ICT-2011-288048

EURECA

Enabling information re-Use by linking clinical Research and CAre

IP
Contract Nr: 288048

Deliverable: D4.1 Requirements analysis and selection of the initial clinical scenarios for core datasets

Due date of deliverable: (09-01-2012)
Actual submission date: (11-15-2012)

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

Responsible WP: UPM

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 Author

Author	Company	E-mail
Cyril Krykwinski	IJB	cyril.krykwinski@bordet.be
Raúl Alonso Calvo	UPM	ralonso@infomed.dia.fi.upm.es
David Pérez Rey	UPM	dperez@infomed.dia.fi.upm.es
Scott Marshall	MAASTRO	scott.marshall@maastro.nl
Norbert Graf	UdS	Norbert.Graf@uniklinikum-saarland.de
Laura Hollink	VUA	l.hollink@vu.nl
Anca Bucur	Philips	anca.bucur@philips.com

0.2 Documents history

Document version #	Date	Change
V0.1	18.06.2012	Starting version, template
V0.2	18.06.2012	Definition of ToC
V0.3	20.09.2012	First complete draft
V0.4	19.10.2012	Integrated version (send to WP members)
V0.5	29.10.2012	Updated version (send PCP)
V0.6	05.11.2012	Updated version (send to project internal reviewers)
Sign off	15.11.2012	Signed off version (for approval to PMT members)
V1.0		Approved Version to be submitted to EU

0.3 Document data

Keywords	
Editor Address data	Name: Cyril Krykwinski Partner: IJB Address: 121 Bd de Waterloo, 1000 Bruxelles Phone: +32 (0)2.541.36.37 Fax: E-mail: cyril.krykwinski@bordet.be
Delivery date	

0.4 Distribution list

Date	Issue	E-mailer

Table of Contents

1 Introduction

1.1 Aim of the project

One of EURECA's main objectives is to provide a semantic interoperability solution to enable seamless, secure, scalable and consistent linkage of healthcare information residing in Electronic Health Record (EHR) systems with information in clinical research information systems, such as clinical trial systems. Its final goal is to enable these systems to efficiently exchange information with meaning in order to support a range of relevant clinical applications that require such a shared meaning (e.g. patient recruitment from healthcare patient records, pharmacovigilance, long-term follow-up).

This semantic interoperability layer relies on a standards-based semantic core dataset that should enable the EURECA environment to capture the meaning of the data with standard terminologies/ontologies and manage the numerous concepts present in the vast amounts of patient data. The semantic core dataset is established by reducing the number of relevant clinical concepts and selecting only those that are relevant to the final goals of the project as defined by selected clinical scenarios. The core dataset will also reduce the number of mappings that need to be built between the terminologies used by the different systems that will be linked to the EURECA interoperability environment. In this way, EURECA will provide a scoped vocabulary basis to future EURECA activities working on extraction tools, like data mining tools or Natural Language Processing tools.

1.2 Structure of the deliverable

To fulfil that purpose, we collected the list of medical terminologies (vocabularies and ontologies) that are already used onsite by clinical partners for their clinical care and clinical research activities, and the terminologies that might be of interest to the whole project in terms of semantic extraction to support these activities. We defined the second set of terminologies according to the initially selected clinically-driven scenarios, so that we could define our need in terms of general domains that should be covered as far as possible.

The whole set of medical terminologies that could be of use within the project is very heterogeneous by definition, as there is no terminology which alone covers such a wide range of clinical domains and concepts. One could override this heterogeneity and the resulting disconnection of concepts by applying some mappings among the data vocabularies of the sources. Some initiatives already exist but are limited to specific terminologies and not always satisfying. That's the reason that the existence of cross mappings could be a quality criterion for the selection of a terminology among all the candidates for the definition of the core dataset.

The selection of terminologies is not accurate enough to correctly define the core dataset. For that matter, subsets of terminologies' domains have to be considered in the context of the goals of the project and the granularity of the scenarios, as well as on the data that will be provided by clinical partners.

This document analyses the scenarios and requirements to define the standards-based semantic core dataset necessary for a semantic interoperability layer.

2 The Semantic Interoperability Approach and the need for a Core Dataset

The EURECA approach with respect to achieving semantic interoperability is described in detail in [1].

Using the SemanticHEALTH report [2] classification of semantic interoperability, we can observe that the current level of semantic interoperability between clinical trials and EHRs is somewhere between level 0 i.e. no interoperability at all, and level 1 i.e. syntactic interoperability. The reason for this is simply the fact that these systems were designed in isolation, not foreseeing the benefits of mutual data exchange and understanding as laid out in section above. Such uncoordinated and isolated development typically results in 'information silos', which are very costly to integrate for applications unanticipated by the original designs. In order to achieve the aforementioned benefits, we have to increase the semantic interoperability level to at least 2b - bidirectional semantic interoperability of meaningful fragments, or even level 3 which requires full semantic interoperability, sharable context. It is, however, also recognized that due to the steep investments needed, the highest level of semantic interoperability should only be sought in specific areas with high potential for improvements.
The essential steps for achieving this semantic interoperability improvement include the definition of sound information models describing the clinical trial systems, building on existing research results when possible [3]. Electronic health records, too, need to be properly modelled; to that end we will adopt the appropriate state-of-the-art representation formalisms such as HL7 CDA, the openEHR Reference Model, and ISO/EN 13606.

Semantic Core Dataset

The foundation of the semantic interoperability layer will be the semantic core dataset comprising well-defined and agreed upon clinical structures consisting of standards-based concepts, their relationships, and quantification (e.g. archetypes using selected terminology concepts) that together sufficiently describe the semantics of the chosen clinical domain.

The semantics of the clinical terms should be captured by standard terminology systems such as SNOMED-CT, ICD, and LOINC. The scalability of the solution needs to be achieved by modularization and scoping, e.g. instead of aiming at inclusion of the complete SNOMED terminology (more than 300 thousand concepts) we identify a core subset that covers the chosen clinical domain. The main rationale here is that only a confined subset of relevant concepts from the clinical ontology will be needed for data extraction and reasoning in a given clinical context/domain while most of the remaining concepts would never be used by reasoning algorithms.

Such a core dataset shall be validated both by clinical and knowledge engineering experts to assure proper coverage and soundness. In the process of identifying the core data set and the corresponding mapping tools, care will be taken to allow for easy extension of the core data set, should the inclusion of new concepts become necessary (e.g. a cross-domain linkage). Relying on well-established and widely used existing terminology standards will facilitate extensible semantic interoperability towards third parties outside of the scope of the project. This approach is in line with the roadmap of

SemanticHEALTH which lists identifying of sound semantic subsets of SNOMED covering a certain clinical domain as one of their priorities [2].

The core semantic data set will be validated in concrete use cases, for the different EHR and clinical trial systems available at the clinical care and clinical trial sites within the consortium. The semantically-aware access to both EHR and Clinical trial data is a machine processable manner. Concepts in the dataset will have their unique identifiers, well understood meaning, as well as a set of synonyms they can be referred as.

Considering the problem of language heterogeneity between the clinical trials and EHRs as primary data capture, we plan to address this issue by offering a gradual approach, semi-automatically translating only those parts of the clinical ontology identified as the core semantic dataset, leveraging existing translations of known terminologies such as SNOMED-CT. When no translation of the relevant standard terminologies exist in that language, we will work out together with the clinical experts a translation of the core dataset into the languages that are used for the primary data capture. Hence, translating (only) the selected semantic core dataset and not the entire clinical coding system enables a modular and scalable approach where the initial translation effort is limited in scope and delivers immediate benefits in increased semantic interoperability.

The identification of a core dataset that sufficiently describes a domain of interest was also a topic for the FP7 INTEGRATE project[1]. However, there we have focused exclusively on the domain of clinical trials in breast cancer. In EURECA the context will be much wider covering a significantly larger number of care- and research-focused scenarios (as described in *Section 3*), several additional domains in oncology that are relevant for our clinical partners, and a much higher language and systems/sources heterogeneity. This challenging context will enable us to validate the feasibility of our approach to achieving semantic interoperability.

[1] INTEGRATE Project, “D3.2 Initial Proposal for the Core Dataset,” available at http://www.fp7-integrate.eu/index.php/downloads

3 User needs, clinical scenarios and data sources based requirements

3.1 Initial clinical scenarios for core datasets

3.1.1 Initial scenarios

Scenarios within the EURECA project were mainly defined by clinical partners (IJB, UdS, UOXF, BIG, Maastro, and GBG) within WP1 and have been presented in deliverable D1.1[2]. It ensures a clinically-driven process so that the subsets of concepts should remain consistent with the general clinical needs that have been collected through the questionnaire on user needs.

Scenario 1 – Information

(UC - Personal medical information recommender)

The goal of this scenario is to provide for patients objective information about treatments and trials about their specific disease.

(UC - Data mining of consultations)

The goal of this scenario is to generate an automatic answer to recurrent questions asked by patients during consultations.

(UC - Similarities of datasets to combine)

The goal of this scenario is to detect and identify similar datasets from different patients.

Scenario 2 – Investigation

Update of guidelines

The goal of this scenario is to update guidelines regularly from data mining of clinical trials and literature.

Protocol & Research investigation

(UC - Opt-out solution for further research)

The goal of this scenario is to provide a platform where patients can select which research they do not like to do with their data or biomaterial.

(UC - Protocol feasibility)

The goal of this scenario is to determine whether a new clinical trial is feasible to start according to the estimation of recruitment potential.

[2] EURECA project, "D1.1 User needs and specifications for the EURECA environment and software services," 2012

The two data sources that need to be linked here are the available patient data on one side (accessible through the EHR system) and the criteria of clinical trials on the other side (captured in the trial descriptions). For the eligibility criteria there is a large public source of data, ClinicalTrials.gov that can be used as input for the selection of the relevant core datasets. We start by selecting and analysing trials in our clinical domains of interest aiming at a modular development of the core dataset with the possibility to easily extend to other clinical domains. The concepts describing the trial criteria need to be linked to actual patient data to evaluate whether a patient matches a trial. To extract the core dataset capturing the content of the patient file, access to sufficient clinical data is needed. Ontologies/terminologies chosen to describe the semantics of the criteria and patient data in a standard form are SNOMED-CT, MedDRA, and LOINC. Other ontologies can be relevant as well, but the goal is to keep the set of concepts usable and maintainable.

(UC - Supporting design of new trials and hypothesis generation)

The goal of this scenario is to support the process of designing new clinical trials.
This scenario is similar with the “protocol feasibility” in terms of sources of data used for determining the core dataset, but it adds a new source: data collected in previous trials. Therefore, next to access to criteria of trials and clinical data here we need to access data collected in the specific context of clinical trials. As clinical research does capture data that is not collected in standard clinical care (e.g. various molecular data), the core dataset needs to be extended beyond the modules developed to capture the semantic content of EHR data to these new datasets.

Scenario 3 – Selection & Recruitment

Choice of treatment

(UC - Outcome prediction)

The goal of this scenario is to learn and validate outcome prediction models from routine patient care data.

Patient recruitment into a trial
In this scenario a clinician needs to select the most suitable trial for a particular patient.
This scenario aims also to support a researcher to identify patients that are eligible for a clinical trial. From the perspective of identifying the core dataset, this scenario has the same requirements as Scenario 1.

Scenario 4 – Reporting

The goal of this scenario is to detect and predict SAEs and SUSARs before a treatment is given to a patient, and then to automatically report SAEs and SUSARs to regulatory authorities.

(UC - Pre-filling of eCRF)

The goal of this scenario is to reuse the clinical data into the trial eCRF systems.

Scenario 5 – Long-term follow-up

The goal of this scenario is to extract the last follow up date and patient status from EHR or the national registry, and to implement a patient diary (or a PHR) to help filling in eCRF.

Scenario 6 – Economic analysis

(UC - Analyse economic data between different procedures)

The goal of this scenario is to analyse economics aspects of different procedures (diagnostic and/or therapeutic) dispensed to a patient.

3.2 Terminologies in the user needs questionnaire

A questionnaire has been designed in WP1 to analyse the user needs. In particular we received answers about the terminologies that are already known and used by clinicians and other users (see *Figure 1*).

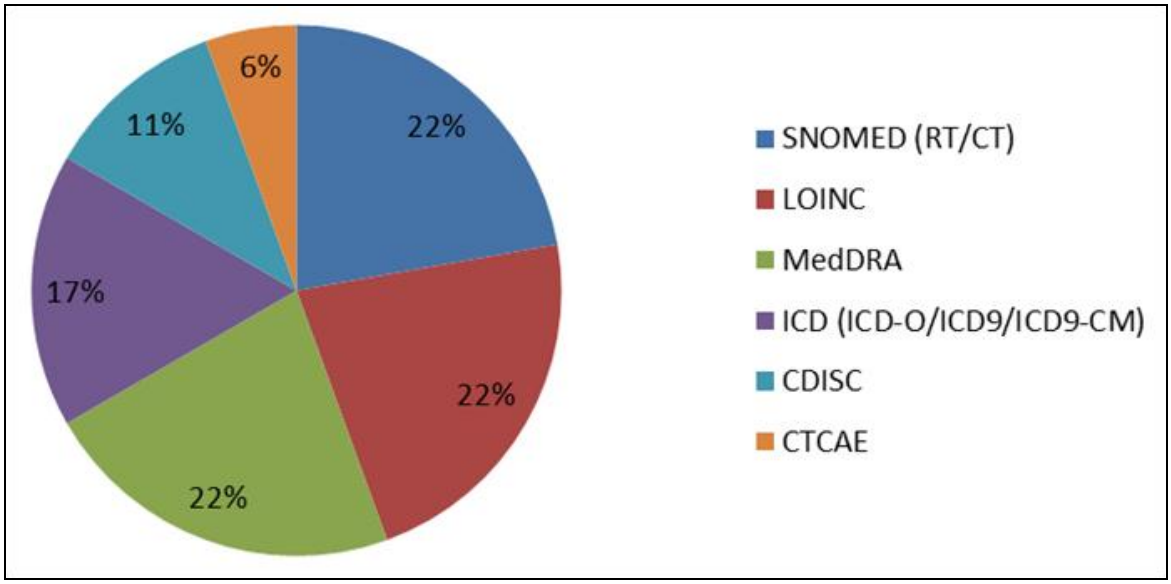

Figure 1 - Use of medical terminologies among interviewed users

The following general interest of using terminologies were given by clinical users:

- To extract more relevant data
- Structuring data in the Medical Record (allergies, infections, contaminations)
- RCM (Résumé Clinique Minimum) - Medical Abstract

And here are the answers concerning the use of all terminologies already used by the same users:

SNOMED (RT/CT): (see *Section 4.3*)

- Management of clinical studies
- Management of MDTs (Multi-Disciplinary Teams)
- Structured data extraction in anatomical pathology reports
- For surgical procedures
- Non-oncological diagnoses and allergies should be coded in SNOMED in the future

LOINC: (see *Section 4.4*)

- For laboratory results
- To categorise CDA sections
- Data exchange with General Practitioners
- Conversion tables between different laboratories

MedDRA: (see *Section 4.2*)

- Medications for the disease recording, treatments
- Adverse Events

ICD (ICD-O/ICD9/ICD9-CM): (see *Section 4.6*)

- ICD-O for topology extraction
- ICD9 for EHR, used for INAMI (National health insurance institute in Belgium)
- ICD9-CM is procedures oriented

CDISC: (see *Section 4.9*)

- For designing CRF terms to have a common language

CTCAE: (see *Section 4.5*)

- Terminology for data recording

4 Ontologies and terminologies

It is very difficult to determine the division between what is referred to as 'vocabularies' and 'ontologies'[3]. Vocabularies define the concepts, terms and relationships to fully describe an area of concern.

Usually the word 'ontology' is used to refer for more formal and complex collection of terms and relationships, while the word 'vocabulary' is used when such strict formalism is not necessary.

Hence, when we are using the word 'ontologies', we are referring to a formal collection of terms, of one or more domains given. Its purpose is to facilitate communication and exchange of information between different systems and entities.

In this section, some well-established and widely used standardized ontologies/ vocabularies from the biomedical area are described.

4.1 National Cancer Institute Thesaurus (NCIt)

National Cancer Institute Thesaurus (NCIt[4]) was developed by the NCI (National Cancer Institute) from EVS[5] (Enterprise Vocabulary Services) Project as a common reference terminology. It was developed to support their efforts in developing a common approach for coding, processing and exchanging cancer related information. Thus, National Cancer Institute Thesaurus is an ontology that has great coverage in the field of cancer, including diseases related to cancer, findings and abnormalities [4].

NCIt is a widely internationally recognized standard for medical coding and reference used by a wide variety of public and private partners as CDISC[6] (Clinical Data Interchange Standards Consortium Terminology), FMT[7] (Federal Medication Terminologies) or FDA[8] (Food and Drug Administration).

Since 1997 NCI has been making a big effort to "*integrate molecular and clinical cancer-related information within a unified biomedical informatics framework, with controlled terminology as its foundational layer*" [5]. What we understand as a thesaurus[9] is a reference work that displays words grouped according to their similarity of meaning, containing synonyms (or even sometimes antonyms).

It differs from a dictionary, because a dictionary contains definitions and pronunciations while a thesaurus represents a list of semantically search keys. NCIt provides a reference terminology for many systems (especially NCI systems). It covers a wide variety of areas such as for clinical care, basic research, and public information.

[3] http://www.w3.org/standards/semanticweb/ontology
[4] http://ncit.nci.nih.gov/
[5] http://evs.nci.nih.gov/
[6] http://www.cdisc.org/
[7] http://www.cancer.gov/cancertopics/cancerlibrary/terminologyresources/fmt
[8] http://www.fda.gov/
[9] http://en.wikipedia.org/wiki/Thesaurus

NCIt intends to integrate molecular and clinical information related to cancer within a unified biomedical system. Some of the main features that NCIt claims to offer are:

- Stable and unique codes for biomedical concepts.
- Synonyms, preferred terms, research codes, definitions, external source codes, and other information.
- Links to other information sources.
- More than 200,000 cross-links between concepts, offering a formal definition based on the logic of many concepts.
- It contains a large amount of information integrated from the NCI and other sources from external partners that are available separately from the NCIt.
- It is frequently updated by experts in the field.

NCIt combines a terminology with term from many domains related to cancer research, and it is able to integrate these terms together through semantic relationships. Currently, it contains over 43,000 concepts, structured into 20 taxonomic trees. NCIt also possesses some tables to track changes in vocabulary over time.

According to user needs (within the biomedical computing environment), NCIt has gone beyond the direct requirements of the terminology needed to create a model of how the key concepts are defined and related to each other, therefore, has become an ontology, since it maintains the integrity of the key concepts and extends its informative power.

NCIt has a major role by providing comprehensive resources that address the requirements of the NCI's terminology and providing semantic-based terminology for the NCI's caCORE[10] (cancer Common Ontologic Representation Environment) biomedical informatics infrastructure. "*The caCORE is an integrated suite of tools and resources supporting data management and application development, encompassing vocabulary, metadata, and biomedical data objects in a public domain technology stack"* [5]. caCORE provides real-time programming interfaces to access NCI Thesaurus and NCI Thesaurus provides much of the semantics that underlie caCORE. In caCORE 3.0.1, EVS gives the semantic base for the metadata component and biomedical data objects. All metadata class and attribute names correspond to concepts in NCIt.

NCIt contains two different models, the Disease model and the Drug model, that are explained below.

Disease model:

This model is structured to support strict definition of cancers and other diseases, specifying features and enforcing logical organization in their classification. These features provide an effective way to:

1. Link molecular findings to cancers
2. Identify diverse disease entities that share common molecular signatures
3. indicate the particular biologic events that characterize and often determine the outcome of a disease and

[10] https://wiki.nci.nih.gov/display/caCORE

4. Provide important information for researchers, health professionals, and the public.

Interdependent characteristics are linked together under specific role groups. This means that they are being assertive as a unit. NCIt disease model represents a prototype that can serve as the basic framework to link diverse cancers that share a unique molecular abnormality to specific drugs.

Drug model:

NCIt drug terminology has over 4000 single agents and about 3000 combination therapies for cancer drugs in clinical treatment and prevention trials. Agents are incorporated in NCIt from multiple primary research and clinical treatment related data sources. "*Drugs are classified on the basis of functional, structural, and therapeutic intent hierarchies, if possible, with text definitions and computable role relationships for mechanism of action, physiologic effects, known effects on gene products as molecular targets, and affected anatomic structures, including subcellular targets, if applicable*" [5].Therapeutic intent, including food and drug administration is represented as properties rather than as roles, to avoid inappropriate inheritance at lower level nodes of drug hierarchies. This model can be used to support targeted drug research but also to support pharmacokinetic and pharmacogenomic research.
Figure 2 shows the drug information model in NCI Thesaurus of the drug called '*Iressa*'. For example, you can see as 'Iressa' is related to the concept "enzyme interaction" through the mechanism of action hierarchy.

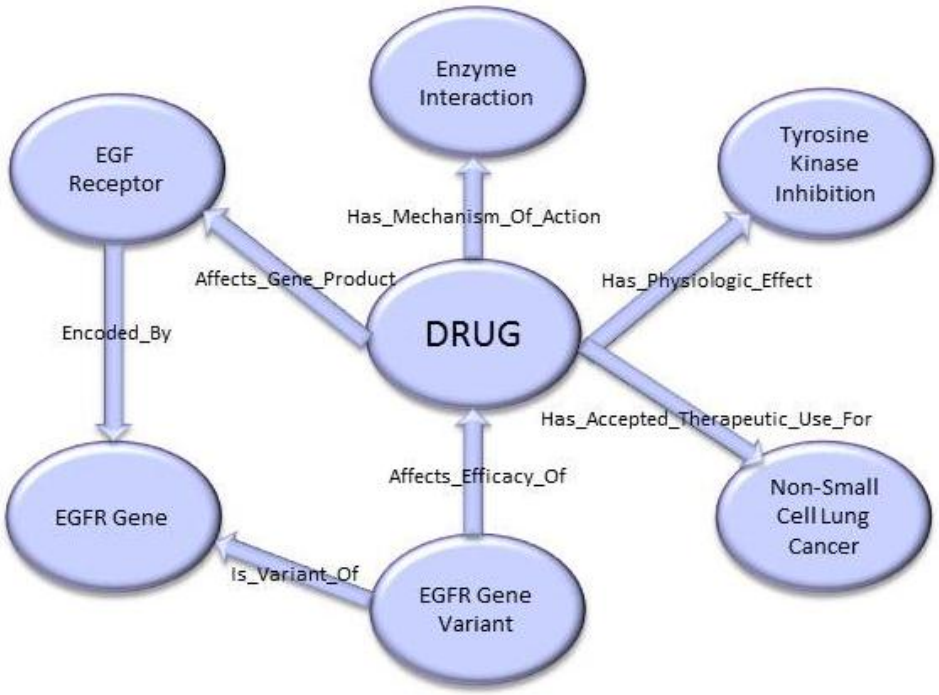

Figure 2 – Drug information model in NCI Thesaurus

NCIt general high-level categories
Abnormal Cell
Activity
Anatomic Structure, System, or Substance
Biochemical Pathway
Biological Process
Chemotherapy Regimen or Agent Combination
Conceptual entity
Diagnostic or Prognostic Factor
Disease, Disorder or Finding
Drug, Food, Chemical or Biomedical Material
Experimental Organism Anatomical Concept
Experimental Organism Diagnosis
Gene
Gene Product
Manufactured Object
I Administrative Concept
Organism
Property or Attribute
Retired Concept

Figure 3 - NCIt general high-level categories

4.2 Medical Dictionary for Regulatory Activities (MedDRA)

Medical Dictionary for Regulatory Activities (MedDRA[11]) is a medical terminology developed to facilitate the sharing of the information about medical products used by humans. It is managed by the MSSO, the Maintenance and Support Services Organization, and owned by the IFPMA[12], International Federation of Pharmaceutical Manufacturers and Associations.

MedDRA terms are referred to diseases, diagnoses, reactions and the results. It is useful to classify information related adverse events associated with the use of biopharmaceuticals and other medical products in humans. However, its use is growing worldwide in many new areas such as clinical research, starting to be a standard for a lot of regulatory scientific authorities.

MedDRA structure is hierarchical. It means that we have terms that are child from a sequence of predecessors' terms. In this hierarchy, one term could be preceded by more than one 'father'.

The hierarchy of the dataset is composed by six levels, shown in *Figure 4*:

[11] http://www.meddramsso.com/
[12] http://www.ifpma.org/

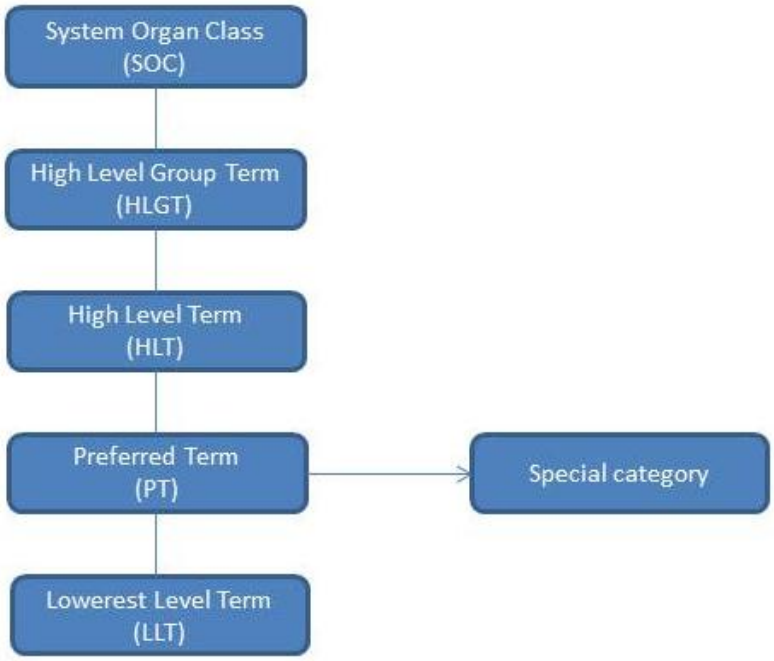

Figure 4 – MedDRA hierarchy levels

SOC
Blood and lymphatic system disorders
Cardiac disorders
Congenital, familial and genetic disorders
Ear and labyrinth disorders
Endocrine disorders
Eye disorders
Gastrointestinal disorders
General disorders and administration site conditions
Hepatobiliary disorders
Immune system disorders
Infections and infestations
Injury, poisoning and procedural complications
Investigations
Metabolism and nutrition disorders
Musculoskeletal and connective tissue disorders
Neoplasms benign, malignant and unspecified (incl cysts and polyps)
Nervous system disorders
Pregnancy, puerperium and perinatal conditions
Psychiatric disorders
Renal and urinary disorders
Reproductive system and breast disorders
Respiratory, thoracic and mediastinal disorders
Skin and subcutaneous tissue disorders
Social circumstances
Surgical and medical procedures
Vascular disorders

Figure 5 – MedDRA primary SOC

In this hierarchy:

- System Organ Class (SOC) represents the broadest concept (see *Figure 5*)
- Preferred Terms (PT) equals a single unique medical concept and
- Lowerest Level Term (LLT) equals a synonym or a lexical variant of a PT.

Each of its terms equals as less a word, and is tagged with a code number that starts with 10000001 alphabetically. For example, the term 'Urticaria' has the code number 100046735. Some examples of terms with code and its hierarchy (six levels) from the bottom up are shown in the following table:

Term	Code
Gastric hemorrhage (LLT)	10017789
Gastric haemorrhage (LLT)	10017788
Gastric and oesophageal haemorrhages (HLT)	10017751
Gastrointestinal haemorrhages NEC (HLGT)	10017959
Gastrointestinal disorders (SOC)	10017947

MedDRA is exactly composed from the union of other terminologies. MedDRA contains terms that come from:

- COSTART (5th edition)
- WHO-ART (98:3)
- J-ARTS (1996)
- HARTS (Release 2.2)
- ICD-9
- ICD-9-CM (4th Revision)

The main drawback of MedDRA is that it is not free, so a license is needed if you want to develop software that is going to be based on it. However, true not-for-profit organizations are offered for a basic subscription, for example an educational institution or a direct patient care provider, as a hospital planning to use MedDRA as a reference tool.

Useful tools offered by MedDRA are Web-Based Browser[13] and MedDRA Desktop Browser[14], which facilitate an easily way to search for terms on the hierarchy.

Figure 6 shows how works the hierarchy of the six levels core dataset. For example, it goes from "blood and lymphatic system disorders" (father-term) to "renal anaemia".

[13] https://meddramsso.com/subscriber_download_tools_wbb.asp
[14] https://meddramsso.com/subscriber_download_tools_browser.asp

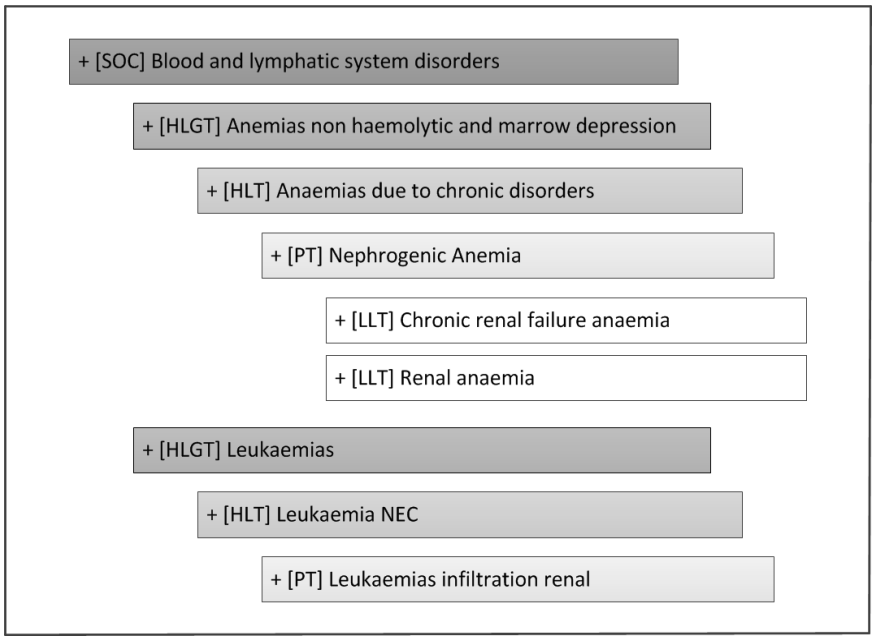

Figure 6 – Real Example of MedDRA hierarchy

4.3 Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT)

Systematized Nomenclature of Medicine – Medical Terms (SNOMED-CT[15]) was born by the combination and expansion of two taxonomies, (i) SNOMED-RT [6], developed by the College of American Pathologies (CAP[16]) and (ii) Clinical Terms Version 3 (CTV3[17]). It was created by the National Health Service (NHS[18]) of the United Kingdom. Currently, SNOMED-CT is property of the International Health Terminology Standards Development Organization (IHTSDO[19]) since 2007.

SNOMED-CT is defined as a systematically organized computer-readable collection of medical terms providing codes, terms, synonyms and definitions. It is considered the most important clinical terminology, thanks to its precision and highly comprehension data.

This terminology allows its users to tag, index and store clinical information; facilitating the proper management of their medical media. Its ease of use has been an important help point for everyone who works with electronic medical record systems (EMRs[20]). And, it has been adopted as the standard clinical terminology for many institutions.

[15] http://www.ihtsdo.org/snomed-ct/
[16] http://www.cap.org/apps/cap.portal/
[17] http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/RCD/
[18] http://www.nhs.uk/Pages/HomePage.aspx/
[19] http://www.ihtsdo.org/
[20] http://www.openclinical.org/emr.html/

SNOMED-CT is composed of concepts and relationships, which formally define the concepts. The concepts in SNOMED-CT have a hierarchical structure. Each top-level hierarchy (see *Figure 7*) contains sub-hierarchies that specify the concept [7]. Also, every concept has associated a few descriptions that describe its different properties. Those descriptions could be:

- **Preferred Term:** Word/phrase used by clinicians to name a clinical concept.
- **Fully Specified Name:** A unique way to name and denominate the concept. It is essentially the Preferred Term, along with a 'semantic tag' as a suffix to indicate the type of concept and to eliminate ambiguity.
- **Synonym:** Additional phrases/terms that could represent that could define the concept at the same level of granularity.

SNOMED-CT general concepts
Body structure
Clinical finding
Environment or geographical location
Event
Linkage concept
Observable entity
Organism
Pharmaceutical/biological product
Physical force
Physical object
Procedure
Qualifier value
Record artifact
Situation with explicit context
Social context
Special concept
Specimen
Staging and scales
Substance

Figure 7 - SNOMED-CT top-level hierarchies

In *Figure 8* is shown an example of a SNOMED concept, "Malignant tumour of breast" specifically:

- ConceptID: ***254837009***
- Fully Specified Name: ***Malignant Tumor of breast (disorder)***
- Preferred Term: ***Malignant Tumor of breast***
- Synonym: ***Breast Cancer***
- Synonym: ***Malignant Tumor of breast***

Figure 8 – Example of a Snomed-CT concept

"SNOMED-CT is multi-hierarchical; a single concept can exist in multiple sub-hierarchies. However, a single concept can exist in more than hierarchy"[21]. As we said before, each concept in SNOMED-CT is defined through is relationships to other concepts. There are two possible types of relationships:

- **IS-A relationships:** Every concept has a defining hierarchical relationship called *IS_A. This IS_A* relationship is, basically, a parent-child relation. Also, a concept can have more than one *IS_A* relationship to other concepts. In that case, the concept will have parent concepts in more than one sub-hierarchy. IS_A relationships connect concepts in a single hierarchy. An example of a concept with two IS_A relations is shown in *Figure 9*.
- **Attribute relationships:** These relationships define the semantics of the elements. They also help to differentiate them from other similar concept definitions, including their own super-types and sub-types. Attribute relationship connect concepts in different hierarchies. *Figure 10* depicts an example of a relationship that connects a concept from the *Disease* hierarchy and other from the *Body structure* hierarchy. As a result, the 'Appendicitis' term that is in the *Disease* hierarchy is connected with the term 'Inflammation' in the *Morphologically abnormal structure* section of the *Body structure hierarchy*.

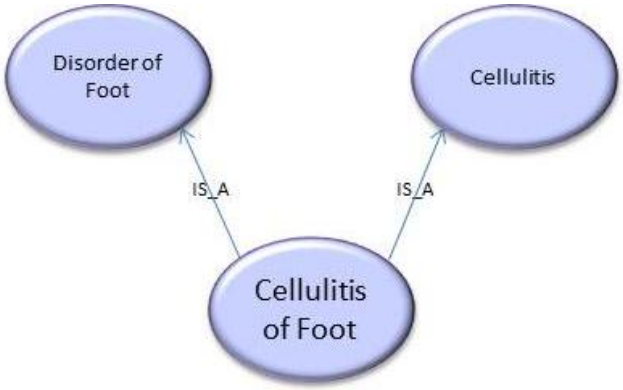

Figure 9 – Example of an IS_A relationship

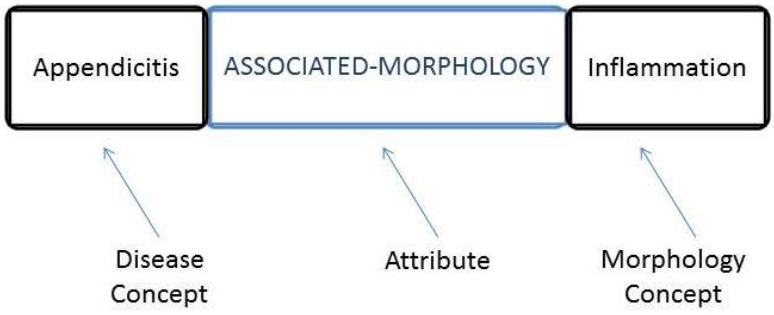

Figure 10 – Example of a concept with two attribute relationships

[21] http://clinfowiki.org/wiki/index.php/SNOMED/

SNOMED-CT is open depending of the country and the purpose of the user. It is allowed to download by previous registration. Also, there are a series of free and useful web SNOMED-CT browsers on the internet, especially interesting are (i) the SNOMED-CT core browser, developed by the Virginia-Maryland Regional College of Veterinary Medicine[22], and (ii) the one facilitated by the NCI. Other browsers of interest are described in [8].

4.4 Logical Observation Identifiers Names and Codes (LOINC)

Logical Observation Identifiers Names and Codes (LOINC[23]) is a standard set of terms for identifying medical laboratory observations. It was developed by the Regenstrief Institute[24] with the intention of providing a definitive standard for identifying clinical information in electronic reports [9]. LOINC's main goal is to facilitate the exchange and pooling of results for clinical care, research and outcomes management.

LOINC database gives a set of universal names and ID codes for identifying laboratory and clinical test results in the context of existing observation for report messages.

LOINC codes identify the results of clinical trials and the results of clinical observations. Other fields that can be transmitted through LOINC codes inside messages are, for example, the identity of the source laboratory, and other special details about the sample.

A formal, distinct and unique six part name is given to each LOINC component term for the test or observation identity.

Each database record has the syntax shown in *Figure 11*:

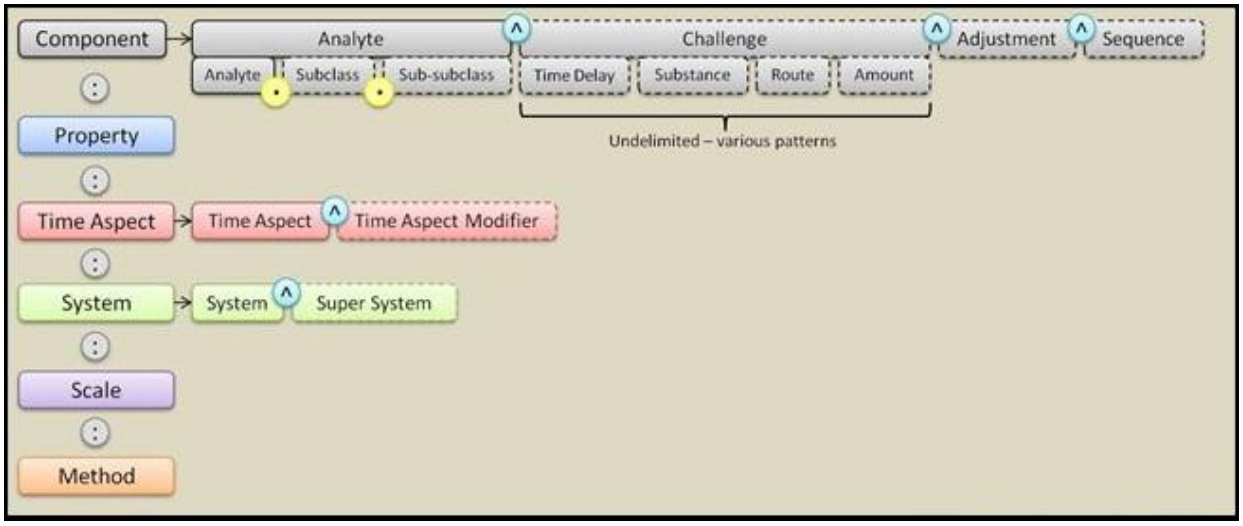

Figure 11 – LOINC Structure

[22] http://www.vetmed.vt.edu/
[23] http://loinc.org/
[24] http://www.regenstrief.org/

Where the six parts are:

1. Name of the component or analytic measured, evaluated or observed (i.e.: glucose, propranolol)
2. Kind of property observed/measured (i.e.: substance concentration, mass, volume)
3. The timing of the measurement (i.e: is it over time or momentary)
4. Type of sample (i.e: urine, serum, blood)
5. Type of scale of measurement (i.e.: qualitative vs. quantitative, ordinal vs. nominal)
6. Type of method of measurement (i.e.: radioimmunoassay, immune blot). This part is optional.

LOINC is available for *free* use as a Microsoft Access database file or as a tab-delimited text file. The Regenstrief Institute provides a Windows-based mapping utility called the **RELMA**[25] **(Re**genstrief **L**OINC **M**apping **A**ssistant**)** to facilitate searches through the LOINC database and to assist efforts to map local codes to LOINC codes. The RELMA package includes also the LOINC data table. Another possibility for exploring LOINC is using the web search application available[26], but it is less effective for a depth use than RELMA full version.

4.5 Common Toxicity Criteria for Adverse Events (CTCAE)

The Common Toxicity Criteria for Adverse Events (CTCAE) terminology – also referred to as Common Toxicity Criteria (CTC) – is developed by the National Cancer Institute (NCI). It is used within clinical trials to report side effects and adverse events. It is also an ontology.

CTCAE are grouped by MedDRA primary System Organ Class (SOC), the highest level of the MedDRA hierarchy, which contains 26 classes (see Figure 5).

It consists of the name of the area of interest together with a grading which refers to the severity of the reaction [10]:

- **Grade 1**: mild adverse event
- **Grade 2**: moderate adverse event
- **Grade 3**: severe adverse event
- **Grade 4**: life threatening adverse event
- **Grade 5**: fatal adverse event (death)

Specific symptoms have values or descriptive comment for each grade (see *Figure 12*).

[25] http://loinc.org/relma
[26] http://search.loinc.org/

Blood and lymphatic system disorders					
	Grade				
Adverse Event	**1**	**2**	**3**	**4**	**5**
Anemia	Hemoglobin (Hgb) <LLN - 10.0 g/dL; <LLN - 6.2 mmol/L; <LLN - 100 g/L	Hgb <10.0 - 8.0 g/dL; <6.2 - 4.9 mmol/L; <100 - 80g/L	Hgb <8.0 g/dL; <4.9 mmol/L; <80 g/L; transfusion indicated	Life-threatening consequences; urgent intervention indicated	Death
Definition: A disorder characterized by an reduction in the amount of hemoglobin in 100 ml of blood. Signs and symptoms of anemia may include pallor of the skin and mucous membranes, shortness of breath, palpitations of the heart, soft systolic murmurs, lethargy, and fatigability.					
Bone marrow hypocellular	Mildly hypocellular or <=25% reduction from normal cellularity for age	Moderately hypocellular or >25 - <50% reduction from normal cellularity for age	Severely hypocellular or >50 - <=75% reduction cellularity from normal for age	Aplastic persistent for longer than 2 weeks	Death
Definition: A disorder characterized by the inability of the bone marrow to produce hematopoietic elements.					
Disseminated intravascular coagulation	-	Laboratory findings with no bleeding	Laboratory findings and bleeding	Life-threatening consequences; urgent intervention indicated	Death
Definition: A disorder characterized by systemic pathological activation of blood clotting mechanisms which results in clot formation throughout the body. There is an increase in the risk of hemorrhage as the body is depleted of platelets and coagulation factors.					
Febrile neutropenia	-	-	ANC <1000/mm3 with a single temperature of >38.3 degrees C (101 degrees F) or a sustained temperature of >=38 degrees C (100.4 degrees F) for more than one hour.	Life-threatening consequences; urgent intervention indicated	Death
Definition: A disorder characterized by an ANC <1000/mm3 and a single temperature of >38.3 degrees C (101 degrees F) or a sustained temperature of >=38 degrees C (100.4 degrees F) for more than one hour.					

Figure 12 - Example of Adverse Events' Grades in CTCAE (e.g. Febrile Neutropenia) [10]

The NCI CTCAE version 4 was modified to be more aligned with the MedDRA terminology. Currently (August 2012) version 4.03 is in use, released in June 2010.

4.6 International Classification of Disease (ICD)

The International Classification of Disease (ICD[27]) is a medical codification classifying diseases and a very wide variety of signs, symptoms, injury, poisoning, social circumstances and external causes of injury or disease. The ICD is published by the World Health Organisation (WHO) and is used worldwide for recording causes of morbidity and mortality related to the field of medicine.
ICD was designed to enable systematic analysis, interpretation and comparison of observational databases on morbidity and mortality in different countries or region at different times. The tenth revision is currently in use (August 2012) as it has emerged in 1983 and was finalised in 1992, and decennial revisions have now been replaced by updates.
Disease, symptoms, injury, poisoning and other grounds of appeal to health services are listed in the ICD with a precision that depends on their importance, i.e. their frequency and the intensity of the public health problem that arise.
ICD is a statistical classification in the sense that an encoding entity can be assigned only to one category of classification. Moreover each disease corresponds to only one code, classification ambiguities being thrown by exclusion rules. ICD attributes to identified entities an alphanumeric code consisting of three to five characters.

[27] http://www.who.int/classifications/icd/en/

4.6.1 ICD-10

ICD-10 brought important changes. It consists in three volumes that have been respectively published in 1993 (Volume 1), 1995 (Volume 2) and 1996 (Volume 3).

ICD-10 identifies 22 chapters (see *Figure 13*), each one defining a set of conditions or diseases. The entire list of codes contains around 14,400 different codes and allows for diagnoses. Using optimal sub classifications, the number of codes can be extended up to 16,000.

Chapter	Blocks	Title
I	A00-B99	Certain infectious and parasitic diseases
II	C00-D48	Neoplasms
III	D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	E00-E90	Endocrine, nutritional and metabolic diseases
V	F00-F99	Mental and behavioural disorders
VI	G00-G99	Diseases of the nervous system
VII	H00-H59	Diseases of the eye and adnexa
VIII	H60-H95	Diseases of the ear and mastoid process
IX	I00-I99	Diseases of the circulatory system
X	J00-J99	Diseases of the respiratory system
XI	K00-K93	Diseases of the digestive system
XII	L00-L99	Diseases of the skin and subcutaneous tissue
XIII	M00-M99	Diseases of the musculoskeletal system and connective tissue
XIV	N00-N99	Diseases of the genitourinary system
XV	O00-O99	Pregnancy, childbirth and the puerperium
XVI	P00-P96	Certain conditions originating in the perinatal period
XVII	Q00-Q99	Congenital malformations, deformations and chromosomal abnormalities
XVIII	R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	S00-T98	Injury, poisoning and certain other consequences of external causes
XX	V01-Y98	External causes of morbidity and mortality
XXI	Z00-Z99	Factors influencing health status and contact with health services
XXII	U00-U99	Codes for special purposes

Figure 13 - List of ICD-10 codes[28]

4.6.2 ICD-O

The International Classification of Diseases for Oncology (ICD-O[29]) [11] is a specific extension to the domain of the ICD for tumour diseases. This classification is widely and mainly used in cancer or tumour registries, for coding the anatomical location (topography) and the histology type (morphology) of tumours, which are usually obtained from pathology reports. The part of ICD-O on morphology represents the sections 8 and 9 of the chapter on morphology in SNOMED-RT.

[28] http://apps.who.int/classifications/icd10/browse/2010/en
[29] http://www.who.int/classifications/icd/adaptations/oncology/en/

To a large extent, the topography classification used in ICD-O is the same that is used for malignant tumour in ICD-10.

4.7 Anatomical Therapeutic Chemical (ATC) classification system

The Anatomical Therapeutic Chemical (ATC[30]) classification system is a pharmaceutical codification developed by the European Pharmaceutical Market Research Association (EPhMRA[31]) and controlled by the World Health Organisation Collaborating Centre for drug statistics methodology (WHOCC[32]). These pharmaceutical codes are used in medical classification to uniquely identify medication and classify therapeutic drugs. It was first published in 1976.
The drugs are divided into different groups according to the organ or system on which they act, and their therapeutic and chemical characteristics [12].

In this system, drugs are classified in groups at five different levels:

- **Level 1**: The first level of the ATC code (see *Figure 14*) is based on a letter (an alphabetical character) for each anatomical group (there are 14 major):

Code	Contents
A	Alimentary tract and metabolism
B	Blood and blood forming organs
C	Cardiovascular system
D	Dermatologicals
G	Genito-urinary system and sex hormones
H	Systemic hormonal preparations, excluding sex hormones and insulins
J	Antiinfectives for immunomodulating agents
L	Antineoplastic and immunomodulating agents
M	Musculo-skeletal system
N	Nervous system
P	Antiparasitic products, insecticides and repellents
R	Respiratory system
S	Sensory organs
V	Various

Figure 14 – Major of the first level of the ATC code

- **Level 2**: The second level of the code defines the therapeutic main group and consists of two digits.
- **Level 3**: The third level of the code defines a therapeutic / pharmacological subgroup and consists of one alphabetical character.
- **Level 4**: The fourth level of the code defines a chemical / therapeutic / pharmacological subgroup and consists of one alphabetical character.
- **Level 5**: The fifth level of the code defines a subgroup for chemical substance and consists of two digits.

Hereinafter is the example of the Fluoxetine, an antidepressant of the selective serotonin reuptake inhibitory class:

[30] http://www.who.int/classifications/atcddd/en/
[31] http://www.ephmra.org/
[32] http://www.whocc.no/

- N**06** Phsychoanaleptics
- N06**A** Antidepressants
- N06A**B** Selective serotonin reuptake inhibitors
- N06AB**03** Fluoxetine

4.8 Medical Subject Headings (MeSH)

Medical Subject Headings (MeSH[33]) is a large controlled vocabulary designed with the goal of indexing the literature in the biomedical field. MeSH is a medical concepts vocabulary. It keeps for each concept a series of equivalent terms. The thesaurus was created by the National Library of Medicine (NLM[34]) of the United States, which continues to manage it. MeSH is used for indexing the articles of over 5,000 medical journals in the database bibliographic PubMed[35] or Medline and in the catalog of the books of the NLM.

MeSH terminology consists basically in a set descriptors, which are often accompanied by:

- a brief description or definition,
- links to related descriptors,
- a list of synonyms or related terms,
- a list of qualifiers which can be used with the term.

These qualifiers, which go together with descriptors, provide semantic information to categorize concepts.

Currently there are 83 qualifiers defined in MeSH. Some examples of qualifiers are “analysis”, “chemistry”, “metabolism” or “radiationeffects”.

Descriptors are organized hierarchically. A given descriptor may appear in several places in the hierarchical tree. MeSH tree is provided in form of XML file in which a particular concept is represented within the file as Figure 15 – Concept in the MeSH tree shows:

The common representation of a concept in an XML file contain:

- The tag <*DescriptorName*> contains the concept which is referred.
- The tag <*QualifiersList*> contains the list of the different qualifiers that apply to the concept.
- The <*TreeList*> tag contain the path of the term in the MeSH hierarchy.
- Finally in the tag <*TermList*> we find different terms that refer to the same concept.

[33] http://www.ncbi.nlm.nih.gov/mesh
[34] http://www.nlm.nih.gov/
[35] http://www.ncbi.nlm.nih.gov/pubmed

```
<DescriptorRecord>
    <DescriptorName>Abattoirs</DescriptorName>
    <QualifiersList>
        <Qualifier>classification</Qualifier>
        <Qualifier>economics</Qualifier>
        <Qualifier>history</Qualifier>
        <Qualifier>instrumentation</Qualifier>
        <Qualifier>legislation &amp;amp; jurisprudence</Qualifier>
        <Qualifier>manpower</Qualifier>
        <Qualifier>standards</Qualifier>
        <Qualifier>statistics &amp;amp; numerical data</Qualifier>
        <Qualifier>ethics</Qualifier>
    </QualifiersList>
    <TreeList>
        <Tree>J01.576.423.200.700.100</Tree>
    </TreeList>
    <TermList>
        <Term>Abattoirs</Term>
        <Term>Abattoir</Term>
        <Term>Slaughterhouses</Term>
        <Term>Slaughterhouse</Term>
    </TermList>
</DescriptorRecord>
```

Figure 15 – Concept in the MeSH tree

4.9 CDISC

CDISC is a non-profit organization that develops and maintains a set of standards for the exchange, submission and archive of clinical research data and metadata. The CDISC standards are different from the vocabularies discussed in this section in the sense that they not only provide a controlled vocabulary of terms but also prescribe how a trial should be reported. The CDISC standards SDTM, CDASH and ADaM are all provided with NCI codes and preferred terms. For a detailed description of these standards we refer to deliverable D2.1[36].

4.10 Radiology Lexicon (RadLex)

RadLex[37] is a comprehensive radiology lexicon initiated by the Radiological Society of North America (RSNA) in 2005 in order to provide radiologists with a unified language to index and retrieve images, imaging reports and medical records. RadLex development is supported both by the National Institute of Biomedical Imaging and Bioengineering

[36] EURECA project, "D2.1 State of the art report on standards," 2012
[37] https://www.rsna.org/RadLex.aspx

(NIBIB) and by the cancer Biomedical Informatics Grid (caBIG) project, a large NIH-sponsored effort to develop unified computing infrastructure for clinical trials. With more than 30,000 terms, RadLex replaces the ACR Index for Radiological Diagnoses and unifies and supplements other lexicons and standards, such as SNOMED-CT and DICOM. RSNA has developed a term browser[38]. RadLex is available for download as well as browsing, visualization, and other ontology access functionality provided by the National Centre for Biomedical Ontology's BioPortal[39] and Web APIs. The RadLex Ontology License permits public access to the Release Version of RadLex® and to use it without charge.

The RadLex Playbook[40] is a special component of RadLex that provides a standard, comprehensive lexicon of radiology orderable and procedure step names. The American College of Radiology (ACR) has become an early adopter of Playbook for use in their CT Dose Index Registry (DIR). The DIR allows facilities to compare their CT dose indices to regional and national values. Using standard procedure names for the data being collected is crucial to establishing national benchmarks. The ACR has begun collecting reporting dose values for the DIR using Playbook procedure names.

[38] http://radlex.org/
[39] http://bioportal.bioontology.org/ontologies/40885/?p=summary
[40] https://www.rsna.org/RadLex_Playbook.aspx

5 Mappings / Links among terminologies / ontologies

A link or mapping is an association between two or more terms in different terminologies or ontologies. Mapping tables are used to, for example, associate diseases in one coding system to diseases in another coding system, or procedures to procedures, organisms to organisms, etc. This association usually needs to be represented with a degree of similarity between the terms. The author of a mapping defines the semantics of that particular mapping. Mappings used to be bi-directional, but it is not necessary.

5.1 SNOMED-CT cross mappings

SNOMED-CT medical ontology provides tables that define cross-mappings with other ontologies (ICD9 and LOINC).

In cross-references (crossmap) provided with LOINC, each LOINC code represents a unique laboratory test distinguished by six main parts. SNOMED-CT provides a plain text file with a table with eleven columns, the first nine related to LOINC and the last two are derived directly from SNOMED-CT. Currently, this table of integration between LOINC and SNOMED-CT is not supported by the IHTSDO and has not been updated for the current version of SNOMED-CT, therefore, this information is only provided for reference purposes only.

An example of the relation between these two ontologies is shown in *Figure 16*. Missing the last two columns refer to LOINC (RELAT_NMS and ANSWERLIST) because they are not always applicable:

LOINC _NUM	COMPONENT	PROPERTY	TIME_ ASPCT	SYSTEM	SCALE_ TYP	METHOD _TYP	Relationship_ Type	ConceptId
5792-7	GLUCOSE	MCNC	PT	UR	QN	TEST STRIP	116684007	123029007
5792-7	GLUCOSE	MCNC	PT	UR	QN	TEST STRIP	116678009	67079006
5792-7	GLUCOSE	MCNC	PT	UR	QN	TEST STRIP	116680003	69376001
5792-7	GLUCOSE	MCNC	PT	UR	QN	TEST STRIP	116686009	122575003
5792-7	GLUCOSE	MCNC	PT	UR	QN	TEST STRIP	116685008	118539007
5792-7	GLUCOSE	MCNC	PT	UR	QN	TEST STRIP	116687000	30766002
5792-7	GLUCOSE	MCNC	PT	UR	QN	TEST STRIP	84203001	117021008

Figure 16 – SNOMED-CT - LOINC Integration table

Where:

- **LOINC_NUM** is the LOINC code.
- **COMPONENT** is one of the six parts of the LOINC fully specified name and maps to the SNOMED-CT RelationshipType *"has measured component".*
- **PROPERTY** is one of the six parts of the LOINC fully specified name and maps to the SNOMED-CT RelationshipType *"has property".*
- **TIME_ASPCT** is one of the six parts of the LOINC fully specified name and maps to the SNOMED-CT RelationshipType *"has time aspect".*
- **SYSTEM** is one of the six parts of the LOINC fully specified name and maps to the SNOMED-CT RelationshipType *"has specimen".*
- **SCALE_TYP** is one of the six parts of the LOINC fully specified name and maps to the SNOMED-CT RelationshipType *"has scale type"*.
- **METHOD_TYP** is one of the six parts of the LOINC fully specified name and maps to the SNOMED-CT RelationshipType *"has method"*.

The two last columns (Relationship_Type and ConceptId) are derived from the SNOMED-CT Concepts Table:

- **Relationship_Type** is the concept identifier (ConceptId) from the SNOMED CT Concepts table. It defines the relationship between the LOINC name and the target SNOMED CT concept.
- **ConceptId** is the target SNOMED CT Concept from the SNOMED CT Concepts table.

SNOMED-CT9 also has cross-mapping tables from clinical concepts to categories listed in ICD-9-CM (International Classification of Diseases 9th Revision Clinical Modification[41]), that is a coding scheme sponsored by the NCHS. It is used for reporting and tracking of mortality, statistical reporting of diseases.

"The SNOMED CT to ICD-9-CM cross-map is updated to reflect the version of ICD-9-CM current as of the date of its release. It is important to note that this mapping table is NOT intended for direct billing or reimbursement without additional authoritative review"[42]. ICD code/s with highest level of specificity has been selected. Terms that cannot be assigned to appropriate ICD-9-CM code are considered 'unmappable'.

The existing mapping between SNOMED-CT and MedDRA is 'hand-made'. Some projects aims to improve this manual mapping through an automatic lexical-based approach. In this mapping there are about 308 direct mappings of MedDRA terms to SNOMED-CT concepts. After segmenting MedDRA terms, it's been identified 535 full mappings associating a MedDRA term with one or more SNOMED-CT concepts [13].

5.2 Unified Medical Language System Metathesaurus

The UMLS[43] Metathesaurus *"is a large, multi-purpose, and multi-lingual thesaurus that contains millions of biomedical and health related concepts, their synonymous names,*

[41] http://www.cdc.gov/nchs/icd/icd9cm.html
[42] http://www.ihtsdo.org/fileadmin/user_upload/doc/tig/trg/trg_app_xmaps_icd_9_cm.html
[43] http://www.nlm.nih.gov/research/umls/

and their relationships"[44]. It is one of the three components of the UMLS project of the National Library of Medicine of the United Stated.

UMLS Metathesaurus is updated twice a year (about May and November). This Metathesaurus incorporate patient care, health services billing, public health statistics and so on. UMLS Metathesaurus transcends the specific thesauri, codes, and classifications it encompasses.

The Metathesaurus has over 1 million biomedical concepts and 5 million concept names, from different vocabularies and classification systems. Some of these vocabularies (ontologies) incorporated are ICD-9-CM, ICD-10, MeSH, SNOMED-CT, LOINC, etc.

The Metathesaurus is organized by concept, and each concept has specific attributes defining its meaning and is linked to the corresponding concept names in the various source vocabularies.

The scope of the source ontology determines the scope of Metathesaurus. The Metathesaurus faithfully represented the different names for the same concept used in the different vocabularies, or the same name for different concepts.

5.3 NCI Metathesaurus

NCI Metathesaurus (NCIm[45]) is a biomedical database that covers most terminologies used by National Cancer Institute (NCI) for clinical care, translational and basic research. One of the most important features of the NCIm is that links to NCI Thesaurus (NCIt) and other related information sources as MedDRA or ICD-10, however the representation is not identical.

NCIm is based on UMLS Metathesaurus complemented with additional cancer vocabulary.

The public current version of the NCIm contains all public domain vocabularies from the UMLS Metathesaurus, and a growing number of NCI specific vocabularies (vocabularies developed by the NCI).

5.4 BioPortal

BioPortal is an open repository of biomedical ontologies created and maintained by the National Centre for Biomedical Ontology (NCBO)[46]. It currently contains 325 ontologies, including NCIt, MedDRA, SNOMED-CT, LOINC, CTCAE, ICD-10 and Mesh, which were discussed in *Section 4* of this deliverable. The portal allows users to search, browse and visualize the ontologies, as well as to add notes and reviews. In addition, BioPortal contains annotations of 27 online biomedical data sources with concepts from the ontologies, including clinicaltrials.gov and PubMed. For a concise description of the design and functionality of BioPortal we refer to [14].

[44] http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html
[45] http://ncimeta.nci.nih.gov/ncimbrowser/
[46] http://www.bioontology.org/

BioPortal provides a large number of links between the concepts of the ontologies it hosts. In July 2010, the total number of links was 3M[47]. Over 1M of these were created by the LOOM algorithm of NCBO. LOOM [15] performs a lexical comparison of names and synonyms of the concepts in the two ontologies. First, all delimiters such as spaces, underscores and parentheses are removed. Then, the strings are compared. Two strings are considered similar if at most one character is a mismatch in strings with length greater than four and none for shorter strings. In [15] the authors evaluate LOOM on the AOEI 2008 anatomy track [16] in which the Mouse adult gross anatomy ontology and the human-anatomy part of the National Cancer Institute (NCI) Thesaurus are mapped. The results show that their simple approach gives a performance that is comparable to other ontology matching algorithms. Another 2M mappings were based on existing shared identifiers in the mapped ontologies. Most of these were UMLS CUI's. Finally, BioPortal provides over 6.000 manually created links. *Figure 17* shows BioPortal terminologies that are currently of interest to EURECA and the numbers of mappings between them. For comparison, *Figure 18* shows the number of concepts in these terminologies.

BioPortal can be accessed in three ways. First, there is a website on which users can search and browse the ontologies, mappings and annotated resources: http://bioportal.bioontology.org/. Second, there are several RESTful services to request or search ontology content, concepts and terms. Third, there is a SPARQL endpoint at http://sparql.bioontology.org through which the data can be queried with the semantic web query language SPARQL. For this purpose, BioPortal has translated the ontologies and data from their original formats (OBO, the Protégé Frame Language, etc.) to RDF where necessary.

	SNOMED-CT	NCIt	LOINC	ICD-10	MedDRA	CTCAE	MeSH
SNOMED-CT							
NCIt	31,983						
LOINC	36,060	9,113					
ICD-10	32,421	1,080	583				
MedDRA	137,840	7,912	4,406	15,001			
CTCAE	953	1,391	95	62	801		
MeSH	76,794	19,461	16,555	3,167	21,284	296	

Figure 17 - BioPortal mapping between different terminologies

Terminology	Number of classes
SNOMED-CT	395,036
NCIt	93,411
LOINC	171,399
ICD-10	12,318
MedDRA	69,389
CTCAE	3,874
MeSH	229,698

Figure 18 - BioPortal teminologies of interest and their number of concepts

[47] Source: http://www.bioontology.org/wiki/index.php/BioPortal_Mappings

6 Requirements analysis and method

6.1 Requirements for automatic text mining for free-text methods

Information to be integrated in the EURECA environment can be represented as free text reports from EHR and from clinical trial eligibility criteria in the CT protocols. In order to extract the information contained in these free text sources, it is necessary to perform syntactic and semantic analysis of textual data. The information extraction process in EURECA project has to be gradual, based on the defined core dataset and supervised by clinicians.

The goal is to structure free text from data sources and store within the EURECA CDM. Initially training data from files annotated from clinicians to structure the free text into the EURECA CDM will be used.

The next step in the free text information extraction is the concept recognition. Concept recognition is very dependent on the chosen terminologies used in the project as the concepts labels. This step is strongly linked with semantic interoperability tasks while the semantic Core Dataset will contain relevant terms from different terminologies from selected domains and relationships among these terms. In order to build the core dataset, concepts from selected vocabularies should be identified within free text documents from data sources and will be used to enrich the core dataset as subset of such vocabularies.

On the other hand, identifying concepts from vocabularies in free text usually is not enough for extracting relevant data from reports. Often, reports contain values and units which have to be found and annotated. For example data as the birth date of a patient; or a numeric value from a lab test.

Other aspect to take into account is that extracted information is going to be stored into the Common Schema defined EURECA. Thus, the information extracted should be structured and using terms of the Core Dataset, so this data will follow the Common Data Model as in depicted in *Figure 19*.

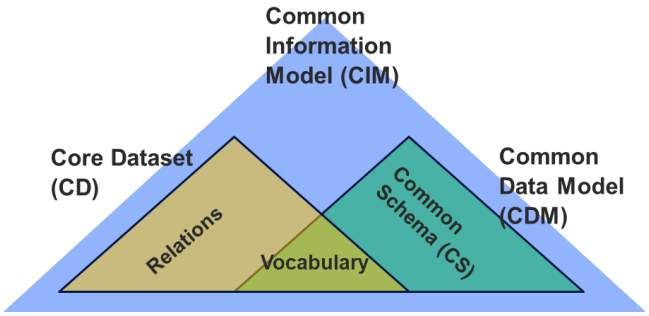

Figure 19 – EURECA Common Information Model

6.2 Terminologies and subset of concepts used on-site

6.2.1 Institut Jules Bordet (IJB)

For several years, IJB has used standard terminologies to standardise the clinical concepts used for its internal applications as well as to be compliant with regulatory authorities.

	DATA		TERMINOLOGIES								
			SNOMED	LOINC	ICD-9-CM	ICD-O	NCI Thesaurus	NCI CTCAE	MedDRA	ATC pharma codes	Others?
IJB	Cancer registry data		Convertible in SNOMED		X	X					ad hoc
	Multidisciplinary Team (MDT) data		Convertible in SNOMED	X		X					ad hoc
	EHR	Consult & discharge reports		for section header codes							
		Anatomical Pathology	X (all that is encoded is SNOMED, few is encoded -> a lot of NLP is necessary)								
		Laboratory data		Convertible in LOINC (part)							ad hoc IRIS network
		Radiology & nuclear medicine data									
	Picture Archiving and Communication System (PACS)										
	Day Hospital (oncological daycare clinic)										
	Chemotherapy prescriptions									Convertible in ATC pharma codes	
	CRF/eCRF		NCI terms partly Convertible in SNOMED				1 Clinical Trial	X	X	X	

Figure 20 - Medical terminologies used at IJB

- SOMED-CT and SNOMED-RT are used internally for IJB's MDT application and for the anatomical pathology application. SNOMED has been chosen at IJB as it is recognised to be one the most complete medical terminologies that allow defining almost all of the clinical concepts, which is well adapted in such a multidisciplinary institute for cancer treatment. One additional reason is that SNOMED is defined as ontology and can thus allow semantic processing.

  For example, *Figure 21* shows the statistical use of SNOMED-RT terms that have been collected in our application for the Anatomical Pathology unit. *Figure 22* and *Figure 23* represent the distribution of terms used in each of the sub-sections of both Morphology and Topography SNOMED-RT sections.
  - *Total terms (with synonyms)* represent the number of total terms existing in each SNOMED-RT sections, by counting also all the synonyms for the same SNOMED-RT code,
  - *Distinct terms* represent the number of distinct existing SNOMED-RT codes.

		SNOMED-RT Sections											TOTAL
		A	C	D	F	G	J	L	M	P	S	T	
Total terms (with synonyms)		1,641	8,254	43,098	19,661	1,544	3,040	23,934	**9,043**	23,655	1,095	**14,200**	**149,165**
Total distinct terms		1,342	6,184	26,449	13,062	1,033	1,639	17,278	**4,336**	18,045	854	**10,796**	**101,018**
Terms used onsite	Distinct terms	0	0	22	0	0	0	4	**1,385**	7	0	**1,144**	**2,562**
	Frequency of use of these terms	0	0	6,481	0	0	0	7,918	**1,183,211**	65,114	0	**1,373,941**	**2,636,665**

Figure 21 – Statistical use of SNOMED-RT sections in the Anatomical Pathology unit
A : Physical Agents; C : Chemical, Drugs and Biological Products; D : Diseases/Diagnoses; F : Function; G : Modifiers; J : Occupations; L : Living Organisms; M : Morphology; P : Procedures; S : Social Context; T : Topography

		Morphology Sub-sections (SNOMED-RT)											TOTAL
		0	1	2	3	4	5	6	7	8-9		A-B		
										8	9	A	B	
Total terms (with synonyms)		657	656	431	980	422	626	782	755	**2,529**	**1,173**	27	5	**9,043**
Total distinct terms		293	328	191	498	155	326	494	398	**1,011**	**617**	20	5	**4,336**
Terms used onsite	Distinct terms	70	25	43	155	69	99	91	206	**420**	**207**	0	0	**1,385**
	Frequency	259,079	928	1,785	23,827	294,879	13,425	324,561	91,229	**145,289**	**28,209**	0	0	**1,183,211**

Figure 22 – Statistical use of SNOMED-RT Morphology sub-sections
(Sub-sections 8 and 9 use ICD-O codes)

		Topography Sub-sections (SNOMED-RT)																TOTAL
		0	1	2	3-4		5-6		7-8-9			A	B	C	D	E	F	
					3	4	5	6	7	8	9							
Total terms (with synonyms)		987	3,565	586	131	1,353	913	249	239	292	204	2,527	129	696	1,264	684	381	**14,200**
Total distinct terms		805	2,826	470	196	1,044	644	190	171	219	159	1,895	93	390	864	558	272	**10,796**
Terms used onsite	Distinct terms	187	160	89	28	69	103	34	21	44	24	79	17	66	158	55	10	**1,144**
	Frequency	160,611	18,985	33,083	4,197	1,363	245,255	17,034	8,001	153,442	12,729	4,098	20,564	43,484	9,497	612,962	28,636	**1,373,941**

Figure 23 - Statistical use of SNOMED-RT Topography sub-sections

Morphology and Topography are clearly the most used sections in our application for anatomical pathology.

For example, for the “Morphology” section, SNOMED-RT defines 9,043 terms from which 4,336 distinct concepts, i.e. which have different codes, exist. 1,385 of these distinct terms have been collected in our application and all of them have been collected 1,183,211 times.

- LOINC has been chosen for standardising IJB’s laboratory results as it is the most well-known vocabulary specialised in this domain. It is also usually imposed by scientific community by general consensus for laboratory data processing.
- ICD-O is imposed by the INAMI – the Belgium national health insurance – for reporting patients’ tumoral morphology in cancer registries, INAMI having then defined a mapping between ICD-O codes and their own codes. ICD-9-CM codes are also used for cancer registry reporting (e.g. for interventional procedures).
- MedDRA is used at IJB as a standard to report clinical trials follow up and response, and it is imposed as a standard for submitting CRFs’ entries to regulatory authorities.

- CTCAE is imposed by scientific community by general consensus, and is being used at IJB for recording the grading of adverse events.
- NCI Thesaurus has been imposed by an investigator within a clinical trial at IJB.
- ATC pharma codes are imposed as a standard for submission to regulatory authorities.

6.2.2 Universität des Saarlandes (UdS)

	DATA	TERMINOLOGIES			
		ICD	ICD-O	NCI CTCAE	MedDRA
UdS	nephroblastoma trial data of SIOP/GPOH			X	partly
	Hospital Information System (HIS)	X	X		
	Clinical cancer registry data	X	X		

Figure 24 - Medical terminologies used at UdS

- ICD/ICD-0 is used in the HIS dataset. This is also used in the Cancer registry of the Saarland. There is no usage of MedDRA nor CTCAE in these two databases. The Cancer registry is just a mortality registry and HIS is not collecting systematically adverse events.
- MedDRA and CTCAE are used in the nephroblastoma trial. But not all cases are coded according to MedDRA as this is not requested.

6.2.3 MAASTRO clinic

	DATA	TERMINOLOGIES		
		ICD	NCI Thesaurus	NCI CTCAE
MAASTRO	EMD (clinical)	X		X
	PACS (imaging)			
	euroCAT (clinical & imaging)		X	
	ZyLAB (clinical - OCR scans)			

Figure 25 - Medical terminologies used at MAASTRO

- NCI-Thesaurus has been chosen for the EU EUROCAT project as well as internally for encoding most clinical terms because it makes a good general medicine alternative to SNOMED and is free. We have found that it provides the most coverage for our domain, although no single terminology provides anything resembling full coverage.
- In our EHR, we use CTCAE specifically to note toxicities. CTCAE is an international standard for oncology and radiation oncology.

- ICD has been adopted as the Dutch national standard vocabulary for referring to diseases. We use ICD in the MAASTRO EHR system.

6.3 Method for the initial core dataset – Perspectives in WP4

To define the semantic core dataset, we need to extract a minimal subset of concepts that are needed to execute all the scenarios described in *Section 2*.

Indeed some of the scenarios already define their own concepts. Some of them are very specific whereas others define general concepts to be extracted (e.g. by Natural Language Processing techniques). The rest of the scenarios need at least a Decision support approach to define these concepts.

6.3.1 Scenarios that define their own concepts

Reporting episodes of febrile neutropenia, Cancer registry reporting

Here the concepts to extract are explicitly detailed in the description of these scenarios, and can be directly used to partly define the core dataset of these scenarios. These scenarios are enough granular to allow a direct determination of domain concepts on their own, by development for example of Natural Language Processing techniques within WP3.

The reporting of episodes of febrile neutropenia needs us to extract the following features (cf. scenario IJB_2):

- Temperature (*UMLS CUI: C0039476*)
- Number of neutrophils (*Neutrophils percentage measurement, C1171400*)
- Clinical and biological documentation of the infection
- Prior episodes of febrile neutropenia (*C0746883*)
- Date of admission (*C1302393*)
- Treatment
- Treatment drugs' name
- Determine whether the patient received antibiotics (*C0003232*) and/or prophylactic (*C0355642*) antifungals (*C0003308*)
- Determine whether the neutropenia is chemotherapy-induced (*C1827687*)
- Outcome of the episode of febrile neutropenia

The reporting in the cancer registry or in the tumour bank needs us to extract the following features (cf. scenario IJB_3):

- Incident tumours
- Recurrent tumours
- Date of incidence (*Date of diagnosis, C2316983*)
- Diagnosis (*C0011900*)
- Topographic (*Body region structure, C0005898*) and morphological (*Morphological type, C0445617*) information of the tumour
- Clinical stage (*Clinical stage finding, C0205563*)
- Pathologic stage (*C1320480*)
- Laterality (*C0332304*)
- Treatment start date (*Date treatment started, C1531783*)
- Nature of treatments, e.g. surgery (*C0543467*), chemotherapy (*C0184613*), radiotherapy (*Therapeutic or Preventive Procedure, C1522449*), hormone therapy (*C0279025*)

Patient recruitment into a trial,
Pre-filling of CRF and AE reports,
Long-term follow-up

The concepts to extract in these scenarios cannot be directly explicited, as they mostly depend on the specificity of the patients considered. Nevertheless the kind of document we have to take into account (protocols, CRF) can be formalised according to the amount of data that will be provided by clinical partners.
It will especially depend on:

- A formalisation of eligibility criteria from clinical trial protocols for patient recruitment into a trial (WP6),
- A formalisation of the fields we have to fill-in in the CRF and AE reports and thus the kind of data to extract for that purpose,
- A formalisation of the fields we have to fill-in in the long-term follow-up CRF and thus the kind of data to extract for that purpose.

6.3.2 Scenarios in which general concepts are defined

Automatic detection and reporting of SAEs/SUSARs

The concepts to extract in these scenarios (SAEs and SUSARs) are general concepts that need to be detected and then reported but that are not clearly explicit in their definition as it depends on the kind of treatments, on the patients and on the clinical trials that will be considered.

6.3.3 Scenarios in which concepts are defined by a Decision Support approach

Update of guidelines,
Hypothesis generation,
Supporting design of new trials,
Protocol feasibility,
Medical information recommender,
Data mining on consultation,
Contextualized overview,
Outcome prediction,
Diagnostic sarcoma classifier,
Analyse economic data between different procedures

The concepts to be considered in these scenarios could not be defined *apriori* like in all the other scenarios presented in the previous sections. They will have to be defined by a Decision Support approach. Most of the time, the key concepts will be extracted by data mining (e.g. e.g. on literature, on EHR/HIS/PHR data).

7 CONCLUSION

The definition of a semantic core dataset is a cornerstone in the development of the interoperability layer within the EURECA project. It will serve as the basis for connecting data between domains and disciplines, as well as between partners from different countries.

Some specific medical vocabularies are already in use by clinical partners for their internal applications and the management of their clinical data, making their data more easily standardised, reusable, and exploitable. In addition, some of these terminologies have been defined as ontologies, which is a strong advantage in the choice of elements for the core dataset because ontologies enable semantic processing of textual data.

The existence of mappings between ontologies is a good reason for not having to reinvent the wheel. It is one of the cornerstones of semantic interoperability between care and clinical trial systems that use different kinds of very specific vocabularies. For example, projects like UMLS and BioPortal are very interesting. They provide a way to aggregate and serve very large clinical terminologies from a single location via a unique identifier, as well as providing the means to create a one-to-one mapping of many vocabularies to each other.

The approach of the EURECA project to the definition of the core dataset must be based first on the scenarios, so that we can consider the effective point of interest to the users and thus the concrete goals and objectives of EURECA. The information provided by the scenarios varies, depending on the amount of precision and on the nature of the features to extract. While some scenarios provide a direct description of the elements to extract, others will need further analysis. In that case, raw data must be annotated according to the information we need to find and extract.

8 REFERENCES

[1] R. Vdovjak, B. Claerhout, and A. Bucur, “Bridging the gap between clinical research and care: approaches to semantic interoperability, security & privacy,” in *HealthInf Conference*, 2012.

[2] V. N. Stroetmann, D. Kalra, P. Lewalle, A. Rector, J.-M. Rodrigues, K. A. Stroetmann, G. Surjan, B. Ustun, M. Virtanen, and P. E. Zanstra, “Semantic Interoperability for Better Health and Safer Healthcare,” 2009.

[3] G. Weiler, M. Brochhausen, N. Graf, F. Schera, A. Hoppe, and S. Kiefer, “Ontology-based data management systems for post-genomic clinical trials within a European grid infrastucture for cancer research,” in *Engineering in Medicine and Biology Society (EMBS 2007), 29th Annual International Conference of the IEEE*, 2007.

[4] J. Golbeck, G. Fragoso, F. W. Hartel, J. Hendler, J. Oberthaler, and B. Parsia, “The National Cancer Institute’s Thésaurus and Ontology,” 2001.

[5] N. Sioutos, S. de Coronado, M. W. Haber, F. W. Hartel, W.-L. Shaiu, and L. W. Wright, “NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information,” *Journal of biomedical informatics*, vol. 40, no. 1, pp. 30–43, Feb. 2007.

[6] K. A. Spackman, K. E. Campbell, and R. A. Côté, “SNOMED RT: a reference terminology for health care,” in *AMIA Annual Fall Symposium*, 1997, pp. 640–4.

[7] T. Benson, *Principles of health interoperability HL7 and SNOMED*. 2010.

[8] J. Rogers and O. Bodenreider, “SNOMED CT : Browsing the Browsers,” *Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED)*, no. En 14463, pp. 30–36, 2008.

[9] C. J. McDonald, S. M. Huff, K. Mercer, J. A. Hernandez, and D. J. Vreeman, “Logical Observation Identifiers Names and Codes (LOINC) - Users’ Guide,” 2011.

[10] National Health Institute - National Cancer Institute, “Common Terminology Criteria for Adverse Events (CTCAE),” 2009.

[11] A. Fritz, C. Percy, A. Jack, K. Shanmugaratnam, L. Sobin, D. M. Parkin, and S. Whelan, *International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3)*. 2000.

[12] World Health Organization, “WHO Drug Dictionary Sample Technical description,” 2006.

[13] F. Mougin, M. Dupuch, and N. Grabar, “Improving the mapping between MedDRA and SNOMED CT,” *Artificial Intelligence in Medicine*, pp. 220–224, 2011.

[14] P. L. Whetzel, N. H. Shah, N. F. Noy, B. Dai, M. Dorf, N. Griffith, C. Jonquet, C. Youn, A. Coulet, C. Callendar, D. L. Rubin, B. Smith, M. Storey, C. G. Chute, and M. A. Musen, “BioPortal : Ontologies and Integrated Data Resources at the Click of a Mouse,” *Nucleic acids research*, vol. 37, pp. 170–3, 2009.

[15] A. Ghazvinian, N. F. Noy, C. Jonquet, N. Shah, and M. A. Musen, “What Four Million Mappings Can Tell You About Two Hundred Ontologies,” in *8th International Semantic Web Conference (ISWC 2009), Washington DC, Springer*, 2009.

[16] C. Caracciolo, J. Euzenat, L. Hollink, R. Ichise, A. Isaac, V. Malaisé, C. Meilicke, J. Pane, P. Shvaiko, H. Stuckenschmidt, O. Šváb-Zamazal, and V. Svátek, “Results of the Ontology Alignment Evaluation Initiative 2008,” in *The Third International Workshop on Ontology Matching at ISWC*, 2008.