FP7-ICT-2007-1
http://www.eu-adr-project.com

D4.4
Report on literature and DB mining

WP4 – Signal substantiation

V1.5
Final

Lead beneficiary: UPF
Date: 30/11/2010
Nature: R/P
Dissemination level: RE
(Restricted to members of the Consortium and Commission Services)

TABLE OF CONTENTS

Document History

Name	Date	Version	Description
Laura I. Furlong, Jan Kors	06.07.2010	1.0	First draft
Erik van Mulligen	08.07.2010	1.1	Description of annotation tool
Paul Avillach	09.07.2010	1.2	MeSH based approach included
Laura I. Furlong	10.07.2010	1.3	Version for internal review
Laura I. Furlong, Jan Kors, Paul Avillach	20.11.2010	1.4	Incorporation of internal reviewers comments
Laura I. Furlong	30.11.2010	1.5	Final version

Definitions

- Partners of the EU-ADR Consortium are referred to herein according to the following codes:

  EMC - Erasmus University Medical Center (Netherlands) – Coordinator
  FIMIM - Fundació IMIM (Spain) – Beneficiary
  UPF - Universitat Pompeu Fabra (Spain) – Beneficiary
  UAVR - University of Aveiro – IEETA (Portugal) – Beneficiary
  NEUROLESI - IRCCS Centro Neurolesi "Bonino-Pulejo" (Italy) – Beneficiary
  UB2 - Université Victor-Segalen Bordeaux II (France) – Beneficiary
  LSHTM - London School of Hygiene & Tropical Medicine (UK) – Beneficiary
  AUH-AS - Aarhus University Hospital, Århus Sygehus (Denmark) – Beneficiary
  AZ - AstraZeneca R&D (Sweden) – Beneficiary
  UNOTT - University of Nottingham (UK) – Beneficiary
  UNIMIB - Università di Milano-Bicocca (Italy) – Beneficiary
  ARS - Agenzia Regionale di Sanità (Italy) – Beneficiary
  PHARMO - PHARMO Coöperation UA (Netherlands) – Beneficiary
  PEDIANET -Societa' Servizi Telematici SRL (Italy) – Beneficiary
  USC - University of Santiago de Compostela (Spain) – Beneficiary
  TAU - Tel-Aviv University (Israel) – Subcontractor
  SIMG - Health Search - Italian College of General Practitioners (Italy) – Subcontractor
  ICL - Imperial College London (UK) – Subcontractor

- **Grant Agreement**: The agreement signed between the beneficiaries and the European Commission for the undertaking of the EU-ADR project (ICT-215847).

- **Project**: The sum of all activities carried out in the framework of the Grant Agreement by the Consortium.
- **Work plan**: Schedule of tasks, deliverables, efforts, dates and responsibilities corresponding to the work to be carried out for the EU-ADR project, as specified in Annex I to the Grant Agreement.
- **Consortium**: The EU-ADR Consortium, conformed by the above-mentioned legal entities.

1. INTRODUCTION

The overall objective of the EU-ADR project is the design, development, and validation of a computerized system that exploits data from electronic healthcare records (EHRs) and biomedical databases for the early detection of adverse drug reactions. To achieve this objective, EU-ADR will exploit clinical data from EHRs of over 30 million patients from several European countries in order to detect 'signals' (combinations of drugs and suspected adverse events that warrant further investigation).
One of the major research issues in EU-ADR is to discriminate between signals that indeed point to an adverse drug reaction and spurious signals. These latter may create unrest and uncertainty in both patients and physicians and may even result in the withdrawal of a useful drug from the market. Also from a commercial and regulatory perspective the cost of a false-positive signal is significant. To discriminate between true signals and spurious signals, previous reporting of the signal in specialized databases and in the biomedical literature will be assessed. This process is referred to as signal filtering. Another important goal of the project is, once a signal is detected, to provide a possible biological explanation for each signal. This process is referred to as signal substantiation, and requires that the signal is placed in the context of current knowledge of biological mechanisms that might explain it.
EU-ADR exploits the currently available databases and other electronic sources that contain information about drug-event associations and biological mechanisms underlying ADR to their best and extends that understanding by means of *in silico* models and simulations of the behaviour of the drugs and the biological systems. In particular, information available from the biomedical literature needs to be identified and extracted for both processes of signal filtering and substantiation. In Work Package 4, task 4.1 is concerned with developing and implementing tools to mine pharmacological databases and scientific literature for relationships between drugs, adverse events, and biological targets (genes, proteins, and their genetic variants). In particular, the Text Mining (TM) activities within WP4 are aimed at developing tools for the extraction of relationships between the entities drug, target and disease, to aid in the filtering and substantiation of the signal (Figure 1).

All the identified signals from WP3 are first checked for being reported before by an analysis of relevant databases and mining of biomedical literature (signal filtering). Then, they are processed by a computational framework that exploits existing biomedical knowledge using a series of bioinformatics approaches, which includes text mining for identifying pair-wise relationships between the entities of interest for the project, *in silico* prediction of drug targets and analysis of biological pathways. Drug metabolism and pharmacogenomic information are also incorporated into the signal substantiation computational framework (Figure 1).

EU-ADR events

The EU-ADR project uses an event-based approach where a limited set of specific clinical events are inspected for their association with all available drugs in the EHR databases

participating in the project. One of the challenges in the event-based approach for signal detection through data mining on EHR databases is the identification of events that are most important in pharmacovigilance and thus warrant priority for monitoring. Based on overall scores, a ranked event list was generated [1]. The top-ranked events were considered as having the highest priority for drug safety monitoring**.** This ranked list comprised 23 adverse events. The top-ranking events were: Cutaneous bullous eruption (BE), Acute renal failure (ARF), Acute Myocardial infarction (AMI), Anaphylactic shock (AS), and Rhabdomyolysis (RHABD). Because of its complexity, an additional event, Upper gastrointestinal bleeding (UGIB), was included to evaluate the method. Each event was defined by medical specialists, and this definition was used to guide the search of UMLS concepts that represent each event. Earlier in the EU-ADR project, a shared semantic foundation was built based on the UMLS concepts grouping together terms from different terminologies with the same medical meaning [2]. The aim was to provide researchers with a standardized list of medical concepts' CUIs (Concept Unique Identifier from the UMLS) and associated terms to be used for: 1) identifying the events investigated in their respective EHR databases for events and drug retrieval, and 2) providing the same definition of the event to filter and substantiate the generated signal. This list of events and their corresponding definitions in terms of UMLS CUIs are used within WP4 activities.

The goal of task 4.1 is to develop TM systems for the identification and extraction of relationships between the entities shown in Figure 1 from textual resources. The ultimate aim of the TM systems is to provide information for the signal filtering and signal substantiation processes.

For simplicity, in the next sections the following terminology is used to refer to the entities of interest for the project:

- Target: gene, protein and sequence variants of genes and proteins
- Disease: disease phenotypes of the adverse drug reactions (events).
- Drug: biologically active chemicals, marketed drugs, drug metabolites

Accordingly, the relationships to be extracted from textual resources are:

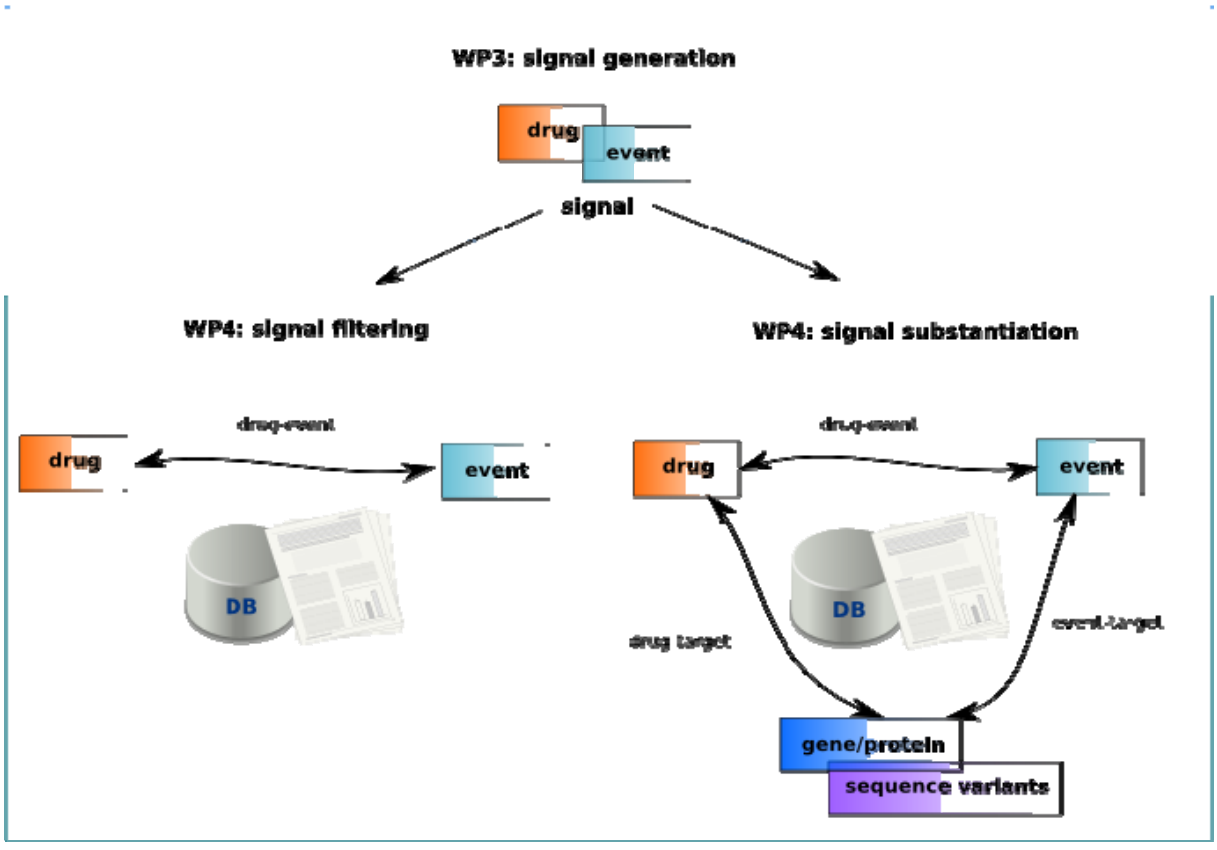
- Target-disease: for example, if the target plays a role in the mechanism underlying the disease, or is a marker of the disease
- Target-drug: binding relationships mainly, but also if drug affects gene expression or modifies in some way the gene or protein function
- Drug-disease: if the drug is associated with the disease (e.g. if a drug produces an adverse effect)

To accomplish these tasks, resources already available from the partners and from the BioNLP community were first evaluated to assess their suitability for the project needs. New tools and resources were developed to accomplish task not covered by available resources. In order to develop the TM system, examples of the relationships encoded in the text are needed for training and evaluation of the system. These examples can be found in a corpus annotated

by domain experts. A significant proportion of the task 4.1 effort was devoted to the development of such a corpus.

The deliverable 4.4 is a report of the work conducted in task 4.1 focused in the development of tools for mining of biomedical literature. Progress on mining of databases has been already reported on D4.1 (drug-event repositories) and 4.3 (drug-target repositories). Work on the integration of text-mining derived information with information derived from other bioinformatic approaches developed in WP4 will be reported in D4.5.

Figure 1: Signals generated by WP3 are processed by WP4 substantiation pipeline in order to find plausible biological explanations for the signals. In particular, text mining activities within WP4 are aimed at extracting information from biomedical literature supporting the different types of relationships that might explain the signal. The WP4 signal filtering process involves the assessment of the association of a drug and an event in the literature and in databases.



2. SYSTEM IMPLEMENTATION

2.1. Systems for Named Entity Recognition (NER)

The first step for the identification and extraction of relevant information from text is the correct identification of entities. In addition to the identification of a span of text as referring to a certain entity (e.g. the term "elastin" refers to an entity of the semantic type "gene"), it is necessary to disambiguate the identified entity to a standard database entry. This process is known as disambiguation, grounding or normalization. This is especially important in order to assign an identity to the information extracted from text and to integrate it with the information extracted, for instance, from the clinical records. In this way, by assigning proper database identifiers, for instance UMLS CUI concepts for disease terms and ATC codes for drugs, we will be able to extract relevant information for a disease and drug mined from the EHRs.
To accomplish named entity recognition with subsequent normalization, dictionary based approaches are common strategies. Thus, part of the effort during the project was devoted to the development and/or adaptation of different dictionaries for the identification of named entities from free text. The dictionaries and NER systems are described below. The recognition of entities used for the MeSH based approach is described in section 2.2.1.

2.1.1. Dictionaries

Drug and disease dictionaries
The dictionaries for drugs and diseases were already described in Deliverable 4.1. Briefly, the drug dictionary is based on ATC and combines multiple resources including DrugBank, UMLS, ChEBI, and ChemIDPlus, building on previous work [3]. The disease dictionary covers diseases in general and adverse events, and is based on a selection of the UMLS Metathesaurus [4]. Both dictionaries were curated by applying a number of re-write and suppress rules on the terms in the dictionary [5] and by manual curation steps. Improvements to these systems include a system to handle disease abbreviations (see section 2.1.2).

Gene/protein dictionary
We constructed a gene/protein dictionary that makes no distinction between genes and proteins for several reasons. Often the gene name is equal to the protein name, with only a difference in case. Also, in literature authors many times do not make a strict distinction between genes and proteins and use gene and protein names interchangeably [6]. We therefore combined gene and protein names in one dictionary, adding proteins as synonyms of the genes that produce them.
For EU-ADR, we limited ourselves to human genes. We combined information on human genes from five genetic databases: Entrez Gene, Swiss-Prot, Genew, GDB, and OMIM. Since these databases are only partially cross-linked, a more elaborate combination algorithm was used [7]. Briefly, for each database entry, identification codes were extracted, including the available links to other databases. Genes and proteins from the different databases with any

matching identification codes were grouped. If there were conflicting identification codes, a procedure was carried out to determine whether there were groups of genes that represent the same gene. Allowance was made for one or two differing codes if there were at least as many identical codes and a certain degree of overlap in terms for the genes to be combined.

Although the databases are curated, they still may contain terms that are less suited for gene and protein identification in text. Therefore several filtering rules were applied. Terms that contain the words "putative", "method", or "similar" were removed. Also EC numbers were removed. We attempted to remove family names, which are often included as gene synonyms. If a term was also found in the dictionary followed by a number, roman numeral, or greek letter, it was removed. For instance, "Zinc finger protein" was also detected as a substring in "Zinc finger protein 51" and was therefore removed as a synonym.

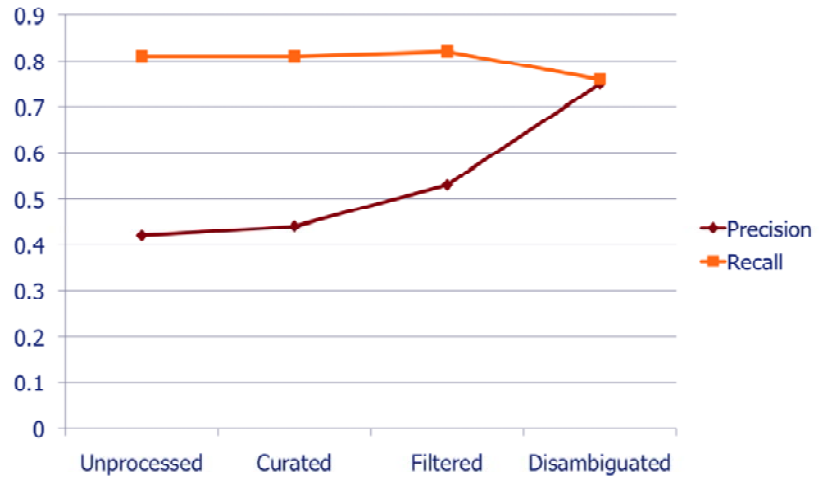
To allow for spelling variations that were not included in the dictionary, we applied two rewrite rules that we previously found to be effective [8]. One rule replaced arabic numbers with roman numerals and vice versa. The other rule inserted a word delimiter (hyphen or space) if the last part of a gene or protein consisted of numbers. For example, "ABC1" becomes "ABC-1". If a word delimiter was present, it was removed.

Finally, a rule was applied to remove highly ambiguous terms. If a term consisted only of tokens that were shorter than three characters, consisted only of numbers or roman numerals, or belonged to a set of stop-words, the term was removed. Examples of terms that were removed, are: "G 4", "2.19", and "And-1".

A manual curation step was also performed. A set of 100,000 randomly selected Medline abstracts were indexed (see section on Peregrine below), and the top 500 most frequent terms found were manually evaluated. If they did not correspond to a gene or protein (e.g., "open reading frame", "alternative splicing", "human") or were very ambiguous (e.g., "CA2", "obesity"), they were removed.

Using Peregrine, the gene/protein dictionary has previously been evaluated on the BioCreative 2 test set (Figure 2) [9]. Progressive inclusion of the different processing steps show increasing precision with only a small decrease in recall.

Figure 2: Effect of progressive processing steps on precision and recall of the gene/protein dictionary, using the Peregrine indexing tool.



Sequence variant dictionary

A dictionary covering gene sequence variations was also developed [10]. In this context the term variation is used to refer to any kind of short range change in the nucleotide sequence of the genome. SNPs are the most studied type of sequence variation, but we can also consider as member of this class short insertions or deletions, named variations as Alu sequences, and other types of variations collected in the dbSNP database [11]. These variations can be mapped to the exonic regions of genes, and produce a change at the protein level, or within introns, untranslated regions or between genes. Some variations may alter protein function, such as non synonymous SNPs, or alter other processes related with the regulation of gene expression. Others may have no functional effect at the protein level, such as synonymous SNPs. However, although synonymous SNPs do not alter the protein sequence, they are currently being regarded as potentially functional in light of reports that document their involvement in other processes such as regulation of mRNA processing or protein folding [12]. From the point of view of a NER system, a variation entity is defined by the combination of tokens that specify the location of the variation in the sequence and the original and altered alleles. It is important to note that this information can be represented as nucleotide sequence or amino acid sequence. For instance, in Figure 3 the term E298D represents the following: E is the original allele of the variation in the protein (residue E in one letter code or Glu in three letter code for Glutamic acid) while D is the altered residue (residue D or Asp or Aspartic acid) and the variation occurs at position 298 of the protein sequence. The same variation could be referred as using the nucleotide representation G894T: in this case the G original allele stands for a guanine residue in the gene sequence at position 894 of the mRNA changes to a thymine residue. In this example a first source of ambiguity in SNP notation is found. How can we know that the G894T term refers to a change in the protein or in the nucleotide sequence? Both possibilities are plausible without any additional information: contextual information extracted from the text or biological information from the sequence. Our strategy focuses in the latter source of information (biological information) as we believe it constitutes the more reliable approach to disambiguate these cases. Thus, the ambiguity is handled by means of the grounding to a sequence database entry (in this case to dbSNP). This process described in detail in [10].

The sequence variation terminology was generated from dbSNP, using a term expansion approach according to a set of patterns manually developed after inspecting the literature for natural language expressions used to refer to sequence variants. For each variant, the collection of amino acid terminology is composed by 286 terms and the nucleotide terminology by 114 terms. In this way, the dictionary is composed by a dbSNP entry identified by an rs number (the dbSNP identifier), and the synonyms are terms formed by combination of tokens representing the alleles and position, as exemplified in Figure 3. Some authors use common names to refer to sequence variations, such as “ACE ID polymorphism”, “ADH1C*2”, especially to refer to sequence variants in drug metabolising enzymes. These terms cannot be derived from sequence databases such as dbSNP. Thus, other sources of information [13] were used to obtain these special synonyms and map them to dbSNP entries (this part of the work was performed in collaboration with David Gurwitz (TAU)).

Figure 3: Examples of mentions of sequence variations in biomedical abstracts. The highlighted terms correspond to mentions of the same sequence variation in different documents.

Article PMID: 11017941
In controls, individuals with the **E298D** mutation in exon 7 (136.1 micromol/L) showed significantly higher (P = 0.001) median plasma NOx than those without this mutation (64.5 micromol/L).

Article PMID: 10205226
Allele frequencies of **298Asp** were concordant across the panels: 8.4% in hypertensive subjects, 8.
The relevance of the **Glu298Asp** polymorphism to hypertension in this population was tested in 2 ways.

Article PMID: 10231340
Ten polymorphisms were detected: three in the 5' flanking sequence at positions -1474, -924 and -788, two in coding sequences 774C --> T (silent) and **G894 --> T** (**Glu-298 --> Asp**) and five in introns 2, 11, 12, 22 and 23.

2.1.2. Named Entity Recognition (NER) systems

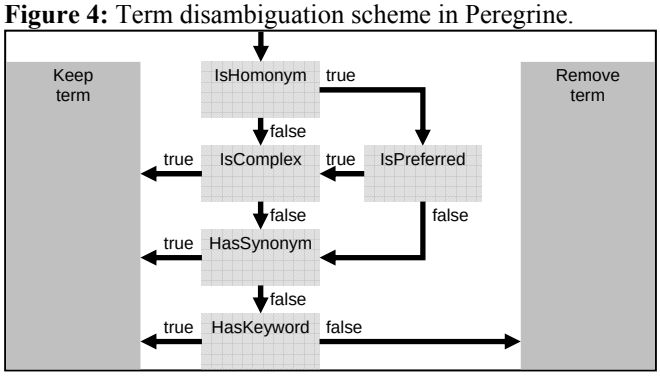
A variety of NER systems are available within the project for the recognition of different types of entities. The NER systems described below are available as standalone programs (Peregrine) and as UIMA modules (all). The UIMA modules and pipeline are described in section 2.4. In addition, the modules or the extracted information are accessible through Web services and Taverna workflows, which will be described in Deliverable 4.5

Peregrine: dictionary-based concept recognition

For the term and concept recognition identification we used our concept recognition software Peregrine [9]. The Peregrine system translates the terms in a dictionary into sequences of tokens (i.e., sequences of words). When such a sequence of tokens is found in a document, the term, and thus the concept associated with that term, is recognized in the text. Terms that are longer than five characters or contain a number are matched case insensitive, otherwise terms are matched case sensitive. Some tokens are completely ignored, since these are considered to be non-informative ("of", "the", "and", and "in"). If the term is considered a 'long form' (i.e., it contains a space and is longer than six characters), the tokens in the thesaurus and in the text are first reduced to their stem using NLM's Lexical Variant Generator program, to allow for small lexical variations. This normalisation does not apply to terms from the gene dictionary or the drug dictionary.
In its default setting, the tokenizer in Peregrine considers everything that is not a letter or a digit to be a word delimiter. To fine tune Peregrine for chemical concept recognition we made

the following adjustments: full stops, commas, plus signs, hyphens, single quotation marks, and all types of parentheses ((, {, []) were excluded from the word-delimiter list. After tokenization, the tokens were stripped of trailing full stops, commas, and non-matching parentheses. Parentheses were also removed if they surrounded the whole token. In addition, a list of common suffixes was used to remove these suffixes at the end of tokens. The suffix list was obtained by scanning the whole UMLS for suffixes that were English verbs or adjectives.
Schuemie and colleagues [8] have evaluated a number of disambiguation rules to disambiguate gene names found in text. These disambiguation rules are potentially also applicable to drugs. Disambiguation of terms found in text was carried out as follows (Figure 4).

Figure 4: Term disambiguation scheme in Peregrine.



We first determine whether a term is a dictionary homonym, i.e., if it refers to more than one term in the dictionary. If the term is not a dictionary homonym it still needs further processing since it can have many meanings in text. Specifically, in order to map a term that is a non-dictionary homonym to an entity in the dictionary it needs to fulfil at least one of the following conditions: (1) it is complex (i.e., longer than five characters or contains a number); (2) another synonym of the entity is found in the same piece of text; (3) a keyword (i.e., a word or "token" that occurs in any of the long-form names of the drug, and appears less than 1000 times in the dictionary as a whole) is found in the same piece of text. If the term is a dictionary homonym, we impose the additional condition that the term needs to correspond to the preferred name (in our case, the entry name in the ATC code list) before it can be evaluated for complexity (test 1). If the term is not the preferred name of the entity, it needs to pass test 2 or test 3 to be mapped to an entity in the dictionary.
The speed of the Peregrine system has been tested by analyzing a random set of 100,000 Medline abstracts, which took less than 4 minutes on a standard PC [9]. The whole of Medline can be indexed in less than 10 hours. Thus, the system can easily be applied to large corpora.

OSIRIS

The OSIRIS system is aimed at the identification of gene sequence variations from text [10]. OSIRIS can be used to link literature references to dbSNP database entries with high accuracy, and is suitable for collecting current knowledge on gene sequence variations and

their relationship to diseases or drug adverse reactions. The system is based on a dictionary of sequence variations and a pattern-based search algorithm for the identification of variation terms and their disambiguation to dbSNP identifiers. The main features of the approach are the following: it covers any type of short range sequence variation (SNPS, short insertions, deletions, etc) that affect not only the sequence at the protein level but also at the nucleotide level (SNPs in non-coding regions of genes like introns, in intergenic regions, etc), is a generic method applicable to all the genes, and not restricted to a gene or protein family, and maps the variation mentions to its database identifier (also known as normalization).
For the EU-ADR project, the system has been adapted in order to be used as a UIMA module and to increase nomenclature coverage of pharmacogenetic relevant sequence variations (see section Sequence variant dictionary). The new version uses a modified version of the "JULIE Lab UIMA Wrapper for Lingpipe Dictionary Chunker" to perform an exact matching of dictionary terms [14].

Abbreviations handling

Initial evaluation of the performance of the NER indicated that disease abbreviations present in the dictionary were a source of false positives. Thus, a modification in the system towards the specific handling of disease abbreviations was introduced which was implemented using the "JULIE Lab Acronym Annotator" [14] in combination with further adaptations to the Lingpipe Annotator. Although the system has been implemented for the handling of disease abbreviations, it can be easily adapted to handle abbreviations of other entities.

Participation in community based efforts

The partners also participated in community challenges in text mining contributing with systems developed and used within the EU-ADR project. In particular, EMC participates as a partner in the CALBC project and both groups (EMC and UPF) contributed with annotations to the CALBC silver standard corpus [15, 16] .

2.2. Relation extraction

Once the entities are identified by the aforementioned systems, the following step is the identification of pair-wise relationships between the entities, as defined in the introduction. The information extracted on the pair-wise relationships is then used for the signal filtering and signal substantiation processes (see Figure 1). The methods used for the extraction of the relationships involve a MeSH based approach, a co-occurrence based approach, and finally a NLP (Natural Language Processing) based approach, all of them described in the next sections.

2.2.1. MeSH based approach

Medline database from the National Library of Medicine (NLM) is a leading source of scientific information. The aim of the MeSH based approach is to automate the search of

publications related to ADRs corresponding to a given drug/adverse event association by: 1) defining a pattern for the queries used to search Medline, and 2) determining the threshold number of extracted publications needed to confirm the knowledge of this association in the literature.

2.2.1.1 Methodology

To determine if an ADR has already been published, we used Medline as a knowledge source. The relevant publications have the drug and the adverse event of interest co-occurring in the same citations. The Medical Subject Headings (MeSH®) thesaurus is a controlled vocabulary produced by the NLM and used for indexing, cataloguing, and searching for biomedical and health-related information and documents. NLM indexers are selecting the most appropriate MeSH descriptors and subheadings (or qualifiers) to resume the full content of an article after reading the full text. This professional indexation enhances the quality of information retrieval. To automate the search of publications related to ADRs corresponding to a given drug/event association, we used the following MeSH-based approach: 1) map the events to MeSH, 2) map the drugs to MeSH, 3) construct the query with MeSH terms and filter the results, and 4) determine a threshold number of publications to confirm the knowledge of this association in the literature by testing the method on an expert-built validation set including true positive and true negative drug/adverse event associations.
We downloaded a subset of Medline via PUBMED to retrieve all the citations with the "adverse effects" MeSH subheading. We then parsed the XML result of the citations to import it in a database. For each citation, we gathered the following information: PMID, MeSH descriptors, major/minor subheadings, substances, date of creation of the citation. We used the 2009AA version of the Unified Medical Language System® (UMLS®), a biomedical terminology integration system handling more than 150 terminologies.

Mapping of events

As already described in the Introduction, the EU-ADR project uses an event-based approach where a limited set of specific events are inspected for their association with all available drugs in the EHR databases participating in the project. For each of the six events a list of UMLS concepts was identified by their CUIs by medical expertts. We used the Metathesaurus of the UMLS to retrieve MeSH codes and the preferred strings in English. If the concept had no direct mapping in MeSH, we used the "restrict to MeSH" algorithm [17] to get the nearest MeSH codes. Topical subheadings (or qualifiers) are used to narrow the specific focus of a main MeSH heading to a particular aspect of the subject. The subheading "chemically induced" (CI) is used to qualify the adverse events in drug safety articles [18].

Drug knowledge in Medline citations

Chemicals and Drugs listed in a citation under "Substances" can consist in either MeSH heading ("Descriptor Records", in which case they are duplicated in the MeSH Term list) or in "Supplementary Concept Records" (SCRs) (n=186,686). Every drug in the SCRs or MeSH

headings has been assigned one or more headings that describe known pharmacological actions of the concerned drug.

Mapping of drugs

In the EU-ADR project, all the databases code their drugs using the Anatomical Therapeutic Chemical (ATC) classification except one (Qresearch uses BNF codes that have been mapped to ATC). Those ATC codes need also to be mapped into MeSH to query Medline. As the ATC classification is not included in the UMLS, we used a mapping from ATC to CUI concept [3] that allowed the mapping from the ATC codes to the MeSH Headings or SCRs.

The subheading "adverse effects" (AE) is used to qualify the drugs in drug safety articles [18] and this can be used only to qualify drugs mentioned in Medline citations using MeSH headings (in the MeSH terms field). SCRs terms do not have any subheading. However, citations with drugs SCRs in the "substances" field can have, in the MeSH terms field, MeSH headings for the drugs pharmacological actions, with the appropriate subheadings.

This situation is illustrated in the following example: Moxifloxacin is a SCR with the MeSH heading "Anti-Infective Agents" as pharmacological action. Aspirin is a MeSH heading with four pharmacological actions: Anti-Inflammatory Agents, Non-Steroidal, Platelet Aggregation, […]. Both these drugs were mentioned in the case report titled: "Drug Points: tachycardia associated with moxifloxacin." (PMID: 11141146). In this, a 49 year-old man was prescribed moxifloxacin for sinusitis and developed tachycardia as an adverse effect of moxifloxacin. It is also reported that he took aspirin for a headache (with no adverse effect). In the citation, two substances are indexed: moxifloxacin, and aspirin. MeSH Terms are:

- **Tachycardia/chemically induced***
- **Anti-Infective Agents/adverse effects***
- Sinusitis/drug therapy
- Anti-Inflammatory Agents, Non-Steroidal/therapeutic use
- Aspirin/therapeutic use
- Headache/drug therapy , […]

Moxifloxacin can only be indexed in the "Substances" field and not the MeSH terms field (it is a SCR) not like aspirin which is a MeSH heading with the "therapeutic use" subheading. The pharmacological action of moxifloxacin, "Anti-Infective Agents", has the subheading "adverse effects" so the adverse effects knowledge can be linked to the appropriate drug (Moxifloxacin and not Aspirin).

Query construction

To retrieve the appropriate publications we used the co-occurrence of four elements in a citation: the drug (from "substances" OR "MeSH terms"), the adverse effect and the two subheadings, AE and CI. We only took into account drugs from the "substances" field if their

pharmacological action was qualified by the subheading "adverse effects" (see the previous example).

Filtering the results by publication type

We considered the following types of publication as non-contributive to describe new ADRs: *Addresses, Bibliography, Biography, Comment, Dictionary, Directory, Duplicate Publication, Editorial, Festschrift, Government Publications, Historical Article, In Vitro, Interactive Tutorial, Interview, Introductory Journal Article, Lectures, Legislation, Patient Education Handout, Periodical Index, Published Erratum and Retracted Publication.*

Constituting the validation sets

The methodology to create the validation set has been described in Deliverable 2.2 [19]. Here, a brief summary of the method is presented. The first drug-event set consisted of true positive associations, defined as adverse events that in the past were recognized as related to drug exposure. These associations constitute well-recognized safety signals. True positive signals consist of drug-event combinations for which a signal was generated and confirmed in the past. The combinations corresponding to this definition were identified through the following two-steps procedure. First step: search performed through the French Pharmacovigilance database for drugs associated with the selected events using Reporting Odds Ratio in the case-non-case method. This step was performed to provide an orientation for the search in the literature and the databases of the regulatory agencies. Second step: Search in the literature and the databases of the regulatory agencies. The confidence in the status of a signal increases with the addition of new evidences and the absence of studies questioning its reliability. Thus, the signals selected for the constitution of the true positive set are mostly referring to historical and very well known associations for which no doubts remain.

The second set consists of true negative signals, defined as drug-event associations for which no signals have been generated so far. True negative signals consist of drug-events combinations for which no signal has ever been generated until the time of the study. The combinations corresponding to this definition were identified through the following four-step procedure: First step: search performed in the French Pharmacovigilance database for drugs that are not associated with the selected events using Reporting Odds Ratio in the case-non-case method. This step was performed to provide an orientation for the search in the literature and the databases of the regulatory agencies. Second step: search in the literature and the databases of the regulatory agencies. The confidence in the status of a signal increases with the addition of new evidences and the absence of studies questioning its reliability. Thus, the signals selected for the constitution of the true negative set are mostly referring to drugs that were marketed for a long time and for which no question remains about a potential association to an event of interest. No data providing more evidence than a single case report had to be found for these combinations in the literature. Third step: the validation of the set was performed using data from the Thomson Reuters Micromedex database. If a signal selected as true negative was the focus of a report in this database, it was rejected from the

set. Fourth step: the FDA AERS spontaneous reporting database was explored as well to cross-check the identified set of true positive and negative signals.

A list of 5 drugs for the true positive set and of 5 drugs for the true negative set was constituted for each of the 6 events. Overall, a list of 60 pairs of drug/adverse effects was available to test our method.

2.2.1.2 Results

Mapping of the drugs

All the ATC codes were mapped successfully to the MeSH (MeSH Descriptors or SCRs) (precision=100%) with the "ATC to CUI" and the "CUI to MeSH" mapping (UMLS). 83% of drugs were MeSH Headings and 17% were SCRs.

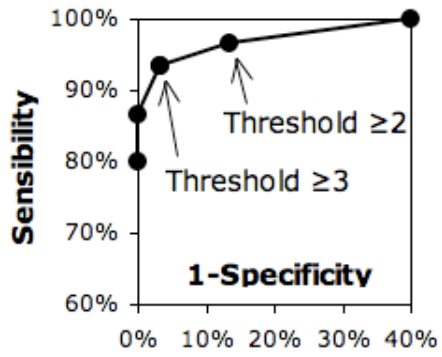
Table 1. Sensibility and specificity for each threshold

Threshold	≥1	≥2	**≥3**	≥4	≥5	≥10
Specificity %	60	87	**97**	97	100	100
Sensitivity %	100	97	**93**	93	87	80

Retrieved publications

The number of retrieved publications for each of the 60 pairs of the validations sets is available for consultation in the Annexe 1. The specificity and sensitivity for each threshold is given in Table 1. The ROC performance of the model is graphically presented in Figure 5.

Figure 5: ROC curve of the model



2.2.1.3 Discussion

The sensibility and specificity measures of publication retrieval in the automated search show a good performance. When using a threshold ≥ 3 extracted publications, the extracting method has a sensibility of 93% and a specificity of 97%. This result is comparable with previous

work with the MeSH and subheading approach [18]. In Garcelon *et al*, the most relevant threshold was ≥3 with a Se=65% and Sp=97%. Their approach didn't use the pharmacological action as an additional knowledge source so their missed drugs that couldn't be mapped to MeSH. Zeng and Cimino carried out an automated disease-chemical knowledge extraction based on the co-occurrence of UMLS concepts [20], obtaining similar results than our method with a Se= 93%.

Our approach offers the opportunity to automatically determine if an ADR (association of a drug and an adverse effect) has already been described in Medline. However, the causality relationship between the drug and an event may be judged only by an expert reading the full text article. Because specific subheadings and keywords are used in the queries that are automatically built, the automated search may be more specific than a manual query. Only publication already indexed with MeSH could be detected by our method. We then exclude all the publications before their indexation.

Errors of classification

Looking at the results within the true negative set, only the couple acute renal failure and prednisolone was misclassified: 5 citations were detected. In this case, however, as prednisone is a corticosteroid that can be used for the therapeutic management in some types of acute renal failure or co-prescribed with drug inducing renal failure. In this case, results provided by our automated search method could be falsely positive.

Validation set

As this definition of true negative signal is based on the existing knowledge at the time the signal is investigated, it has to be understood that signals that are currently considered as true negative could become positive signals in the future based on new evidence on drug safety. Thus, this set of true negative signals can only be considered as such at the time and date of its constitution.

2.2.2. Co-occurrence based approach

The co-occurrence-based approach for extracting drug-event relationships from DrugBank and DailyMed has been described in Deliverable 4.1. Briefly, from each DrugBank entry, a field specifying ATC codes and a field listing potential adverse events were extracted and processed by Peregrine. The Peregrine output was then processed to link the ATC codes of the drugs to the UMLS concept identifiers of the adverse events. DailyMed contains Summary Product Characteristics (SPCs) of drugs. Each SPC was parsed to extract the "title" field (containing the drug name) and the "adverse reaction" and "boxed warning" fields (containing the adverse events). These fields were subsequently indexed by Peregrine and the output was processed to link ATC codes to UMLS concept identifiers of adverse events.

In addition we also extracted drug-event relationships from the whole of Medline database. Medline abstracts were split into sentences, the sentences were indexed by Peregrine and the output was processed to check if a sentence contained a drug (a corresponding ATC code was

found in the drug thesaurus) and an adverse event (corresponding with one of the UMLS concept identifiers for the adverse events studied in EU-ADR). A chi-square test was performed to check if the probability of the drug and the adverse event co-occurring together in a sentence was significantly different than would be expected by chance. To account for multiple testing, the significance level was set at 0.0001. Only if the P-value was lower, the co-occurrence of drug and adverse event was accepted as a drug-event relationship.

2.2.3. NLP based approach

Co-occurrence based approaches are suitable for large-scale processing but often suffer of lack of specificity and a high rate of false positives. In addition, they do not provide details on the directionality or on the type of association between entities. In this regard, other approaches that use more detailed information from the text are required. NLP based approaches are a suitable solution if more details about the relationships are required.

A system for the extraction of relationships from text is currently under development. At the current stage we cannot report on results since the development of the EU-ADR corpus was finished at the time of writing of this deliverable. Meanwhile, the work was devoted to the development of the tools that will be required once the corpus is ready. The system combines the use of NLP approaches and machine learning. The UIMA pipeline (described in section 2.4) is used to extract morpho-syntactic and semantic features required to train the classifiers. In this regard, the MST parser required for obtaining dependency parse trees was re-trained on a domain specific corpus (the Genia corpus [21]) in order to be able to cope with syntactic structures specific for the biomedical domain. In addition, Collection Readers and CAS Consumers that provide the output files required for the models to be trained were developed.

A relation between two entities is represented as a label on a sentence which can contain at most one binary relationship. Thus, the problem of relation extraction is mapped to a simpler and better known classification problem. Currently, we are exploring with SVM and case-based reasoning algorithms using different combinations of linguistic features. In addition, association clue words will be used as an additional feature.

For SVM we use the JSRE implementation [22]. The system is based on two kernels that use shallow linguistic information. We have implemented a third kernel that considers deep linguistic parsing, in the form of dependency parse trees. In addition, it can consider as additional features key words that denote a given relationship (referred as association key words). For initial development of the system, since the annotation of the EU-ADR corpus is still in progress (see below), we used a data set restricted to one of the association types of interest to the project (SNP-disease association) provided by one of the partners (UPF). The annotation on this corpus follows the same schema as the EU-ADR corpus, distinguishing between sentences that contain an association between the entities of interest (“positive association”), sentences that state that there is no association between the entities (“negative association”) and sentences in which both entities co-occur but are not related by any means (“false association”). Preliminary results on the SNP-disease corpus for a binary classification problem (to distinguish positive association from the rest of sentences) are the following: 0.85

Precision, 0.96 Recall, 0.9 F-score. The baseline for this dataset is 0.61 Precision, 1 Recall and 0.76 F-score. These results were obtained using the 3 kernels (using features such as POS, stemming, and dependency parsing) and association keywords, which were obtained from the corpus and the literature. Results on the performance of the relation extraction system will be available in the near future. Moreover, once the development of relationship extraction system is finished, it will be incorporated as a module in the UIMA pipeline.

2.3. EU-ADR corpus

2.3.1. Corpus development

The Text Mining (TM) activities in WP4 are aimed at developing tools for the extraction of relationships between the entities drug, target and disease, to aid in the substantiation of the signal. Thus, the final aim of the TM activities is to extract from the biomedical literature the following relationships: target-disease, target-drug and drug-disease.

During the first year an evaluation on publicly available corpora was conducted (see Annexe 2). The evaluation indicated that none of the already available corpora fitted our needs, and therefore it was decided that a collection of documents annotated by experts with the above mentioned relationships was required for the text mining activities described above (the EU-ADR corpus from now on). The EU-ADR corpus consists of Medline abstracts containing semantic annotations on the entities and their relationships. The annotations were performed by domain experts who are capable of deciding if a text describes a relationship, and are therefore considered to provide a "gold standard" data set. A detailed description of the EU-ADR corpus is provided in the Annexe 2. This document also served as annotation guidelines.

Another important issue for the development of the EU-ADR corpus was the availability of a suitable annotation tool to aid the experts' work. An annotation tool that performs automatic annotation of entities and helps the experts in the annotation of relationship has been developed (described in section 2.3.2).

In addition, detailed guidelines for the annotation of the documents were developed. A team of experts from the project partners were selected and trained on the task of annotation of documents for text mining activities. Two pilot annotation experiments were performed and the experiences were used to further refine the annotation guidelines using the feedback provided by the domain experts and to fine tune the annotation tool to cope with annotators' needs.

After completing the pilot phase, the annotation started on a set of 100 abstracts for each relationship type (drug-target, drug-disease, target-disease). Each abstract was annotated independently by three annotators, yielding a total of 900 (300x3) abstract annotations.

To harmonize the different annotations, a simple majority voting scheme was applied: if two of the three annotators agreed on a given annotation, it became part of the final EU-ADR

corpus. If an annotation was given by only one annotator, it was discarded. To allow for slightly differing entities marked by different annotators, we used a word-based voting scheme rather than a scheme that requires terms to match exactly. For example, one annotator may have marked "diabetes" whereas the other marked "severe diabetes". With exact match, this would be considered a disagreement, whereas one may argue that the annotators agreed on the term "diabetes". A relationship was only included in the corpus if there was agreement both on the entities involved and on the type of the relationship (positive, negative, and speculative, see description of the types in the "EU-ADR corpus development" document). The analysis of the inter-annotator agreement on the relationships indicated that there was not a clear decision regarding positive and speculative associations among the annotators. Thus, speculative and positive relationship were considered equivalent when assessing inter-annotator agreement. If one annotator marked a relationship type as positive and another annotator marked the same entities as the first but the relationship type as speculative, we also considered this as an agreement with either positive or speculative relation type.

Table 2 shows the number of entities and relationships on which there was agreement and the total number of annotations, for the abstracts corresponding with each of the three relationships.

Table 2: Agreement on entity and relationship annotation.

Abstract set	Entity[a]	Relationship[b]
Drug-disorder	1542/1888 (82%)	222/508 (44%)
Drug-target	2231/2541 (88%)	219/634 (35%)
Target-disorder	1829/2257 (81%)	291/714 (41%)

[a]Number of entities on which there was agreement/Total number of annotated entities.

[b]Number of relationships on which there was agreement/Total number of annotated relationships.

The agreement on the annotated entities is high (>80%) for all abstract sets. Regarding the relationships, the agreement is much lower. In part this is to be expected because the annotators have to agree on the two entities involved in the relationship and on the relationship type before the relation is included in the corpus. In particular, it may be difficult to distinguish between a speculative relationship and the absence of a relationship.

2.3.2. Development of an annotation tool

For the annotation of entities and relationships for drugs, targets and disorders in abstracts we explored existing annotation tools. The requirements that the computer already should provide an initial markup using the entities defined within EU-ADR and that the system should be available through the web and easy to use were taken into consideration when looking for

annotation tools. None of the systems analysed met these requirements. Development of an EU-ADR online annotation tool has been done to support the expert annotators with the easiest way of providing this information. The tool is based on software from Knewco to show in-text highlights of terms recognized by Peregrine [9]. The relations between the terms are automatically derived using the semantic types of the terms (concepts) and proposed to the annotator. The annotator can revise the entities marked up and the relations found indicating the presence of the relationship and its nature (positive association, negative association, and speculative association, see definition in document "EU-ADR corpus annotation guidelines"). In addition, the annotators can add new entities if required. Annotations are stored server-side and can later be retrieved and corrected.

The tool is available online at the following URL:
http://aneurist.erasmusmc.nl/sda/annotate.py
Username: sda
Password: eu-adr

2.4. EU-ADR text mining workflow

A workflow for biomedical text processing was developed in the context of the UIMA framework [23]. The workflow contains, in addition to the modules required for semantic processing (NER modules) described above, multiple modules for morphosyntactic (i.e. Natural Language Processing of text) processing (Figure 6). For that, modules available from other research groups as well as modules developed within the consortium were integrated in the pipeline (Table 3).

Figure 6: EU-ADR text mining pipeline based on UIMA. Details on available Analysis Engines (AEs) are provided in Table 3. The workflow retrieves input text (MEDLINE citations) from a local database from the web by the Collection Reader module, processes the text by a variety of mosphosyntactic and semantic AEs, and provides the extracted in formation in different formats by the Cas Consumer modules.

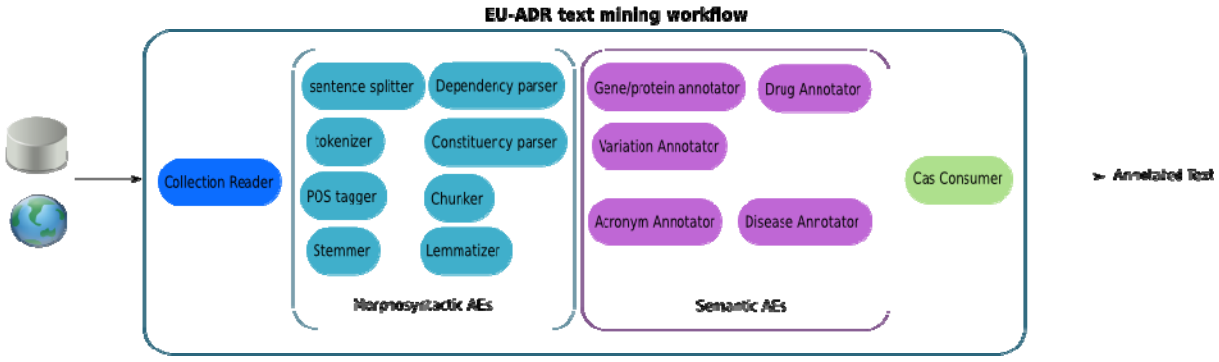


Table 3: Analysis engines modules available for the EU-ADR text mining workflow. Collection readers and Cas consumers are not shown.

Component	Description	Type
JULIE Sentence Splitter	Based on CRF	morphosyntactic
JULIE Tokenizer	Based on CRF	morphosyntactic
OpenNLP POS tagger	Based on MaxEnt	morphosyntactic
OpenNLP Chunker	Based on MaxEnt	morphosyntactic
OpenNLP Constituency Parser	Based on MaxEnt	morphosyntactic
MST Dependency Parser	Maximum spanning trees	morphosyntactic
Acronym Annotator	Pattern-based rules	semantic
Disease Annotator	NER for diseases based on LingPipe	semantic
Peregrine Wrapper	NER for genes/proteins, drugs	semantic
OSIRISv1.3	NER for SNPs	semantic

Among the latter, we developed the following modules: a UIMA wrapper for Peregrine (for the recognition of the entities "chemicals & drugs" and "genes & proteins"), a NER system for diseases based on LingPipe [24], and a NER system for the entity "SNPs & sequence variants" based on the OSIRIS approach.
In addition, the LingPipe module within the EU-ADR UIMA pipeline can be used with different dictionaries (drug, disease). The availability of different NER for the recognition of named entities in the text offers the possibility of annotation of entities with different system, which has been reported to improve the precision and recall of automatically annotated texts [25].
The modules for the morphosyntatic processing are required to extract features from text that can be used to train the machine learning classifiers that will be used to extract the relationships.

Figure 7: EU-ADR UIMA type system

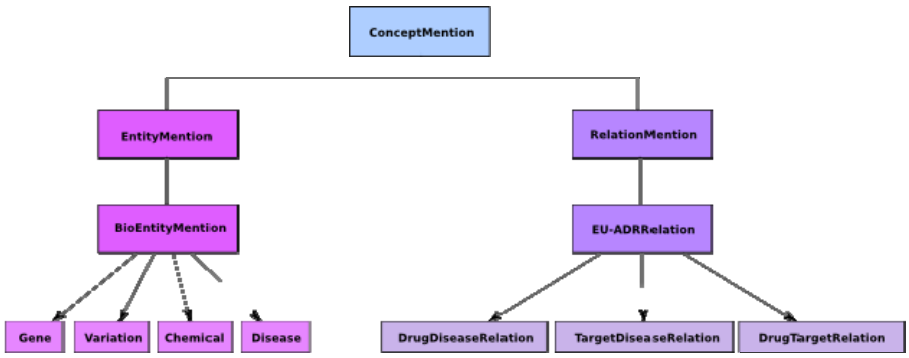
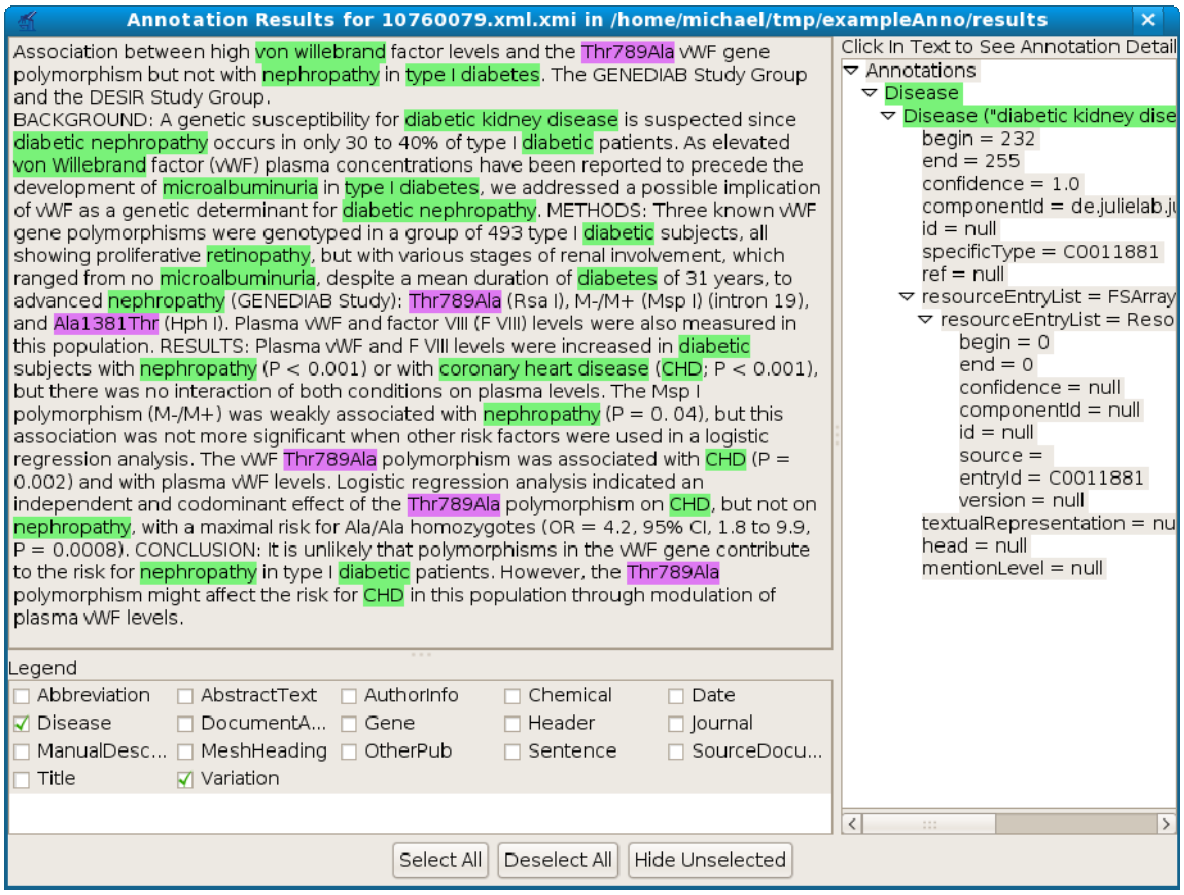


Figure 8: Exemplary abstract annotated with the entities variation and disease by the EU-ADR text mining workflow.



The UIMA Type System from Julie lab was used as a basis for the EU-ADR text mining workflow [26]. More specifically, we use the following Type System layers: julie-basic-types, julie-morpho-syntax-types, julie-semantics-concept-types and julie-semantics-biology-types. These layers contain many useful types, but they do not cover all the entities and relationship types required in our project. Therefore, we extended the Type System with specific types covering sequence variations (e.g. SNPs) and a number of specific relation types, as shown in Figure 7.

Figure 8 shows an exemplary output of the annotation of diseases and variation with the EU-ADR text mining workflow. The annotation details show the UMLS concept identifier assigned to the text “diabetic kidney disease”.

3. DISCUSSION

Tools for mining relationships between drugs, targets and disease phenotypes were developed. During the development of this text mining system, our philosophy was to re-use

already available tools within the consortium or the text mining community, and adapt or develop new tools if required.

We decomposed the problem of relation extraction in pair-wise relationships that were mined independently from each other from the text, and are later combined within signal substantiation pipeline in a later stage. This allows simplification of the tasks, but also allows the integration of other pieces of information that support a given relationship during substantiation (for example, evidence coming from other databases or different bioinformatic approaches). This integration process will be presented in Deliverable 4.5.

The systems developed extract information that can be used for signal filtering (drug-disease associations) and for signal substantiation (drug-disease, drug-target, target-disease associations). To accomplish these tasks, a variety of approaches were used, such as co-occurrence of entities in Medline abstracts as well as co-occurrence of MeSH annotations, and relations expressed in natural language at the sentence level between a pair of entities. Since each approach has its own advantages and limitations, we believe that by combining the information extracted from the different systems we can overcome the limitations of each individual approach. The best way of combining these methodologies is work in progress.

One of the main difficulties in the work reported is the development of an annotated corpus. Our experience is that a great effort is needed to accomplish the task since it is very time consuming and laborious. We believe that this effort is of value since it will provide a resource to the community that is not already available, in addition to allow the development of the text mining system required for the project.

The information extracted by the systems here described is integrated with information generated by other bioinformatic approaches developed within WP4. The integrated WP4 system will be described in Deliverable 4.5.

ANNEXES

Annex 1: MeSH based approach evaluation with the TP and TN validation sets

Table 4: Number of MEDLINE notices retrieved on the true positive set. AMI: Acute Myocardial infarction, ARF: Acute renal failure, AS: Anaphylactic shock, BE: Cutaneous bullous eruption, RHABD: Rhabdomyolysis, UGIB: Upper gastrointestinal bleeding

Event	Drug	NB
UGIB	Aspirin	490
RHABD	Simvastatin	95
ARF	Cisplatin	64
AMI	Rofecoxib	63
RHABD	Gemfibrozil	45
AMI	Doxorubicin	36
UGIB	Ticlopidine	34
AMI	Fluorouracil	32
UGIB	Prednisone	27
AMI	Sildenafil	25
ARF	Amphotericin b	25
BE	Trimethoprim-sulfamethoxazole combination	22
RHABD	Atorvastatin	21
AMI	Sumatriptan	19
UGIB	Piroxicam	18
RHABD	Bezafibrate	17
ARF	Ampicillin	17
BE	Allopurinol	17
ARF	Acyclovir	16
BE	Nevirapine	16
AS	Diclofenac	15
RHABD	Procetofen	15
BE	Lamotrigine	15
ARF	Vancomycin	12
AS	Acetylcysteine	8
UGIB	Dipyridamole	6
ARF	Prednisone	5
AS	Ceftriaxone	4
BE	Piroxicam	4
AS	Abacavir	2
AS	Moxifloxacin	1

Table 5: Number of MEDLINE notices retrieved on the true negative set. AMI: Acute Myocardial infarction, ARF: Acute renal failure, AS: Anaphylactic shock, BE: Cutaneous bullous eruption, RHABD: Rhabdomyolysis, UGIB: Upper gastrointestinal bleeding

Event	Drug	NB
ARF	Prednisone	5
UGIB	Furosemide	2
AS	Cetirizine	2
AMI	Omeprazole	2
UGIB	Amlodipine	1
UGIB	Ramipril	1
AS	Clarithromycin	1
AMI	Itraconazole	1
RHABD	Amoxicillin	1
RHABD	Pindolol	1
ARF	Acenocoumarol	1
BE	Metformin	1
UGIB	Albuterol	0
UGIB	Zolpidem	0
AS	Clofibrate	0
AS	Citalopram	0
AS	Haloperidol	0
AMI	Acyclovir	0
AMI	Fluvoxamine	0
AMI	Lorazepam	0
RHABD	Clobazam	0
RHABD	Ketoprofen	0
RHABD	Perindopril	0
ARF	Clobazam	0
ARF	Levodopa	0
ARF	Salmeterol	0
BE	Digoxin	0
BE	Bacitracin	0
BE	Perindopril	0
BE	Spironolactone	0

Annex 2: EU-ADR corpus development

See document "EU-ADR corpus development" for a complete description on the development of the corpus.

References

1. Trifiro G, Fourrier-Reglat A, Sturkenboom MC, Diaz Acedo C, Van Der Lei J: **The EU-ADR project: preliminary results and perspective**. *Stud Health Technol Inform* 2009, **148**:43-49.
2. Avillach P, Mougin F, Joubert M, Thiessard F, Pariente A, Dufour JC, Trifiro G, Polimeni G, Catania MA, Giaquinto C *et al*: **A semantic approach for the homogeneous identification of events in eight patient databases: a contribution to the European eu-ADR project**. *Stud Health Technol Inform* 2009, **150**:190-194.
3. Hettne KM, Stierum RH, Schuemie MJ, Hendriksen PJ, Schijvenaars BJ, Mulligen EM, Kleinjans J, Kors JA: **A dictionary to identify small molecules and drugs in free text**. *Bioinformatics* 2009, **25**(22):2983-2991.
4. **The Unified Medical Language System** [http://www.nlm.nih.gov/research/umls/]
5. Hettne KM, van Mulligen EM, Schuemie MJ, Schijvenaars BJ, Kors JA: **Rewriting and suppressing UMLS terms for improved biomedical term identification**. *J Biomed Semantics*, **1**(1):5.
6. Hatzivassiloglou V, Duboue PA, Rzhetsky A: **Disambiguating proteins, genes, and RNA in text: a machine learning approach**. *Bioinformatics* 2001, **17**(suppl_1):S97-106.
7. Kors J, Schuemie M, Schijvenaars B, Weeber M, Mons B: **Combination of genetic databases for improving identification of genes and proteins in text.** *BioLINK, Detroit* 2005.
8. Schuemie MJ, Mons B, Weeber M, Kors JA: **Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification**. *Journal of biomedical informatics* 2007, **40**(3):316-324.
9. Schuemie M, Jelier R, Kors J: **Peregrine: Lightweight gene name normalization by dictionary lookup.** *Proc of the Second BioCreative Challenge Evaluation Workshop* 2007:131 - 133.
10. Furlong LI, Dach H, Hofmann-Apitius M, Sanz F: **OSIRISv1.2: a named entity recognition system for sequence variants of genes in biomedical literature**. *BMC Bioinformatics* 2008, **9**(1):84.
11. Sherry ST, Ward M, Sirotkin K: **dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation**. *Genome Res* 1999, **9**(8):677-679.
12. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM: **A "silent" polymorphism in the MDR1 gene changes substrate specificity**. *Science* 2007, **315**(5811):525-528.
13. Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, Klein TE: **PharmGKB: the Pharmacogenetics Knowledge Base**. *Nucleic Acids Res* 2002, **30**(1):163-165.
14. **Julie Lab** [http://www.julielab.de/]

15. Rautschka M, Sanz F, Furlong L: **Towards a system for the recognition of disease terms in biomedical texts.** In: *First CALBC Workshop: 17-18 June 2010; European Bioinformatics Institute, Hinxton, Cambridgeshire, UK.*: EMBL-EBI; 2010.
16. Rebholz-Schuhmann D, Jimeno A, Li C, Kafkas S, Lewin I, Kang N, Corbett P, Milward D, Buyko E, Beisswanger E *et al*: **Assessment of NER solutions against the first and second CALBC Silver Standard Corpus.** In: *Fourth International Symposium on Semantic Mining in Biomedicine (SMBM): 25-26 Ocotber 2010; European Bioinformatics Institute, Hinxton, Cambridgeshire, UK.*; 2010.
17. Bodenreider O: **The Unified Medical Language System (UMLS): integrating biomedical terminology**. *Nucleic Acids Res* 2004, **32**(Database issue):D267-270.
18. Garcelon N, Mougin F, Bousquet C, Burgun A: **Evidence in pharmacovigilance: extracting adverse drug reactions articles from MEDLINE to link them to case databases**. *Stud Health Technol Inform* 2006, **124**:528-533.
19. Fourrier-Reglat A, Pariente A, Miremont-Salame G, Polimeni G, David A, Catania MA, Salvo F, Moore N, Trifiro G: **EU-ADR Deliverable 2.2: Two validation sets with supplementary information**. In: *EU-ADR deliverables.* 2010.
20. Zeng Q, Cimino JJ: **Automated knowledge extraction from the UMLS**. *Proc AMIA Symp* 1998:568-572.
21. Kim J: **GENIA corpus-semantically annotated corpus for bio-textmining**. *Bioinformatics* 2003, **19**(Suppl 1):i180 - i182.
22. Giuliano C, Lavelli A, Romano L: **Exploiting shallow linguistic information for relation extraction from biomedical literature**. In: *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006): 2006*; 2006: 5-7.
23. **UIMA** [http://incubator.apache.org/uima/]
24. **LingPipe** [[http://www.alias-i.com/lingpipe]]
25. Leitner F, Krallinger M, Rodriguez-Penagos C, Hakenberg J, Plake C, Kuo C, Hsu C, Tsai R, Hung H, Lau W: **Introducing meta-services for biomedical information extraction**. *Genome Biology* 2008, **9**(Suppl 2):S6.
26. Hahn U, Buyko E, Landefeld R, Mühlhausen M, Poprat M, Tomanek K, Wermter J: **An overview of JCoRe, the JULIE lab UIMA component repository**. In: *Proceedings of the LREC: 2008*; 2008: 1–7.