ICT-2011-288048

EURECA

Enabling information re-Use by linking clinical Research and CAre

IP
Contract Nr: 288048

Deliverable: D3.1 Initial prototype for concept extraction out of EHR free text

Due date of deliverable: (01-31-2013)
Actual submission date: (14-02-2013)

Start date of Project: 01 February 2012 Duration: 42 months

Responsible WP: Xerox

Revision: <outline **,**draft, **proposed**, accepted>

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Service	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (excluding the Commission Services)	

0 DOCUMENT INFO

0.1 Authors

Author	Company	E-mail
Salah Ait-Mokhtar	Xerox	Salah.Ait-Mokhtar@xrce.xerox.com
Caroline Hagege	Xerox	Caroline.Hagege@xrce.xerox.com
Pajolma Rupi	Xerox	Pajolma.Rupi@xrce.xerox.com

0.2 Internal reviewer

Raúl Alonso	UPM

0.3 Reviewers

Raúl Alonso	UPM
Berry de Bruijn	NRC
Cyril Krykwinski	IJB

0.4 Documents history

Document version #	Date	Change
V0.1	01.15.2013	Starting version, template
V0.2	01.18.2013	Definition of ToC
V0.3	01.23.2013	First draft
V0.4	01.30.2013	Complete draft
V0.5	01.31.2013	Integrated version (send to WP members)
V0.6	02.10.2013	Updated version (send to project internal reviewers)
Sign off		Signed off version (for approval to PMT members)
V1.0		Approved Version to be submitted to EU

0.5 Document data

Keywords	
Editor Address data	Name: Caroline Hagege Partner: Xerox Address: 6 chemin de Maupertuis, 38240 Meylan Phone: +33 (0)4.76.61.51.32 Fax: E-mail: Caroline.Hagege@xrce.xerox.com
Delivery date	14.02.2013

0.6 Distribution list

Date	Issue	E-mailer
14.02.2013	1.0	Benoit.ABELOOS@ec.europa.eu
		INFSO-ICT-288048@ec.europa.eu

Table of Contents

1 Introduction

This deliverable is the first one of WP3 and consists in the first version of the prototype for term extraction out of free text. Information extraction from free medical text using Natural Language Processing (NLP) is currently an important field considering the huge amount of unstructured textual document in the medical domain (patient data, clinical trials and guidelines, medical literature). The ability to process automatically the information expressed in these documents can help to bridge gaps between patient information and clinical literature which is the main goal of the EURECA project.
At this stage of the project, patient data is not yet available for all technical partners. As a result we built this first prototype for medical concept identification using Clinical Trials texts which are freely available, having in mind that as soon as patient data become available, some adaptations of the tool will be necessary. The Semantic core dataset (WP4) which is also not completely defined at this stage of the project has also to be integrated in the annotation tool in a simple and efficient way. For these reasons, we try to adopt a general approach that is modular enough to enable easy adaptation to different document types and domains.

One of our concerns for the development of the annotation tool was also to facilitate the integration of specific terminological information that can be requested for specific scenarios and use-cases defined in the project. As stated in the DOW, information extraction process from free text should be modular and based on the identified core dataset for that domain. This led us to choose as information source the UMLS (Unified Medical Language System) framework ([Bodenreider 07]) as it encompasses a wide range of terminologies/ontologies that are considered to be relevant by our medical partners (see deliverable D4.1 section 3.2).

The global analysis for the general user needs regarding terminological resources has concluded that the terminologies that are widely used by the partners are the following:

- SNOMED (SNOMED-RT/SNOMED-CT)
- LOINC
- MedDRA
- ICD (ICD-O/ICD9/ICD9-CM/ICD10)
- CDISC
- CTCAE
- NCI Thesaurus
- Radlex

Almost all of them are integrated within UMLS. As a result, we can take advantage of the general format provided by UMLS Metathesaurus and Semantic Network having a single access point to exploit all (or part) of these terminologies, without losing the specific information coming from a specific source.

In this document, we are using the following naming conventions:

Term: Any string that denotes a medical concept or entity.
Ontology: Formal description of concepts and relations holding between these concepts.

Terminology: A term repository (structured or not).
Lexicon: A set of strings associated with linguistic information, each string being a normalized form of a word (single or compound) that can appear in a text.
Medical lexicon: A set of strings (associated with linguistic information) corresponding to normalized forms of single or compound words that can appear in a text, when these words correspond to a medical term.
General lexicon: A set of strings (associated with linguistic information) corresponding to normalized forms of single or compound words belonging to the general domain language.
Lexical item: A member of the lexicon.

In the following chapter 2, we quickly explain how to use the Java-based concept identifier in a command-line interface or through its Java API. We then present in details the UMLS resource on which the tool is based, and the work we did in order to select from and transform the raw UMLS terminological data and to generate a UMLS-based NLP lexicon (chapter 3). The next chapter (4) describes the integration of this NLP lexicon to the NLP processing chain and how the effects on POS tagging and lexical ambiguities are handled. The last chapter (5) is a brief conclusion and presentation of future work in NLP-related tasks of WP3.

2 Using the term annotation tool

The term annotation tool is delivered as a zipped package named “Xmedlan.zip”. When unzipped, the main folder Xmedlan contains a jar file, Xmedlan.jar, an NLP resource folder, named “GRM”, and a basic configuration file.

Java 6+ is required. The tool can be used either from a command line interface or through its java API. It can take an input text file (raw text format, UTF-8 encoding), or a collection of such texts, and produce a set of medical term occurrences matching UMLS concepts and one or more term specifications in source terminologies. Currently, available terminologies are SNOMEDCT, NCI, LNC (LOINC), and ICD9-CM.

2.1 Command line usage

Run the following command in a console:

```
java -jar <xmedlan_pathname>/Xmedlan.jar –doc:<text_pathname>
```

where `<xmedlan_pathname>` is a pathname to the main directory where the package has been installed, and `<text_pathname>` is the pathname of the text document to process, or the root directory of the collection of text documents to process.

For an example of the output produced by the command-line version of the annotation tool, please refer to section 4.2 (“Free text annotation”).

2.2 Using the Java library

The pathname to the Xmedlan main folder, which contains the jar file, needs to be added to the java CLASSPATH.

In order to process a text or collection of texts, you need to create a `TextAnalyzer` object:

```
TextAnalyzer textAnalyzer = new TextAnalyzer();
```

You can then call method `findTerms(String pathname)` of the `TextAnalyzer` class, passing the pathname of the text file to process, or the pathname of the directory where the collection of text files is. In the latter case, the tool will go recursively through all the subdirectories.

```
List<TermOccurrence> terms = textAnalyzer.findTerms(text_pathname);
```

It is also possible to run the term identifier on a text string instead of a text file:

```
List<TermOccurrence> terms = textAnalyzer.findTermsInString(text);
```

The **`findTerms(String text_pathname)`** and **`findTermsInString(String text)`** methods return a list of term occurrences found in the input text(s). Each `TermOccurence` object comes with information that can be accessed through the following methods:

class **TermOccurrence**	
`String` **`getTextPathname()`**	Returns a string representing the absolute pathname of the input text file where the term occurrence has been found, and `null` when the `findTermsInString()` method is used.
`int` **`getStart()`**	Returns an integer representing the start offset, i.e. index of the first character of this term occurrence in the input text
`int` **`getLength()`**	Returns the length, in terms of characters, of this term occurrence
`String` **`getForm()`**	Returns the string form of this term occurrence
`String` **`getLemma()`**	Returns a lemmatized (normalized) form of this term occurrence
`List<UmlsReading>` **`getUmlsReadings()`**	Returns a list of possible readings, provided in UMLS for this term

Therefore, each term occurrence has a set of possible readings, which UMLS provides for the term. In the current version of the term identifier, a UMLS reading (instance of the `UmlsReading` class) consists of a concept unique identifier (CUI), a list of terminology names where the term is present (e.g. “SNOMEDCT”), the original codes of the term in the source terminologies, and the semantic types assigned to the term in UMLS. This information can be accessed with the following methods:

class **UmlsReading**	
`String` **`getCUI()`**	Returns a string representing the UMLS concept unique identifier (CUI) assigned to the term
`String[]` **`getSourceTerminologies()`**	Returns the names of terminologies where the term is present
`String` **`getSourceCodes(`**`Terminology sourceTerminology`**`)`**	Returns the original code of the term in a given source terminology. Possible terminologies are defined in the enum structure `Terminology`. In the current version, possible terminology values are: Terminology. SNOMEDCT, Terminology.NCI, Terminology.LNC (i.e. LOINC), and Terminology.ICD9CM (i.e. ICD9-CM). The returned value is a string representing the code(s) of the term in the selected source terminology (a term can have multiple distinct codes in a terminology).
`String[]` **`getTypes()`**	Returns the UMLS semantic types assigned to the term

3 Creating medical lexicons from existing terminologies

The annotation tool relies on medical lexicons that were created using and adapting existing resources. More precisely, we rely on existing terminologies (and also in a near future availability of semantic core dataset dedicated to given scenarios) and as said in introduction, we adopted UMLS as source of terminological information for building lexicons.

3.1 UMLS

UMLS is developed by the U.S. National Library of Medicine (NLM) and it is updated twice a year.
It combines a variety of source vocabularies, ontologies and terminologies by integrating them into its three knowledge sources: **Metathesaurus**, **Semantic Network** and SPECIALIST lexicon. This integration results in a very large medical knowledge base, covering numerous themes in medical domain. The availability of UMLS is subject to several restrictions carried by the respective original source terminologies. These vocabularies are organized into categories depending on the licensing constraints applied to them. Category 0 is free of charge, while different constraints are applied to the other categories.

3.1.1 Connecting terminologies through mapping

The enormous number of biomedical resources available and their variations result in having different terms defining the same concept or having the same concepts defined differently in different source terminologies/ontologies. The UMLS handles these cases by providing a mapping structure between these terminologies and the possibility to translate a term among the various terminologies. The following figure illustrates this mapping structure.

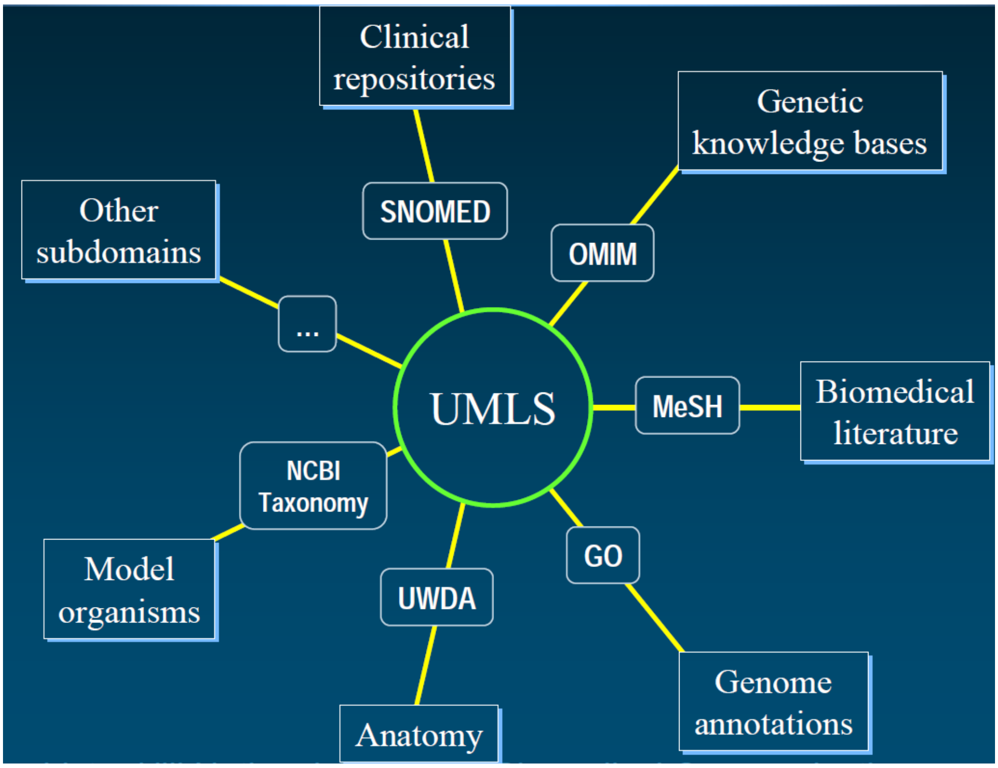


Figure 1: An overview of UMLS merging the different terminologies

3.1.2 UMLS Knowledge Sources

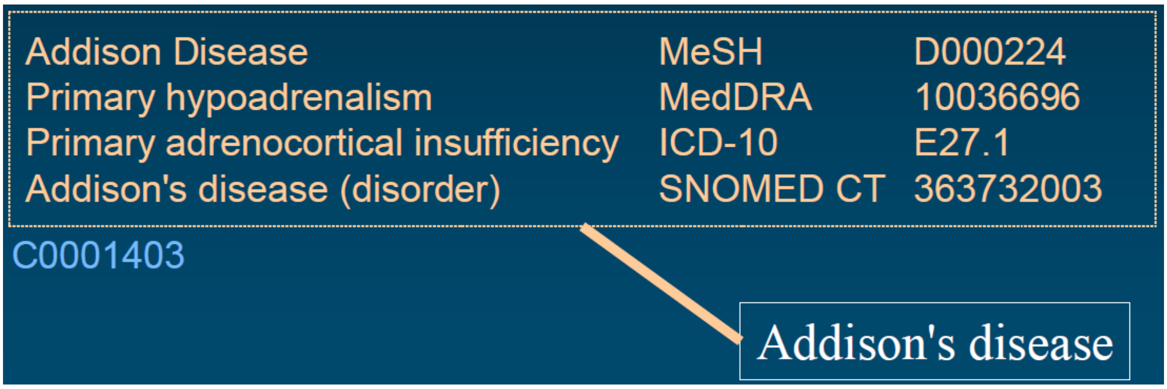
As specified above, there are three knowledge sources constituting the UMLS: Metathesaurus, which contains a collection of biomedical concepts and relationships between them, Semantic Network, which contains semantic types that characterize the terms present in the Metathesaurus and links between them, and SPECIALIST Lexicon which contains information about common English vocabulary, biomedical terms, terms found in MEDLINE and terms found in the UMLS Metathesaurus.

The knowledge sources are delivered as machine readable files. Our work is focused on the usage of Metathesaurus and Semantic Network knowledge sources which cover a broad range of biomedical information. Meanwhile, the SPECIALIST Lexicon is a syntactic lexicon of biomedical and general English. Only a small number of Metathesaurus terms are represented in SPECIALIST Lexicon (According to [Xu 10], only 1% of the UMLS data is enclosed in the SPECIALIST Lexicon).

3.1.2.1 Metathesaurus

The Metathesaurus is the base of UMLS. It is a very large, multi-purpose, and multilingual vocabulary database that contains information about biomedical and health related concepts, their various names and the relationships among them. The 2012AB release of UMLS includes more than 2.8 million concepts and 11.2 million terms from over 160 source terminologies. All the information present in Metathesaurus is labelled as to its source terminologies, by preserving the meanings, concept names, and relationships from these sources.

The Metathesaurus is organized by concept or meaning. The main knowledge the Metathesaurus represents, consists in linking the various names and views of the same concept and in identifying relationships between different concepts in different source terminologies. For each Metathesaurus concept a unique identifier (CUI) is assigned. Metathesaurus concept structure includes concept names, the identifiers of concept names and other useful characteristics of them. In the next figure, we show an example of representing the concept *Addison's disease* (CUI= C0001403) in the UMLS Metathesaurus.



Addison Disease	MeSH	D000224
Primary hypoadrenalism	MedDRA	10036696
Primary adrenocortical insufficiency	ICD-10	E27.1
Addison's disease (disorder)	SNOMED CT	363732003

C0001403

Addison's disease

Figure 2: The representation of the concept Addison's disease in the UMLS Metathesaurus

3.1.2.2 Semantic Network

Semantic Network offers a representation of the semantic types, semantic relation types and the relationships between them. Two examples of entries in Semantic Network are: "*Sign or Symptom*" which represents a semantic type and "*diagnoses*" which represents a semantic relation type.
The relationships between semantic types can be hierarchical or non-hierarchical and the ones between semantic relation types can be only hierarchical. The hierarchy of types (either semantic types or semantic relation types) is established by using the primary link "*isa*". The semantic types can be seen as high level categories, which provide a consistent categorization of all concepts present in UMLS Metathesaurus. They are represented in a tree structure and consist of two major hierarchies:

- Entity
- Event

The semantic relation types can be seen as a set of normalized properties and they are used for expressing the non-hierarchical relationships between the semantic types. For example, the relationship represented as: "*Sign or Symptom **diagnoses** Pathologic Function"* infers that the semantic relation type ***diagnoses*** holds between the semantic types *Sign or Symptom* and *Pathologic Function*.

3.1.3 Linking Metathesaurus to Semantic Network

In the picture below we introduce a schematic representation of Metathesaurus and Semantic Network. Each Metathesaurus concept is categorized to at least one semantic type from Semantic Network (blue continuous lines in the next figure).

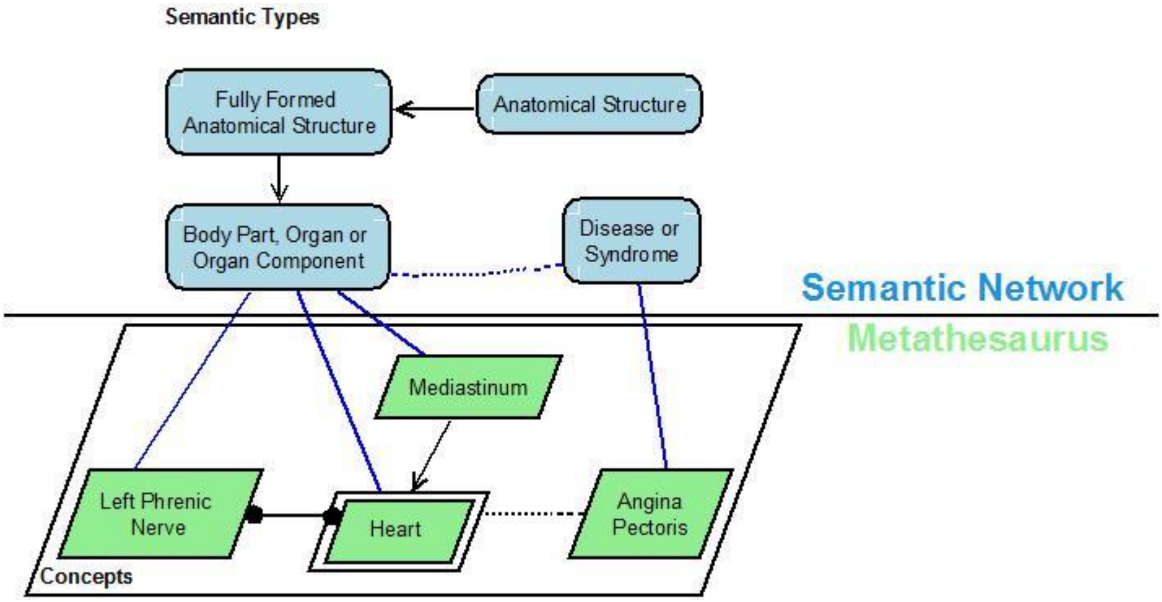


Figure 3: Schematic representation of Metathesaurus and Semantic Network

We refer to the non-hierarchical relationships between two semantic types in Semantic Network as *relation patterns* (blue non-continuous lines). These patterns are possible links between concepts (black non-continuous lines) that have been assigned to those semantic types. This means that the relationship between concepts, which represents an instantiation of the respective relation pattern towards the Metathesaurus concepts, may or may not hold.

It is needed to distinguish between the different types of properties used in the different types of relations UMLS provides. For the relation instances[1], the properties relating the concepts are either defined locally in the terminology or they are defined by the UMLS. We refer to the properties defined in the source terminology as source properties and to the properties defined in UMLS as generalized properties because they are a generalization of the source properties. For the relation patterns, we refer to their properties as normalized properties because they can be seen as a normalization of the source properties.

Examples of generalized properties are:

- PAR_OF which stands for "parent of"
- SY which stands for "source-asserted synonymy"

Examples of source properties are:

- has_ingredient
- may_treat

Examples of normalized properties are:

- diagnoses
- treats

[1] Relations between the Metathesaurus concepts which can be either hierarchical or non-hierarchical represented respectively by continuous directed black lines or continuous non-directed black lines in Figure 4

The source properties in Metathesaurus can be aligned to the normalized properties in Semantic Network. For example, the source properties “*treats*” and “*may treat*” can be aligned to the normalized property “*treats*”. This knowledge is not yet provided by UMLS.

3.2 Choice of UMLS terminologies subset

For this deliverable, the annotator we provide uses a UMLS subset (from release 2012AB) which is compliant with the current uses of medical partners. The source terminologies we selected are:
SNOMED-CT
LOINC
ICD9-CM
NCI

Note that NCI contains almost all the data enclosed in CDISC (that has been pointed out to be an important vocabulary source by the project partners) and has the advantage of being part of UMLS category 0 vocabularies. NCI also contains CTCAE (version 3) which was published both separately and integrated into NCI. MEDRA is not part of this subset being not a licence-free vocabulary. As for SNOMED, we acquired a licence since this resource, according to informal conversation with clinical partners, seems to be extremely relevant and commonly used. Radlex was also not included in this first deliverable.

One important concern is the fact that medical terminologies/ontologies are **not** specialized medical lexicons. According to our naming conventions in introduction, strings representing concepts in the ontology/terminology do not have necessarily a real lexical correspondence (see [Hirst 03]). As a result, an important adaptation work is necessary to select from the terminology potential lexical elements. One step further consists in transforming some terms with the objective of making them lexical item candidates in order to enrich the lexicons that will enable term annotation from free text.

The UMLS subset generated from the source terminologies listed above includes more than 0.5 million of UMLS concepts and more than 1.6 million of terms.

3.3 Adapting existing terminologies for building medical lexical resources

The methodology we adopted to adapt existing medical terminologies to medical lexicons that can be used by Natural Language Processing (NLP) tools is the following.

1- Selection of the terminologies/ontologies we want to integrate
2- Populating a KB whose ontological schema corresponds to the information structure provided by UMLS Metathesaurus and Semantic Network
3- Apply rules for the deletion/modification/suppression of terms (strings representing concepts) contained in the terminological resources

Once these three operations are performed, the resulting list of terms can then be integrated into lexical tools that are used in NLP processors.

3.3.1 Selection of terminologies

This selection is done in a declarative way using UMLS dump files and filtering the sources we want to keep.

Metathesaurus data files are provided in two different formats:

- Rich Release Format (RRF)
- Original Release Format (ORF)

The RRF format provides a better representation of detailed semantics of each source terminology. For this reason, we chose this format to work with.
More than 20 files represent the data in each Metathesaurus entry. They are organized in five different groups:

- Concepts, Concept Names, and their sources
- Attributes
- Relationships
- Data about the Metathesaurus
- Indexes

The main files we worked with fall in the first three groups. Their content is explained as following:
MRCONSO.RRF: For each unique term in a given source terminology, there is exactly one entry in this file.
MRSAB.RRF: It stores the information about the fully specified version for the current release of the source terminology.
MRSTY.RRF: This file contains one row for each assignment of a semantic type to a Metathesaurus concept.
MRREL.RRF: This is the main file of the relations between different concepts of Metathesaurus. One row of this file represents only one direction of the relation. The relations should be read from right to left, which means that the second concept of the row is in a relation with the first concept of the row.
MRDOC.RRF: It keeps track of the allowed values of selected data elements or attributes that have a finite number of abbreviations as allowed values.

Semantic Network data files are provided in two different formats:

- Unit record format
- Relational table format

We chose to work with the Relational Table format because it is easier to be parsed.
The basic files are:
SRDEF: Contains the definitions of all semantic types and semantic relation types
SRSTR: Contains the information about the structure of the Semantic Network

The filtering can be done either by specifying the source name and/or by specifying the source category. As said before, for this deliverable we provide an annotator based on the following vocabularies: SNOMED-CT, LOINC, ICD9-CM and NCI.

3.3.2 Populating a KB with UMLS information

One of the goals of the term annotation tool, is to be able to provide together with the terms recognition in the text, any kind of information that is enclosed in the UMLS Metathesaurus and Semantic Network. For the different use-cases and scenarios, users and application needs will help to specify which annotations they want to see together with the extracted term. In order to explicit the kind of information that can be associated with the terms, Figure 5 shows the ontological model of UMLS data that we defined. This ontological model keeps the full main useful information present in both Metathesaurus and Semantic Network.

As stated in section 3.1.2.1, Metathesaurus is organized by concept. We represent it by the Concept Element node in our ontology. The Semantic Network is represented by a categorization of Metathesaurus concepts into semantic types and by relationships between these concepts. Semantic types are considered as concepts in our model. The structure of the Metathesaurus concepts is well preserved by storing the information related to terms, strings and atoms associated to each concept.
The properties present in the UMLS constitute an important knowledge in terms of relationships between Metathesaurus concepts. The hierarchy of properties that we are introducing in our model consists of the source properties, normalized properties and generalized properties which are explained in section 3.1.3. The source properties in Metathesaurus are mapped to the generalized properties according to the knowledge present in the relationships between concepts (we refer to this mapping by using the property mapsGeneralizedProperty). Our ontology model reflects also the mapping between the source properties and the normalized properties (we refer to this alignment by using the property mapsNormalizedProperty) in order to support the inclusion of this knowledge once it will be provided by UMLS (it is one of the UMLS future works).

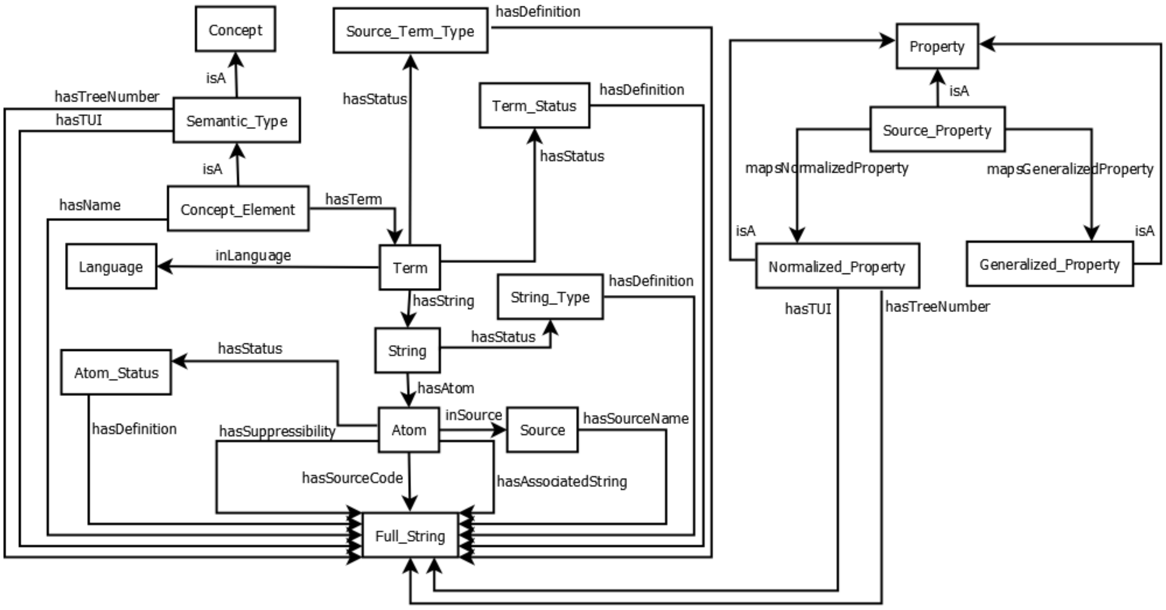


Figure 4: Ontological information in UMLS

3.3.3 From terminological resources to lexical resources

The third step for the building of a medical lexicon from terminological resources is a set of cleaning operations that will:

1- Remove terms that will not appear as such in the texts
2- Perform some transformations to existing terms creating new terms that can potentially appear in texts
3- Add new terms relying on the resource

These transformations have been tested and experienced before in previous work.
[Hettne 10] and [Wu 12] have shown that term filtering operations are necessary for obtaining an adequate medical lexicon. More precisely, [Hettne 10] conducted experiments for the building of a medical lexicon using UMLS Metathesaurus. Term suppression techniques are applied for filtering out terms which are considered as irrelevant lexical items. Complementary, rewriting techniques are applied for generating new medical terms which are not present originally in the resource. These new terms can be the basis of building a more consolidated medical lexicon and the aim is to use them for medical concept annotation.
In [Wu 12] the UMLS Metathesaurus terms characteristics are exploited for discovering which of them are generalizable across data sources. The statistics obtained from the analysis of a training corpus from Mayo clinic were used to generate filters to be applied on a testing corpus from i2b2/VA. The original Metathesaurus lexicon used for this analysis was reduced significantly.

We implemented term transformations that have been mentioned and successfully evaluated in the state-of-the-art on our selected set of terms[2]. Here follow the details of these transformations.

3.3.3.1 Term deletion rules

The following term deletion rules (in bold-italic) have been applied on the dataset.

Short token removes a term if after tokenization and removal of stop words the whole term consists of single characters, or arabic or roman numbers.
For example: the term “1011” will be removed.

Dosages removes all the terms that contain a dosage in percent, gram, microgram or milliliter.
For example: the term “Trimipramine 10mg tablet” will be removed.

At-sign removes the terms that contain the @-character.
For example: the term “SMA@ gene” will be removed.

EC (enzyme classification number) removes the terms that contain enzyme classification numbers.

[2] More details about this work can be found in: “A Framework to generate Sets of Terms from Large Scale Medical Vocabularies for Natural Language Processing” – to be published in Computational Semantics in Clinical Text (CSCT 2013) Workshop – March, 19th 2013, Potsdam, Germany

For example: the term “EC 2.7.7.7” will be removed because it corresponds to an enzyme classification number.

Any classification removes the terms containing the following properties: “NEC” at the end of a term and preceded by a comma, “NEC” within parentheses or brackets at the end of a term and preceded by a space, “not elsewhere classified”, “unclassified”, “without mention”.
For example: the term “Unclassified Hepatocellular Adenoma” will be removed.

Any underspecification removes the terms containing the following properties: “not otherwise specified”, “not specified”, or “unspecified”; “NOS” at the end of a term and preceded by a comma, or “NOS” within parentheses or brackets at the end of a term and preceded by a space.
For example: the term “Unspecified lymphadenopathy” will be removed.

Miscellaneous removes the terms containing the following properties: “other” at the beginning of a term and followed by a space character or at the end of a term and preceded by a space character; “deprecated”, “unknown”, “obsolete”, “miscellaneous”, or “no” at the beginning of a term and followed by a space character.
For example: the term “Other emphysema” will be removed.

Special characters removes the terms if they begin with “[” end with “)” .
For example: the term “[M]Brenner tumors (morphologic abnormality)” will be removed.

Maximum number of words removes the terms with 7 or more words.

Maximum number of characters removes the terms with more than 55 characters

The following table shows the impact of the term deletion rules on the UMLS subset we selected.

Rule	# deleted terms
Short token	197
Dosages	37042
At-sign	65
EC (enzyme classification number)	221
Any classification	2312
Any underspecification	2404
Miscellaneous	19888
Special characters	15955
Maximum number of words	327836
Maximum number of characters	292288

Table 1: Impact of term deletion rules

3.3.3.2 Term modification rules

These rules modify a term from the resource that is unlikely to correspond to a string in a text and give a new string corresponding to the same concept. Unlike the original string, the new generated string may appear as a term in a text.
Here follows the list of the implemented term modification rules:

Angular brackets removes expressions within angular brackets appearing anywhere in a term.
For example: the term "Hemoglobin A<sub>2</sub> Sphakia" will be modified into "Hemoglobin A2 Sphakia"[3].
Semantic type removes the expressions within parentheses that match the list of semantic types in the UMLS.
For example: the term "Genus Scleroderma (fungus) (organism)" will be modified into "Genus Scleroderma".

The following table summarizes the impact of the term modification rules on our UMLS subset.

Rule	# matches	# resulting terms
Angular brackets	3915	3895
Semantic type	200	182

Table 2: Impact of term modification rules

3.3.3.3 Term addition rules

These rules create new terms out of existing terms. Here follows the list of these rules

Syntactic inversion adds the syntactic inversion of a term if the term contains a comma followed by a space and does not contain a preposition or conjunction (the pattern of a comma followed by a space should appear only once)
For example: the term "HIV infection, pediatric" will produce the new term "pediatric HIV infection".

Possessives removes the possessive "'s" at the end of a word and adds the new term. For example: the term "Alzheimer's Disease" will produce the new term "Alzheimer Disease"

Short form/long form adds the short form (i.e. the abbreviation) and the long form (normal form) of the term.

[3] With the UMLS subset used for this deliverable the application of "Angular brackets" rule resulted mainly in the elimination of superscripts and subscripts.

For example: the term “Idiopathic normal pressure hydrocephalus (INPH)” will produce the new terms: “Idiopathic normal pressure hydrocephalus” (which is the long form of the original term) and “INPH” (which is the short form of the original term).

The following table shows the impact of term addition rules on our subset.

Rule	# matches	# new terms
Syntax inversion	54230	54164
Possessives	9141	9135
Short/long form	3662	3693

Table 3: Impact of term modification rules

3.3.3.4 Other term deletion rules

In addition to the term transformation rules that have been mentioned and tested in the literature, we extend the deletion rules in order to exclude terms from given UMLS semantic types that are already enclosed in general purpose lexicons and are of general interest.

This is the case of all terms that are covered by general purpose Named Entity Recognition systems usually included in NLP analysis tools. More precisely, we suppress all terms belonging to the following UMLS semantic types:

"Temporal_Concept, "Idea_or_Concept", "Quantitative_Concept", "Spatial_Concept", "Manufactured_Object", "Conceptual_Entity", "Geographic_Area, "Language", "Organization", "Social_Behavior, "Functional_Concept, “Intellectual_Product” and "Classification".

In order to transform this set of terms into a medical lexicon, some linguistic information has to be associated to this term list (see definition of lexicon in the introduction).

By default, we associate the noun part-of-speech value to these items. Section 4.1.1 explain how this by-default value can be modified.

The medical lexicon we finally obtain is exported into a specific text format which is then compiled into finite-state transducers (FST) (see [Beesley 03]). For each lexical element we also associate information about corresponding UMLS semantic type, name of the source terminology and the CUI.

Here follows an example of the textual format of a medical lexical entry that is compiled into FST.

Atrazine +Noun +UMLSC +Hazardous_or_Poisonous_Substance
+Laboratory_or_Test_Result +Pharmacologic_Substance +Organic_Chemical
atrazine

The first part corresponds to the lemma of the medical lexical element, followed by the part-of-speech category (noun here) followed by an information stating that this entry comes from UMLS and followed by four features giving all the possible UMLS semantic types that are associated to this entry.

4 Integrating medical lexical resources into NLP tools

The medical FST can then be merged with the general purpose lexicon (also in FST format) and the final net is then integrated into NLP analyser in order to perform the annotation. The merging of the two lexical sources (medical and general lexicons) forces to make some adaptations that we describe here.

4.1 Adaptations for the merging of medical lexical resources with general purpose lexical resources

4.1.1 Recovering correct part-of-speech for medical terms

Lexical elements contain linguistic information (minimally a part-of-speech) which is necessary for using this lexicon in more refined NLP tasks (such as part-of-speech tagging, chunking or syntactic/semantic analysis)
The annotator we provide should be able to be integrated into a general purpose linguistic analysis tool in order to perform linguistic analysis using medical information (deliverable D3.5). For a refined linguistic analysis, linguistic tools usually rely (among other) in part-of-speech information which is assigned to the lexical items. By default, we consider that the new medical lexical entries extracted from UMLS terminologies have the part-of-speech noun. It's true that the vast majority of the terms correspond to nouns. However, some medical lexical entries we extracted have other part-of-speech values. This is for example the case of “abdominal” which is enclosed in UMLS (via NCI resource) with CUI C0000726. For further linguistic processing, we want this lexical entry (which is also a term) to have “adjective” as part-of-speech. As a result, a textual string like “abdominal surgery” appearing in a free text description will be annotated on one side by the annotator as the concatenation of two terms (abdominal + surgery), and on the other side, the linguistic analysis will provide the information that the adjective “abdominal” is the modifier of the noun “surgery” which is the basis to provide a semantic representation for the whole expression “abdominal surgery”.
Because of the size of the medical lexicon, relying on the fact that most of lexical elements are nouns and having in mind that the integration of new medical lexical resources has to be an easy process, we decided not to perform a manual review of each lexical entry and adopt instead some heuristics to recover the correct part-of-speech of medical lexical items. These heuristics take advantage of the linguistic properties of the lexical items and also of the general purpose pre-existing lexicon.

4.1.1.1 Recovering adjectives

If a medical lexical item is ambiguous with an element of general lexicon which has adjective part-of-speech, and if this lexical item terminates by “al”, then we consider this lexical item as an adjective only and keep the UMLS terminological information associated to it.
Example: “abdominal”, “medical”, “surgical”, “external” etc.

If a medical lexical item is ambiguous with an element of the general lexicon which has adjective part-of-speech and if this lexical item terminates by “an”, then we consider this lexical item as an adjective only and keep the terminological information associated to it.
Example: “ovarian”, “median”, “human” etc.

If a medical lexical item is ambiguous with an element of general lexicon which has adjective part-of-speech, and if this lexical item terminates by “ant”, then we consider this lexical item as an adjective only and keep the terminological information associated to it.
Example: “pregnant”, “significant”, “odorant”, etc.

4.1.1.2 Recovering verbal progressive forms

If a medical lexical item is ambiguous with a verbal progressive form from the general lexicon, we give to this medical lexical item the corresponding part-of-speech (verb, noun or adjective) that will be disambiguated in further processing.
Example: “poisoning”, “participating”, etc.

4.1.1.3 Recovering past-participles

If a medical lexical item is ambiguous with a verbal past-participle form from the general lexicon, we give to this medical lexical item the corresponding part-of-speech (past-participle or verb)
Examples: “delayed”, “wanted”, “estimated”

4.1.2 Dealing with real ambiguity between medical and non-medical lexical items

The combination of medical lexical elements with general lexical elements creates new classes of lexical ambiguity that were not present in the general lexicon alone.
As a general approach, once the recovering of part-of-speech (see 4.1.1) has been performed, we decided for the other ambiguous lexical items to favour the medical lecture **if both lexical items have the same part-of-speech from open linguistic categories** (for instance, noun).
In case of real ambiguity (different part-of-speech tags) for the medical and non-medical lexical item, we decided to favour the non-medical lecture **if the general lexical item belongs to a linguistic close category** (e.g. preposition, sub-classes of adverbs, punctuation marks, conjunctions).

We have not at this stage performed a formal evaluation of both the coverage of the part-of-speech recovering (not having an annotated gold standard to do it) and the correctness of this a-priori disambiguation. One possibility would be to use the Specialist lexicon provided by UMLS (see [UmlsRef 09]) to compare the outputs provided by our annotator with the output that a lookup with this lexicon will provide. However, as the Specialist lexicon only covers a very small subset of UMLS data, we will not have the same lexical coverage that the one provided by our annotator.

We intend to make a complete evaluation of disambiguation after the implementation of the disambiguation of concepts that we have to perform for more use-case dedicated adaptations of this general purpose annotator.

4.2 Free text annotation

As patient recruitment for clinical trials has been considered as an important use-case and as we do not have access to patient data, the first experiments we performed for free text annotation were conducted on clinical trials and more precisely in the analysis of eligibility criteria.

Here follows an example taken from the eligibility criteria section of a clinical trials coming from www.clinicaltrials.gov.

INPUT TEXT

Inclusion Criteria:

- Patients with a histologically or cytologically proven metastatic breast cancer.
- Patients with at least one bidimensionally measurable lesion (diameter > 1 cm), or an evaluable bone lesion that will not undergo biopsy.
- Age > 18 years.
- Life expectancy of at least 6 months.
- ECOG performance status 0-3.

…

ANNOTATION

```
Inclusion Criteria  8      C1512693     NCI:C25532    TYPES:Qualitative_Concept

metastatic breast cancer   97     C0278488      NCI:C3995
       TYPES:Neoplastic_Process

Patients      42     C0030705      NCI:C16960    TYPES:Patient_or_Disabled_Group

biopsy 280    C0005558      SNOMEDCT:86273004    NCI:C15189
       LNC:MTHU028106|LP68311-7|LP20669-5      TYPES:Diagnostic_Procedure

evaluable     236    C1516986      NCI:C8503     TYPES:Disease_or_Syndrome

bone lesion   246    C0238792      NCI:C43260    TYPES:Disease_or_Syndrome

lesion 191    C0221198      SNOMEDCT:52988006|49755003 NCI:C3824     TYPES:Finding

Patients      137    C0030705      NCI:C16960    TYPES:Patient_or_Disabled_Group

Age    302    C0001779      SNOMEDCT:424144002|397659008|397669002|102518004
       NCI:C25150    LNC:MTHU010047|LP28815-6    TYPES:Organism_Attribute
Age    302    C1114365      LNC:30525-0   TYPES:Clinical_Attribute

Life expectancy     332    C0023671      LNC:MTHU021387|LP75025-4
       TYPES:Group_Attribute

ECOG performance status    384    C1520224      SNOMEDCT:423740007   NCI:C25400
       TYPES:Clinical_Attribute
```

The first field of the tabular output corresponds to the term that was recognized in the input text. The second field corresponds to the character position (i.e. offset) of the term in the input text. Here the string “Inclusion Criteria” that has been recognized as a term starts at character 8 relatively to the beginning of the document. The third field corresponds to the UMLS CUI (concept unique identifier) of the term. The following fields indicate in which specific terminology(ies) the term appears. The terminology name comes first followed by the corresponding identifier code(s) in this terminology. For instance “ECOG performance status” appears both in SNOMED and NCI. In SNOMED the identifier of this term is “423740007” while in NCI the identifier is “C25400”. Note that a term can have more than one code in its source terminology although the term denotes a single UMLS concept: in such cases all the source codes are provided (separated with the vertical bar character “|”). Finally, the last field gives the list of UMLS semantic types for the extracted term. For instance, the term “metastatic breast cancer” appearing in NCI has a UMLS semantic type, which is “Neoplastic_Process”. Note that ambiguity for term annotation is preserved by the annotator. For instance “Age” corresponds to two different UMLS concepts (C0001779 and C1114365) from semantic types “Organism_Attribute” and “Clinical_Attribute” respectively. Disambiguation will occur in a later stage, considering the application of the annotation tool in specific scenarios.

5 Conclusions and next steps

We developed in the context of this deliverable a term annotation tool that recognizes and annotates the occurrences of medical terms in free texts based on a subset of terminological resources included in UMLS Metathesaurus (namely SNOMED-CT, LOINC, ICD9-CM and NCI). The construction of the annotator has been guided by two main concerns:

- The term annotation tool must be easy to adapt to new terminological resources (for different scenarios some terminologies or part of terminologies are more adapted than others) and it has to be flexible enough to integrate the semantic core dataset that is currently under definition.
- The term annotation tool will be part of a larger framework of natural language processing of medical free text. As a result, we developed some methodology to bridge the existing gap between medical terms and medical lexical entries (implementing term modification rules) and we applied some general heuristics to facilitate the integration of medical lexicon into a general purpose NLP lexicon.

We intend to specifically adapt this annotation tool for the following scenarios: Reporting episodes of febrile neutropenia, cancer registry reporting, update of guidelines and trial recruitment. The adaptation to these scenarios will guide the next steps of the work which are:

- Semantic disambiguation of terms
- Integration of semantic core dataset
- Annotation of relations between terms

Finally, an important aspect of this deliverable is that the first annotation tool deals with English terminologies (producing thus an annotator for English medical language). Different project clinical partners have however needs for other languages (French, Dutch and German). The architecture we adopted is flexible enough to be adapted for other languages once we have terminological resources for these different languages transformed in a specific format. Possible multilingual adaptations are also part of the next steps.

6 References

[Beesley 03] Beesley K. R, Karttunen L. (2003) *Finite State Morphology*, Stanford CA. CSLI Publications.

[Bodenreider 07] Bodenreider, O. (2007). *The Unified Medical Language System (UMLS) and the Semantic Web*. ”http://www.nettab.org/2007/slides/Tutorial_Bodenreider.pdf”.

[Browne 00] Browne, A. C., A. T. McCray, and S. Srinivasan (2000). *The SPECIALIST LEXICON*. Lister Hill National Center for Biomedical Communications, National Library of Medicine.

[Demner-Fushman 10] Demner-Fushman, D., J. G. Mork, S. E. Shooshan, and A. R. Aronson (2010). *UMLS content views appropriate for NLP processing of the biomedical literature vs. clinical text.* Journal of Biomedical Informatics 43(4), 587–594.

[Harkema 04] Harkema, H., R. Gaizauskas, M. Hepple, A. Roberts, I. Roberts, N. Davis, and Y. Guo (2004). *A large scale terminology resource for biomedical text processing*. In Proceedings of the NAACL/HLT 2004 Workshop on Linking Biological Literature, Ontologies and Databases: Tools for Users, Boston

[Hettne 10] Hettne, K., E. van Mulligen, M. Schuemie, B. Schijvenaars, and J. Kors (2010). *Rewriting and suppressing UMLS terms for improved biomedical term identification.* Journal of Biomedical Semantics 1(1), 5.

[Hirst 03] Hirst G (2003). *Ontology and the Lexicon*. In Handbook on Ontologies in Information Systems, Springer. 209—230

[Savova 10] Savova, G. K., J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute (2010). *Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications*. Journal of the American Medical Informatics Association 17(5), 507–513.

[Schwartz 03] Schwartz, A. S. and M. A. Hearst (2003). *A simple algorithm for identifying abbreviation definitions in biomedical text*. In Proceedings of the Pacific Symposium on Biocomputing, pp. 451–462.

[Wu 12] Wu, S. T.-I., H. Liu, D. Li, C. Tao, M. A. Musen, C. G. Chute, and N. H. Shah (2012). *Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis*. Journal of the American Medical Informatics Association 19(e1), e149–e156.

[Xu 10] Xu R, Musen M. A. Shah N. H. (2010) *A Comprehensive Analysis of Five Million UMLS Metathesaurus Terms Using Eighteen Million MEDLINE Citations* AMIA Annual Symposium Proceedings. 2010. 907—911.

[UmlsRef 09] UMLS® Reference Manual [Internet]. Bethesda (MD): National Library of Medicine (US); 2009 Sep-. *6, SPECIALIST Lexicon and Lexical Tools*. Available from: http://www.ncbi.nlm.nih.gov/books/NBK9680/