the results for *Run 2* with respect to *Run 1*, this is not prominent. When the version of the ontology including inferred DDIs is used (*Run 2*), the number of correctly identified DDIs increases lightly. However, the number of false positives – non-interacting drug pairs incorrectly identified by the system to be interacting pairs – increases considerably.

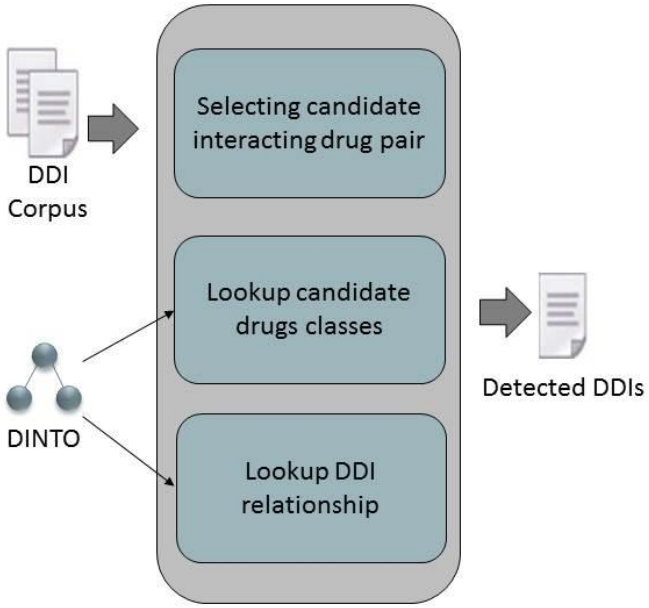

Figure 10.3. Architecture of the DINTO-based-RE system

The reason for this is that our system cannot take into account context. In our approach, we consider that if a candidate drug pair in text is described to interact in the ontology, then it is a DDI description. However, two interacting drugs can be mentioned in a sentence that is not discussing the DDI between them. Therefore, they should not be identified as a DDI. For example, the following sentence has two interacting drug pairs between the entities *ibuprofen* and *cyclosporin A*:

« Although **ibuprofen**$_{e1}$ can be administered with **cyclosporin A**$_{e2}$ in patients with normal renal function, caution should be used when administering **ibuprofen**$_{e3}$ concurrently with **cyclosporin A**$_{e4}$ to patients with mild to moderate renal insufficiency. » (i)

In this example, both drugs appear twice in the sentence, but only their last two mentions are involved in the description of a DDI. As *ibuprofen* and *cyclosporin A* are described to interact in DINTO, the system recognized both drug interacting pairs (*e1-e2* and *e3-e4*) as two DDIs.

Another reason for the increase in false positives for *Run 2* is the system's incapability to distinguish between drugs constituting a coordinate structure and therefore, to recognize that they are not describing a DDI. Coordinate structures are usually used in pharmacological texts to show specialization. Therefore, they do not describe a DDI. For example, the following sentence uses a coordinate structure to numerate drugs belonging to the same group.

« The absorption of **fluoroquinolone antibiotics**$_{e1}$ such as **moxifloxacin**$_{e2}$, **ciprofloxacin**$_{e3}$ or **lomefloxacin**$_{e4}$ could be reduced in patients in treatment with **aluminium**$_{e5}$. » (ii)

This coordinate structure shows a specialization of the group of drugs *fluoroquinolone antibiotics* in three of its members: *moxifloxacin, ciprofloxacin,* and *lomefloxacin.* The sentence is describing the interaction of *moxifloxacin*, *ciprofloxacin* and *lomefloxacin* with *aluminium*. However, *moxifloxacin* is described in DINTO to interact with *ciprofloxacin* and *lomefloxacin*, too. Therefore our system identifies them as interacting pairs, although this information is not provided in the sentence. These false positives could be resolved with a simple rule that prevents mentions of drugs referring to the same drug to be considered as a candidate DDI. Coordinate structures are generally used to indicate specialization. In the example, *moxifloxacin*, *ciprofloxacin,* and *lomefloxacin* are members of the group of drugs *fluoroquinolone antibiotics*. This information is represented in DINTO with the '*has role*' relationship. We believe that a more advance system could avoid these false positives by using this information (e.g., a combination of pattern recognition identifying the coordinate structure and the '*has role*' information to discard those drug pairs showing the specialization).

The lack of evidence to confirm the existence of a DDI is another source of false positives. An example is the following sentence:

« There are no clinical data on the use of **amantadine**$_{e1}$ with **ziprasidone**$_{e2}$. » (iii)

The interaction between *amantadine* and *ziprasidone* is represented in the ontology. Therefore, the system incorrectly recognizes this drug pair as a DDI mention. These false positives could be avoided integrating negation recognition (Chowdhury & Lavelli, 2013a). Moreover, identification of sentences denying or showing lack of certainty for a DDI with a plausible underlying mechanism – and therefore inferred in DINTO – could be an useful tool in the study of new DDIs, where documentation of existing literature about the DDI is an important fact to establish its relevance (Tatro, 2010).

10.2.2 Comparison in the framework of the *SemEval-2013 DDIExtraction shared task*

In the previous section, we have presented and discussed the results achieved by a simple system exploiting DINTO for DDI extraction from pharmacological texts. Here, we compare them with the results obtained by participants on the second *DDIExtraction task*.

The substask for DDI extraction from pharmacological texts attracted the attention of eight different teams, each of them submitting a maximum of three different runs. Gold-standard annotations − correct, human-created annotations − of pharmacological substances were provided to participants for both train and test data. Each participant system must output an ASCII list including all pairs of drugs in each sentence and its prediction (1 if the pair is a DDI and 0 otherwise). Evaluation was relation-oriented and based on the standard precision, recall and F-score metrics.

A detailed description of this task, the participating systems, and their results are described in **Section 4.2.2**. We can summarize that, in general, approaches based on kernels methods achieved better results than the classical feature-based methods. It is important to emphasize, as well, that most systems used primarily syntactic information, while semantic information was poorly used. The best results were submitted by the team from **FBK-irst** (Chowdhury & Lavelli, 2013b) who applied a novel hybrid kernel-based RE approach and exploited the scope of negations and semantic roles for negative instance filtering. The second best results were obtained by the **WBI team** (Thomas et al., 2013) combining several kernel methods. Regarding the different datasets, results on the DDI-DrugBank dataset were much better than on the DDI-MEDLINE dataset for all systems.

As expected based on results in **Table 10.5** and **Table 10.6,** our system based on DINTO obtains lower results than those systems specifically created to participate on the *DDIExtraction task* (**Figure 10.4**). As mentioned before, our main limitations are related to the lack of a system able to exploit ontological knowledge. Another important issue is that the ontology only covers fifty per cent of entities annotated in the test-dataset corpus. Therefore, it is not possible to identify DDIs for the other fifty per cent of entities. Most absent entities belong to the entity types group of drugs (`group`) and not-approved drugs (`drug_n`). The strategy to avoid the low coverage for `drug_n` entities has been exposed previously in **Section 10.1.1**. Regarding the first ones, the main problem is that DDIs between two groups of drugs, or between a group of drugs and an individual drug, have not been represented in the ontology. Therefore, a high number of DDIs have not been processed by our system. A future more sophisticated system could recognize group of drugs names in text using the '*role*' classes in the ontology (for a better description of this class, see **Section 7.1.3**) and use the information within the ontology to infer DDIs between groups of drugs. For example, we could identify as a group of drugs all those classes in the ontology with a *'has role'* relationship with the class *'beta-blockers.'* If all – or almost all − members have a *'may interact with'* relationship with the drug *'rifampicin'*, we could infer that there is an interaction between members of the *'beta-blockers'* group and *'rifampicin'*.

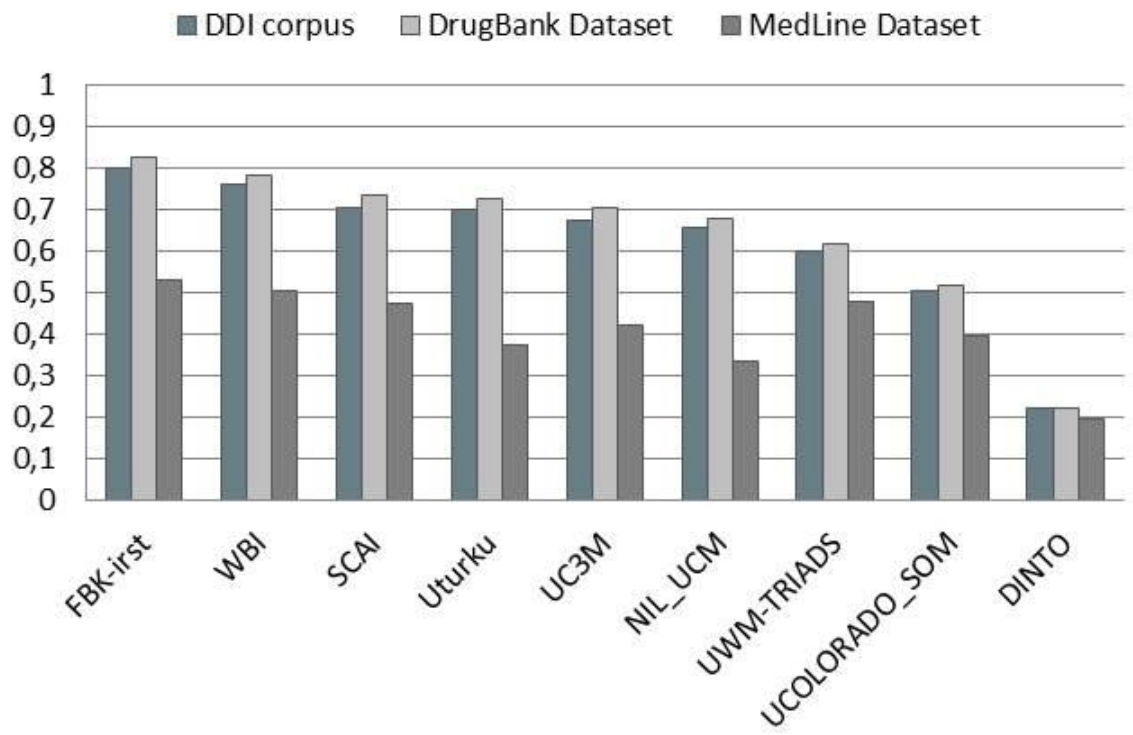

Figure 10.4. Results for the DDI extraction task for participants' best runs and for DINTO-based system (*Run 2*)

10.2.3 Ensemble *SemEval-2013 DDIExtraction task* participants and DINTO

Hybrid approaches combine different methods with the final aim to attain the best of each method and to exploit knowledge from different source. In order to investigate the impact of using DINTO on the performance obtained by participants on the *SemEval-2013 DDIExtraction shared task*, we have built different ensemble systems by combining each one of the participants' best runs with DINTO. The aim of this experiment is to identify the strengths and weaknesses of our ontology in a RE task.

The ensemble system is very simple. We take the output for each one of the eight best participants' runs. Those DDIs detected as non-interacting pairs by their systems are looked up in the ontology. If the drug pair has an interaction in DINTO, then it is confirmed as describing a DDI. Thus, the ensemble method would improve the recall of each system in the hopes of not adversely affecting precision. Since this approach was similar to that performed previously, it could be expected that drug coverage and identification of false positives would again be limited.

We compare the results achieved by each one of the original systems with two different versions of the ontology. The first one is an ensemble using a version of DINTO with only known DDIs (*Baseline*), while the second one is an ensemble using an extended version of DINTO with known and inferred DDIs (*Extended*). Again, our aim is

to investigate how the increase in the number of DDIs represented in the ontology would influence the results.

Figure 10.5 and **Figure 10.6** show the comparison of F1 scores in the DDI-DrugBank and the DDI-MEDLINE datasets, respectively. For all participants' systems, the combination with DINTO (both *Baseline* and *Extended* versions) leads to a decrease in performance for the DDI-DrugBank dataset. The reason is the previously discussed increase in false positives deriving from the use of DDI relationships without any additional technique to detect contextual information. Therefore, it is not possible to detect which sentence describes a DDI and which one does not. Therefore, although the ensemble with DINTO leads to the detection of new DDIs in the DDI-DrugBank dataset − as manifested by recall results (**Figure 10.7**) − the increase in false positives leads to a decrease in precision (**Figure 10.8**). Therefore, F1 final scores are lower for ensembles of the original system with the ontology. Regarding the two different versions tested, we observe that results are lower when the number of DDIs included in the ontology increases (*Extended* version). The reason is that as the number of possible interacting pairs increases, the number of false positives rises. Therefore, the use of DINTO as resource for IE should be combined with a mechanism that takes into account the context in which the pair drug is described in the text.

In contrast, all systems obtain better results for the DDI-MEDLINE dataset when they are combined with the first version of DINTO (*Baseline*). Regarding the second version (*Extended*), most systems achieve results similar to those obtained by the original system. Moreover, three of them improve their performances (**UTurku** (Björne, Kaewphan, & Salakoski, 2013), **NIL_UC3M** (Bokharaeian & Diaz, 2013), and **UWM_TRIADS** (Rastegar-Mojarad, Boyce, & Prasad, 2013)). Therefore, DINTO has proven to be more useful with scientific texts such as those from the DDI-MEDLINE dataset, which describe new DDIs not described elsewhere but that might have been inferred in DINTO, than with DrugBank descriptions, which collect known DDIs already represented in the ontology. As for the comparison of the two different versions, the reason for obtaining better results with the *Baseline* than with the *Extended* version is the weight of the decrease in precision derived from the increase in possible DDIs. As can be seen in **Figure 10.7**, there is an increase in Recall using *Extended* versus using *Baseline* for all systems in the DDI-MEDLINE dataset. However, the considerable number of false positives leads to a decrease in Precision (**Figure 10.8**) and therefore, a decrease in the overall F1 values.

From the results of this evaluation, we can conclude and anticipate that the creation of a new system able to exploit ontological knowledge from DINTO could increase the performance of current systems for the extraction of DDIs from texts. Specifically, we believe that the combination of advanced techniques, specifically approaches based on kernel methods that exploit lexical and syntactic information from sentences, and ontologies would demonstrate to be especially useful in complex scientific texts such as those conforming the DDI-MEDLINE dataset.

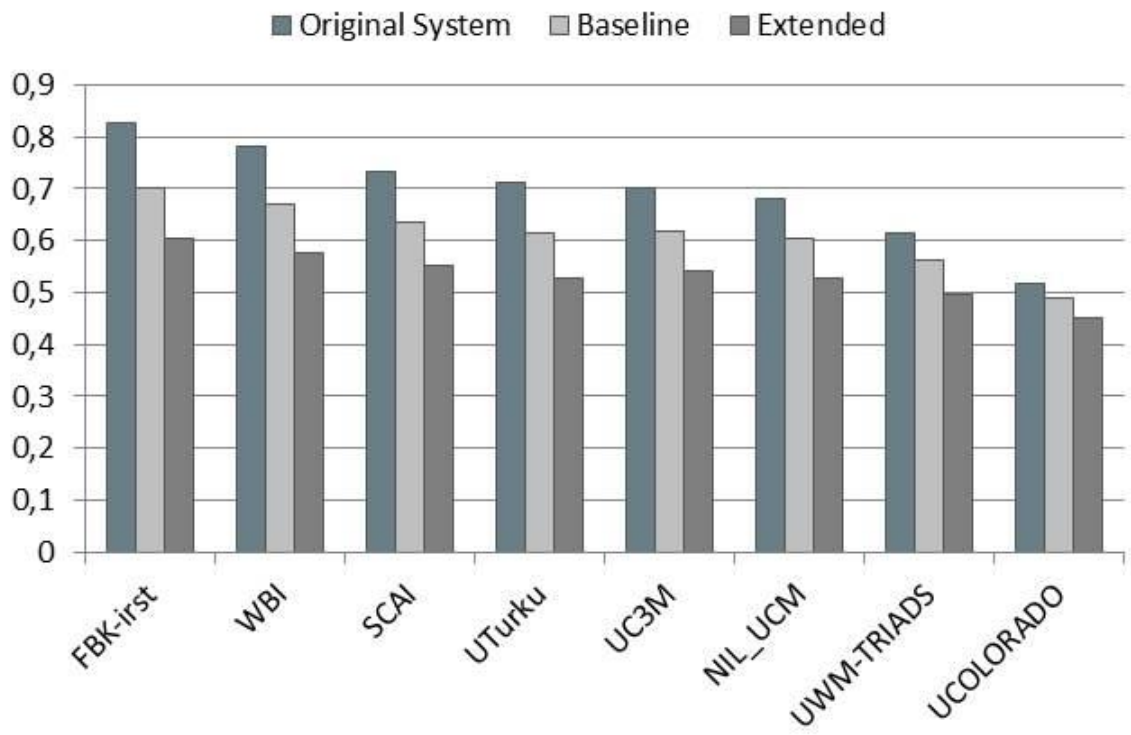

Figure 10.5. Results for the ensemble systems with DINTO for the DDI-DrugBank dataset

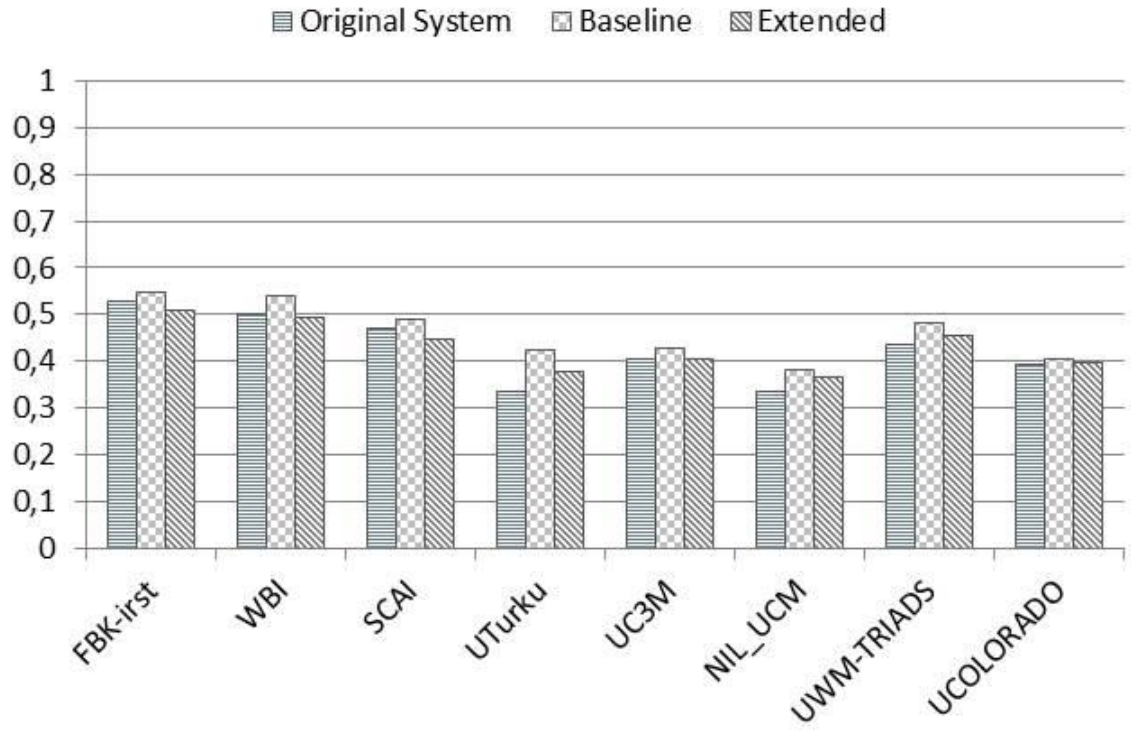

Figure 10.6. Results for the ensemble systems with DINTO for the DDI-MEDLINE dataset

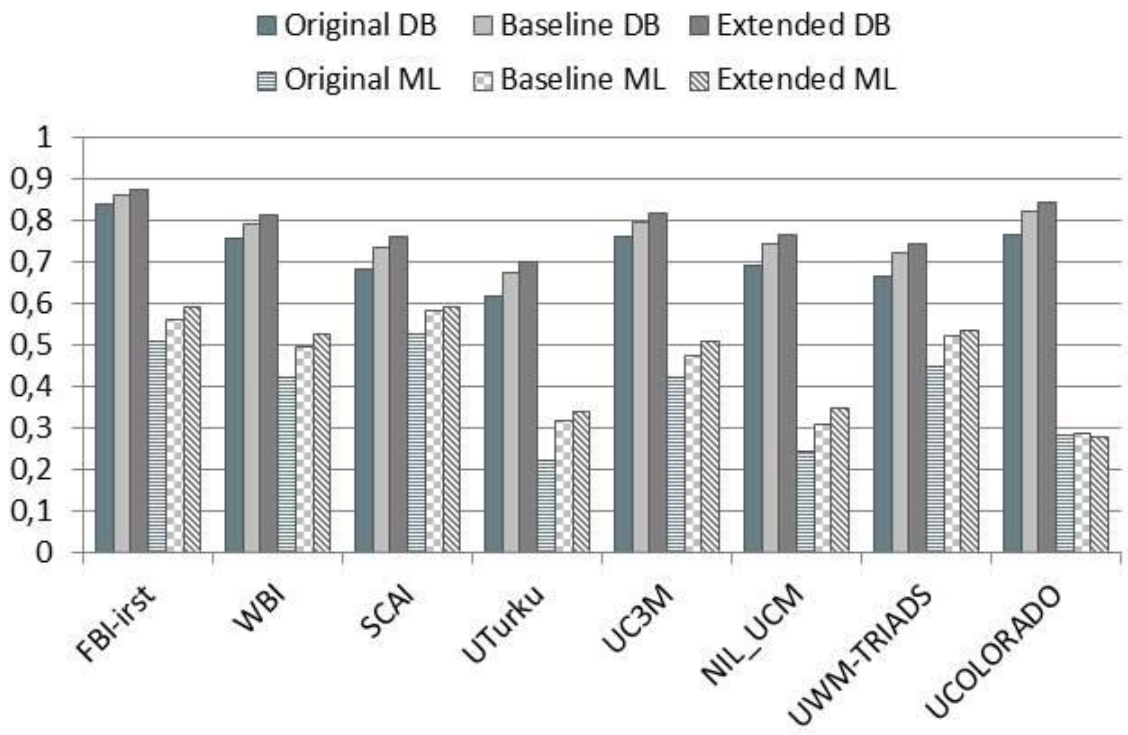

Figure 10.7. Recall results for the ensembles with DINTO for both DDI-DrugBank (solid colours) and DDI-MEDLINE (shaded colours) datasets

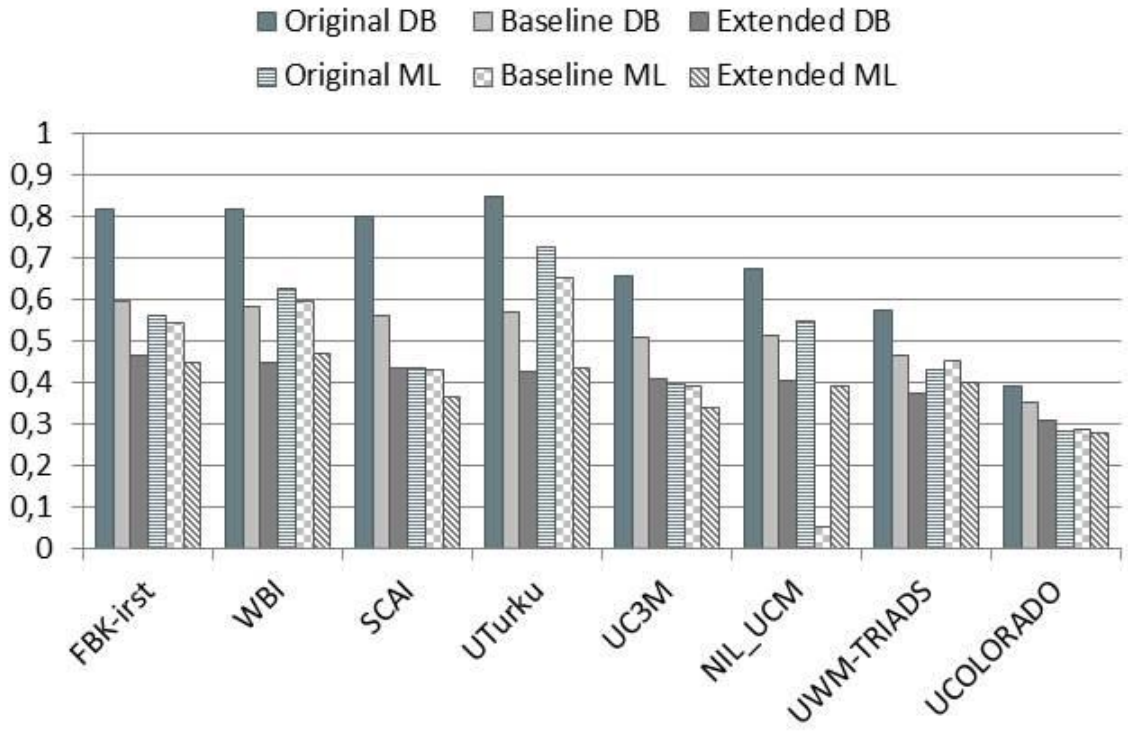

Figure 10.8. Precision results for the ensembles with DINTO for both DDI-DrugBank (solid colours) and DDI-MEDLINE (shaded colours) datasets.

10.3 Discussion and conclusions

The main objective of this evaluation is to provide a proof-of-concept of the usefulness of DINTO for IE. The content of the ontology has been previously evaluated from a technical point of view, showing that DINTO is consistent and that it adheres to ontological recommendations and standards (see **Section 8.1**). However, we want to provide insights into the potential role that our ontology could play in NER and RE. Moreover, we are interested in highlighting the strengths but also the weaknesses of DINTO, in a way that we can identify future work directions. Our main limitation is, however, the lack of a robust system able to exploit ontology knowledge for IE. Nevertheless, we have created two very simple systems to perform this evaluation, which we expect would draw some preliminary results of the potential usefulness of applying DINTO for NER and RE.

On the one hand, NER is a prior and essential step for mining useful knowledge from the biomedical literature and therefore, the extraction of drug named entities is crucial for the development of a DDI extraction system. The evaluation of a simple system based on DINTO for drug NER in text obtains comparable results to that obtained by the dictionary-based system in the *SemEval-2013 DDIExtraction shared task* (Segura-Bedmar et al., 2013). We have observed that using an extended version of the ontology – integrating information from two different resources ChEBI and DrugBank – leads to better results than using a baseline version – including only information from ChEBI.

In this evaluation, we have identified that the coverage of DINTO for two types of entities − groups of drugs (`group`) and not-approved pharmacological substances (`drug_n`) − is low in comparison with `drug` and `brand` type entities. Although the first ones are covered by the ontology as subclasses of the top-class '*role*', they have not been included in the current system. Therefore, the coverage of DINTO for group of drugs has not been evaluated. To address this limitation, we plan to create in our future work a new *OntoGazetteer* including terms describing group of drugs that are presented in the ontology. The second ones, `drug_n` substances, present a greater variation and complexity on their naming, which makes them challenging to be recognized for all different participants on the *DDIExtraction task*. Even when several of these substances have been imported from our information resources (e.g., more than five thousand substances imported in DINTO are classified as non-approved in DrugBank) the coverage is still low. The inclusion of other resources, such as the PubChem database (Bolton et al., 2008), could contribute to an even larger representation of these substances in DINTO.

Ambiguous terms and spelling variations are another important source of errors in this evaluation. Our current system is unable to overcome these problems. While ambiguity resolution is still challenging (Segura-Bedmar et al., 2013), we believe that a future system based on DINTO and integrating approximate string matching, lemmatization, or stemming techniques could overcome the spelling variation-related problem.

On the other hand, the goal of the RE task is to detect semantic relations between entities in texts. Therefore, semantic resources such as ontologies could be useful resources (Müller et al., 2004; Wimalasuriya, 2010; Zhang et al., 2012). For example,

Percha & Altman (2012) used an ontology to normalize relationships extracted from text through text mining techniques. Similarly, Coulet et al. (2011) used an ontology to map, normalize, and compare heterogeneous biomedical relationships extracted via text mining, to identify those having similar semantics but different syntax. In our case, we have evaluated DINTO for DDI extraction on the basis of a simple premise. It relies on the fact that the task of understanding a natural language sentence can be seen as identifying possible situations that the sentence in question might describe (Cimiano et al., 2014). If there were a plausible mechanism that could lead to a DDI between two drugs, it would be possible that the interaction between them might be described in the literature.

Therefore, we use a version of DINTO with known DDIs and another one with known and inferred DDIs and establish that, if a candidate interacting pair in the text is described to interact in the ontology, then the sentence is likely to be describing a DDI. The main limitation of this approach is that it does not take into account any contextual features. Therefore, the number of false positives increases considerably with the number of potential DDIs represented in the ontology. We use the results of the best systems presented by the eight different participant teams of the *SemEval DDIExtraction task* and ensemble their systems with DINTO. Although the number of correctly identified DDIs (Recall) increases for all systems in both the DDI-DrugBank and the DDI-MEDLINE datasets, overall F1 results decrease due to the increase in the number of false negatives.

During the analysis of these results, we have identified different characteristics that might contribute to the development of an IE system that could exploit the information in DINTO. For example, the use of relationships other than *'may interact with'* – the only one used in this experiment – could contribute to reducing the number of false positives due to coordinate structures or the identification of interactions between groups of drugs. For example, the system could rule out those interacting pairs of drugs appearing in a coordinate structure and having a relationship '*has role'* between them (e.g., '*moxifloxacin*' '*has role*' some '*fluoroquinolone*') or with the same group of drugs (e.g., '*moxifloxacin*' '*has role*' some '*fluoroquinolone*', and '*ciprofloxacin*' '*has role*' some '*fluoroquinolone*'). Another strategy could be to employ negation detection to identify sentences denying or showing lack of certainty for a given DDI (Bokharaeian, Diaz, Neves, & Francisco, 2014). We believe, furthermore, that the creation of a system combining DINTO with advanced techniques for RE such as kernel-based methods, which integrate contextual information from sentences, could achieve better results than those obtained in this evaluation.

In spite of the modest results obtained in this preliminary evaluation, we believe that DINTO will be a useful resource for the NLP research community. Approaches using domain knowledge have been recently applied with success to the pharmacological domain (Garten et al., 2010; Kang et al., 2014). The use of knowledge resources can reduce the number of false positives generated by the current DDI extraction systems because these resources can help to distinguish between those pairs of drugs that are DDIs from those are not. The information required for a semantic-based IE system can be taken, for example, from pharmacological databases such as DrugBank, PharmGKB (Hewett et al., 2002), SIDER (Kuhn et al., 2010), or KEGG (Kanehisa et al., 2012), among others. Some of them describe specific pairs of interacting drugs. For example, for the creation of DINTO we have used the database DrugBank. Moreover, as we have demonstrated in this thesis, a larger number of DDIs can be deduced indirectly by exploiting, for example, the drug-protein relationships. Thus, the relationships of two

different drugs with the same protein can be used to infer the mechanism leading to a DDI (Hage & Tweed, 1997). A similar approach could consists in using the relationships of two different drugs with the same ADR to infer possible DDIs (Campillos et al., 2008). For example, *morphine* is related to the side effect CNS depression. Therefore, other drugs producing the same ADR, such as *oxycodone*, could interact with *morphine*.

Up to now, the main limitation for the development of semantic-based approaches has been the availability of appropriate knowledge bases in a machine-readable format. However, the creation of these knowledge bases is becoming more feasible and common in the pharmacological domain (Khelashvili et al., 2010; Whirl-Carrillo et al., 2012). This is due to the increasing number of databases and web servers providing structured and semi-structured pharmacological information, such as Drug- Bank or KEGG. Moreover, there are different community projects such as Bio2RDF (Belleau et al., 2008) or LODD (Samwald et al., 2011), which work to link the various sources of biological and pharmacological data together, enabling the integration of several pharmacological aspects described in different databases (Pathak et al., 2013). Another important factor is the proliferation of biomedical ontologies to store and formally represent domain knowledge. Ontologies enable the integration of the information disperse through different and heterogeneous databases and provide resources that can be exploited by IE systems (Cimiano et al., 2014; Wimalasuriya, 2010).

To conclude, the results obtained in this evaluation outline the potential of DINTO as a useful resource for IE from pharmacological texts. We believe that future directions for DDI extraction might entail the combination of syntactic and semantic information and we believe that DINTO will be a useful and practical resource. We propose several specific ideas for future works, both regarding the ontology itself and with a future ontology-based IE system:

- To evaluate the coverage for groups of drugs of DINTO.
- To increase the number of substances non-approved for human use (`drug_n`) through the integration of the PubChem database.
- To design a new information extraction system able to exploit the information from DINTO. This system will have the following characteristics:
  - It should be able to recognize different types of entities (including group of drugs names).
  - It should integrate approximate string matching algorithms, lemmatization, or stemming techniques.
  - The system should exploit relationships in the ontology others than *'may interact with',* such as the *'has role'* relationship.
  - It should integrate CRF techniques for NER and kernel-based methods exploiting lexical and syntactic information for RE.

Chapter 11

Conclusions

Drug-drug interactions (DDIs) are common adverse drug reactions (ADRs) representing an important risk to patients safety, and an increase in healthcare costs (Pirmohamed et al., 2004). Their early detection is, therefore, a main concern in the clinical setting (Kulkarni et al., 2013; Moura, Acurcio, & Belo, 2009; Patel et al., 2014). Although there are different databases supporting healthcare professionals in the detection of DDIs, the quality of these databases is very uneven and the consistency of their content is limited (Clauson, Seamon, Clauson, & Van, 2004; Fulda et al., 2000; Olvey et al., 2010a). On the other hand, these databases do not scale well to the large and growing number of pharmacovigilance literature in recent years (Paczynski et al., 2012). In addition, large amounts of current and valuable information are hidden in published articles, scientific journals, books and technical reports (Aronson, 2007). Thus, the large number of DDI information sources has overwhelmed most healthcare professionals because it is not possible to remain up to date on everything published about DDIs.

Computational methods can play a key role in the identification, explanation and prediction of DDIs on a large scale, since they can be used to collect, analyze and manipulate large amounts of biological and pharmacological data (Percha & Altman, 2013). Natural language processing (NLP) techniques can be used to retrieve and extract DDI information from pharmacological texts, supporting researchers and healthcare professionals on the challenging task of searching DDI information among different and heterogeneous sources (Hansten, 2003). However, these methods rely on the availability of specific resources providing the domain knowledge, such as databases, terminological vocabularies, corpora, ontologies, and so forth, which are necessary to address the Information Extraction (IE) tasks. The main objective of this thesis is to contribute to the early detection of DDIs by developing semantic resources useful for the development of

these IE systems. In this chapter, we review the research objectives and discuss how they have been accomplished in this work, describe the dissemination of our work, and propose the future work on both the DDI corpus and DINTO as a direction for further related research.

11.1 Evaluation of research objectives

Our first objective (**Objective 1**) is the study of existing corpora relevant to the DDI domain, in a way that we can analyze their strengths and weaknesses. Since our aim is to provide a benchmark for the development and evaluation of IE systems, we review those characteristics defining a gold-standard corpus. By reviewing the literature, we establish that a gold-standard corpus is a manually annotated corpus, whose quality has been proven by the measurement of the agreement between different annotators and the creation and publication of detailed annotation guidelines. We also find that, to be useful, a gold-standard corpus must be rich in information and include a large variety of documents and annotated instances representing the diversity of document types and instances and relationships relevant to the intended task. Finally, another important characteristic of a gold-standard is its acceptance and usage by the research community. Based on these characteristics, we review and analyze the different available corpora annotated *i*) with pharmacological substances, and *ii*) with DDIs. From this analysis, we conclude that there is not existing corpora that can be used as gold-standard for training and evaluation of NLP techniques for the DDI domain.

Consequently, we pursue our second objective: the creation of the DDI corpus (**Objective 2**). To do this, we try to overcome the main limitations found in previous corpora, and attempt to fulfil the characteristics of a gold-standard corpus. The starting point is the DrugDDI corpus, on which several improvements have been made. Firstly, new documents are included in the corpus, in a way that the DDI corpus consists of DrugBank texts and MEDLINE abstracts. With these two types of documents, the size of the corpus is significantly increased and we provide a source of texts with different complexity and linguistic patterns. Thus, NLP systems can be trained to extract information from different types of document. Secondly, a new annotation process is defined. In the DDI corpus, two experts in pharmacology and two text miners with background in pharmacovigilance participate in the annotation task. A new annotation schema is created during the annotation of a training set, and annotation guidelines are developed in an iterative process. They provide a detailed description of different types of pharmacological entities, along with different types of DDIs, too. In this way, the resulting corpus is the most richly semantically annotated resource for pharmacological text processing built to date. Thirdly, the agreement between annotators is measured in order to assess the level of difficulty of the annotation task, along to the quality of the corpus. The annotated corpus and the annotation guidelines are made freely available for the research community at http://labda.inf.uc3m.es/ddicorpus.

With the aim to validate the usefulness of the DDI corpus to the intended task, we use it as a benchmark for training and evaluation of NLP systems devoted to the NER and RE tasks in the DDI domain (**Objective 3**). The second edition of the DDIExtraction shared task series provides a benchmarking framework for the evaluation of IE systems.

These systems are trained and evaluated using the DDI corpus for two different tasks: *i*) drug named entity recognition and classification, and *ii*) extraction and classification of DDIs. The task has attracted a great deal of attention from several research groups, which have shown a significant improvement with respect to the previous 2011 edition. The increase in size of the corpus, the inclusion of different types of documents and the quality of their annotations might have been a significant contribution to this improvement. The participating systems and their results are described in detailed in this thesis.

In addition to the *SemEval DDIExtraction task* participants, we believe that the DDI corpus will be used by other NLP research groups working on the pharmacological domain. Indeed, our research group keeps encouraging research on this field and has organized the "*Special Issue Special Issue on Mining the Pharmacovigilance Literature*" in the *Journal of Biomedical Informatics*, a special issue on automatic extraction of relationships between biomedical entities relevant to the pharmacovigilance field launched on June 2014 and that will be published in 2015. In the call for papers, submissions that use the DDI corpus are especially welcomed because their results can be compared with those reported in the second *DDIExtraction task.*[58] Therefore, we anticipate that the DDI corpus will be further used by different research groups in different settings.

From this, we conclude that one of the main objectives of this thesis, the creation of a corpus that contributes to the research and development by the NLP research community of automatic tools for the early detection of DDIs, has been achieved.

The DDI corpus is used to study the different linguistic phenomena in texts describing DDI information (**Objective 4**). Linguistic aspects of drug names and common syntactic phenomena in pharmacological text, such as hypernymic propositions or coordinate structures, make the manual annotation of documents difficult and, therefore, might represent obstacles for automatic IE systems. We believe that the review of these annotation issues provided in this thesis can be a useful guide for their developers. In addition to this, we have used the DDI corpus to perform a linguistic pattern analysis, which shows how the different relevant concepts are described in pharmacological texts, and keyword and concordance analyses, which show how these concepts relate to each other in this domain. The gained knowledge is applied to the development of the CM of the ontology for drug-drug interactions ontology DINTO.

However, before creating this new resource, we have studied the existing semantic resources relevant in pharmacovigilance, with special interest on the representation of DDI-related information (**Objective 5**). We have focused on their scope and level of granularity for the representation of pharmacological substances and DDIs, and the representation of other relevant concepts, such as ADRs, too. Furthermore, we have reviewed and compared the different approaches that, to the best of our knowledge, have dealt in some way with the task of conceptualization of DDI knowledge, and have identified and studied their unresolved issues (**Objective 6**). From these reviews, we conclude that no existing ontology or conceptual model represents in a comprehensive way all the DDI-related information necessary for its application to IE or computational prediction of DDIs.

[58] http://www.sciencedirect.com/science/journal/15320464/49

Based on this analysis, we have created an ontology for the representation of all DDI-related knowledge (**Objective 7**). To the best of our knowledge, DINTO is the first ontology representing both PK and PD DDIs, and all their related mechanisms. The conceptualization in DINTO provides the most comprehensive CM for the DDI domain, including crucial aspects such as DDI mechanism, effect, or recommendations. It reuses information from related ontological and non-ontological resources, decreasing the costs of ontology population and increasing its coverage. It has been evaluated from a technical point of view in a three-step strategy to ensure its consistency, completeness and quality. DINTO adheres to the OBO Foundry principles and other ontology development standards. Specifically, it has been developed in collaboration with the ChEBI ontology, one of the most important OBO Foundry ontologies. Moreover, DINTO has been reviewed by the OBO Foundry community and has been accepted and listed as one of the *"OBO Foundry candidate ontologies and other ontologies of interest."*[59] In addition to this, we have been invited to collaborate in the development of a new resource named the Drug-drug International and Drug-drug Interaction Evidence Ontology (DIDEO), a project involving people from different Institutions and leaded at the *Division of Biomedical Informatics* at the University of Arkansas and the *Department of Biomedical Informatics* at the University of Pittsburgh (Brochhausen et al., 2014).

In addition to this technical evaluation, we have tested the robustness of DINTO in two different applications (**Objective 8**). Firstly, we have used it to infer new DDIs and their mechanisms. The combination of OWL drug-protein relationships and SWRL rules representing DDI mechanisms has led to the first experiment inferring both PK and PD DDIs in the same representation framework and in a large scale. Moreover, it has provided the most detailed inference of different DDI mechanisms. Therefore, this evaluation has shown that DINTO is a good resource for inference of DDIs and DDI-related knowledge. Secondly, we have evaluated the usefulness of DINTO in an IE system for NER and DDI extraction, and have compared the results with those obtained by the DDIExtraction-2013 task participants. The preliminary results are modest due to the simplicity of the IE method used, because the development of an IE system is out of the scope of this thesis. Nevertheless, we think that a more sophisticated method using our ontology may obtain better results. Furthermore, in this evaluation we have identified the main characteristics that should fulfil a future IE system based on DINTO, which might serve as a roadmap for future work in this field.

From this, we can conclude that the twofold objective of this thesis has been achieved. Firstly, we have contributed to the improvement of the early detection of DDIs from scientific literature through the development of two different resources, an annotated corpus and a comprehensive ontology, which have enabled the development, training and evaluation of automatic NLP systems for pharmacological texts in the field of DDIs. And secondly, we have applied the ontology to the inference of new knowledge, and in particular, to the inference of new DDIs and their mechanisms that could not have been reported in biomedical publications.

[59] http://www.obofoundry.org/

11.2 Publications

As a result of this work, several publications have been presented in workshops, conferences, and specific journals.

In *The DDI corpus: an annotated corpus with pharmacological entities and drug-drug interactions* (Journal of Biomedical Informatics 46 (2013), pp. 914-920*)*, we have described the construction of the DDI corpus, including how the text were collected and processed, the annotation guidelines and the annotation process, and the evaluation through the measurement of the agreement between different annotators.

In *Annotation issues in pharmacological texts* (Procedia – Social and Behavioral Sciences. 5th International Conference on Corpus Linguistics (CILC2013) 2013, vol. 95, pp. 211–219), we have reviewed the main sources of annotation problems that affect in general the manual annotation process, the linguistic phenomena that complicate the manual annotation of drug named entities, and the different syntactic phenomena that should be considered during the annotation of DDIs.

In *Lessons learnt from the DDIExtraction-2013 Shared Task* (Journal of Biomedical Informatics 51 (2014), pp.152–164), we have described the second edition of the *DDIExtraction Shared Task series*, which provided the DDI corpus as a benchmark for the implementation and comparative assessment of IE systems for drug NER and DDI extraction. The paper focuses on the latter one, providing a description and comparison of the systems and their results, an analysis of the major sources of their errors, and proposes an ensemble system combining the top three methods using majority and union voting strategies.

In *An ontology for drug-drug interactions* (6th International Workshop on Semantic Web Applications and Tools for Life Sciences (SWAT4LS) 2014, Edinburgh, UK), we have described the first steps in the development of DINTO that were carried out during the internship at the European Bioinformatics Institute (EMBL-EBI).

All the work described here has been supported by three research projects: the Research Network MA2VICMR [S2009/TIC-1542] of the Regional Government of Madrid, the project MULTIMEDICA [TIN2010-20644-C03-01] supported by the Spanish Ministry of Education, and the TrendMiner project [FP7-ICT287863] in the European Commission Seventh Framework Programme.

11.3 Future work

From the experience gained in the construction of the DDI corpus, the identification of annotation issues in the pharmacological literature (**Section 3.4**), the quantitative features of the annotation process (**Section 3.5**), and the IAA (**Section 4.1**), as well as from the lessons learnt from the *SemEval-2013 DDIExtraction shared task* (**Section 4.2**), we have identified directions for future improvements of the DDI corpus:

1. Inclusion of a larger number of MEDLINE abstracts, in order to obtain a more balanced corpus respecting to the number of DrugBank texts, as well as other information resources such as package inserts, patient records, case studies, and discharge summaries, among others.

2. More detailed description of `drug_n` entities, which can be divided into different groups in order to reduce the bias in the annotation of this entity type.

3. Multiple-annotation process involving at least three annotators. In this strategy, all documents are annotated by more than one annotator (Jagannathan et al., 2009; Wilbur et al., 2006).

4. Selection of a more advanced annotation tool, facilitating the annotation by multiple-annotators (Neves & Leser, 2014).

5. Annotation of linguistic phenomena required for a better understanding of the text, such as negation, modality, cataphora, or anaphora.

6. Annotation of DDIs at the document level instead of sentence level, to capture those interactions spanning several sentences.

7. Annotation of relevant pharmacological information, including quantitative information (drug dosage, time interval between administration of the drugs, alterations in PK parameters, drug concentration, etc.), and qualitative information (adverse drug reactions, indication, pharmaceutical form, administration route, etc.). This information can be useful for the development of new IE systems for both drug NER and DDI extraction.

8. Annotation of relevant DDI-related information, such as protein entities involved in the DDI mechanism and their relationships with other instances in the text.

9. Inclusion of metadata in the annotated corpora, such as provenance of the documents, alternative IDs for the pharmacological entities, annotator identification, etc.

10. Integration of the DDI corpus with other related corpora, such as the PK and PK-DDI corpus. One of the approach proposed to achieve this interoperability between corpora is the conversion of corpora and the systems output to appropriate formats, such as the Resource Description Framework (RDF) format to represent the annotations, and the use of OWL ontologies to model meaning of annotations (Klein et al., 2014).

Regarding DINTO, the evaluations in inference of DDIs and their mechanisms (**Chapter 9**) and ontology-based IE (**Chapter 10**) and have brought future lines of research that could increase the usefulness of the ontology.

1. Evaluation of the coverage of group of drugs in DINTO.

2. Inclusion of a larger number of substances not approved for human use (`drug_n` type entities), integrating for example information from the PubChem database (Bolton et al., 2008).

3. Inclusion of ADR information, integrating the OAE (He et al., 2011) or AERO (Courtot et al., 2011) ontologies, for example, and translating drug-ADR relationships from the SIDER database (Kuhn et al., 2010).

4. Inclusion of therapeutic index of drugs, which can be used to identify clinically significant DDIs (Boyce et al., 2007)

5. Inclusion of drug bioactivity data for the inference of the principal DDI mechanism (Gaulton et al., 2012).

6. Inclusion of physicochemical properties of drugs, which can be used to identify DDIs occurring by chelation.

7. Inclusion of new drug-protein relationships, which can lead to the prediction of DDIs that cannot be explained by current knowledge.

8. Maintenance of DINTO and collaboration with other related efforts, such as the Drug-drug Interaction and Drug-drug Interaction Evidence Ontology (DIDEO) (Brochhausen et al., 2014).

Glossary

ADE	Adverse Drug Event
ADR	Adverse drug reaction
AE	Adverse Effect
AERO	Adverse Event Reporting Ontology
ARM	Adverse Reactions and Mechanism Ontology
ASP	Answer Set Programming
ATC	Anatomical, Therapeutic and Chemical
BFO	Basic Formal Ontology
BRO	Biomedical Resource Ontology
CDSS	Clinical Decision Support System
ChEBI	Chemical Entities of Biomedical Interest
CHEMDNER	Chemical compound and Drug Name Recognition
CLEF	Clinical E-Science Framework
CM	Conceptual Model
CNS	Central nervous system
COSTART	Coding Symbols for a Thesaurus of Adverse Reaction Terms
CPRS	Computerized Patient Record System
CQ	Competency Question
CRF	Conditional Random Fields
CS	Classification Scenario
CUI	Concept Unique Identifier
DDI	Drug-Drug Interaction
DIDEO	Drug-drug Interaction and Drug-drug Interaction Evidence Ontology
DIKB	Drug Interaction Knowledge Base
DINTO	Drug-drug Interactions Ontology
DIO	Drug Interaction Ontology

DL	Description Logics
DOE	*Denominación Oficial Española*
DrOn	Drug Ontology
DTD	Document Type Definition
EHR	Electronic Health Record
EMA	European Medicines Agency
EMBL-EBI	European Bioinformatics Institute
EU-ADR	Exploring and Understanding Adverse Drug Reactions
FDA	Food and Drugs Administration
FMA	Foundational Model Anatomy
FOL	First Order Logic
GO	Gene Ontology
IAA	Inter-annotator agreement
ICH	International Conference on Harmonisation
IE	Information Extraction
IHTSDO	International Health Terminology Standards Development Organisation
INN	International Nonproprietary Name
IR	Information Retrieval
IUPAC	International Union of Pure and Applied Chemistry
Jochem	Joint Chemical Dictionary
KA	Knowledge Acquisition
MedDRA	Medical Dictionary for Regulatory Activities
MEMM	Maximum Entropy Markov Model
MeSH	Medical Subject Headings
MMTx	MetaMap Transfer Tool
NCI	National Cancer Institute
NCIT	National Cancer Institute Thesaurus
NDC	National Drug Code
NDF	National Drug File
NDF-RT	National Drug File Reference Terminology
NER	Named Entity Recognition
NLM	National Library of Medicine
NLP	Natural Language Processing
OAE	Ontology of Adverse Events

OBO Foundry	Open Biological and Biomedical Ontologies Foundry
OSCAR	Open-Source Chemistry Analysis Routines
ORSD	Ontology Requirements Specification Document
OWA	Open World Assumption
OWL	Web Ontology Language
PD	Pharmacodynamic
PDO	Pharmacodynamics Ontology
PI	Package Insert
PK	Pharmacokinetic
PKO	Pharmacokinetics ontology
POS	Part-of-speech
RDF	Resource Description Framework
RE	Relation Extraction
RN	Registry Number
RO	Relation Ontology
SADL	Semantic Application Design Language
SIO	Semanticscience Integrated Ontology
SNOMED CT	SNOMED Clinical Terms
SPC	Summary of Product Characteristics
SVM	Support Vector Machines
SWRL	Semantic Web Rules Language
ITI TXM	Tissue Expressions and Protein-Protein Interactions
UML	Unified Modeling Language
UMLS	Unified Medical Language System
UNII	Unique Ingredient Identifier
URI	Unique Resource Identifier
USAN	United States Adopted Name
VA	Veterans Affairs
VHA	Veterans Health Administration
W3C	World Wide Web Consortium
WHO	World Health Organization
WHO-ART	WHO's Adverse Reaction Terminology
XML	Extensible Markup Language

Bibliography

Agence française de sécurité sanitaire des produits de santé. (2006). *Thésaurus des interactions médicamenteuses* (p. 165). Paris.

Alex, B., Grover, C., Haddow, B., & Kabadjov, M. (2008). The ITI TXM corpora: Tissue expressions and protein-protein interactions. In *Proceedings of the LREC Workshop on Building and Evaluating Resources for Biomedical Text Mining* (pp. 11–8).

Alexandrou, D. Al, Pardalis, K. V, Bouras, T. D., Karakitsos, P., & Mentzas, G. N. (2012). SEMPATH Ontology: modeling multidisciplinary treatment schemes utilizing semantics. *IEEE Transactions on Information Technology in Biomedicine : A Publication of the IEEE Engineering in Medicine and Biology Society*, *16*(2), 235–40. doi:10.1109/TITB.2011.2161588

Ananiadou, S., Kell, D. B., & Tsujii, J. (2006). Text mining and its potential applications in systems biology. *Trends in Biotechnology*, *24*(12), 571–9. doi:10.1016/j.tibtech.2006.10.002

Andronis, C., Sharma, A., Virvilis, V., Deftereos, S., & Persidis, A. (2011). Literature mining, ontologies and information visualization for drug repurposing. *Briefings in Bioinformatics*, *12*(4), 357–68. doi:10.1093/bib/bbr005

Antonelli, D., Atar, S., Freedberg, N. A., & Rosenfeld, T. (2005). Torsade de Pointes in Patients on Chronic Amiodarone Treatment : Contributing Factors and Drug Interactions. *The Israel Medical Association Journal: IMAJ*, *7*(March), 2–4.

Aramaki, E., Miura, Y., Tonoike, M., Ohkuma, T., Mashuichi, H., & Ohe, K. (2009). TEXT2TABLE: Medical Text Summarization System based on Named Entity Recognition and Modality Identification. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing* (pp. 185–192). Association for Computational Linguistics.

Arikuma, T., Yoshikawa, S., Azuma, R., Watanabe, K., Matsumura, K., & Konagaya, A. (2008). Drug interaction prediction using ontology-driven hypothetical assertion framework for pathway generation followed by numerical simulation. *BMC Bioinformatics*, *9 Suppl 6*(Suppl 6), S11. doi:10.1186/1471-2105-9-S6-S11

Aronson, A. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings / AMIA ... Annual Symposium. AMIA Symposium* (pp. 17–21).

Aronson, A. R., & Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, *17*(3), 229–36. doi:10.1136/jamia.2009.002733

Aronson, J. K. (2004). Drug interactions-information, education, and the British National Formulary. *British Journal of Clinical Pharmacology*, *57*(4), 371–2. doi:10.1111/j.1365-2125.2004.02125.x

Aronson, J. K. (2007). Communicating information about drug interactions. *British Journal of Clinical Pharmacology*, *63*(6), 637–639. doi:10.1111/j.1365-2125.2007.02948.x

Artstein, R., & Poesio, M. (2005). Bias decreases in proportion to the number of annotators. In *Proceedings of FG-MoL* (pp. 141–150).

Back, D., & Else, L. (2013). The importance of drug–drug interactions in the DAA era. *Digestive and Liver Disease*, *45*(5), S343–S348.

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., … Hunter, L. E. (2012). Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, *13*(1), 161. doi:10.1186/1471-2105-13-161

Baker, P., Hardie, A., & McEnery, T. (2006). *A Glossary of Corpus Linguistics* (p. 192). Edinburgh University Press.

Barillot, M. J., Sarrut, B., & Doreau, C. G. (1997). Evaluation of drug interaction document citationin nine on-line bibliographic databases. *Annals of Pharmacotherapy*, *31*(1), 45–49.

Bashyam, V., Divita, G., Bennet, D. B., Browne, A. C., & Taira, R. K. (2007). A Normalized Lexical Lookup Approach to Identifying UMLS Concepts in Free Text. *Studies in Health Technology and Informatics*, *129*(1), 545–9.

Baxter, K. (2013). *Stockley's Drug Interactions*. (K. Baxter & C. L. Preston, Eds.) (10th ed., p. 1680). London: Pharmaceutical Press.

Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., & Morissette, J. (2008). Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, *41*(5), 706–16. doi:10.1016/j.jbi.2008.03.004

Benet, L. Z. (1984). Pharmacokinetic parameters: which are necessary to define a drug substance? *European Journal of Respiratory Diseases*, *134*, 45–61.

Bergk, V., Haefeli, W. E., Gasse, C., Brenner, H., & Martin-Facklam, M. (2005). Information deficits in the summary of product characteristics preclude an optimal management of drug interactions: a comparison with evidence from the literature. *European Journal of Clinical Pharmacology*, *61*(5-6), 327–35. doi:10.1007/s00228-005-0943-4

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc.

Bisgin, H., Liu, Z., Fang, H., Xu, X., & Tong, W. (2011). Mining FDA drug labels using an unsupervised learning technique--topic modeling. *BMC Bioinformatics*, *12 Suppl 1*(Suppl 10), S11. doi:10.1186/1471-2105-12-S10-S11

Björne, J., Kaewphan, S., & Salakoski, T. (2013). UTurku : Drug Named Entity Recognition and Drug-Drug Interaction Extraction Using SVM Classification and Domain Knowledge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (pp. 651–659). Atlanta, USA.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, *32*(Database issue), D267–70. doi:10.1093/nar/gkh061

Bodenreider, O. (2008). Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearbook of Medical Informatics*, *3841*, 67–79.

Bodenreider, O., & Stevens, R. (2006). Bio-ontologies: current trends and future directions. *Briefings in Bioinformatics*, *7*(3), 256–74. doi:10.1093/bib/bbl027

Bokharaeian, B., & Diaz, A. (2013). NIL_UCM : Extracting Drug-Drug interactions from text through combination of sequence and tree kernels. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (Vol. 2, pp. 644–650). Atlanta, USA.

Bokharaeian, B., Diaz, A., Neves, M., & Francisco, V. (2014). Exploring Negation Annotations in the DrugDDI Corpus. In *Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BIOTxtM 2014)*. Reykjavik.

Bolton, E. E., Wang, Y., Thiessen, P. A., & Bryant, S. H. (2008). PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Annual Reports in Computational Chemistry* (Volume 4.). Washington, DC: American Chemical Society.

Bonatti, P., Calimeri, F., Leone, N., & Ricca, F. (2010). Answer Set Programming. In *A 25-year perspective on logic programming* (pp. 159–182). Springer-Verlag.

Bonnabry, P., Sievering, J., Leemann, T., & Dayer, P. (1999). Quantitative drug interactions prediction system (Q-DIPS): a computer-based prediction and management support system for drug metabolism interactions. *European Journal of Clinical Pharmacology*, *55*(5), 341–7.

Bontcheva, K., Tablan, V., Maynard, D., & Cunningham, H. (2004). Evolving GATE to meet new challenges in language engineering. *Natural Language Engineering*, *10*(3-4), 349–373. doi:10.1017/S1351324904003468

Boring, D. (1997). The development and adoption of nonproprietary, established, and proprietary names for pharmaceuticals. *Drug Inf J*, *31*, 621–34.

Boyce, R., Collins, C., Horn, J., & Kalet, I. (2007). Modeling drug mechanism knowledge using evidence and truth maintenance. *IEEE Transactions on Information Technology in Biomedicine*, *11*(4), 386–97.

Boyce, R., Collins, C., Horn, J., & Kalet, I. (2009). Computing with evidence Part I: A drug-mechanism evidence taxonomy oriented toward confidence assignment. *Journal of Biomedical Informatics*, *42*(6), 979–89. doi:10.1016/j.jbi.2009.05.001

Boyce, R., Collins, C., Horn, J., & Kalet, I. (2010a). Computing with evidence part I: A drug-mechanism evidence taxonomy oriented toward confidence assignment. *Journal of Biomedical Informatics*, *42*(6), 979–989. doi:10.1016/j.jbi.2009.05.001.

Boyce, R., Collins, C., Horn, J., & Kalet, I. (2010b). Computing with evidence part II: an evidential approach to predicting metabolic drug-drug interactions. *Journal of Biomedical Informatics*, *42*(6), 990–1003. doi:10.1016/j.jbi.2009.05.010.

Boyce, R., Collins, C., Horn, J., & Kalet, I. J. (2004). Qualitative pharmacokinetic modeling of drugs. *AMIA Annual Symposium Proceedings*, 71–5.

Boyce, R., Gardner, G., & Harkema, H. (2012). Using Natural Language Processing to Identify Pharmacokinetic Drug-Drug Interactions Described in Drug Package Inserts. In *Proceedings of the 2012 Workshop on BioNLP* (pp. 206–213).

Brank, J., Grobelnik, M., & Mladeni, D. (2005). A survey of ontology evaluation techniques. In *Data Mining and Data Warehouses (SiKDD)*.

Brenninkmeijer, C. Y. A., Dunlop, I., Goble, C., Gray, A. J. G., Pettifer, S., & Stevens, R. (2013). Computing Identity Co-Reference Across Drug Discovery Datasets. In *Semantic Web Applications and Tools for Life Sciences*.

Brochhausen, M., Schneider, J., Malone, D., Empey, P. E., Hogan, R., & Boyce, R. D. (2014). Towards a foundational representation of potential drug- drug interaction knowledge. In *Drug-Drug Interaction Knowledge Representation Workshop (DIKR). In International Conference on Biomedical Ontologies (icbo14)*. Houston, US.

Brown, E., Wood, L., & Wood, S. (1999). The medical dictionary for regulatory activities (MedDRA). *Drug Safety : An International Journal of Medical Toxicology and Drug Experience*, *20*(2), 109–17.

Bryant, A. D., Fletcher, G. S., & Payne, T. H. (2014). Drug interaction alert override rates in the meaningful use era: no evidence of progress. *Applied Clinical Informatics*, *5*(3), 802–13. doi:10.4338/ACI-2013-12-RA-0103

Cami, A., Manzi, S., Arnold, A., & Reis, B. Y. (2013). Pharmacointeraction network models predict unknown drug-drug interactions. *PloS One*, *8*(4), e61468. doi:10.1371/journal.pone.0061468

Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J., & Bork, P. (2008). Drug target identification using side-effect similarity. *Science*, *321*(5886), 263–6. doi:10.1126/science.1158140

Cano, C., Monaghan, T., Blanco, A., Wall, D. P., & Peshkin, L. (2009). Collaborative text-annotation resource for disease-centered relation extraction from biomedical text. *Journal of Biomedical Informatics*, *42*(5), 967–77. doi:10.1016/j.jbi.2009.02.001

Carter, J. S., Brown, S. H., Erlbaum, M. S., Gregg, W., Elkin, P. L., Speroff, T., & Tuttle, M. S. (2002). Initializing the VA Medication Reference Terminology Using UMLS Metathesaurus Co-Ocurrences. In *AMIA Symposium* (pp. 116–120).

Charniak, E., & Johnson, M. (2005). Coarse-to-fine n -best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (Vol. 1, pp. 173–180). Association for Computational Linguistics.

Cheng, F., & Zhao, Z. (2014). Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *Journal of the American Medical Informatics Association : JAMIA*, *21*(e2), e278–86. doi:10.1136/amiajnl-2013-002512

Chinchor, N., & Sundheim, B. (1993). MUC-5 EVALUATION METRICS. In *5th conference on Message Understanding* (pp. 69–78). Association for Computational Linguistics.

Chowdhury, F. M., & Lavelli, A. (2013a). Exploiting the Scope of Negations and Heterogeneous Features for Relation Extraction : A Case Study for Drug-Drug Interaction Extraction. In *2013 Conference of the North American Chapter of the Associa- tion for Computational Linguistics: Human Language Technology (NAACL 2013)* (Vol. 2004, pp. 765–771). Atlanta, USA.

Chowdhury, F. M., & Lavelli, A. (2013b). FBK-irst : A Multi-Phase Kernel Based Approach for Drug-Drug Interaction Detection and Classification that Exploits Linguistic Information. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (Vol. 2, pp. 351–355).

Cimiano, P., Unger, C., & McCrae, J. (2014). Ontology-Based Interpretation of Natural Language. *Synthesis Lectures on Human Language Technologies*, *7*(2), 1–178. doi:10.2200/S00561ED1V01Y201401HLT024

Cimino, J. J. (1998). Desiderata for Controlled Medical Vocabularies in the Twenty-First Century. *Methods Inf Med*, *37*(4-5), 394–403.

Cimino, J. J. (2006). In defense of the Desiderata. *Journal of Biomedical Informatics*, *39*(3), 299–306. doi:10.1016/j.jbi.2005.11.008

Cimino, J. J., York, N., Huff, S., Care, I. H., City, S. L., Broverman, C., … Library, N. (1998). Development of a Standard Terminology to Support Medication Messages. *Journal of the American Medical Informatics Association*.

Clauson, K. A., Seamon, M. J., Clauson, A. S., & Van, T. B. (2004). Evaluation of drug information databases for personal digital assistants. *American Journal of Health-System Pharmacy*, *61*(10), 1015–24. doi:1079-2082/04/0502-1015$06.00

CodePlex. (2014). *XML Notepad*. *Microsoft*. Retrieved December 09, 2014, from https://xmlnotepad.codeplex.com/

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cohen, K. B., Ogren, P. V, Fox, L., & Hunter, L. (2005). Corpus design for biomedical natural language processing. In K. Cohen, L. Hirschman, H. Shatkay, & C. Blaschke (Eds.), *ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics* (pp. 38–45). Detroit. Retrieved from http://dl.acm.org/citation.cfm?id=1641490

Cone, E. J., Fant, R. V, Rohay, J. M., & Associates, P. (2004). Oxycodone Involvement in Drug Abuse Deaths . II . Evidence for Toxic Multiple Drug-Drug Interactions. *Journal of Analytical Toxicology*, *28*(June), 217–225.

Consortium, T. U. (2008). The universal protein resource (UniProt). *Nucleic Acids Research*, *36*(Database issue), D190–5. doi:10.1093/nar/gkm895

Corbett, P., Batchelor, C., House, T. G., & Teufel, S. (2007). Annotation of Chemical Named Entities. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing* (pp. 57–64).

Coulet, A., Garten, Y., Dumontier, M., Altman, R. B., Musen, M. a, & Shah, N. H. (2011). Integration and publication of heterogeneous text-mined relationships on the Semantic Web. *Journal of Biomedical Semantics*, *2 Suppl 2*(Suppl 2), S10. doi:10.1186/2041-1480-2-S2-S10

Coulet, A., Smaïl-Tabbone, M., Napoli, A., & Devignes, M. D. (2006). Suggested Ontology for Pharmacogenomics (SO-Pharm): Modular Construction and Preliminary Testing. In *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops* (pp. 648–657). Springer Berlin Heidelberg.

Courtot, M., Brinkman, R. R., & Ruttenberg, A. (2011). Reporting Adverse Events: Basis for a Common Representation. In *International Conference on Biomedical Ontology. Representing Adverse Events Workshop* (Vol. 02, pp. 3–10). Buffalo, NY, USA.

Cristani, M., & Cuel, R. (2004). A comprehensive guideline for building a domain ontology from scratch. In *International Conference on Knowledge Management (I-KNOW'04)*. Graz, Austria.

Croset, S., Hoehndorf, R., & Rebholz-Schuhmann, D. (2012). Integration of the Anatomical Therapeutic Chemical Classification System and DrugBank using OWL and text-mining. In *4 th WORKSHOP OF THE GI WORKGROUP "ONTOLOGIES IN BIOMEDICINE AND LIFE SCIENCES" (OBML)* (pp. 23–26). Dresden, Germany.

Cuenca, B. (2011). Comparison of reasoners for large ontologies in the OWL 2 EL profile. *Semantic Web*, *2*, 71–87. doi:10.3233/SW-2011-0034

Cunningham, H., Tablan, V., Roberts, A., & Bontcheva, K. (2013). Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Computational Biology*, *9*(2), e1002854. doi:10.1371/journal.pcbi.1002854

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. (1992). A Practical Part-of-Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing* (pp. 133–140).

Day-Richter, J., Harris, M. a, Haendel, M., & Lewis, S. (2007). OBO-Edit--an ontology editor for biologists. *Bioinformatics (Oxford, England)*, *23*(16), 2198–200. doi:10.1093/bioinformatics/btm112

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., … Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, *36*(Database issue), D344–50. doi:10.1093/nar/gkm791

Deléger, L., Grouin, C., & Zweigenbaum, P. (2010). Extracting medical information from narrative patient records: the case of medication-related information. *Journal of the American Medical Informatics Association : JAMIA*, *17*(5), 555–8. doi:10.1136/jamia.2010.003962

Deleger, L., Li, Q., Lingren, T., Kaiser, M., Molnar, K., Stoutenborough, L., … Marsolo, K. (2012). Building Gold Standard Corpora for Medical Natural Language Processing Tasks. *AMIA Annual Symposium Proceedings Archive*, *2012*, 144–153.

Deleger, L., Lingren, T., Ni, Y., Kaiser, M., Stoutenborough, L., Marsolo, K., … Solti, I. (2014). Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research. *Journal of Biomedical Informatics*, *50*, 173–83. doi:10.1016/j.jbi.2014.01.014

Dipper, S., Götze, M., & Skopeteas, S. (2004). Towards user-adaptive annotation guidelines. In *Proceedings of the COLING 2004 5th International Workshop on Linguistically Interpreted Corpora* (pp. 23–30). Geneva, Switzerland.

Doan, S., Kawazoe, A., Conway, M., & Collier, N. (2009). Towards role-based filtering of disease outbreak reports. *Journal of Biomedical Informatics*, *42*(5), 773–780. doi:10.1016/j.jbi.2008.12.009

DuBuske, L. M. (2005). The role of P-glycoprotein and organic anion-transporting polypeptides in drug interactions. *Drug Safety : An International Journal of Medical Toxicology and Drug Experience*, *28*(9), 789–801.

Duda, S., Aliferis, C., Miller, R., Statnikov, A., & Johnson, K. (2005). Extracting drug-drug interaction articles from MEDLINE to improve the content of drug databases. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, 216–20.

Duke, J. D., Han, X., Wang, Z., Subhadarshini, A., Karnik, S. D., Li, X., … Li, L. (2012). Literature based drug interaction prediction with clinical assessment using electronic medical records: novel myopathy associated drug interactions. *PLoS Computational Biology*, *8*(8), e1002614. doi:10.1371/journal.pcbi.1002614

Dumontier, M., Baker, C. J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., … Hoehndorf, R. (2014). The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *Journal of Biomedical Semantics*, *5*(1), 14. doi:10.1186/2041-1480-5-14

Edwards, D. J. (2012). Beneficial Pharmacokinetic Drug Interactions. *Advances in Pharmacoepidemiology & Drug Safety*, *01*(S1), 1–5. doi:10.4172/2167-1052.S1-002

Eltyeb, S., & Salim, N. (2014). Chemical named entities recognition: a review on approaches and applications. *Journal of Cheminformatics*, *6*(1), 17. doi:10.1186/1758-2946-6-17

Fahmi, O. A., Hurst, S., Plowchalk, D., Cook, J., Guo, F., Youdim, K., … Obach, R. S. (2009). Comparison of Different Algorithms for Predicting Clinical Drug- Drug Interactions , Based on the Use of CYP3A4 in Vitro Data : Predictions of Compounds as Precipitants of Interaction □ ABSTRACT : *Drug Metabolism and Disposition*, *37*(8), 1658–1666. doi:10.1124/dmd.108.026252.

Fernández-López, M., Gómez-Pérez, A., & Juristo, N. (1997). Methontology: from ontological art towards ontological engineering. In *AAAI Symposium on Ontological Engineering* (pp. 33–40). Stanford.

Fort, K., & Sagot, B. (2010). Influence of Pre-annotation on POS-tagged Corpus Development. In *Proceedings of the Fourth Linguistic Annotation Workshop* (pp. 56–63).

Friedman, C., Kra, P., & Rzhetsky, A. (2002). Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, *35*(4), 222–235. doi:10.1016/S1532-0464(03)00012-1

Fromm, M. F., & Kim, R. B. (Eds.). (2011). Drug Transporters. In *Handbook of Experimental Pharmacology, Volume 201* (p. 454). London: Springer. doi:10.1007/978-3-642-14541-4

Fulda, T. R., Valuck, R. J., Zanden, J. Vander, & Parker, S. (2000). Disagreement among drug compendia on inclusion and ratings of Drug-Drug Interactions. *Current Therapeutic Research*, *61*(8), 540–548.

Ganeva, M., Gancheva, T., Troeva, J., Kiriyak, N., & Hristakieva, E. (2013). Clinical relevance of drug-drug interactions in hospitalized dermatology patients. *Advances in Clinical and Experimental Medicine : Official Organ Wroclaw Medical University*, *22*(4), 555–63.

Garten, Y., Coulet, A., & Altman, R. B. (2010). Recent progress in automatically extracting information from the pharmacogenomic literature. *Pharmacogenomics*, *11*(10), 1467–1489. doi:10.2217/pgs.10.136.Recent

Gaulton, A., Bellis, L. J., Bento, a P., Chambers, J., Davies, M., Hersey, A., … Overington, J. P. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, *40*(Database issue), D1100–7. doi:10.1093/nar/gkr777

Gebser, M., Ostrowski, M., & Schaub, T. (2009). Constraint Answer Set Solving. In *Proceedings of the Twenty-fifth ICLP'09* (pp. 235–249). Pasadena, California, USA.

Gelfond, M., & Lifschitz, V. (1991). Michael Gelfond Vladimir Lifschitz. *New Generation Computing*, *9*, 365–387.

Gerner, M., Nenadic, G., & Bergman, C. M. (2010). LINNAEUS: a species name identification system for biomedical literature. *BMC Bioinformatics*, *11*, 85. doi:10.1186/1471-2105-11-85

Glimm, B., Horrocks, I., Giorgos, M., & Zhe, S. (2014). HermiT : An OWL 2 Reasoner. *Journal of Automated Reasoning*.

Golbreich, C. (2005). Combining Rule and Ontology Reasoners for the Semantic Web. In A. Adi, S. Stoutenburg, & S. Tabet (Eds.), *First Interanational Conference in Rules and Rule Markup Languages for the Semantic Web.* (pp. 6–22). Galway, Ireland: Springer Berlin Heidelberg.

Gómez-Pérez, A. (1999). Evaluation of taxonomic knowledge in ontologies and knowledge bases. In *Banff Knowledge Acquisition for Knowledge-Based Systems Workshop. KAW'99*. Banff, Alberta, Canada.

Gómez-Pérez, A., Fernández-López, M., & Corcho, O. (2004). *Ontological Engineering*. Springer.

Gómez-Pérez, A., & Manzano-Macho, D. (2003). A survey of ontology learning methods and techniques OntoWeb Consortium. *Deliverable 1.5, OntoWeb Project*.

Gómez-Pérez, A., Martínez-Romero, M., Rodríguez-González, A., Vázquez, G., & Vázquez-Naya, J. M. (2013). Ontologies in medicinal chemistry: Current status and future challenges. *Current Topics in Medicinal Chemistry*, *13*, 576–590.

Gottlieb, A., Stein, G. Y., Oron, Y., Ruppin, E., & Sharan, R. (2012). INDI: a computational framework for inferring drug interactions and their associated recommendations. *Molecular Systems Biology*, *8*(592), 1–12. doi:10.1038/msb.2012.26

Grego, T., & Couto, F. M. (2013). LASIGE : using Conditional Random Fields and ChEBI ontology. In *7th International Workshop on Semantic Evaluation (SemEval 2013)* (Vol. 2, pp. 660–666). Atlanta, USA.

Groppe, S. (2011). *Data Management and Query Processing in Semantic Web Databases* (p. 279). Springer Science & Business Media.

Grüninger, M., & Fox, M. S. (1995). Methodology for the Design and Evaluation of Ontologies. In *Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95*.

Guimerà, R., & Sales-Pardo, M. (2013). A network inference method for large-scale unsupervised identification of novel drug-drug interactions. *PLoS Computational Biology*, *9*(12), e1003374. doi:10.1371/journal.pcbi.1003374

Gurulingappa, H., Rajput, A. M., Roberts, A., Fluck, J., Hofmann-Apitius, M., & Toldo, L. (2012). Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, *45*(5), 885–92. doi:10.1016/j.jbi.2012.04.008

Hage, D. S., & Tweed, S. a. (1997). Recent advances in chromatographic and electrophoretic methods for the study of drug-protein interactions. *Journal of Chromatography. B, Biomedical Sciences and Applications*, *699*(1-2), 499–525.

Hansten, P. D. (2003). Drug interaction management. *Pharmacy World & Science : PWS*, *25*(3), 94–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12840961

Hansten, P. D., & Horn, J. R. (2014). *The Top 100 Drug Interactions: A Guide to Patient Management* (15th ed., p. 178). H & H Publications LLP.

Hansten, P. D., Horn, J. R., & Hazlet, T. K. (2001). ORCA: OpeRational ClassificAtion of drug interactions. *Journal of the American Pharmaceutical Association (Washington,D.C. : 1996)*, *41*(2), 161–5.

Harpaz, R., Chase, H. S., & Friedman, C. (2010). Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC Bioinformatics*, *11*(Suppl 9), S7. doi:10.1186/1471-2105-11-S9-S7

Harpaz, R., Haerian, K., Chase, H. S., & Friedman, C. (2010). Statistical Mining of Potential Drug Interaction Adverse Effects in FDA's Spontaneous Reporting System. In *AMIA Annual Symposium Proceedings* (Vol. 2010, pp. 281–5).

Harris, Z. (1982). *A grammar of english on mathematical principles*. New York: Wiley.

Harris, Z. (1991). *A theory of language and information: a mathematical approach*. Oxford: Clarendon Press.

Hassanpour, S., O'Connor, M. J., & Das, A. K. (2011). Evaluation of semantic-based information retrieval methods in the autism phenotype domain. *AMIA Annual Symposium Proceedings*, *2011*, 569–77.

Hassanzadeh, O., Zhu, Q., Freimuth, R., & Boyce, R. (2013). Extending the " Web of Drug Identity " with Knowledge Extracted from United States Product Labels. In *Proceedings of the 2013 AMIA Summit on Translational Bioinformatics* (pp. 64–68). San Francisco.

Hastings, J., Brass, A., Caine, C., Jay, C., & Stevens, R. (2014). Evaluating the Emotion Ontology through use in the self-reporting of emotional responses at an academic conference. *Journal of Biomedical Semantics*, *5*(38), 1–17.

Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., … Steinbeck, C. (2013). The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research*, *41*(Database issue), D456–63. doi:10.1093/nar/gks1146

He, Y., & Kayaalp, M. (2006). A Comparison of 13 Tokenizers on MEDLINE. TECHNICAL REPORT LHNCBC-TR-2006-003. *The Lister Hill National Center for Biomedical Communications*, (December).

He, Y., Road, N. P., & Ex, E. (2008). Ontology-Based Protein-Protein Interactions Extraction from Literature using the Hidden Vector State Model. In *Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on.* (pp. 736–743).

He, Y., Xiang, Z., Sarntivijai, S., Toldo, L., & Ceusters, W. (2011). AEO: A Realism-Based Biomedical Ontology for the Representation of Adverse Events. In *International Conference on Biomedical Ontology. Representing Adverse Events Workshop*. Buffalo, NY, USA.

Hepple, M. (2000). Independence and Commitment: Assumptions for Rapid Training and Execution of Rule-based POS Taggers. In *38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)* (Vol. 7, pp. 278–285). Hong Kong.

Herrero-Zazo, M., Hastings, J., Segura-Bedmar, I., Croset, S., Martinez, P., & Steinbeck, C. (2013). An ontology for drug-drug interactions. In *6th International Workshop on Semantic Web Applications and Tools for Life Sciences (SWAT4LS)*. Edinburgh, UK.

Herrero-Zazo, M., Segura-Bedmar, I., & Martínez, P. (2013). Annotation Issues in Pharmacological Texts. In *5th International Conference on Corpus Linguistics (CILC2013). Procedia - Social and Behavioral Sciences.* (Vol. 95, pp. 211–219). Alicante, Spain: Elsevier Ltd. doi:10.1016/j.sbspro.2013.10.641

Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P., & Declerck, T. (2013). The DDI corpus : An annotated corpus with pharmacological substances and drug – drug interactions. *Journal of Biomedical Informatics*, *46*(5), 914–920. doi:10.1016/j.jbi.2013.07.011

Hettne, K. M., Stierum, R. H., Schuemie, M. J., Hendriksen, P. J. M., Schijvenaars, B. J. a, Mulligen, E. M. Van, … Kors, J. a. (2009). A dictionary to identify small molecules and drugs in free text. *Bioinformatics (Oxford, England)*, *25*(22), 2983–91. doi:10.1093/bioinformatics/btp535

Hewett, M., Oliver, D. E., Rubin, D. L., Easton, K. L., Stuart, J. M., Altman, R. B., & Klein, T. E. (2002). PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Research*, *30*(1), 163–5.

Hitzler, P., Krötzsch, M., & Rudolph, S. (2009). *Foundations of Semantic Web Technologies* (p. 455). Chapman & Hall/CRC.

Hochman, M., Hochman, S., Bor, D., & McCormick, D. (2008). News media coverage of medication research: reporting pharmaceutical company funding and use of generic medication names. *JAMA : The Journal of the American Medical Association*, *300*(13), 1544–50. doi:10.1001/jama.300.13.1544

Hoehndorf, R., Dumontier, M., & Gkoutos, G. V. (2012). Evaluation of research in biomedical ontologies. *Briefings in Bioinformatics*. doi:10.1093/bib/bbs053

Hogan, W. R., Hanna, J., Joseph, E., & Brochhausen, M. (2011). Towards a Consistent and Scientifically Accurate Drug Ontology. In *International Conference on Biomedical Ontology*. (pp. 1–6).

Hojo, Y., Echizenya, M., Ohkubo, T., & Shimizu, T. (2011). Drug interaction between St John's wort and zolpidem in healthy subjects. *Journal of Clinical Pharmacy and Therapeutics*, *36*(6), 711–5. doi:10.1111/j.1365-2710.2010.01223.x

Holford, M. E., Khurana, E., Cheung, K.-H., & Gerstein, M. (2010). Using semantic web rules to reason on an ontology of pseudogenes. *Bioinformatics (Oxford, England)*, *26*(12), i71–8. doi:10.1093/bioinformatics/btq173

Horrocks, I., Patelschneider, P., Bechhofer, S., & Tsarkov, D. (2005). OWL rules: A proposal and prototype implementation. *Web Semantics: Science, Services and Agents on the World Wide Web*, *3*(1), 23–40. doi:10.1016/j.websem.2005.05.003

Huang, J., Niu, C., Green, C. D., Yang, L., Mei, H., & Han, J.-D. J. (2013). Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network. *PLoS Computational Biology*, *9*(3), e1002998. doi:10.1371/journal.pcbi.1002998

Huang, M., Zhu, X., Ding, S., Yu, H., & Li, M. (2006). ONBRIRES: Ontology-Based Biological Relation Extraction System. In *APBC* (pp. 327–336).

Huang, Strong, J., Zhang, L., Reynolds, K., Nallani, S., Temple, R., … Lesko, L. (2008). New Era in Drug Interaction Evaluation: US Food and Drug Administration Update on CYP Enzymes, Transporters, and the Guidance Process. *Journal of Clinical Pharmacology*, *48*, 662–670.

Huang, Z., Bao, Y., Dong, W., Lu, X., & Duan, H. (2014). Online treatment compliance checking for clinical pathways. *Journal of Medical Systems*, *38*(10). doi:10.1007/s10916-014-0123-0

Huber, T., Linden, U. Den, & Rockt, T. (2013). WBI-NER : The impact of domain-specific features on the performance of identifying and classifying mentions of drugs. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (Vol. 2, pp. 356–363). Atlanta, USA.

Hunter, L., & Cohen, K. (2006). Biomedical Language Processing: Perspective What's Beyond PubMed? *Molecular Cell*, *21*(5), 589–594.

Ibáñez, A., Alcalá, M., García, J., & Puche, E. (2008). Interacciones medicamentosas en pacientes de un servicio de medicina interna. *Farmacia Hospitalaria*, *32*(5), 293–297. doi:10.1016/S1130-6343(08)75950-6

Imai, T., Hayakawa, M., & Ohe, K. (2013). Development of Description Framework of Pharmacodynamics Ontology and Its Application to Possible Drug-drug Interaction Reasoning. *Studies in Health Technology and Informatics*, *192*, 567–571. doi:10.3233/978-1-61499-289-9-567

Ito, K., Brown, H. S., & Houston, J. B. (2004). Database analyses for the prediction of in vivo drug-drug interactions from in vitro data. *British Journal of Clinical Pharmacology*, *57*(4), 473–86. doi:10.1111/j.1365-2125.2003.02041.x

Iyer, S. V, Harpaz, R., Lependu, P., Bauer-Mehren, A., & Shah, N. H. (2013). Mining clinical text for signals of adverse drug-drug interactions. *Journal of the American Medical Informatics Association*, *21*, 353–362. doi:10.1136/amiajnl-2013-001612

Jagannathan, V., Mullett, C. J., Arbogast, J. G., Halbritter, K. A., Yellapragada, D., Regulapati, S., & Bandaru, P. (2009). Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. *International Journal of Medical Informatics*, *78*(4), 284–91. doi:10.1016/j.ijmedinf.2008.08.006

Jankel, C., McMillan, J., & Martin, B. (1994). Effect of drug interactions on outcomes of patients receiving warfarin or theophylline. *American Society of Health-System Pharmacists*, *51*(5), 661–666.

Jessop, D. M., Adams, S. E., Willighagen, E. L., Hawizy, L., & Murray-Rust, P. (2011). OSCAR4: a flexible architecture for chemical text-mining. *Journal of Cheminformatics*, *3*(1), 41. doi:10.1186/1758-2946-3-41

Jonquet, C., Shah, N. H., Cherie, H., Musen, M. A., Callendar, C., Storey, M., & Vw, B. C. C. (2009). NCBO Annotator : Semantic Annotation of Biomedical Data. In *International Semantic Web Conference, Poster and Demo session.*

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, *40*(Database issue), D109–14. doi:10.1093/nar/gkr988

Kang, N., Singh, B., Bui, C., Afzal, Z., van Mulligen, E. M., & Kors, J. a. (2014). Knowledge-based extraction of adverse drug events from biomedical text. *BMC Bioinformatics*, *15*(1), 64. doi:10.1186/1471-2105-15-64

Karnik, S., Subhadarshini, A., Wang, Z., Rocha, L. M., & Li, L. (2011). Extraction of drug-drug interactions using all paths graph kernel. In *Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction* (pp. 83–88). Huelva, Spain.

Kenakin, T. P. (2012). *Pharmacology in Drug Discovery: Understanding Drug Response* (p. 247). Academic Press.

Khelashvili, G., Dorff, K., Shan, J., Camacho-Artacho, M., Skrabanek, L., Vroling, B., … Filizola, M. (2010). GPCR-OKB: the G Protein Coupled Receptor Oligomer Knowledge Base. *Bioinformatics (Oxford, England)*, *26*(14), 1804–5. doi:10.1093/bioinformatics/btq264

Kim, J., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). GENIA corpus--a semantically annotated corpus for bio-textmining. *Bioinformatics*, *19*(Suppl 1), i180–i182. doi:10.1093/bioinformatics/btg1023

Kim, J.-D., Ohta, T., & Tsujii, J. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, *9*(10), 10. doi:10.1186/1471-2105-9-10

Klarin, K. J. I. (2007). The Relationship between Number of Drugs and Potential Drug-Drug Interactions in the Elderly: A Study of Over 600 000 Elderly Patients from the Swedish Prescribed Drug Register. *Drug Safety*, *30*(10), 911–918.

Klein, A., Riazanov, A., Hindle, M. M., & Baker, C. J. O. (2014). Benchmarking infrastructure for mutation text mining. *Journal of Biomedical Semantics*, *5(1)*(11), 1–13.

Klein, D., & Manning, C. D. (2003). Accurate Unlexicalized Parsing. In *Proceedings of the 41st annual meeting on association for computational linguistics. Volumen 1* (pp. 423–430). Association for Computational Linguistics.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., … Wishart, D. S. (2011). DrugBank 3.0: a comprehensive resource for “omics” research on drugs. *Nucleic Acids Research*, *39*(Database issue), D1035–41. doi:10.1093/nar/gkq1126

Ko, Y., Abarca, J., Malone, D. C., Dare, D. C., Geraets, D., Houranieh, A., … Wilhardt, M. (2007). Practitioners ’ Views on Computerized Drug – Drug Interaction Alerts in

the VA System. *Journal of the American Medical Informatics Association*, *14*, 56–64. doi:10.1197/jamia.M2224.Introduction

Kolárik, C., Hofmann-Apitius, M., Zimmermann, M., & Fluck, J. (2007). Identification of new drug classification terms in textual resources. *Bioinformatics (Oxford, England)*, *23*(13), i264–72. doi:10.1093/bioinformatics/btm196

Kolárik, C., Klinger, R., Friedrich, C. M., Hofmann-Apitius, M., & Fluck, J. (2008). Chemical Names: Terminological Resources and Corpora Annotation. In *Proceedings of the Workshop on Building and Evaluating Resources for Biomedical Text Mining* (pp. 51–58). Marrakech, Morocco.

Konagaya, A. (2012). Towards an in silico approach to personalized pharmacokinetics. In Prof. Aurelia Meghea (Ed.), *Molecular Interactions* (pp. 263–282).

Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., & Valencia, A. (2013). Overview of the chemical compound and drug name recognition (CHEMDNER) task. In *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop* (pp. 6–37). Fundación CNIO Carlos III.

Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J., & Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*, *6*(343), 343. doi:10.1038/msb.2009.98

Kulkarni, V., Bora, S. S., Sirisha, S., Saji, M., & Sundaran, S. (2013). A study on drug-drug interactions through prescription analysis in a South Indian teaching hospital. *Therapeutic Advances in Drug Safety*, *4*(4), 141–6. doi:10.1177/2042098613490009

Lafferty, J., Mccallum, A., & Pereira, F. C. N. (2001). Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. In *The International Conference on Machine Learning (ICML - 2001)* (Vol. 2001, pp. 282–289).

Lamurias, A., Grego, T., & Couto, F. M. (2013). Chemical compound and drug name recognition using CRFs and semantic similarity based on ChEBI. In *BioCreative Challenge Evaluation Workshop* (p. 75).

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., … Wishart, D. S. (2014). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research*, *42*(Database issue), D1091–7. doi:10.1093/nar/gkt1068

Lea, M., Rognan, S. E., Koristovic, R., Wyller, T. B., & Molden, E. (2013). Severity and Management of Drug–Drug Interactions in Acute Geriatric Patients. *Drugs & Aging*, *30*(9), 721–727.

Leech, G. (1993). Corpus Annotation Schemes. *Literary and Linguistic Computing*, *8*(4), 275–281.

Lependu, P., Iyer, S. V, Bauer-Mehren, a, Harpaz, R., Mortensen, J. M., Podchiyska, T., … Shah, N. H. (2013). Pharmacovigilance Using Clinical Notes. *Clinical Pharmacology and Therapeutics*, (October 2012), 1–9. doi:10.1038/clpt.2013.47

Levine, R. R., Walsh, C. T., & Schwartz-Bloom, R. D. (2005). *Levine's Pharmacology: Drug Actions and Reactions* (Seventh Ed.). CRC Press.

Levy, G., & Reuning, R. H. (1964). Effect of Complex Formation on Drug Absorption I. Complexes of Salicylic Acid with Absorbable and Nonabsorbable Compounds. *Journal of Pharmaceutical Sciences*, *53*, 1471–75.

Lezcano, L., Sicilia, M.-A., & Rodríguez-Solano, C. (2011). Integrating reasoning and clinical archetypes using OWL ontologies and SWRL rules. *Journal of Biomedical Informatics*, *44*(2), 343–53. doi:10.1016/j.jbi.2010.11.005

Lipscomb, C. (2000). Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, *88*(July), 265–266.

Liu, X., Xu, Y., Li, S., Wang, Y., Peng, J., Luo, C., … Jiang, H. (2014). In Silico target fishing: addressing a "Big Data" problem by ligand-based similarity rankings with data fusion. *Journal of Cheminformatics*, *6*(1), 33. doi:10.1186/1758-2946-6-33

Lönneker, B. (2003). Ontology Aspects in Relation Extraction. In *Workshop on Ontologies and Information Extraction. EUROLAN-2003 summer school on the Semantic Web and Language Technology-Its Potential and Practicalities.* Bucharest, Romania.

Lozano-Tello, A., & Gómez-Pérez, A. (2004). ONTOMETRIC : A Method to Choose the Appropriate Ontology. *Journal of Database Management*, *15*(2), 1–18.

Lu, Z., Bada, M., Ogren, P. V, Cohen, K. B., & Hunter, L. (2006). Improving Biomedical Corpus Annotation Guidelines. In *Proceedings of the Joint BioLINK and 9th Bio-Ontologies Meeting* (pp. 89–92). Fortaleza, Brazil.

Mack, R., & Hehenbergerb, M. (2002). Text-based knowledge discovery: search and mining of life-sciences documents. *Drug Discovery Today*, *7*(11), S89–S98.

Martínez-Romero, M., Vázquez-Naya, J. M., Pereira, J., Pereira, M., Pazos, A., & Baños, G. (2013). The iOSC3 system: using ontologies and SWRL rules for intelligent supervision and care of patients with acute cardiac disorders. *Computational and Mathematical Methods in Medicine*, *2013*, 650671. doi:10.1155/2013/650671

Martiny, V. Y., & Miteva, M. A. (2013). Advances in molecular modeling of human cytochrome P450 polymorphism. *Journal of Molecular Biology*, *425*(21), 3978–92.

McClosky, D. (2010). *Any domain parsing: automatic domain adaptation for natural language parsing*. Brown University.

McCray, A. T., Aronson, A. R., Browne, A. C., & Rindflesch, T. C. (1993). UMLS knowledge for biomedical language processing. *Bulletin of the Medical Library Association*, *81*(2), 184–94.

Mccray, A. T., Srinivasan, S., & Browne, A. C. (1994). Lexical Methods for Managing Variation in Biomedical Terminologies. In *Annual Symposium on Computer Application in Medical Care* (pp. 235–239).

McNaught, A. D., & Wilkinson, A. (1997). *IUPAC Compendium of Chemical Terminology* (2nd ed.). Oxford, UK: Blackwell Scientific Publications.

Mendonça, F., Coelho, K., Andrade, A., & Almeida, M. (2012). Knowledge Acquisition in the construction of ontologies: a case study in the domain of hematology. In *3rd International Conference on Biomedical Ontology* (pp. 2–6).

Meystre, S., & Haug, P. J. (2006). Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *Journal of Biomedical Informatics*, *39*(6), 589–99. doi:10.1016/j.jbi.2005.11.004

Mille, F., Degoulet, P., & Jaulent, M. (2007). Modeling and Acquisition of Drug-Drug Interaction Knowledge. In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics* (pp. 900–904). IOS Press.

Moitra, A., Palla, R., Tari, L., & Krishnamoorthy, M. (2014). Semantic Inference for Pharmacokinetic Drug-Drug Interactions. *2014 IEEE International Conference on Semantic Computing*, *2*, 92–95. doi:10.1109/ICSC.2014.36

Morton, T., & LaCivita, J. (2003). Wordfreak: an open tool for linguistic annotation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology Demonstrations - NAACL '03* (Vol. 4, pp. 17–18). Morristown, NJ, USA: Association for Computational Linguistics. doi:10.3115/1073427.1073436

Moura, C., Acurcio, F., & Belo, N. (2009). Drug-Drug Interactions Associated with Length of Stay and Cost of Hospitalization. *Journal of Pharmacy & Pharmaceutical Sciences*, *12*(3), 266–272.

Müller, C., & Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, *3*, 197–214.

Müller, H. M., Kenny, E. E., & Sternberg, P. W. (2004). Textpresso : An Ontology-Based Information Retrieval and Extraction System for Biological Literature. *PLoS Biology*, *2*(11), 1984–1998. doi:10.1371/journal.pbio.0020309

Naqvi, S. (2000). The challenge of posttransplant tuberculosis. *Transplantation Proceedings*, *32*(3), 650–1. doi:http://dx.doi.org/10.1016/S0041-1345(00)00931-3

Nelson, S. J., Zeng, K., Kilbourne, J., Powell, T., & Moore, R. (2011). Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association : JAMIA*, *18*(4), 441–8. doi:10.1136/amiajnl-2011-000116

Neves, M. (2014). An analysis on the entity annotations in biological corpora. *F1000Research*, *3*, 96. doi:10.12688/f1000research.3216.1

Neves, M., & Leser, U. (2014). A survey on annotation tools for the biomedical literature. *Briefings in Bioinformatics*, *15*(2), 327–40. doi:10.1093/bib/bbs084

Niemelä, I. (1999). Logic programs with stable model semantics as a constraint programming paradigm. *Annals of Mathematics and Artificial Intelligence*, *25*(3-4), 241–273. doi:10.1023/A:1018930122475

Nikfarjam, A., & Gonzalez, G. H. (2011). Pattern mining for extraction of mentions of Adverse Drug Reactions from user comments. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, *2011*, 1019–26.

Nikolić, B., & Ilić, M. (2013). Assessment of the consistency among three drug compendia in listing and ranking of drug-drug interactions. *Bosnian Journal of Basic Medical Sciences*, *13*(4), 253–8.

NLM. (2014). *National Library of Medicine: Fact Sheet Medline*. *U.S. National Library of Medicine*. Retrieved December 09, 2014, from http://www.nlm.nih.gov/pubs/factsheets/medline.html

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., … Musen, M. a. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, *37*(Web Server issue), W170–3. doi:10.1093/nar/gkp440

Obreli Neto, P. R., Nobili, A., de Lyra, D. P., Pilger, D., Guidoni, C. M., de Oliveira Baldoni, A., … Nakamura Cuman, R. K. (2012). Incidence and predictors of adverse drug reactions caused by drug-drug interactions in elderly outpatients: a prospective cohort study. *Journal of Pharmacy & Pharmaceutical Sciences : A Publication of the Canadian Society for Pharmaceutical Sciences, Société Canadienne Des Sciences Pharmaceutiques*, *15*(2), 332–43. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22579011

Olivié, A. (2007). *Conceptual modeling of information systems* (p. 455). Springer Berlin Heidelberg New York.

Olvey, E. L., Clauschee, S., & Malone, D. C. (2010a). Comparison of critical drug-drug interaction listings: the Department of Veterans Affairs medical system and standard reference compendia. *Clinical Pharmacology and Therapeutics*, *87*(1), 48–51. doi:10.1038/clpt.2009.198

Olvey, E. L., Clauschee, S., & Malone, D. C. (2010b). Comparison of critical drug-drug interaction listings: the Department of Veterans Affairs medical system and standard reference compendia. *Clinical Pharmacology and Therapeutics*, *87*(1), 48–51. doi:10.1038/clpt.2009.198

Otero-López, M. J., Alonso-Hernández, P., Maderuelo-Fernández, J. Á., Garrido-Corro, B., Domínguez-Gil, A., & Sánchez-Rodríguez, Á. (2006). Acontecimientos adversos prevenibles causados por medicamentos en pacientes hospitalizados. *Medicina Clínica*, *126*(3), 81–87. doi:10.1157/13083875

Paczynski, R. P., Alexander, G. C., Chinchilli, V. M., & Kruszewski, S. P. (2012). Quality of evidence in drug compendia supporting off-label use of typical and atypical antipsychotic medications. *The International Journal of Risk and Safety in Medicine*, *24*(3), 137–146.

Paolillo, S., Pellegrino, R., Salvioni, E., Contini, M., Iorio, A., Bovis, F., … Agostoni, P. (2013). Role of Alveolar β2-Adrenergic Receptors on Lung Fluid Clearance and Exercise Ventilation in Healthy Humans. *PloS One*, *8*(4), e61877. doi:10.1371/journal.pone.0061877

Patel, P. S., Rana, D. a, Suthar, J. V, Malhotra, S. D., & Patel, V. J. (2014). A study of potential adverse drug-drug interactions among prescribed drugs in medicine outpatient department of a tertiary care teaching hospital. *Journal of Basic and Clinical Pharmacy*, *5*(2), 44–8. doi:10.4103/0976-0105.134983

Pathak, J., Kiefer, R. C., & Chute, C. G. (2013). Using Linked Data for Mining Drug-Drug Interactions in Electronic Health Records. *Studies in Health Technology and Informatics*, *192*, 682–686.

Payne, P. R. O., Mendonça, E. a, Johnson, S. B., & Starren, J. B. (2007). Conceptual knowledge acquisition in biomedicine: A methodological review. *Journal of Biomedical Informatics*, *40*(5), 582–602. doi:10.1016/j.jbi.2007.03.005

Percha, B., & Altman, R. B. (2012). Discovery and Explanation of Drug-Drug Interactios Via Text Mining. *Pacific Symposium on Biocomputing*, 410–421.

Percha, B., & Altman, R. B. (2013). Informatics confronts drug-drug interactions. *Trends in Pharmacological Science*, *34*(3), 178–184. doi:10.1016/j.tips.2013.01.006.Informatics

Pérez-Nueno, V. I., Karaboga, A. S., Souchet, M., & Ritchie, D. W. (2014). GES polypharmacology fingerprints: a novel approach for drug repositioning. *Journal of Chemical Information and Modeling*, *54*(3), 720–34. doi:10.1021/ci4006723

Peters, L., Bodenreider, O., & Bahr, N. (2014). Evaluating drug-drug interaction information in NDF-RT and DrugBank. In *Proceedings of the International Conference on Biomedical Ontology (ICBO) 2014:(in press).*

Pirmohamed, M., James, S., Meakin, S., Green, C., Scott, A. K., Walley, T. J., … Breckenridge, A. M. (2004). Adverse drug reactions as cause of admission to hospital: prospective analysis of 18820 patients. *British Medical Journal*, *329*(July), 15–19.

Plessers, P., & De Troyer, O. (2006). Resolving Inconsistencies in Evolving Ontologies. In Y. Sure & J. Domingue (Eds.), *The Semantic Web: Research and Applications.*

3rd European Semantic Web Conference, ESWC 2006 (pp. 200–2014). Budva, Montenegro: Springer Science & Business Media.

Poesio, M., Barbu, E., Giuliano, C., & Romano, L. (2008). Supervised relation extraction for ontology learning from text based on a cognitively plausible model of relations. In *ECAI 2008 3rd Workshop on Ontology Learning and Population* (pp. 1–5).

Pustejovsky, J., & Stubbs, A. (2012). *Natural Language Annotation for Machine Learning* (p. 342). O'Reilly Media, Inc.

Quinney, S. K., Zhang, X., Lucksiri, A., Gorski, J. C., Li, L., & Hall, S. D. (2010). Physiologically Based Pharmacokinetic Model of Mechanism-Based Inhibition of CYP3A by Clarithromycin. *Drug Metabolism and Disposition*, *38*(2), 241–248. doi:10.1124/dmd.109.028746.)

Rajpurohit, N., Aryal, S., Khan, M., Stys, A., & Stys, T. (2014). Propafenone associated severe central nervous system and cardiovascular toxicity due to mirtazapine: a case of severe drug interaction. *South Dakota Medicine: The Journal of the South Dakota State Medical Association*, *67*(4), 137–139.

Rastegar-Mojarad, M., Boyce, R. D., & Prasad, R. (2013). UWM-TRIADS : Classifying Drug-Drug Interactions with Two-Stage SVM and Post-Processing. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (Vol. 2, pp. 667–674). Atlanta, USA.

Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., … Wheeldin, B. (2007). The CLEF corpus: semantic annotation of clinical text. In *Proceedings of the 2007 American Medical Informatics Association Annual Symposium* (pp. 625–9). Chicago, IL, USA.

Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., & Setzer, A. (2009). Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics*, *42*(5), 950–66. doi:10.1016/j.jbi.2008.12.013

Rocktäschel, T., Weidlich, M., & Leser, U. (2012). ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics (Oxford, England)*, *28*(12), 1633–40. doi:10.1093/bioinformatics/bts183

Rodríguez-Terol, a., Caraballo, M. O., Palma, D., Santos-Ramos, B., Molina, T., Desongles, T., & Aguilar, a. (2009). Quality of interaction database management systems. *Farmacia Hospitalaria (English Edition)*, *33*(3), 134–146. doi:10.1016/S2173-5085(09)70079-6

Rosario, B., & Hearst, M. (2004). Classifying semantic relations in bioscience texts. In *Proceedings of the 42nd Annual Meeting on the Association for Computational Linguistics (ACL 2004)* (pp. 430–437).

Rosse, C., & Mejino, J. L. V. (2003). A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *Journal of Biomedical Informatics*, *36*(6), 478–500. doi:10.1016/j.jbi.2003.11.007

Rubrichi, S., & Quaglini, S. (2012). Summary of Product Characteristics content extraction for a safe drugs usage. *Journal of Biomedical Informatics*, *45*(2), 231–9. doi:10.1016/j.jbi.2011.10.012

Rubrichi, S., Quaglini, S., Spengler, A., Russo, P., & Gallinari, P. (2013). A system for the extraction and representation of summary of product characteristics content. *Artificial Intelligence in Medicine*, *57*(2), 145–54. doi:10.1016/j.artmed.2012.08.004

Samwald, M., Freimuth, R., Luciano, J. S., Lin, S., Powers, R. L., Marshall, M. S., … Boyce, R. D. (2013). An RDF / OWL Knowledge Base for Query Answering and Decision Support in Clinical Pharmacogenetics. *Studies in Health Technology and Informatics*, *192*, 539–542.

Samwald, M., Jentzsch, A., Bouton, C., Kallesøe, C. S., Willighagen, E., Hajagos, J., … Stephens, S. (2011). Linked open drug data for pharmaceutical research and development. *Journal of Cheminformatics*, *3*(1), 19. doi:10.1186/1758-2946-3-19

Sanchez-Cisneros, D., & Aparicio, F. (2013). UEM-UC3M: An Ontology-based named entity recognition system for biomedical texts. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (Vol. 2, pp. 622–627). Atlanta, USA.

Schuemie, M. J. (2007). Peregrine: lightweight gene name normalization by dictionary lookup. In *Proceedings of the Biocreative 2 workshop* (pp. 131–133). Madrid.

Segura-Bedmar, I. (2010). *Application of information extraction techniques to pharmacological domain*. University Carlos III of Madrid.

Segura-Bedmar, I., Crespo, M., de Pablo-Sánchez, C., & Martínez, P. (2010). Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. *BMC Bioinformatics*, *11 Suppl 2*(Suppl 2), S1. doi:10.1186/1471-2105-11-S2-S1

Segura-Bedmar, I., Martínez, P., & de Pablo-Sánchez, C. (2010). Extracting drug-drug interactions from biomedical texts. *BMC Bioinformatics*, *11*(Suppl 5), P9. doi:10.1186/1471-2105-11-S5-P9

Segura-Bedmar, I., Martínez, P., & de Pablo-Sánchez, C. (2011a). A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. *BMC Bioinformatics*, *12 Suppl 2*(Suppl 2), S1. doi:10.1186/1471-2105-12-S2-S1

Segura-Bedmar, I., Martínez, P., & de Pablo-Sánchez, C. (2011b). Using a shallow linguistic kernel for drug-drug interaction extraction. *Journal of Biomedical Informatics*, *44*(5), 789–804. doi:10.1016/j.jbi.2011.04.005

Segura-Bedmar, I., Martínez, P., & Herrero-Zazo, M. (2013). SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (pp. 341–350). Atlanta, USA.

Segura-Bedmar, I., Martínez, P., & Herrero-Zazo, M. (2014). Lessons learnt from the DDIExtraction-2013 Shared Task. *Journal of Biomedical Informatics*, *51*, 152–164. doi:10.1016/j.jbi.2014.05.007

Segura-Bedmar, I., Martínez, P., & Sánchez-Cisneros, D. (2011). The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts. In *Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction* (pp. 1–9). Huelva, Spain.

Segura-Bedmar, I., Martínez, P., & Segura-Bedmar, M. (2008). Drug name recognition and classification in biomedical texts. A case study outlining approaches underpinning automated systems. *Drug Discovery Today*, *13*(17-18), 816–23. doi:10.1016/j.drudis.2008.06.001

Shaban-Nejad, A., & Haarslev, V. (2009). Bio-Medical Ontologies Maintenance and Change Management. In A. Shidu & T. Dillon (Eds.), *Biomedical Data and Applications* (pp. 143–168). Springer Berlin Heidelberg.

Shojanoori, R., & Juric, R. (2013). Semantic remote patient monitoring system. *Telemedicine and E-Health*, *19*(2), 129–136. doi:10.1089/tmj.2012.0128

Simperl, E. (2009). Reusing ontologies on the Semantic Web: A feasibility study. *Data & Knowledge Engineering*, *68*(10), 905–925. doi:10.1016/j.datak.2009.02.002

Simpson, M. S., & Demner-Fushman, D. (2012). Biomedical Text Mining : A Survey Of Recent Progress. In *Mining Text Data* (pp. 465–517).

Sioutos, N., de Coronado, S., Haber, M. W., Hartel, F. W., Shaiu, W.-L., & Wright, L. W. (2007). NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, *40*(1), 30–43. doi:10.1016/j.jbi.2006.02.013

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., & Katz, Y. (2007). Pellet: A practical OWL-DL reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web*, *5*(2), 51–53. doi:10.1016/j.websem.2007.03.004

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., … Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, *25*(11), 1251–5. doi:10.1038/nbt1346

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., … Rosse, C. (2005). Relations in biomedical ontologies. *Genome Biology*, *6*(5), R46. doi:10.1186/gb-2005-6-5-r46

Solomon, W. D., Wroe, C. J., Rector, a L., Rogers, J. E., Fistein, J. L., & Johnson, P. (1999). A reference terminology for drugs. In *Annual Fall Symposium of American Medical Informatics Association* (pp. 152–5). Washington, DC: Philidelphia, PA: Hanley & Belfus.

Spasic, I., Ananiadou, S., McNaught, J., & Kumar, A. (2005). Text mining and ontologies in biomedicine: making sense of raw text. *Briefings in Bioinformatics*, *6*(3), 239–51.

Stearns, M. Q., Price, C., Spackman, K. A, & Wang, A. Y. (2001). SNOMED clinical terms: overview of the development process and project status. In S. Bakken (Ed.), *Proceedings of the AMIA Symposium, American Medical Informatics Association* (pp. 662–6). Philadelphia: Hanley and Belfus.

Steinman, M. A., Chren, M. M., & Landefeld, C. S. (2007). What's in a name? Use of brand versus generic drug names in United States outpatient practice. *Journal of General Internal Medicine*, *22*(5), 645–8. doi:10.1007/s11606-006-0074-3

Stenetorp, P., & Topi, X. G. (2011). BioNLP Shared Task 2011 : Supporting Resources. In *Proceedings of BioNLP Shared Task 2011 Workshop* (pp. 112–120). Portland, Oregon, USA: Association for Computational Linguistics.

Stevens, R., Goble, C. a, & Bechhofer, S. (2000). Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics*, *1*(4), 398–414.

Stojanovic, L. (2004). *Methods and Tools for Ontology Evolution*. Universitaet Fridericiana zu Karlsruhe.

Suárez-Figueroa, M. C., Gómez-Pérez, A., & Fernández-López, M. (2012). The NeOn Methodology for Ontology Engineering. In *Ontology Engineering in a Networked World* (pp. 9–34). Springer Berlin Heidelberg.

Suárez-Figueroa, M. C., Gómez-Pérez, A., & Villazón-Terrazas, B. (2009). How to Write and Use the Ontology Requirements Specification Document. In *On the Move to Meaningful Internet Systems: OTM 2009* (pp. 966–982).

Sutton, N., Wojtulewicz, L., Mehta, N., & Gonzalez, G. (2012). Automatic Approaches for Gene-Drug Interaction Extraction from Biomedical Text : Corpus and Comparative Evaluation. In *BioNLP '12 Proceedings of the 2012 Workshop on Biomedical Natural Language Processing* (pp. 214–222).

Sweetman, E., & Martindale, S. (2006). *Martindale: the complete drug reference.* (Thirty six.). London: Pharmaceutical Press.

Taboada, M., Martínez, D., Pilo, B., Jiménez-Escrig, A., Robinson, P. N., & Sobrido, M. J. (2012). Querying phenotype-genotype relationships on patient datasets using semantic web technology: the example of Cerebrotendinous xanthomatosis. *BMC Medical Informatics and Decision Making*, *12*, 78. doi:10.1186/1472-6947-12-78

Takarabe, M., Shigemizu, D., Kotera, M., Goto, S., & Kanehisa, M. (2011). Network-based analysis and characterization of adverse drug-drug interactions. *Journal of Chemical Information and Modeling*, *51*(11), 2977–2985.

Tari, L., Anwar, S., Liang, S., Cai, J., & Baral, C. (2010). Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. *Bioinformatics (Oxford, England)*, *26*(18), i547–53. doi:10.1093/bioinformatics/btq382

Tatonetti, N. P., Fernald, G. H., & Altman, R. B. (2011). A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. *Journal of the American Medical Informatics Association : JAMIA*, *19*(1), 79–85. doi:10.1136/amiajnl-2011-000214

Tatro, D. (2010). *Drug interaction facts 2010: The Authority on Drug Interactions*. St. Louis, MO: Wolters Kluwer Health.

Tenenbaum, J. D., Whetzel, P. L., Anderson, K., Borromeo, C. D., Dinov, I. D., Gabriel, D., … Lyster, P. (2011). The Biomedical Resource Ontology (BRO) to enable resource discovery in clinical and translational research. *Journal of Biomedical Informatics*, *44*(1), 137–45. doi:10.1016/j.jbi.2010.10.003

Thakrar, B. T., Grundschober, S. B., & Doessegger, L. (2007). Detecting signals of drug-drug interactions in a spontaneous reports database. *British Journal of Clinical Pharmacology*, *64*(4), 489–95. doi:10.1111/j.1365-2125.2007.02900.x

Thomas, P., Neves, M., Rocktäschel, T., & Leser, U. (2013). WBI-DDI : Drug-Drug Interaction Extraction using Majority Voting. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (pp. 628–635). Atlanta, USA.

Tipton, K., & Boyce, S. (2000). History of the enzyme nomenclature system. *Bioinformatics (Oxford, England)*, *16*(1), 34–40.

Tirmizi, S. H., Aitken, S., Moreira, D. a, Mungall, C., Sequeda, J., Shah, N. H., & Miranker, D. P. (2011). Mapping between the OBO and OWL ontology languages. *Journal of Biomedical Semantics*, *2 Suppl 1*(Suppl 1), S3. doi:10.1186/2041-1480-2-S1-S3

Tsarkov, D., & Harrocks, I. (2006a). FaCT++ Description Logic Reasoner: System Description. In *Third International Joint Conference, IJCAR. Automated Reasoning.* (pp. 292–297).

Tsarkov, D., & Harrocks, I. (2006b). FaCT++ Description Logic Reasoner: System Description. In *3th International Joint Conference, IJCAR. Automated Reasoning.* (pp. 292–297).

U. S. Department of Veterans Affairs, V. H. A. (2012). National Drug File – Reference Terminology (NDF-RT ™) Documentation. Retrieved from http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT Documentation.pdf

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargasvera, M., Motta, E., & Ciravegna, F. (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web*, *4*(1), 14–28. doi:10.1016/j.websem.2005.10.002

Uschold, M. (1996). *Building Ontologies : Towards a Unified Methodology*. *Expert Systems '96, the 16th Annual Conference of the British Computer Society Specialist Group on Expert Systems*.

Usié, A., Alves, R., Solsona, F., Vázquez, M., & Valencia, A. (2014). CheNER: chemical named entity recognizer. *Bioinformatics (Oxford, England)*, *30*(7), 1039–40. doi:10.1093/bioinformatics/btt639

Van Mulligen, E. M., Fourrier-Reglat, A., Gurwitz, D., Molokhia, M., Nieto, A., Trifiro, G., … Furlong, L. I. (2012). The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships. *Journal of Biomedical Informatics*, *45*, 879–884. doi:10.1016/j.jbi.2012.04.004

Van Puijenbroek, E. P., Egberts, A. C. G., Heerdink, E. R., & Leufkens, H. G. M. (2000). Detecting drug-drug interactions using a database for spontaneous adverse drug reactions: an example with diuretics and non-steroidal anti-inflammatory drugs. *European Journal of Clinical Pharmacology*, *56*(9-10), 733–8. doi:10.1007/s002280000215

Vargas, E., Navarro, M. I., Laredo, L., García-Arenillas, M., García-Mateo, M., & Moreno, A. (1997). Effect of Drug Interactions on the Development of Adverse Drug Reactions. *Clinical Drug Investigation*, *13*(5), 282–289.

Vázquez-Naya, J. M., Martínez-Romero, M., Porto-Pazos, A. B., Novoa, F., Valladares-Ayerbes, M., Pereira, J., … Dorado, J. (2010). Ontologies of drug discovery and design for neurology, cardiology and oncology. *Current Pharmaceutical Design*, *16*(24), 2724–36.

Vilar, S., Harpaz, R., Uriarte, E., Santana, L., Rabadan, R., & Friedman, C. (2012). Drug-drug interaction through molecular structure similarity analysis. *Journal of the American Medical Informatics Association : JAMIA*. doi:10.1136/amiajnl-2012-000935

Vilar, S., Uriarte, E., Santana, L., Lorberbaum, T., Hripcsak, G., Friedman, C., & Tatonetti, N. P. (2014). Similarity-based modeling in large-scale prediction of drug-drug interactions. *Nature Protocols*, *9*(9), 2147–63. doi:10.1038/nprot.2014.151

Vrandečić, D. (2010). *Ontology Evaluation*. Karlsruhe Institute of Technology.

Warrer, P., Hansen, E. H., Juhl-Jensen, L., & Aagaard, L. (2012). Using text-mining techniques in electronic patient records to identify ADRs from medicine use. *British*

Journal of Clinical Pharmacology, *73*(5), 674–84. doi:10.1111/j.1365-2125.2011.04153.x

Wermeling, D. P., Feild, C. J., Smith, D. A., Chandler, M. H., Clifton, G. D., & Boyle, D. A. (1994). Effects of Long-Term Oral Carvedilol on the Steady-State Pharmacokinetics of Oral Digoxin in Patients With Mild to Moderate Hypertension. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, *14*(5), 600–606. doi:10.1002/j.1875-9114.1994.tb02857.x

Whirl-Carrillo, M., McDonagh, E. M., Hebert, J. M., Gong, L., Sangkuhl, K., Thorn, C. F., … Klein, T. E. (2012). Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology and Therapeutics*, *92*(4), 414–7. doi:10.1038/clpt.2012.96

WHO. (n.d.). *The Anatomical Therapeutic Chemical (ATC) Classification System*. Retrieved January 20, 2014, from http://www.whocc.no/atc_ddd_index/

WHO. (1992). International monitoring of adverse reactions to drugs: adverse reaction terminology. *WHO Collaborating Centre for International Drug Monitoring*.

WHO. (2002). *The Importance of Pharmacovigilance: Safety Monitoring of Medicinal Products* (p. 52). Retrieved from http://apps.who.int/medicinedocs/en/d/Js4893e/

WHO. (2006). *The use of stems in the selection of International Nonproprietary Names (INN) for pharmaceutical substances*. *World Health Organization Press*. Retrieved December 05, 2014, from http://www.who.int/medicines/services/inn/en/

Wilbur, W. J., Rzhetsky, A., & Shatkay, H. (2006). New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, *7*, 356. doi:10.1186/1471-2105-7-356

Wimalasuriya, D. C. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, *36*(3), 306–323. doi:10.1177/0165551509360123

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., … Hassanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, *36*(Database issue), D901–6. doi:10.1093/nar/gkm958

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., … Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, *34*(Database issue), D668–72. doi:10.1093/nar/gkj067

Wissler, L., Almashraee, M., Monett, D., & Paschke, A. (2014). The Gold Standard in Corpus Annotation. In *IEEE Student Conference*. Passau.

Wu, H.-Y., Karnik, S., Subhadarshini, A., Wang, Z., Philips, S., Han, X., … Li, L. (2013). An integrated pharmacokinetics ontology and corpus for text mining. *BMC Bioinformatics*, *14*(1), 35. doi:10.1186/1471-2105-14-35

Xu, R., & Wang, Q. (2013). Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. *BMC Bioinformatics*, *14*, 181. doi:10.1186/1471-2105-14-181

Ye, Y., Jiang, Z., Diao, X., Yang, D., & Du, G. (2009). An ontology-based hierarchical semantic modeling approach to clinical pathway workflows. *Computers in Biology and Medicine*, *39*(8), 722–32. doi:10.1016/j.compbiomed.2009.05.005

Yip, V., Mete, M., Topaloglu, U., & Kockara, S. (2010). Concept Discovery for Pathology Reports using an N-gram Model. In *Summit on translational medicine* (pp. 43–47).

Yoshikawa, S., Satou, K., & Konagaya, A. (2004). Drug interaction ontology (DIO) for inferences of possible drug-drug interactions. *Studies in Health Technology and Informatics*, *107*(Pt 1), 454–8.

Zaccara, G., Gangemi, P. F., Bendoni, L., Menge, G. P., Schwabe, S., & Monza, G. C. (1993). Influence of single and repeated doses of oxcarbazepine on the pharmacokinetic profile of felodipine. *Therapeutic Drug Monitoring*, *15*(1), 39–42.

Zhang, C., Hoffmann, R., & Weld, D. S. (2012). Ontological Smoothing for Relation Extraction. In *AAAI Conference on Artificial Intelligence*.

Zhang, L., Reynolds, K. S., Zhao, P., & Huang, S.-M. (2010). Drug interactions evaluation: an integrated part of risk assessment of therapeutics. *Toxicology and Applied Pharmacology*, *243*(2), 134–45. doi:10.1016/j.taap.2009.12.016

Zhang, L., Zhang, Y. D., Zhao, P., & Huang, S.-M. (2009). Predicting drug-drug interactions: an FDA perspective. *The AAPS Journal*, *11*(2), 300–6. doi:10.1208/s12248-009-9106-3

Zhang, Y., Lin, H., Yang, Z., Wang, J., & Li, Y. (2012). A single kernel-based approach to extract drug-drug interactions from biomedical literature. *PloS One*, *7*(11), e48901. doi:10.1371/journal.pone.0048901

Zhou, S., Yung Chan, S., Cher Goh, B., Chan, E., Duan, W., Huang, M., & McLeod, H. L. (2005). Mechanism-based inhibition of cytochrome P450 3A4 by therapeutic drugs. *Clinical Pharmacokinetics*, *44*(3), 279–304.

Annexes

Annex 1

Comparison of concepts included in the different CMs

		Mille	Rubrichi	NDF-RT	DIO	DIKB	PDO	PKO	DINTO
Active Ingredient			x	x	x	x	x	x	x
Drug Product			x	x			x	x	
Drug Class			x	x					x
DDI		x	x	x					x
DDI participant		x		x					x
Precipitant						x			x
Object						x			x
Risk factor	patient-related	x	x						x
	drug-related	x	x			x			x
DDI mechanism		x							x
DDI effect		x	x						x
Recommendations		x	x						x
ADR		x	x	x					x
Indication			x	x					
Significance				x					x
PK process	absorption			x	x				x
	distribution			x	x		x		x
	metabolism			x	x	x			x
	excretion			x	x				x
Molecular interaction				x	x		x		x
Metabolite				x	x				x
Protein	target								x
	enzyme				x	x		x	x
	transporter				x			x	x
	carrier				x				x
Drug-protein relation	inhibits					x	x		x
	activates						x		x
	…								x
Physiological effect							x		
PK parameter								x	x
PK study								x	x
Drug dose				x				x	x
Pharmaceutical form			x	x					x
Administration route				x					x
Adm. frequency								x	
Excipient			x						x
Study subject								x	x

Annex 2

Ontology Requirements Specification Document (ORSD) for DINTO

Ontology Requirements Specification Document (ORSD) for DINTO	
1	Purpose
	To create a formal representation of DDI knowledge that covers all relevant aspects related with the domain and that allows both humans and machines consulting and obtaining established information and/or inferred knowledge about specific DDIs.
2	Scope
	The DDI domain and all relevant information that is used to describe DDIs in texts.
3	Implementation Language
	OWL2
4	Intended End-Users
	User 1. Text miner (working in the development of an information extraction (IE) system). **User 2**. Developer working on clinical decision support or signal detection systems. **User 3**. Curator (trying to identify DDI-related information quickly and easily). **User 4**. Researcher (looking for information about a suspected DDI). **User 5**. Clinician (looking for information about a suspected DDI).
5	Intended Uses
	Use 1. Creation of a system that automatically recognizes and extracts DDIs and related information from texts. **Use 2**. Development of alert systems supporting the detection of DDIs in clinical practice. **Use 3**. Annotation of scientific literature with known and inferred DDIs to support database curators. **Use 4**. Description of known and inferred DDIs that can be searched and consulted by researchers at clinical institutions. **Use 5**. Description of possible mechanisms underlying known and inferred DDIs to support signal detection in pharmacovigilance.
6	Ontology Requirements
	a. Non-Functional Requirements
	NFR 1. The ontology must cover standard nomenclatures as well as synonyms and term variants. **NFR 2**. The ontology must cover all DDI-related information that can be used to describe the existence and characteristics of a DDI. **NFR 3**. The ontology must follow the OBO Foundry recommendations for building ontologies. **NFR 4**. The ontology must include known DDIs and allow the inference of new ones on the basis of their pharmacokinetic and pharmacodynamic mechanisms.
	b. Functional Requirements: Groups of Competency Questions
	CQs related to the EXISTENCE of a DDI **CQ1**. Is there an interaction between DrugA and DrugB? **CQ2**. Is the effect of DrugA modified by DrugB?

Annex 2 (continuation 2)

Ontology Requirements Specification Document (ORSD) for DINTO

	Ontology Requirements Specification Document (ORSD) for DINTO (Continuation)
	b. Functional Requirements: Groups of Competency Questions
	CQs regarding the MECHANISM of the DDI **CQ3**. What is the mechanism of the interaction between DrugA and DrugB? **CQ4**. Is a PK parameter of DrugA modified by DrugB? **CQ5**. What is the type of the interaction between DrugA and DrugB? **CQ6**. What is the type of PD interaction between DrugA and DrugB? **CQ7**. Is there a PK interaction between DrugA and DrugB? ***CQs regarding the EFFECT of the DDI*** **CQ8**. Is the toxicity of DrugA exacerbated by DrugB? **CQ9**. Is the effect of the DDI an adverse effect of DrugA or DrugB? **CQ10**. Is the effect EFFECT1 produced by an interaction between DrugA and DrugB? **CQ11**. What is the effect of the interaction between DrugA and DrugB? ***CQs regarding the MANAGE of a DDI*** **CQ12**. Can be DrugA and DrugB used concomitantly? **CQ13**. Should be discontinued the treatment with DrugA when DrugB is administered? **CQ14**. Should be modified the dosage of DrugA when DrugB is administered? **CQ15**. How could be avoided the effect of the interaction between DrugA and DrugB? **CQ16**. Is there any minimal elapse of time to administer DrugA after stopping the administration of DrugB? **CQ17**. Is there any recommendation in order to avoid the DDI between DrugA and DrugB? ***CQs related to FACTORS affecting to the DDI*** **CQ18**. What patient's characteristics affecting the DDI are present? **CQ19**. What is the age of the patient? **CQ20**. What is the sex of the patient? **CQ21**. What is the race or ethnic of the patient? **CQ22**. What administration form of DrugA interacts with DrugB? **CQ23**. Is there any alternative administration form of DrugA that does not produce the interaction with DrugB? ***CQs related to the interaction of a drug and a PROTEIN*** **CQ24**. What protein does DrugA interacts with? **CQ25**. Is DrugA metobolized by any of the cytochrome P450 isoenzymes? Which ones? **CQ26**. Is DrugA transported by P-glycoprotein? **CQ27**. Has DrugA got any effect (inhibitor or inductor) of any of the cytochrome P450 isoenzymes? **CQ28**. Has DrugA got a narrow therapeutic index?

Annex 2 (continuation 3)

Ontology Requirements Specification Document (ORSD) for DINTO

<table>
<tr><td></td><td colspan="6">Ontology Requirements Specification Document (ORSD) for DINTO (Continuation)</td></tr>
<tr><td></td><td colspan="6">b. Functional Requirements: Groups of Competency Questions</td></tr>
<tr><td></td><td colspan="6">CQs related to the SIGNIFICANCE of the DDI
CQ29. How frequently the DDI has been described?
CQ30. What type of study describes the DDI?
CQ31. What is the certainty of the DDI?
CQ32. Has been the DDI established in well performed and controlled studies?
CQ33. What is the severity of the DDI?</td></tr>
<tr><td>7</td><td colspan="6">Pre-glossary of terms</td></tr>
<tr><td></td><td colspan="6">a. Terms from CQs and Frequency</td></tr>
<tr><td></td><td>DrugA/DrugB 43
effect 9
interaction 10
modified 3
mechanism 2
PK parameter 1
PK interaction 1
PD interaction 1
toxicity 1
exacerbated 1
adverse effect 1
concomitantly 1
discontinued 1</td><td colspan="2">treatment 1
administered 6
avoid* 2
elapse of time 1
recommendation 1
patient 4
age 1
sex 1
ethnic 1
race 1
administration form 1
alternative 1
protein 1</td><td colspan="3">cytochrome P450 2
P-glycoprotein 1
transported 1
metabolized 1
inhibitor 1
inductor 1
narrow therapeutic 1
index 1
frequently 1
study 1
controlled studies 1
certainty 1
established 1
severity</td></tr>
<tr><td></td><td colspan="6">b. Relevant terms identified during manual annotation of the DDI corpus</td></tr>
<tr><td></td><td colspan="2">pharmacokinetics
absorption
area under the curve
bioavailability
half life
concentration
maximum concentration
steady-state
distribution volume
protein binding
free fraction</td><td colspan="2">biotransformation
metabolism
first pass metabolism
enzyme induction
enzyme inhibition
enzyme
cytochromeP450 enzyme
system
P-glycoprotein
metabolite
prodrug</td><td colspan="2">elimination
elimination constant
excretion
clearance
pharmacodynamics
drug synergism
additive effect
potentiation
antagonism
monitorization</td></tr>
</table>

Annex 3

Labels used in the linguistic pattern analysis

Type	Description
certainty	Those words indicating the degree of confidence or certainty in the information given
coadministration-pattern	Refers to those concepts describing WHEN the two drugs are or should be administered.
concentration-location	
consequence	A term or group of terms showing a relation between different statements.
control-group	
cyp-role	
dosage	
final-effect	Decrease in effectiveness, increase in toxicity (or increase in effectiveness, if this is a beneficial DDI, etc.)
int-assertion	
metabolite	
monitorization	
object	The drug 'suffering' the interaction
object-ad-effect	
object-pharma-effect	The pharmacological effect of the object interacting drug
pd-effect	
pd-mechanism	
pharmaceutical-form	
phisyological-process	
pk-effect	The alteration in a pk-parameter due to a DDI
pk-effect-value	A value describing the extent of the alteration of the pk-parameter
pk-mechanism	The alteration in a pk-process due to a DDI
pk-parameter	
pk-process	
precipitant	The drug 'causing' the interaction
recommendation	
recommendation-degree	This feature refers to those concepts describing the encouragement for the realization of the recommendation previously made
significance	
study-subject	
study-subject-type	The type of subject: animal, dog, rat, human, patient, volunteer…
study-subject-condition	Some characteristic distinguishing the subjects, such as renal function.
study-subject-number	The number of subjects included in the study.
study-type	
unkown-role	The interacting drug which role as precipitant or object cannot be known.

Annex 4

Analysis of terms expressing modification in the DDI corpus

Main groups of terms					
↑	**increases**	**induces**	↓	**decreases**	**inhibits**
accelerating	increase	induce	antagonize	decrease	inhibit
add	increased	induced	blocks	decreased	inhibited
added	increases	inducer	decrease	decreases	inhibiting
elevate	increasing	(inducers)	decreased	decreasing	inhibition
elevated	elevate	inducing	decreases	reduce	inhibitor
elevation	elevated	induction	decreasing	reduced	inhibits
elevations	elevation		delay	reduces	antagonize
enhance	elevations		delayed	delaying	reduction
enhanced	accelerating		depresses		
enhancing			discontinuation		
greater			inhibit		
high			inhibited		
higher			inhibiting		
increase			inhibition		
increased			inhibitor		
increases			inhibits		
increasing			less		
induce			lowered		
induced			reduce		
inducer			reduced		
(inducers)			reduces		
inducing			reduction		
induction			depresses		
potentiate			shortened		
potentiated					
prolong					
prolongation					
prolonged					
twice					
double					

Annex 4 (continuation 2)

Analysis of terms expressing modification in the DDI corpus

Main groups of terms (continuation)					
→	causes	↔	changes	influences	is associated with
cause	cause	affect	change	influence	interact
caused	caused	alter	changes		
causes	causes	altered	alter		
causing	causing	avoided	altered		
leading	leading	binding			
produce		change			
produce		changes			
produced		indicate			
promotes		influence			
release		interact			
result		interfere			
resulted		interferes			
resulting		occur			
results		reported			
		showed			

Analysis Results:

↑
acceler* : It is used related to a PK mechanism. Accelerating refers to the rate of the process. Therefore, it is a term that describes **an increase in the rate (velocity) of a PK process.**
add* : add refers to **an increase** (in this case, in the risk of toxicity). However, **added** and **addition** are terms related with a **coadministration pattern**.
elevat* : This term is related with the levels or concentrations of one drug or substance. Therefore, if we consider levels and concetrations as a PK parameter, this term shows **an increase in PK parameter**.
enhanc* : It is used to indicate **an increase in PK parameter, PK process, effect and toxicity**.

Annex 4 (continuation 3)

Analysis of terms expressing modification in the DDI corpus

↑ (continuation)
high*: It is used related to doses, levels or concentrations and *"risk of toxicity"*. **Higher** is used related to concentration or levels and PK parameters. **Higher** and **greater** can be associated with a numerical value. **Highly** is used mainly related to the affinity of a drug or substance for plasmatic proteins (therefore, it is related with an specific PK process).
increas*: The word **increase** is mostly associated with a PK parameter and less frequently with a PK process, concentration or levels, dosage of a drug (in this case, it seems to refer to a recommendation), an effect, toxicity or toxic effect, a risk of an effect. The word "**increase**" usually has a value (%), although, sometimes it doesn't have it (a generic reference to *"an increase"*). When it is preceded by a modal verb (e.g., *may*) it does not refer to a PK parameter (an exception occurs with bioavailability). This fact could be related with the certainty of the described observation. The word "**increase**" can be related with the term *"significant"*, which could provide important information related to the significance of the interaction. An increase is caused by *something*, which sometimes is described in the texts. We have not found other important patterns with the rest of words **increased**, **increases** or **increasing**.
induc*: It is an important term, too. The different variation of the lemma **induc*** present interesting characteristics: - **induce**: The word "**induce**" is used referring to the PK process ***metabolism*** (e.g., *to induce the metabolism of*) or the activity of a specific enzyme (e.g., *induce CYP3A4 activity*). - **induced**: in this study we found that this word is used only to describe the **effect caused by the inducting drug**, for example: *an increase in DRUGLABEL-induced toxicity*. - **inducer**: this word is used referring to the "*inducting*" drug. - **inducing**: this word is used only in the following way: **with enzyme-inducing GROUPLABEL**. - **induction**: refers the process.
potentiat*: **potentiate** and **potentiated** refer to the activity, effect or toxicity of one drug.
prolong*: This concept is related with "*time*", since it means "to lengthen in duration or space; extend". The term **prolong** is used with both, effects of a drug and PK parameters (including drug concentrations).
twice: this word is mainly used to refer to the posology of the drugs (treatment) and not to referring to a modification in some value.
double*: It is used to show an increase (with a specific extend) of a PK parameter.

Annex 4 (continuation 4)

Analysis of terms expressing modification in the DDI corpus

↓
antagon*: The mentions refer to an effect, so this term is not used regarding PK features in this corpus.
block*: **blocks** and **blocking**, both referring to mechanism.
decreas*: A high frequency in the corpus. The word "**decrease**" is used in the same way that has been described for "**increase**". An important difference is that **decrease** is used to describe a ↓ in the effectiveness of the drug that can be related with therapeutic failure. The word **decreased** is used in the same way that "**decrease"** and it can be used as well regarding to the efficacy or effectiveness of the drug (e.g., *resulting in decreased contraceptive effectiveness*). This can be an interesting pattern, since it should not be frequently the use of the word "**increase**" related to the efficacy of one drug. **"decreasing"** seems to relate a cause and a consequence. Therefore, the study of gerund verbs could be useful for the study of relationships between different types of outputs.
delay*: is used with PK processes, although there is a mention with an effect.
discontinu*: these terms refers to the interruption of the treatment of one drug. Therefore, it seems to be related with the label "*co-administration pattern*".
inhibit*: High frequency in the corpus. The word "**inhibit"** is used in two main ways: - Is used to describe a characteristic of one drug (*drugs that inhibit CYP3A4*). This can be an indirect way to describe a DDI. - Is used to express a modification in a PK process (mainly with metabolism, but with other processes too), in the activity of an enzyme (study the relation of this with the metabolic process), in the effect or activity of one drug, but NOT with PK parameters (at least in the sentences included in the analysis). **"inhibited"** is used in a similar way, although it is important to remark the example "*DRUGLABEL is inhibited by GROUPLABEL"*. Here we can see that the term is used referring to the drug (in general). **"inhibitor"** is used as a characteristics of drugs and to express the inhibition of the metabolism process (no others processes have been identified in the analysis). **"inhibitory"** refers to the activity or effect of a drug. The other terms do not present interesting characteristics.
less: This term does not describe a modification.
lower*: the words "**lowered"**, "**lowering"** and "**lowers"** (not "**lower"**) are the ones that describe a modification. Is used with PK parameter (including concentration or levels).
reduc*: These terms are broadly used related to a PK process, a PK parameter, drug concentration or levels, an effect or efficacy of one drug, and with the drug requirements (the latter one is related with dosages or recommendations, and not with showing a modification).
short*: This term does not present interesting characteristics.
depress*: This term does not present interesting characteristics.

Annex 4 (continuation 5)

Analysis of terms expressing modification in the DDI corpus

→
caus*: This term is not used to express a modification. Moreover, it usually precedes a *"modification term"*, such as "**increase"** or "**decrease"**. It is used (without a *modification term*) with an adverse effect.
lead*: the terms **"lead"** and **"leads"** are not very frequent and are used in the same way that cause. The term **"leading to"** is interesting, since it describes a relationship between two outputs. As well as the previous one, most times it is followed by a modification term.
produc*: The same use than the previous one: it refers to a *modification term.*
promot*: It is not very frequent. It is used in the corpus with a PK process.
releas*: It is not important in this context. Mostly related with *"pharmaceutical forms"*.
result*: It is quite frequent. The same characteristics than the previous ones.

↔
affect*: This term is does not describe the type of modification produced (↓,↑ or →) . It is frequently used with negation and referring to concentrations, drug effects, PK processes (most frequently) and PK parameters.
alter*: It is used with PK processes, concentrations and PK parameters.
avoid*: It is mostly related with recommendation sentences. **bind***: This term is directly related with two processes (plasma proteins binding and physicochemical binding).
chang*: This term is used with PK parameters, concentrations, effects (e.g., *menstrual changes*) and PK processes.
indicat*: It is related with a type of study describing the DDI.
influenc*: It is used with PK parameters and concentrations.
interact*: This term does not show a modification.
interfer*: This term are mainly used with PK processes, although it is used with PK parameters too.
occur*: It refers to an output (*something that occurred or something that may occur*). Therefore, it is not relevant to express a modification.
report*: More relevant regarding the type of study.
show*: Although "**shown**" is frequently used in the corpus, it is not very important to express modification. It is more relevant regarding the type of study.

Annex 5

Reasons and solutions for duplicated classes in DINTO

Reasons and solution for duplicated classes in DINTO		
label	**Error reason**	**Manual change**
abiraterone	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
aclidinium	There is not link in ChEBI to DrugBank. The link in DrugBank to ChEBI is not to *aclidinium* but to *aclidinium bromide.* CASRN between them are not coincident. There is an error in the cross-reference from DrugBank to ChEBI.	We keep the one from DrugBank, since it is the one that has the DDI-related information and eliminate the one from ChEBI.
alogliptin	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
aminophylline	This class is no longer presented in ChEBI (CHEBI_2659).	It is eliminated from DINTO.
amphetamines	DrugBank uses the plural from *amphetamines* while ChEBI uses the singular form *amphetamine.* There is not DDI information for *amphetamines* in DrugBank.	We eliminate the new class.
asenapine	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
avanafil	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
axitinib	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
azilsartan medoxomil	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. However, CASRN do not match. The correct one is ChEBI's, while the one in DrugBank corresponds to *azilsartan.*	We eliminate the CASRN from the class imported from DrugBank and assign the URI from ChEBI.
bedaquiline	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
bicalutamide	There is link in ChEBI to DrugBank and in DrugBank to ChEBI. However, the latter one is an old (alternative) ChEBI ID. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
boceprevir	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
bosutinib	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
cabozantinib	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.

Annex 5 (continuation 2)

Reasons and solutions for duplicated classes in DINTO

Reasons and solution for duplicated classes in DINTO		
label	**Error reason**	**Manual change**
calcium carbonate	There is not link in ChEBI to DrugBank, neither in DrugBank to ChEBI. There is not CASRN or KEGG IDs in DrugBank.	We study and compare all the information provided in ChEBI and DrugBank entries and decide that they refer to the same substance. We merge both classes keeping the ChEBI URI.
canagliflozin	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
carfilzomib	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
ceftaroline fosamil	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. However, CASRN do not match. The correct one is ChEBI's, while the one in DrugBank corresponds to *ceftaroline.*	We eliminate the CASRN from the class imported from DrugBank and assign the URI from ChEBI.
dabigatran etexilate	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
dabrafenib	There is not link in ChEBI to DrugBank, neither in DrugBank to ChEBI. There is not CASRN in DrugBank, but there is KEGG DRUG ID, which matchs with the one in ChEBI.	We merge both clases keeping the ChEBI URI.
deferiprone	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
enzalutamide	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
eribulin	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
ezogabine	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
fidaxomicin	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
fluticasone furoate	There is not link in ChEBI to DrugBank, neither in DrugBank to ChEBI. However, both CASRN and KEGG IDs match.	We merge both classes keeping the ChEBI URI.
gabapentin enacarbil	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
geldanamycin	There is not link in ChEBI to DrugBank, neither in DrugBank to ChEBI. However, both CASRN and KEGG IDs match.	We merge both classes keeping the ChEBI URI.

Annex 5 (continuation 3)

Reasons and solutions for duplicated classes in DINTO

Reasons and solution for duplicated classes in DINTO		
label	**Error reason**	**Manual change**
iloperidone	There is not link in ChEBI to DrugBank, neither in DrugBank to ChEBI. However, both CASRN and KEGG IDs match.	We merge both classes keeping the ChEBI URI.
ivacaftor	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
linagliptin	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
lomitapide	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
lorcaserin	There is not link in ChEBI to DrugBank, neither in DrugBank to ChEBI. However, both CASRN and KEGG IDs match.	We merge both classes keeping the ChEBI URI.
lurasidone	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
lysergic acid diethylamide	There is not link in ChEBI to DrugBank, neither in DrugBank to ChEBI. However, both CASRN and KEGG IDs match.	We merge both classes keeping the ChEBI URI.
monastrol	There is not link in ChEBI to DrugBank, neither in DrugBank to ChEBI. There is not CASRN or KEGG IDs in DrugBank.	We study and compare all the information provided in ChEBI and DrugBank entries and decide that they refer to the same substance. We merge both classes keeping the ChEBI URI.
monothioglycerol	There is not link in ChEBI to DrugBank, neither in DrugBank to ChEBI. There is not CASRN or KEGG IDs in DrugBank.	We study and compare all the information provided in ChEBI and DrugBank entries and decide that they refer to the same substance. We merge both classes keeping the ChEBI URI.
mycophenolate mofetil	There is not link in DrugBank to ChEBI but there is link in ChEBI to DrugBank. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
octreotide	This class is no longer presented in ChEBI (CHEBI_7726).	It is eliminated from DINTO.
ospemifene	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
oxcarbazepine	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.

Annex 5 (continuation 4)

Reasons and solutions for duplicated classes in DINTO

Reasons and solution for duplicated classes in DINTO		
label	**Error reason**	**Manual change**
parecoxib	There is not link in ChEBI to DrugBank, neither in DrugBank to ChEBI. There is not CASRN or KEGG IDs in DrugBank.	We study and compare all the information provided in ChEBI and DrugBank entries and decide that they refer to the same substance. We merge both classes keeping the ChEBI URI.
pasireotide	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
pazopanib	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
perampanel	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
pitavastatin	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
pomalidomide	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
pralatrexate	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
regorafenib	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
roflumilast	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
ruxolitinib	There is not link in ChEBI to DrugBank but there is link in DrugBank to ChEBI. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
salicylamide	There is not link in ChEBI to DrugBank, neither in DrugBank to ChEBI. However, CASRN match.	We merge both classes keeping the ChEBI URI.
saxagliptin	There is link in ChEBI to DrugBank and in DrugBank to ChEBI. However, the latter one is an old (alternative) ChEBI ID. Match between the CASRN.	We merge both classes keeping the ChEBI URI.
sibutramine	There is not link in DrugBank to ChEBI. This class is no longer presented in ChEBI (ChEBI_9137).	We eliminate the class from ChEBI and keep the new class from DrugBank.
teduglutide	There is not link in ChEBI to DrugBank, neither in DrugBank to ChEBI. However, CASRN match.	We merge both classes keeping the ChEBI URI.
teleprevir	There is not link in ChEBI to DrugBank, neither in DrugBank to ChEBI. However, CASRN match.	We merge both classes keeping the ChEBI URI.

Annex 5 (continuation 5)

Reasons and solutions for duplicated classes in DINTO

Reasons and solution for duplicated classes in DINTO		
label	**Error reason**	**Manual change**
telavancin	There is not link in ChEBI to DrugBank. The link in DrugBank to ChEBI is not to *telavancin* but to *telavancin hydrochloride.* CASRN between them are not coincident. There is an error in the cross-reference from DrugBank to ChEBI.	We eliminate the CASRN from the class imported from DrugBank and assign the URI from ChEBI corresponding to ChEBI_71226.
teriflunomide	There is not link in ChEBI to DrugBank, neither in DrugBank to ChEBI. However, CASRN match.	We merge both classes keeping the ChEBI URI.
thioproline	There is not link in ChEBI to DrugBank, neither in DrugBank to ChEBI. CASRN do not match, because the one in DrugBank is an older one.	We study and compare all the information provided in ChEBI and DrugBank entries and decide that they refer to the same substance. We merge both classes keeping the ChEBI URI.
ticagrelor	There is not link in ChEBI to DrugBank, neither in DrugBank to ChEBI. However, both CASRN and KEGG IDs match.	We merge both classes keeping the ChEBI URI.
valacyclovir/valaciclovir	There is a spelling variation between the both labels. Son en el mismo. Llamo a valaciclovir DBName	We merge both classes keeping the ChEBI URI. We manually change the label *valaciclovir* to the annotation property DBName
vandetanib	There is not link in DrugBank to ChEBI, but there is link in ChEBI to DrugBank. However, it corresponds with a different substance (DB08764)	We merge both classe keeping the ChEBI URI and manually eliminate the incorrect cross-reference from ChEBI.
vanoxerine	There is not link in ChEBI to DrugBank, neither in DrugBank to ChEBI. There is not CASRN or KEGG IDs in DrugBank.	We study and compare all the information provided in ChEBI and DrugBank entries and decide that they refer to the same substance. We merge both classes keeping the ChEBI URI.
vemurafenib	There is not link in ChEBI to DrugBank, neither in DrugBank to ChEBI. However, CASRN match.	We merge both classes keeping the ChEBI URI.
vismodegib	There is not link in ChEBI to DrugBank, neither in DrugBank to ChEBI. However, CASRN match.	We merge both classes keeping the ChEBI URI.
ximelagatran	There is not link in DrugBank to ChEBI but there is link in ChEBI to DrugBank. Match between the CASRN.	We merge both classes keeping the ChEBI URI.

Annex 6

Description of object properties in DINTO

Annotation property	Definition	Inverse of	Sub Property of	Domain	Range	Sub property chain
adsorbs	A relationship between two chemicals in which one of them is retained by the other one.	is adsorbed by				No
alters	A relationship between an entity x (process or continuant) and a process y, which bears a quality that is altered by x, leading to a change compare to normal or previous value.	is altered by				No
decreases	A relationship between an entity x (process or continuant) and an entity y, which bears a quality that is decreased by x, leading to a change that is decreased compare to normal or previous value.	is decreased by	alters			Yes (8)
increases	A relationship between an entity x (process or continuant) and an entity y, which bears a quality that is increased by x, leading to a change that is increased compare to normal or previous value.	is increased by	alters			Yes (7)
binds	A relationship between a pharmacological entity and a protein.	is binded by				No
has pharmacological target	A relationship of a pharmacological entity that interacts with a protein - denominated target – producing the pharmacological effect of the entity.	is pharmacological target of	binds	pharmacological entity	target	No
induces	A relationship between a pharmacological entity and a protein where the entity increases the activity of the protein.	is induced by	binds	pharmacological entity	protein entity	No

Annex 6 (continuation 2)

Description of object properties in DINTO

Annotation property	Definition	Inverse of	Sub Property of	Domain	Range	Sub property chain
inhibits	A relationship between a pharmacological entity and a protein where the entity decreases the activity of the protein.	is inhibited by	binds	pharmacological entity	protein entity	No
is substrate of	A relationship between a chemical entity and a protein where a substrate interaction can occur between them. Substrate Interaction involves temporary non-covalent binding through intermolecular physical forces of attraction and spatial complementarity between biologically-active molecules and their target molecule or between a biological molecule and an underlaying surface.	has substrate	binds			
is carried by	A relationship between a pharmacological entity and a protein which carries the entity.	carries	is substrate of	pharmacological entity	carrier	
is metabolised by	A relationship between a chemical entity and an enzyme involved in the metabolism of the entity.	metabolizes	is substrate of	pharmacological entity	enzyme	
is transported by	A relationhip between a pharmacological entity and a protein which transports the entity.	transports	is substrate of	pharmacological entity	transporter	
modulates	A relationhip between a pharmacological entity which binds a protein, triggering a response or blocking it.	is modulated by	binds	pharmacological entity	protein entity	

Annex 6 (continuation 3)

Description of object properties in DINTO

Annotation property	Definition	Inverse of	Sub Property of	Domain	Range	Sub property chain
activates	A relationship between a pharmacological entity and a protein, where the chemical entity selectively binds to and activates the protein.	is activated by	modulates			
blocks	A relationship between a pharmacological entity and a protein, where the entity binds to but does not activate the protein, thereby blocking the actions of endogenous or exogenous agonists.	is blocked by	modulates			
related with	A relationship between a pharmacological entitiy and a protein that are known to interact in some way, although more specific details about the interaction are still ignored.	symmetric	binds			
chelates	A relationship between two chemicals describing 'the formation or presence of bonds between two or more separate binding sites within the same ligand and a single central atom' that leads to the formation of a compound called chelate.	is chelated by				
describes	A relationship between an information resource describing a DDI and this DDI.	is described by	information resource	DDI		
determines	A relationship between an entity a and an entity b, which existence or some characteristic of b depends directly in the entity a.	is determined by				

Annex 6 (continuation 4)

Description of object properties in DINTO

Annotation property	Definition	Inverse of	Sub Property of	Domain	Range	Sub property chain
has agent	A relationship between a DDI mechanism and the precipitant entity that triggers the process.	is agent in	DDI mechanism	pharmacological entity and (has role) some precipitant)		
has DDI effect	A relationship between a DDI and the altered physiological effect that occurs as a consequence of the DDI.	is DDI effect of	DDI	DDI effect		
has effect	A relationship between a pharmacological entity and its physiological effect.	is effect of		pharmacological entity	physiological effect	
has metabolite	A relationship between a pharmacological entity and the metabolite produced as a consequence of the metabolism of the entity.	is metabolite of		pharmacological entity	pharmacological entity	
has parameter	A relationship between a pharmacokinetic process and the pharmacokinetic parameter describing it.	is parameter of		pharmacokinetic process	pharmacokinetic parameter	
has participant	A relationship between a DDI and the two drugs - one object and one precipitant - participating in it.	is participant in		DDI	pharmacological entity	
has object	A functional relationship between a DDI and an object.	is object in	has participant	DDI	pharmacological entity	
has precipitant	A functional relationship between a DDI and a precipitant.	is precipitant in	has participant	DDI	pharmacological entity	

Annex 6 (continuation 5)

Description of object properties in DINTO

Annotation property	Definition	Inverse of	Sub Property of	Domain	Range	Sub property chain
has product	A relationship between the pharmacokinetic process 'drug metabolism', by which a drug is metabolized, and the metabolite that is formed as a consequence.	is product of		drug metabolism		
has quality	Has quality is a relation between an entity and the quality that it bears.	is quality of				
has role	The relationship between a chemical and the role that it can exhibit in a chemical or biological context.	is role of		pharmacological entity	role	
has study entity	A relationship between an information resource describing some aspect of a DDI and the two pharmacological entities that are studied as interacting drugs.			information resource	pharmacological entity and (is participant in some (DDI and is described in some information resource))	
has study subject	A relationship between an information resource describing some aspect of a DDI and the individual that is subject in the study.			information resource	study subject	
is adsorbed by	A relationship between two chemicals in which one of them is retained by the other one.	adsorbs				

Annex 6 (continuation 6)

Description of object properties in DINTO

Annotation property	Definition	Inverse of	Sub Property of	Domain	Range	Sub property chain
is agent in	A relationship between a pharmacological entity and a DDI mechanism that leads to a DDI where the entity is the precipitant or perpetrator of the interaction.	has agent		pharmacological entity and (has role some precipitant)	DDI mechanism	
is altered by	A relationship between a process y that bears a quality that is altered by an entity x (process or continuant), leading to a change with compare to normal or previous value.	alters				
is decreased by	A relationship between an entity y that bears a quality that is decreased by an entity x (process or continuant), leading to a change that is decreased compare to normal or previous value.	decreases				Yes (8)
is increased by	A relationship between an entity y that bears a quality that is increased by an entity x (process or continuant), leading to a change that is increased compare to normal or previous value.	increases				Yes (7)
is binded by	A relationship between a protein and the pharmacological entity that interacts with the protein.	binds				
carries	A relationhips between a protein and a pharmacological entity that is carried by the protein.	is carried by	has substrate	carrier	pharmacological entity	

Annex 6 (continuation 7)

Description of object properties in DINTO

Annotation property	Definition	Inverse of	Sub Property of	Domain	Range	Sub property chain
metabolizes	A relationship between an enzyme and a pharmacological entity which is metabolized by it.	is metabolised by	has substrate	enzyme	pharmacological entity	
transports	A relationship between a protein and a pharmacological entity that is transported in the body by that protein.	is transported by	has substrate	transporter	pharmacological entity	
has substrate	A relationship between a protein and a chemical entity when a substrate interaction can occur between them. Substrate Interaction involves temporary non-covalent binding through intermolecular physical forces of attraction and spatial complementarity between biologically-active molecules and their target molecule or between a biological molecule and an underlying surface.	is substrate of	is binded by			
is induced by	A relationship between a protein and a pharmacological entity which increases the activity of this protein.	induces	is binded by	protein entity	pharmacological entity	
is inhibited by	A relationship between a protein and a pharmacological entity which reduces the activity of this protein.	inhibits	is binded by	protein entity	pharmacological entity	
is modulated by	A relationhip between a protein and a pharmacological entity, which binds to it triggering a response or blocking it.	modulates	is binded by	protein entity	pharmacological entity	

Annex 6 (continuation 8)

Description of object properties in DINTO

Annotation property	Definition	inverse of	Sub Property of	domain	range	Sub property chain
is activated by	A relationship between a protein and a pharmacological entity, where the protein is selectively binded and activated by the chemical entity.	activates	is modulated by			
is blocked by	A relationship between a protein and a pharmacological entity, where the protein is binded, but not activated, by the entity, thereby blocking the actions triggered by endogenous or exogenous agonists.	blocks	is modulated by			
is pharmacolgoical target of	A relationship between a protein denominated target and a pharmacological entity that interact producing the pharmacological effect of the entity.	has pharmacological target	is binded by	target	pharmacological entity	
is chelated by	A relationship between two chemicals describing 'the formation or presence of bonds between two or more separate binding sites within the same ligand and a single central atom' that leads to the formation of a compound called chelate.	chelates				
is DDI effect of	A relationship between an altered physiological effect and the DDI responsible of the alteration	has DDI effect		DDI effect	DDI	

Annex 6 (continuation 9)

Description of object properties in DINTO

Annotation property	Definition	Inverse of	Sub Property of	Domain	Range	Sub property chain
is described in	A relationship between a DDI and an information resource describing some aspect of the interaction.	describes		DDI	information resource	
is determined by	A relationship between an entity b and an entity a, when the existence or some characteristic of b depends directly in the entity a.	determines				
is effect of	A relationship between a physiological effect and the chemical entity that produces the effect.	has effect		physiologic effect	pharmacological entity	
is metabolite of	A relationship between a metabolite and a pharmacological entity, where the metabolite is produced as a consequence of the metabolism of the entity.	has metabolite			pharmacological entity	
is parameter of	A relationship between a pharmacokinetic parameter and the pharmacokinetic process that it describes.	has parameter		pharmacokinetic parameter	pharmacokinetic process	
is participant in	A relationship between a drug, precipitant or object, and the DDI in which it is one of the two participating drugs.	has participant		pharmacological entity	DDI	
is object in	A relationship between a drug and the DDI in which it is the object participating drug.	has object		pharmacological entity	DDI	
is precipitant in	A relationship between a drug and the DDI in which it is the precipitant participating drug.	has precipitant		pharmacological entity	DDI	

Annex 6 (continuation 10)

Description of object properties in DINTO

Annotation property	Definition	inverse of	Sub Property of	domain	range	Sub property chain
is preceded by	A transitive, temporal relation in which one process is preceded (has occurred later than) another process.	precedes				
is product of	A relationship between a metabolite and the pharmacokinetic process drug metabolism in which a drug is metabolized into this metabolite.	has product		drug metabolism		
is quality of	Is quality of is a relation between a quality and the entity that it is a property of.	has quality				
is regulated by	A relationship between a process and a physiological entity (continuant or process), where the entity modulates a measurable attribute of the process, quality or function.	regulates				
is facilitate by	A relationship between a process y and an entity x (continuant or process), where the occurrence of y depends directly or is heavily dependent on x.	facilitates	is regulated by			
is impaired by	A relationship between a process y and an entity x (continuant or process), where x avoids the occurrence or reduces the magnitude of process y.	impairs	is regulated by			
is role of	The relationship between a role and the chemical that can exhibit it in a chemical or biological context.	has role		role	pharmacological entity	

Annex 6 (continuation 11)

Description of object properties in DINTO

Annotation property	Definition	Inverse of	Sub Property of	Domain	Range	Sub property chain
is undergone by	A relationship between a pharmacokinetic process and the chemical entity that undergoes the process.	undergoes		pharmacokinetic process	pharmacological entity	
may interact with	A relationship between two pharmacological entities, y and x, where the levels or effects of one of them (y) are altered by the other (x).	Symmetric				
precedes	A transitive, temporal relation in which one process precedes (has occurred earlier than) another process.	is preceded by				
regulates	A relationship between a physiological entity (continuant or process) and a process, where the entity modulates a measurable attribute of the process, quality or function.	is regulated by				
facilitates	A relationship between an entity x (continuant or process) and a process y, which occurrence depends directly or is heavily dependent on x.	is facilitated by	regulates			
impaires	A relationship between an entity x (continuant or process) and a process y, where x avoids the occurrence or reduces the magnitude of process y.	is impaired by	regulates			
undergoes	A relationship between a chemical entity and the pharmacokinetic process that the chemical entity undergoes in the body.	is undergone by		pharmacological entity	pharmacokinetic process	

Annex 7

CQs, answers and axioms for the DDI between *rifampicin* and *cyclosporin A*

	Question	Answer	Axioms
1	Is there an interaction between ***rifampicin*** and ***cyclosporin A***?	*Yes, there is an interaction between both drugs. This information is explicitly represented in the ontology at the class level. As well, the interaction can be asserted between individuals.*	**rifampicin** *may_interact_with* some **cyclosporin A**
			cyclosporin A *may_interact_with* some **rifampicin**
			cyclosporin A/rifampicin DDI equivalent to ((*has_participant* some **cyclosporin A**) and (*has_participant* some **rifampicin**))
			Rifampicin *may_interact_with* Cyclosporin A
			Cyclosporin A *may_interact_with* Rifampicin
			Cyclosporin A/Rifampicin DDI *has_participant* CyclosporinA
			Cyclosporin A/Rifampicin DDI *has_participant* Rifampicin
2	Is the effect of ***cyclosporin A*** modified by ***rifampicin***?	*Cyclosporin A's effects represented in this example (immunosuppressant effect and transplant rejection) are modified by rifampicin. The first one is decreased, while the second one is increased.*	Cyclosporin A *has_effect* ImmunosuppresantEffect
			Cyclosporin A *has_effect* TransplantRejection
			Rifampicin *decreases* ImmunosuppresantEffect
			Rifampicin *increases* TransplantRejection
3	What is the mechanism of the interaction between ***cyclosporin A*** and ***rifampicin***?	*It is a pharmacokinetic mechanism: specifically the induction of the enzyme cytochrome P450 3A4.*	Cyclosporin A/rifampicin DDI *is_preceded_by* CytochromeP4503A4Induction
			CytochromeP4503A4Induction *is_a* **enzyme activity induction**
			Cyclosporin A/rifampicin DDI *is_preceded_by* CytochromeP4503A4Induction
4	Is a PK parameter of ***cyclosporin A*** modified by ***rifampicin***?	*Yes, rifampicin alters Cmax of cyclosporin A.*	Rifampicin *alters* Cmax
			Cmax *is_a* **pharmacokinetic parameter**

classes; *object_properties*; Individuals; *attributes*

Annex 7 (continuation 2)

CQs, answers and axioms for the DDI between *rifampicin* and *cyclosporin A*

	Question	Answer	Axioms
5	What is the type of the interaction between ***cyclosporin A*** and ***rifampicin***?	*The DDI is classified in two different ways in the ontology. Firstly, it is classified as an enzyme induction DDI. Therefore, it is a PK DDI. Secondly, it is classified in basis of the consequence for the patient as a potentially harmful DDI.*	CyclosporinA/Rifampicin DDI *is_a* **enzyme induction DDI**
			CyclosporinA/Rifampicin DDI *is_a* **potentially harmful DDI**
6	What is the type of PD interaction between ***cyclosporin A*** and ***rifampicin***?	*There is not a PD DDI between these two drugs.*	
7	Is there a PK interaction between ***cyclosporin A*** and **rifampicin**?	*Yes, the interaction between these two drugs is a PK DDI.*	CyclosporinA/Rifampicin DDI *is_a* **enzyme induction DDI**
			enzyme induction DDI *is_a* **enzyme alteration DDI**
			enzyme alteration DDI *is_a* **pharmokinetic DDI**
8	Is the toxicity of ***cyclosporin A*** exacerbated by ***rifampicin***?	*No, the effect of this DDI is a decreased in the therapeutic effect of cyclosporin A, not an increase in the toxicity.*	CyclosporinA/Rifampicin DDI *has_DDI_effect* TransplantRejection
			TransplantRejection *is effect of* CyclosporinA
			TransplantRejection *is_a* **decreased therapeutic effect**
			decreased therapeutic effect *is_a* **altered therapeutic effect**

classes; *object_properties*; Individuals; *attributes*

Annex 7 (continuation 3)

CQs, answers and axioms for the DDI between *rifampicin* and *cyclosporin A*

	Question	Answer	Axioms
9	Is the effect of the DDI an adverse effect of ***cyclosporin A*** or ***rifampicin***?	*No, the effect of this DDI is a decreased in the therapeutic effect of cyclosporin A, not an increase in the adverse effect.*	CyclosporinA/Rifampicin DDI *has_DDI_effect* TransplantRejection
			TransplantRejection *is_effect_of* CyclosporinA
			TransplantRejection *is_a* **decreased therapeutic effect**
			decreased therapeutic effect *is_a* **altered therapeutic effect**
10	Is the effect ***transplant rejection*** produced by an interaction between ***cyclosporin A*** and ***rifampicin***?	*Yes, the effect of the interaction between cyclosporin A and rifampicin is transplant rejection.*	CyclosporinA/Rifampicin DDI *has_DDI_effect* TransplantRejection
11	What is the effect of the interaction between ***cyclosporin A*** and ***rifampicin***?	*The effect of the interaction between cyclosporin A and rifampicin is transplant rejection, the consequence of the decrease in the therapeutic effect of cyclosporin A.*	CyclosporinA/Rifampicin DDI *has_DD_ effect* TransplantRejection
			TransplantRejection *is_effect_of* CyclosporinA
			TransplantRejection *is_a* **decreased therapeutic effect**
			decreased therapeutic effect *is_a* **altered therapeutic effect**
12	Can be ***cyclosporin A*** and ***rifampicin*** used concomitantly?	*No, it is necessary to change the dose schedule in order to not administrate both drugs concomitantly.*	CyclosporinA/Rifampicin DDI *has_dose_recommendation change dose schedule*
13	Should be discontinued the treatment with ***cyclosporin A*** when ***rifampicin*** is administered?	*It is necessary to change the dose schedule in order to not administrate both drugs concomitantly.*	CyclosporinA/Rifampicin DDI *has_dose_recommendation change dose schedule*

classes; *object_properties*; Individuals; *attributes*

Annex 7 (continuation 4)

CQs, answers and axioms for the DDI between *rifampicin* and *cyclosporin A*

	Question	Answer	Axioms
14	Should be modified the dosage of ***cyclosporin A*** when ***rifampicin*** is administered?	*Yes, it is necessary to increase the dose of cyclosporin A.*	CyclosporinA/Rifampicin DDI *has_dose_recommendation increase from baseline*
15	How could be avoided the effect of the interaction between ***cyclosporin A*** and ***rifampicin***?	*To avoid the DDI it is necessary to change the dose schedule in order to not administrate both drugs concomitantly and to to increase the dose of cyclosporin A.*	CyclosporinA/Rifampicin DDI *has_dose_recommendation change dose schedule* CyclosporinA/Rifampicin DDI *has_dose_recommendation increase from baseline*
16	Is there any minimal elapse of time to administer ***cyclosporin A*** after stopping the administration of ***rifampicin***?	*Yes, it is necessary to change the dose schedule in order to not administrate both drugs concomitantly. The numerical value is not represented in the ontology.*	CyclosporinA/Rifampicin DDI *has_dose_recommendation change dose schedule*
17	Is there any recommendation in order to avoid the DDI between ***cyclosporin A*** and ***rifampicin***?	*Yes, to avoid the DDI it is necessary to change the dose schedule in order to not administrate both drugs concomitantly and to increase the dose of cyclosporin A.*	CyclosporinA/Rifampicin DDI *has_dose_recommendation change dose schedule* CyclosporinA/Rifampicin DDI *has_dose_recommendation increase from baseline*
18	What patient's characteristics affecting the DDI are present?	*Information related to patient characteristics is not included in the description of the DDI in the text source.*	

classes; *object_properties*; Individuals; *attributes*

Annex 7 (continuation 5)

CQs, answers and axioms for the DDI between *rifampicin* and *cyclosporin A*

	Question	Answer	Axioms
19	What is the age of the patient?	*Information related to patient characteristics is not included in the description of the DDI in the text source.*	
20	What is the sex of the patient?	*Information related to patient characteristics is not included in the description of the DDI in the text source.*	
21	What is the race or ethnic?	*Information related to patient characteristics is not included in the description of the DDI in the text source.*	
22	What administration form of ***cyclosporin A*** interacts with ***rifampicin***?	*Information related to drugs' pharmaceutical forms is not included in the description of the DDI in the text source.*	
23	Is there any alternative administration form of ***cyclosporin A*** that does not produce the interaction with ***rifampicin***?	*This information is not represented in the ontology.*	

classes; *object_properties*; Individuals; *attributes*

Annex 7 (continuation 6)

CQs, answers and axioms for the DDI between *rifampicin* and *cyclosporin A*

	Question	Answer	Axioms
24	What proteins interact with ***cyclosporin A***?	*This information is explicitly represented in the ontology at the class level: calcineurin subunit b isoform 2, calcium signal-modulating cyclophilin ligand, cytochrome p450 3a4, cytochrome p450 3a7, cytochrome p450 3a5, cytochrome p450 2c8, cytochrome p450 2c9, cytochrome p450 2c19, cytochrome p450 2d6, multidrug resistance protein 1, atp-binding cassette sub-family g member 2, bile salt export pump, canalicular multispecific organic anion transporter 2, canalicualr multispecific organic anion transporter 1, ilial sodium/bile acid cotransporter, multidrug resistance-associated protein 7, multidrug resistance-associated protein 1, sodium/bile acid cotransporter, solute carrier family 22 member 6, solute carrier organic anion transporter family member 1a2 and solute carrier organic anion transporter family member 1b1.*	**cyclosporin A …** *has_pharmacological_target* some **calcineurin subunit b isoform 2** *has_pharmacological_target* some **calcium signal-modulating cyclophilin ligand** *is_metabolized_by* some **cytochrome p450 3a4** *is_metabolized_by* some **cytochrome p450 3a7** *is_metabolized_by* some **cytochrome p450 3a5** *is_transported_by* some **multidrug resistance protein 1** *binds* some **calcium signal-modulating cyclophilin ligand** *induces* some **cytochrome p450 3a7** *induces* some **cytochrome p450 3a5** *induces* some **multidrug resistance protein 1** *inhibits* some **atp-binding cassette sub-family g member 2** *inhibits* some **bile salt export pump** *inhibits* some **calcineurin subunit b isoform 2** *inhibits* some **canalicular multispecific organic anion transporter 2** *inhibits* some **canalicualr multispecific organic anion transporter 1** *inhibits* some **cytochrome p450 2c8** *inhibits* some **cytochrome p450 2c9** *inhibits* some **cytochrome p450 2c19** etc.

Annex 7 (continuation 7)

CQs, answers and axioms for the DDI between *rifampicin* and *cyclosporin A*

	Question	Answer	Axioms
25	Is ***cyclosporin A*** metobolized by any of the cytochrome P450 isoenzymes? Which ones?	*Yes, it is metabolized by several CYP450 isoenzymes: cytochrome p450 3a4, cytochrome p450 3a7 and cytochrome p450 3a5.*	*is_metabolized_by* some **cytochrome p450 3a4** *is_metabolized_by* some **cytochrome p450 3a7** *is_metabolized_by* some **cytochrome p450 3a5**
26	Is ***cyclosporin A*** transported by P-glycoprotein?	*Yes, it is transported by P-glycoprotein (which synonym is multidrug resistance protein 1).*	*is_transported_by* some **multidrug resistance protein 1**
27	Has ***cyclosporin A*** got any effect (inhibitor or inductor) of any of the cytochrome P450 isoenzymes?	*Yes, cyclosporin A inhibits several CYP450 isoenzymes (2C8, 2C9, 2C19, 2D6 and 3A4) and induces others, as well (3A7, 3A5)*	*induces* some **cytochrome p450 3a7** *induces* some **cytochrome p450 3a5** *inhibits* some **cytochrome p450 2c8** *inhibits* some **cytochrome p450 2c9** *inhibits* some **cytochrome p450 2c19** *inhibits* some **cytochrome p450 2d6** *inhibits* some **cytochrome p450 3a4**
28	Has **cyclosporin A** got a narrow therapeutic index?	*This information is not represented in the ontology.*	

classes; *object_properties*; Individuals; *attributes*

Annex 7 (continuation 8)

CQs, answers and axioms for the DDI between *rifampicin* and *cyclosporin A*

	Question	Answer	Axioms
29	How frequently has the DDI been described?	*This DDI have not been described frequently.*	CyclosporinA/Rifampicin DDI *has_incidence low*
30	What type of study describes the DDI?	*The DDI is described in a clinical study carried out in patients.*	CyclosporinA/Rifampicin DDI *is_described_in* Study1
			Study1 *is_a* **patient study**
31	What is the certainty of the DDI?	*The DDI is well-known and established.*	CyclosporinA/Rifampicin DDI *has_documentation_level established*
32	Has been the DDI established in well performed and controlled studies?	*The DDI is described in a clinical study carried out in patients*.*	CyclosporinA/Rifampicin DDI *is_described_in* Study1
			Study1 *is_a* **patient study**
33	What is the severity of the DDI?	*It is a major and clinically relevant DDI.*	CyclosporinA/Rifampicin DDI *has_severity major*
			CyclosporinA/Rifampicin DDI *has_relevance clinical relevance*

classes; *object_properties*; Individuals; *attributes*

Annex 8

CQs, answers and axioms for the DDI between *naloxone* and *morphine*

	Question	Answer	Axioms
1	Is there an interaction between ***morphine*** and ***naloxone***?	*The interaction between morphine and naloxone is not explicitly described in the ontology, since it is not described in the original information source (the database DrugBank). However, the information can be represented in the ontology[60].*	**naloxone** *may_interact_with* some **morphine**
			morphine *may_interact_with* some **naloxone**
			naloxone/morphine DDI equivalent to ((*has_participant* some **morphine**) and (*has_participant* some **naloxone**))
			Naloxone *may_interact_with* Morphine
			Morphine *may_interact_with* Naloxone
			Naloxone/Morphine DDI *has_participant* Morphine
			Naloxone/Morphine DDI *has_participant* Naloxone
2	Is the effect of ***morphine*** modified by ***naloxone***?	*The effect of morphine represented in this example (CNS depression) is decreased by naloxone.*	Morphine *has_effect* CNSDepression
			Naloxone *decreases* CNSDepression
3	What is the mechanism of the interaction between ***naloxone*** and ***morphine***?	*It is a pharmacodynamic mechanism, specifically the antagonism of the mu-opioid receptor.*	Naloxone/Morphine DDI *is_preceded_by* MuReceptorAntagonism
			MuReceptorAntagonism *is_a* **antagonistic DDI mechanism**
4	Is a PK parameter of ***morphine*** modified by ***naloxone***?	*No, in this interaction there is not a PK parameter altered.*	

classes; *object_properties*; Individuals; *attributes*

[60] We manually create the class 'naloxone/morphine DDI'.

Annex 8 (continuation 2)

CQs, answers and axioms for the DDI between *naloxone* and *morphine*

	Question	Answer	Axioms
5	What is the type of the interaction between ***morphine*** and ***naloxone***?	*The DDI is classified in two different ways in the ontology. Firstly, it is classified as an antagonistic DDI. Therefore, it is a PD DDI. Secondly, it is classified in basis of the consequence for the patient as a potentially beneficial DDI.*	Naloxone/Morphine DDI *is_a* **antagonistic DDI**
			Naloxone/Morphine DDI *is_a* **potentially beneficial DDI**
6	What is the type of PD interaction between ***morphine*** and **na*l*oxone**?	*The type of interaction is antagonistic DDI.*	Naloxone/Morphine DDI *is_a* **antagonistic DDI**
7	Is there a PK interaction between ***morphine*** and ***naloxone***?	*No, the interaction between these drugs in the type pharmacodynamic.*	
8	Is the toxicity of ***morphine*** exacerbated by ***naloxone***?	*No, the effect of morphine is decreased by naloxone.*	Naloxone/Morphine DDI *has_DDI_effect* CNSDepressionRecovery
			CNSDepressionRecovery *is_a* **decreased adverse effect**
9	Is the effect of the DDI an adverse effect of ***morphine*** or ***naloxone***?	*No, the effect of the DDI is a decrease in the adverse effect of morphine.*	Naloxone/Morphine DDI *has_DDI_effect* CNSDepressionRecovery
			CNSDepressionRecovery *is_a* **decreased adverse effect**
10	Is the effect ***CNS depression recovery*** produced by an interaction between ***naloxone*** and ***morphine***?	*Yes, the effect of the interaction between naloxone and morphine is the recovery of the CNS depression induced by morphine.*	Naloxone/Morphine DDI *has_DDI_effect* CNSDepressionRecovery

classes; *object_properties*; Individuals; *attributes*

Annex 8 (continuation 3)

CQs, answers and axioms for the DDI between *naloxone* and *morphine*

	Question	Answer	Axioms
11	What is the effect of the interaction between ***morphine*** and ***naloxone***?	*The effect of the interaction between naloxone and morphine is the recovery of CNS depression induced by morphine.*	Naloxone/Morphine DDI *has_DDI_effect* CNSDepressionRecovery
12	Can be ***morphine*** and ***naloxone*** used concomitantly?	*Since the interaction between these drugs in beneficial, this question is not relevant in this example.*	
13	Should be discontinued the treatment with ***naloxone*** when ***morphine*** is administered?	*Since the interaction between these drugs in beneficial, this question is not relevant in this example.*	
14	Should be modified the dosage of ***morphine*** when ***naloxone*** is administered?	*Since the interaction between these drugs in beneficial, this question is not relevant in this example.*	
15	How could be avoided the effect of the interaction between ***morphine*** and ***naloxone***?	*Since the interaction between these drugs in beneficial, this question is not relevant in this example.*	
16	Is there any minimal elapse of time to administer ***morphine*** after stopping the administration of ***naloxone***?	*Since the interaction between these drugs in beneficial, this question is not relevant in this example.*	
17	Is there any recommendation in order to avoid the DDI between ***morphine*** and ***naloxone***?	*Since the interaction between these drugs in beneficial, this question is not relevant in this example.*	

classes; *object_properties*; Individuals; *attributes*

Annex 8 (continuation 4)

CQs, answers and axioms for the DDI between *naloxone* and *morphine*

	Question	Answer	Axioms
18	What patient's characteristics affecting the DDI are present?	*Information related to patient characteristics is not included in the description of the DDI in the text source.*	
19	What is the age of the patient?	*Information related to patient characteristics is not included in the description of the DDI in the text source.*	
20	What is the sex of the patient?	*Information related to patient characteristics is not included in the description of the DDI in the text source.*	
21	What is the race or ethnic?	*Information related to patient characteristics is not included in the description of the DDI in the text source.*	
22	What administration form of ***morphine*** interacts with ***naloxone***?	*Information related to drugs' pharmaceutical forms is not included in the description of the DDI in the text source.*	
23	Is there any alternative administration form of ***morphine*** that does not produce the interaction with ***naloxone***?	*This information is not represented in the ontology.*	

classes; *object_properties*; Individuals; *attributes*

Annex 8 (continuation 5)

CQs, answers and axioms for the DDI between *naloxone* and *morphine*

	Question	Answer	Axioms
24	What proteins ***morphine*** interacts with?	*This information is explicitly represented in the ontology at the class level. Morphine interacts with several proteins: delta-type opioid receptor, kappa-type opioid receptor, mu-type opioid receptor, cytochrome p450 1a2, cytochrome p450 2c8, cytochrome p450 2d6, cytochrome p450 3a4, udp-glucuronosyltransferase 1-1,udp-glucuronosyltransferase 1-3, udp-glucuronosyltransferase 1-8, udp-glucuronosyltransferase 2b15, udp-glucuronosyltransferase 2b4, udp-glucuronosyltransferase 2b7 and multidrug resistance protein 1.*	**morphine …** *has_pharmacological_target* some **delta-type opioid receptor** *has_pharmacological_target* some **kappa-type opioid receptor** *has_pharmacological_target* some **mu-type opioid receptor** *is_transported_by* some **multidrug resistance protein 1** *activates* some **delta-type opioid receptor** *activates* some **kappa-type opioid receptor** *activates* some **mu-type opioid receptor** *inhibits* some **multidrug resistance protein 1** *is_metabolized_by* some **cytochrome p450 1a2** *is_metabolized_by* some **cytochrome p450 2c8** *is_metabolized_by* some **cytochrome p450 2d6** *is_metabolized_by* some **cytochrome p450 3a4** *is_metabolized_by* some **udp-glucuronosyltransferase 1-1** *is_metabolized_by* some **udp-glucuronosyltransferase 1-3** *is_metabolized_by* some **udp-glucuronosyltransferase 1-8** etc.
25	Is ***morphine*** metobolized by any of the cytochrome P450 isoenzymes? Which ones?	*Yes, it is metabolized by several CYP450 isoenzymes: cytochrome p450 1a2, cytochrome p450 2c8, cytochrome p450 2d6 and cytochrome p450 3a4.*	**morphine** *is_ metabolized_by* some **cytochrome p450 1a2**
			morphine *is_ metabolized_by* some **cytochrome p450 2c8**
			morphine *is_ metabolized_by* some **cytochrome p450 2d6**
			morphine *is_ metabolized_by* some **cytochrome p450 3a4**

classes; *object_properties*; Individuals; *attributes*

Annex 8 (continuation 6)

CQs, answers and axioms for the DDI between *naloxone* and *morphine*

	Question	Answer	Axioms
26	Is ***morphine*** transported by P-glycoprotein?	*Yes, it is transported by P-glycoprotein (which synonym is multidrug resistance protein 1)*	**morphine** *is_transported_by* some **multidrug resistance protein 1**
27	Has ***morphine*** got any effect (inhibitor or inductor) of any of the cytochrome P450 isoenzymes?	*No, morphine does not induce or inhibit any CYP 450 isoenzyme.*	
28	Has ***morphine*** got a narrow therapeutic index?	*This information is not represented in the ontology.*	
29	How frequently the DDI has been described?	*This is a well-known DDI.*	Naloxone/Morphine DDI *has_incidence high*
30	What type of study describes the DDI?	*The DDI is described in a pharmacological information resource.*	Naloxone/Morphine DDI *is_described_in* PharmacologicalLiterature
31	What is the certainty of the DDI?	*The DDI is described in different sources.*	Naloxone/Morphine DDI *has_documentation_level established*
32	Has been the DDI established in well performed and controlled studies?	*This information is not included in the description of the DDI in the text source.*	
33	What is the severity of the DDI?	*Since the interaction between these drugs in beneficial, this question is not relevant in this example.*	

classes; *object_properties*; Individuals; *attributes*

Annex 9

CQs, answers and axioms for the DDI between *propafenone* and *mirtazapine*

	Question	Answer	Axioms
1	Is there an interaction between ***propafenone*** and ***mirtazapine***?	*The interaction between propafenone and mirtazapine is not explicitly described in the ontology, since this is a new interaction not described in the original information source (the database DrugBank). However, the information can be represented in the ontology.*[61]	**propafenone** *may_interact_with* some **mirtazapine**
			mirtazapine *may_interact_with* some **propafenone**
			propafenone/mirtazapine DDI equivalent to (*has_participant* some **mirtazapine**) and (*has_participant* some **propafenone**)
			Propafenone *may_interact_with* Mirtazapine
			Mirtazapine *may_interact_with* Propafenone
			Propafenone/Mirtazapine DDI *has_participant* Mirtazapine
			Propafenone/Mirtazapine DDI *has_participant* Propafenone
2	Is the effect of ***mirtazapine*** modified by ***propafenone***?	*The effect of mirtazapine represented in this example (bradycardia) is increased by propafenone.*	Mirtazapine *has_effect* Seizure
			Propafenone *increases* Seizure
3	What is the mechanism of the interaction between ***mirtazapine*** and ***propafenone***?	*It is a pharmacokinetic mechanism, specifically the inhibition of the enzyme Cytochrome P450 A 2D6*	Propafenone/Mirtazapine DDI *is_preceded_by* CytochromeP4502D6Inhibition
			CytochromeP4502D6Inhibition *is_a* **enzyme activity inhibition**
4	Is a PK parameter of ***mirtazapine*** modified by ***propafenone***?	*There is not a direct relationship between propafenone and a PK parameter in the ontology.*	

classes; *object_properties*; Individuals; *attributes*

[61] We manually created the class 'propafenone /mirtazapine DDI'.

Annex 9 (continuation 2)

CQs, answers and axioms for the DDI between *propafenone* and *mirtazapine*

	Question	Answer	Axioms
5	What is the type of the interaction between **mirtazapine** and **propafenone**?	The DDI is classified in two different ways in the ontology. Firstly, it is classified as an enzyme inhibition DDI. Therefore, it is a PK DDI. Secondly, it is classified in basis of the consequence for the patient as a potentially harmful DDI.	Propafenone/Mirtazapine DDI *is_a* **enzyme inhibition DDI**
			Propafenone/Mirtazapine DDI *is_a* **potentially harmful DDI**
6	What is the type of PD interaction between **mirtazapine** and **propafenone**?	There is not a PD DDI between these two drugs.	
7	Is there a PK interaction between **mirtazapine** and **propafenone**?	Yes, the interaction between these two drugs is a PK DDI.	Propafenone/Mirtazapine DDI *is_a* **enzyme inhibition DDI**
			enzyme inhibition DDI *is_a* **enzyme alteration DDI**
			enzyme alteration DDI *is_a* **pharmokinetic DDI**
8	Is the toxicity of **mirtazapine** exacerbated by **propafenone**?	Yes, the effect of this DDI is an increase in the toxic effect of mirtazapine.	Propafenone/Mirtazapine DDI *has_DDI_effect* Bradycardia
			Bradycardia *is_effect_of* Mirtazapine
			Bradycardia *is_a* **increased toxic effect**
			increased toxic effect *is_a* **altered toxic effect**
9	Is the effect of the DDI an adverse effect of **mirtazapine** or **propafenone**?	Yes, the effect of this DDI is an increase in the adverse effect of mirtazapine.	Propafenone/Mirtazapine DDI *has_DDI_effect* Bradycardia
			Bradycardia *is_effect_of* Mirtazapine
			Bradycardia *is_a* **adverse effect**
			Bradycardia *is_increased_by* Propafenone

classes; *object_properties*; Individuals; *attributes*

Annex 9 (continuation 3)

CQs, answers and axioms for the DDI between *propafenone* and *mirtazapine*

	Question	Answer	Axioms
10	Is the effect ***bradycardia*** produced by an interaction between ***mirtazapine*** and ***propafenone***?	*Yes, the effect of the interaction between mirtazapine and propafenone is bradycardia.*	Propafenone/Mirtazapine DDI *has_DDI_effect* Bradycardia
11	What is the effect of the interaction between ***mirtazapine*** and ***propafenone***?	*The effect of the interaction between mirtazapine and propafenone is bradycardia, the consequence of the increase in the toxic effect of mirtazapine.*	Propafenone/Mirtazapine DDI *has_DDI_effect* Bradycardia Bradycardia *is_effect_of* Mirtazapine Bradycardia *is_a* **increased toxic effect** **increased toxic effect** *is_a* **altered toxic effect**
12	Can be ***mirtazapine*** and ***propafenone*** used concomitantly?	*Information related to recommendation is not included in the description of the DDI in the text source.*	
13	Should be discontinued the treatment with ***mirtazapine*** when ***propafenone*** is administered?	*Information related to recommendation is not included in the description of the DDI in the text source.*	
14	Should be modified the dosage of ***mirtazapine*** when ***propafenone*** is administered?	*Information related to recommendation is not included in the description of the DDI in the text source.*	

classes; *object_properties*; Individuals; *attributes*

Annex 9 (continuation 4)

CQs, answers and axioms for the DDI between *propafenone* and *mirtazapine*

	Question	Answer	Axioms
15	How could be avoided the effect of the interaction between ***mirtazapine*** and ***propafenone***?	*Information related to recommendation is not included in the description of the DDI in the text source.*	
16	Is there any minimal elapse of time to administer ***mirtazapine*** after stopping the administration of ***propafenone***?	*Information related to recommendation is not included in the description of the DDI in the text source.*	
17	Is there any recommendation in order to avoid the DDI between ***mirtazapine*** and ***propafenone***?	*Information related to recommendation is not included in the description of the DDI in the text source.*	
18	What patient's characteristics affecting the DDI are present?	*The patient is a 69 years old Caucasian male.*	Patient 1 *has_age 69* Patient 1 *has_race_or_ethnic caucasian* Patient 1 *has_gender male*
19	What is the age of the patient?	*The patient is 69 years old.*	Patient 1 *has_age 69*
20	What is the sex of the patient?	*The patient is male.*	Patient 1 *has_gender male*
21	What is the race or ethnic?	Patient is Caucasian.	Patient 1 *has_race_or_ethnic caucasian*
22	What administration form of ***mirtazapine*** interacts with ***propafenone***?	*Mirtazapine is administered as tablets*.*	Mirtazapine *has_pharmaceutica_form tablet*

classes; *object_properties*; Individuals; *attributes*

Annex 9 (continuation 5)

CQs, answers and axioms for the DDI between *propafenone* and *mirtazapine*

	Question	Answer	Axioms
23	Is there any alternative administration form of ***mirtazapine*** that does not produce the interaction with ***propafenone***?		
24	What proteins ***mirtazapine*** interacts with?	*This information is explicitly represented in the ontology at the class level. Mirtazapine interacts with several proteins: 5-hydroxytryptamine 2a, 5-hydroxytryptamine 3 receptor, alpha-2a adrenergic receptor, cytochrome p450 1a2, cytochrome p450 2c8, cytochrome p450 2c9, cytochrome p450 2d6, cytochrome p450 3a4, kappa-type opioid receptor, 5-hydroxytryptamine 2c receptor and histamine h1 receptor.*	**mirtazapine…** *has_pharmacological_target* some **5-hydroxytryptamine 2a** *has_pharmacological_target* some **alpha-2a adrenergic receptor** *activates* some **kappa-type opioid receptor** *blocks* some **5-hydroxytryptamine 2a receptor** *blocks* some **5-hydroxytryptamine 2c receptor** *blocks* some **5-hydroxytryptamine 3 receptor** *induces* some **cytochrome p450 2d6** *inhibits* some **cytochrome p450 2d6** *inhibits* some **cytochrome p450 3a4** *is_metabolized_by* some **cytochrome p450 1a2** *etc.*
25	Is ***mirtazapine*** metobolized by any of the cytochrome P450 isoenzymes? Which ones?	*Yes, it is metabolized by several CYP450 isoenzymes: cytochrome p450 1a2, cytochrome p450 2c8, cytochrome p450 2c9 and cytochrome p450 3a4.*	**mirtazapine** *is_metabolized_by* some **cytochrome p450 1a2** **mirtazapine** *is_metabolized_by* some **cytochrome p450 2c8** **mirtazapine** *is_metabolized_by* some **cytochrome p450 2c9** **mirtazapine** *is_metabolized_by* some **cytochrome p450 2d6** **mirtazapine** *is_metabolized_by* some **cytochrome p450 3a4**

classes; *object_properties*; Individuals; *attributes*

Annex 9 (continuation 6)

CQs, answers and axioms for the DDI between *propafenone* and *mirtazapine*

	Question	Answer	Axioms
26	Is ***mirtazapine*** transported by P-glycoprotein?	*No, it is not.*	
27	Has ***mirtazapine*** got any effect (inhibitor or inductor) of any of the cytochrome P450 isoenzymes?	*Yes, mirtazapine inhibits cytochrome p450 isoenzymes 3a4 and 2d6 and induces 2d6.*	**mirtazapine** *induces* some **cytochrome p450 2d6**
			mirtazapine *inhibits* some **cytochrome p450 2d6**
			mirtazapine *inhibits* some **cytochrome p450 3a4**
28	Has ***mirtazapine*** got a narrow therapeutic index?	*This information is not represented in the ontology.*	
29	How frequently the DDI has been described?	*This is the first description of the DDI.*	Propafenone/Mirtazapine DDI *has_incidence rare*
30	What type of study describes the DDI?	*The DDI is described in case report in a unique patient.*	Propafenone/Mirtazapine DDI *is_described_in* PMID:24791374
			PMID:24791374 *is_a* Individual Human Data
			PMID:24791374 *has_subject_ number 1*
31	What is the certainty of the DDI?	*The DDI has only been observed in a case report.*	Propafenone/Mirtazapine DDI *has_documentation_level suspected*
32	Has been the DDI established in well performed and controlled studies?	*The DDI has only been observed in a case report.*	Propafenone/Mirtazapine DDI *is_described_in* PMID:24791374
			PMID:24791374 *is_a* Individual Human Data
33	What is the severity of the DDI?	*It is a major and clinically relevant DDI.*	Propafenone/Mirtazapine DDI *has_severity major*
			Propafenone/Mirtazapine DDI *has_relevance clinical relevance*

classes; *object_properties*; Individuals; *attributes*

Annex 10

Evaluation template for DINTO

Version: 1.0.0
Evaluation date: 23/06/2014

OBO FOUNDRY PRINCIPLES		
Open		✓
Format		OWL
URIs	Every class and relation have a unique identifier	✓
	The URI is constructed from a base URI, a unique prefix and a numerical identifier	✓[1]
Versioning	The ontology provider has procedures for identifying distinct successive versions.	✓
Delineated content	Coherent natural language definitions of top-level term(s)	✓
	Cross-product links to other OBO Foundry ontologies	✓
Textual Definitions	Textual definitions (SOP) for a substantial and representative fraction of terms	✓
	Equivalent formal definitions for at least a substantial number of terms	✓
	There is evidence of implementation of a strategy to provide definitions for all remaining undefined terms	✓
Relations	The ontology uses relations which are unambiguously defined following the pattern of definitions laid down in the OBO Relation Ontology.	✓
Documented	The ontology is well-documented (e.g., in a published paper describing the ontology or in manuals for developers and users)	✓
Plurality of users	The ontology has a plurality of independent users	✓
Collaboration	The ontology is developed collaboratively with other OBO Foundry members.	✓
Locus of authority	There is a single person who is responsible for the ontology, ensuring its maintenance and prompt response to users feedback.	✓
Naming conventions	Explicit and concise names	✓
	Context independent names	✓
	Avoiding taboo words	✓
	Avoiding encoding administrative metadata in names	✓
	Unequivocal names, avoiding homonyms	✓[7]
	Avoiding conjunctions	✓

Annex 10 (continuation 2)

Evaluation template for DINTO

Naming conventions (continuation)	Singular nominal form	✓[2]
	Using positive names	×[1]
	Avoiding catch-all names	✓
	Recycling strings	✓
	Using genus-differentia style names	✓
	Using space as word separators	✓[4]
	Expanding abbreviations and acronyms	✓[5]
	Expanding special symbols to words	✓
	Lower case beginnings	✓[6]
	Avoiding character formatting	✓
	Naming conventions for relations is consistent along the ontology	✓
Maintenance		✓

CIMINO'S DESIDERATA FOR CONTROLLED MEDICAL VOCABULARIES		
Content	The content of the ontology increased over previous versions	-[8]
	The ontology allows compositional extensibility	✓
	A methodology for further expanding content has been established	✓
Concept orientation	Each concept in the ontology has a single, coherent meaning ('nonvagueness', 'nonambiguity', 'nonredundacy')	×[7]
Concept permanence	None meaning of concepts have been deleted	-[8]
	Number of preferred terms that have been changed	-[8]
Non-semantic Concept Identifier	Concepts have a unique non-hierarchical numerical identifier	✓[1]
	Concepts have a preferred term	✓
	Different names are included as synonyms	✓
Polyhierarchy	The controlled vocabulary is s strict hierarchy or a polyhierarchy	P
Formal definitions	Concepts have formal definitions	✓
"Not Elsewhere Classified"	The controlled vocabulary rejects the use of "not elsewhere classified" terms.	✓[3]

Annex 10 (continuation 3)

Evaluation template for DINTO

CIMINO'S DESIDERATA FOR CONTROLLED MEDICAL VOCABULARIES		
Multiple granularities	Concepts are described at multiple levels of granularity	✓
Multiple consistent views	Multiple views of the vocabulary, suitable for different purposes, are provided	✓
Representing concepts	The controlled vocabulary contains context representation through formal information	✓
	The controlled vocabulary contains context representation through natural language information	✓
Evolving Gracefully	Clear and detailed descriptions of what changes occur and why are documented	-[8]
Recognize redundancy	There is a mechanism enabling redundancy recognition	YES

GÓMEZ-PÉREZ EVALUATION			
Analysis of inconsistences	The ontology is consistent[10]	using FaCT++	12897219 ms
		using HermiT 1.3.8.	230723 ms
		using Pellet	-
	There is not circularity errors at distance zero		✓
	There is not circularity errors at distance 1		✓
	There is not circularity errors at distance *n*		✓
	There is not partition error at the class level		✓
	There is not partition error at the instance level		✓
	There is not semantic inconsistency		✓
Analysis of incompleteness	There is not concepts existing in the domain that have been overlooked in the classification		✓
	Disjoint classes have been identified and represented		✓[9]
	There is not omission of the completeness constraint between subclasses and the base class.		✓
Analysis of redundancies	There are not redundancies of *Subclass-Of* relations		✓
	There are not redundancies of *Instance-Of* relations		-
	There are not identical formal definitions of classes		✓
	There are not identical formal definitions of classes		✓

Annex 10 (continuation 4)

Evaluation template for DINTO

1. Numerical identifiers.
All those entities created in DINTO have an URI constructed from a base URI (http://purl.obolibrary.org/), a unique prefix (DINTO_) and a numerical identifier.
However, imported classes from the PKO have URIs that do not follow this structure. They have a base URI (www.owl-ontologies.com/2009/11/5/), a unique prefix (PKO.owl#) and a non-numerical identifier (e.g., AUC). In the same way, those classes imported from BRO have URIs with non-numerical identifiers. They have a base URI (www.bioontology.org/ontologies/), a unique prefix (BiomedicalResourceOntology.owl/) and a non-numerical identifier (e.g., Structured_Knowledge_Resource).
We decided to maintain these URIs in order to preserve the original source.
Another exception to this recommendation is annotation properties. Following the example of annotation properties used in other resources, such as the ChEBI ontology (e.g., http://purl.obolibrary.org/obo#Definition) new annotation properties in DINTO have not numerical identifiers. In this way, the annotation property 'ATCcode' URI is 'http://purl.obolibrary.org/obo#ATCcode'.

2. Naming conventions: Singular nominal form:
There are two classes imported from the PKO that used plural forms: 'PK_Parameters_invivo' and 'PK_Parameters_invitro'.
We decided to maintain the plural nominal form in order to preserve the original source.

3. Naming conventions: Using positive names:
There are exceptions to this naming convention in DINTO, since we have created the classes:

- 'non-clinically relevant DDI'
- 'non-clinically relevant DDI effect'
- 'uncertain DDI'
- 'uncertain DDI effect'
- 'unobservable DDI effect'

Although one of the OBO Foundry recommendations is to avoid use of negation, which could decrease precision in the interpreted meaning, we decided to maintain these names. The main reason is that these terms are commonly used in DDIs texts, as we observed during the study of the DDI corpus and other pharmacological resources. To ensure a correct interpretation of these concepts all of them have natural language definitions in the ontology.

4. Naming conventions: Using space as word separators:
We have used space as word separators for all those new entities created in DINTO. However, classes imported from the PKO and BRO use symbol '_' as word separator. We decided to maintain it in order to preserve the original source.

5. Naming conventions: Expanding abbreviations and acronyms.
We have expanded abbreviations and acronyms in DINTO with the exception of the term drug-drug interaction (DDI) since it is a widely known acronym. Moreover, it is frequently used in the ontology. Therefore, we believed that the use of the acronym would increase readability.
Another exception is the use of the acronym PK and PD in the PKO's classes 'PK_Parameters_invivo' and 'PK_Parameters_invitro'. In this case, we decided to maintain the acronyms in order to preserve the original source.

Annex 10 (continuation 5)

Evaluation template for DINTO

It is important to note that pharmacokinetic parameters imported from the PKO are named using abbreviations, as well (e.g., 'AUC' instead of 'area under the curve'). We believe that this naming procedure is correct, since the abbreviated form is the most commonly used for these concepts and they are unequivocal representations of the corresponding concept in the pharmacological domain.

6. Naming conventions: Lower case beginnings.
We have used lower case beginnings for all new classes created in DINTO. However, classes imported from the PKO and BRO use upper case beginnings. We decided to maintain them in order to preserve the original sources.

7. Unequivocal names, avoiding homonyms.
We identified that some classes of 'pharmacological entity' class shared the same label. The reason is that there was an unexpected problem during the automatic mapping between our two information resources (ChEBI ontology and DrugBank database). We manually reviewed all classes and found 66 duplicate classes. We merged them and eliminated the incorrect ones (see **Section 7.1.6**).

8. Previous versions of the ontology.
This is the first version of the ontology. Therefore, it does not proceed to describe changes related to previous versions.

9. Disjoint classes:
We identified disjoint classes in DINTO. However, it should be taken into account that many times it is not possible to establish disjointness in this domain. For example, we have explained in this thesis that DDIs are commonly classified in two main groups as pharmacodynamic (PD) and pharmacokinetic (PK) DDIs in basis of their mechanisms. However, different mechanisms can contribute to the apparition of a DDI, leading to the possibility that a PK DDI could be classified, at the same time, as a PD DDI. Another example affects to roles; the same pharmacological entity can have several roles, including both 'precipitant' and 'object'. Therefore, in some cases, we have not establishes disjoint classes.

10. Analysis of consistency using a reasoner.
DINTO is a large ontology with more than 25 thousand classes and 377 thousand axioms. Using a reasoner in such a large ontology leads to issues (see **Section 9.2**). Therefore, to check the consistency of the ontology we had to apply different strategies in order to reduce its size. This procedure has been proposed by other authors, such as Brank et al. (2005), who explains: *«An ontology is a fairly complex structure and it is often more practical to focus on the evaluation of different levels of the ontology separately rather than trying to directly evaluate the ontology as a whole»*, or Holford et al. (2010), who had to reduce the complexity of its ontology in order to classify it using a reasoner.
The first strategy is to reduce the number of classes and relationships between them. To do this we delete most pharmacological substances and proteins, obtaining a much shorter ontology,[62] making possible to check the consistency of the whole conceptual model.

[62] This version is named DINTO 0.0.6 and is available under request to the authors.

Annex 11

Inference results in *IExp3*

DDI		DINTO														DrugBank			
		TR	Ag	Ant	ER	ES	EId	EIn	Tn	TS	TId	TIn	CR	CS	CId	CIn	Tg	E	T	N
1	abciximab tirofiban		x														x			
2	acenocoumarol fluoxetine					x		x					x					x		
3	alfacalcidol calcipotriol			x													x			
4	aluminium ciprofloxacin																			x
5	alvimopan dezocine		x	x													x			
6	alvimopan morphine			x													x			
7	amiodarone cisapride	x				x		x									x	x		
8	amiodarone phenytoin					x	x	x			x	x						x	x	
9	amprenavir diazepam					x		x			x	x						x	x	
10	atorvastatin eltrombopag					x		x				x						x	x	
11	attapulgite doxycycline																			x
12	bepridil astemizole					x												x	x	
13	bismuth subsalicylate minocycline																			x
14	bromazepam itraconazole					x		x										x		
15	calcium enoxacin																			x
16	calcium ibandronate																			x
17	cerivastatin rifampicin					x	x	x		x	x	x						x	x	
18	chlorpropamide acetylsalicylic acid					x	x	x				x						x	x	
19	cholecalciferol alfacalcidol		x										x				x			
20	colchicine clotrimazole						x	x			x	x						x	x	
21	dabigatran etexilate verapamil									x	x	x							x	
22	dabrafenib midazolam					x	x	x		x	x	x						x	x	
23	desipramine ziprasidone		x			x		x									x	x		
24	devenlafaxine fluoxetine	x				x		x									x	x		
25	dextromethorphan desvenlafaxine	x				x											x	x		
26	diazoxide bendroflumethiazide	x															x			

Annex 11 (continuation 2)

Inference results in *IExp3*

DDI		DINTO														DrugBank			
		TR	Ag	Ant	ER	ES	EId	EIn	Tn	TS	TId	TIn	CR	CS	CId	CIn	Tg	E	T	N
27	diclofenac cyclosporin a					x		x				x						x	x	
28	digoxin quinidine					x	x	x		x	x	x						x	x	
29	dihydroergotamine clarithromycin					x	x	x				x						x	x	
30	dofetilide clarithromycin					x	x	x										x		
31	dolasetron artemether					x	x											x		
32	dolasetron ziprasidone		x			x		x									x	x		
33	donepezil chlorpomazine	x				x		x									x	x		
34	doxepin cimetidine		x					x				x					x	x	x	
35	doxepin phenylephrine	x					x										x	x		
36	eltrombopag cimetidine							x										x		
37	ergotamine acebutolol											x							x	
38	ergotamine carvedilol		x			x		x									x	x	x	
39	ergotamine propranolol	x				x		x				x					x	x	x	
40	erythromycin vinorelbine					x		x				x						x	x	
41	estropipate raloxifeno		x			x		x									x	x		
42	fesoterodine ketoconazole					x		x		x		x						x	x	
43	fluconazole phenytoin							x				x						x	x	
44	fosamprenavir cyclosporin a					x		x										x		
45	gemfibrozil rosuvastatin					x	x	x				x						x	x	
46	glucosamine glyburide					x												x		
47	grepafloxacin chlorpromazine					x				x		x						x	x	
48	grepafloxacin iron dextran																			x
49	halofantrine trimipramine					x		x										x		
50	ibuprofen carvedilol					x		x				x						x	x	
51	imipramine fluvoxamine	x				x		x				x					x	x	x	
52	imipramine ziprasidone		x			x		x									x	x		
53	indinavir dantrolene					x		x										x		

Annex 11 (continuation 3)

Inference results in *IExp3*

DDI		DINTO														DrugBank			
		TR	Ag	Ant	ER	ES	EId	EIn	Tn	TS	TId	TIn	CR	CS	CId	CIn	Tg	E	T	N
54	iron demeclocycline																			x
55	isoproterenol acebutolol	x															x			
56	itraconazole trazodone					x		x										x	x	
57	ketoprofen acetylsalicylic acid	x				x		x				x					x	x	x	
58	labetalol fenoterol			x													x			
59	lopinavir darunavir	x				x		x									x	x		
60	magnesium temafloxacin																			x
61	maprotiline ziprasidone		x			x		x									x	x		
62	mefloquine halofantrine		x			x		x									x	x		
63	midazolam clobazam	x				x		x		x	x	x					x	x	x	
64	midrodine dexamethasone																	x		
65	mirabegron desipramine					x		x												
66	moclobemide selegiline	x				x		x									x	x		
67	nelfinavir cyclosporin a					x	x	x		x	x	x						x	x	
68	nortriptyline cisapride			x		x		x									x	x		
69	ondansetron asenapine	x				x		x									x	x		
70	paliperidone amantadine			x													x			
71	pancuronium quinine									x		x							x	
72	pimozide mesoridazine		x					x									x	x		
73	posaconazole vinorelbine					x		x										x	x	
74	probenecid cefixime																		x	
75	promethazine benzphetamine	x				x											x	x		
76	quetiapine donepezil	x				x		x									x	x		
77	regorafenib irinotecan					x		x				x						x	x	
78	rifabutin bromazepam					x	x											x		
79	saquinavir diazepam					x		x		x	x	x						x	x	
80	sertraline phenytoin					x	x	x				x						x	x	

Annex 11 (continuation 4)

Inference results in *IExp3*

DDI		DINTO														DrugBank			
		TR	Ag	Ant	ER	ES	EId	EIn	Tn	TS	TId	TIn	CR	CS	CId	CIn	Tg	E	T	N
81	sucralfate etidronic acid																			x
82	sulindac bumetanide	x																	x	
83	telaprevir ketoconazole					x		x		x		x						x	x	
84	timolol chlorpropamide					x						x						x	x	
85	toremifene chlorpromazine					x				x		x						x	x	
86	trazodone nefazodone		x	x		x		x									x	x	x	
87	trimipramine venlafaxine	x				x		x				x					x	x	x	
88	trospium dicyclomine		x														x			
89	valproic acid acetylsalicylic acid					x	x	x		x								x	x	
90	verapamile carbamezepine	x				x	x	x		x	x	x					x	x	x	
91	vincristine ketoconzale					x		x		x	x	x						x	x	
92	vismodegib ivacaftor					x		x				x		x				x	x	
93	zinc norfloxacin																			x

TR: Target-related mechanism
Ag: Agonistic mechanism
Ant: Antagonistic mechanism

ER: Enzyme-related mechanism
ES: Enzymatic saturation mechanism
EId: Enzyme induction mechanism
EIn: Enzyme inhibition mechanism

Tn: Transporter-related mechanism
TS: Transporter saturation mechanism
TId: Transporter induction mechanism
TIn: Transporter inhibition mechanism

CR: Carrier-related mechanism
CS: Carrier saturation mechanism
CId: Carrier induction mechanism
CIn: Carrier inhibition mechanism

Tg: Target-related mechanism in DrugBank
E: Enzyme-related mechanism in DrugBank
T: Transporter-related mechanism in DrugBank
N: Non-absorbable complex formation mechanism in DrugBank