ICT-2011-288048

EURECA

Enabling information re-Use by linking clinical REsearch and CAre

IP
Contract Nr: 288048

Deliverable: 4.5
Extension of the core data sets

Due date of deliverable: (31-01-2015)
Actual submission date: (17-03-2015)

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

Responsible WP: WP 4

Revision: final

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
Kerstin Rohm	FhG-IBMT	Kerstin.rohm@ibmt.fraunhofer.de
Gabriele Weiler	FhG-IBMT	Gabriele.weiler@ibmt.fraunhofer.de
Ahmed Ibrahim	Philips	Ahmed.a.ibrahim@philips.com
Sergio Paraíso Medina	UPM	sparaiso@infomed.dia.fi.upm.es
Santiago Aso	UPM	saso@infomed.dia.fi.upm.es
Scott Marshall	MAASTRO	m.scott.marshall@maastro.nl
Sheng Yu	UOXF	sheng.yu@oncology.ox.ac.uk

0.2 Documents history

Document version #	Date	Change
V0.1	23/09/2014	TOC
V0.2	05/11/2014	Improved TOC
V0.3	18/12/2014	Version with contribution from partners
V0.5	26/01/2015	Improved version within serval comments and questions
V0.6	20/02/2015	Complete clinical datasets
V0.7	04/03/2015	Improved version for internal review
V1.0	17/03/2015	Final version after internal review

0.3 Document data

Keywords	
Editor Address data	Name: Kerstin Rohm Partner: Fraunhofer-IBMT Address: Ensheimer Str. 48, 66386 St. Ingbert, Germany E-Mail: Kerstin.rohm@ibmt.fraunhofer.de

0.4 Distribution list

Date	Issue	E-mailer
		fp7-eureca-all@listas.fi.upm.es
		INFSO-ICT-288048@ec.europa.eu
		Benoit.abeloos@ec.europa.eu

Table of Contents

1 Introduction

During the EURECA project activities, a specific platform has been developed to enable a harmonized and secure access to data from different clinical domains and to accommodate the re-use of all available data. The semantic data processing of this platform is realised by the so-called “Semantic Interoperability Platform”.
The Semantic Interoperability Platform hosts the clinical data and describes them by a consistent data model – the EURECA Common Information Model (CIM). The CIM consists of the EURECA Common Data Model (CDM) and the EURECA Core Dataset. The CDM is a HL7 RIM based data repository, which is terminology linked to the EURECA Core Dataset (see Figure 1). [2]

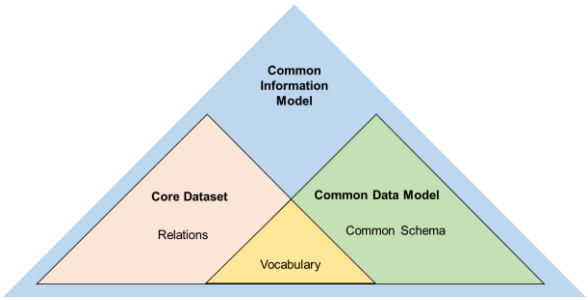


Figure 1: Common Information Model

The Core Dataset represents the required medical vocabulary as well as the relationships among these concepts through the standard vocabularies:

- SNOMED-CT (Systematized Nomenclature Of Medicine Clinical Terms) [1]
- LOINC (Logical Observation Identifiers Names and Codes) [2]
- HGNC (HUGO Gene Nomenclature Committee) [3]

The composition of the Core Dataset’s ontology concepts results from the evaluation of the clinical datasets in the domains of breast and lung cancer as well as sarcoma nephroblastoma. Deliverable 4.2 describes these evaluations in detail. It was identified, that SNOMED-CT covers most of the evaluated medical concepts, but specific laboratory analyses and radiotherapy concepts demonstrated the requirement for additional concepts from HGNC and LOINC in the Core Dataset. Deliverable 4.2 also concluded that a relatively small group of concepts are occurring in a large number of clinical trials. [1]

Several reasons, such as the inclusion of new medical domains within its specific datasets, require an extension of the Core Dataset from time to time. For each of these reasons, it has to be evaluated, if the Core Dataset covers the concerning new datasets and its concepts sufficiently. Otherwise, the Core Dataset needs to be updated.

[1] Systematized Nomenclature Of Medicine Clinical Terms (SNOMED-CT), http://www.ihtsdo.org/snomed-ct/
[2] Logical Observation Identifiers Names and Codes (LOINC®) - http://www.loinc.org
[3] HUGO Gene Nomenclature Committee, http://www.genenames.org/

This deliverable describes in detail the EURECA specific technologies, which facilitate a Core Dataset extension with a relatively small effort.

As the clinical partners discovered the EURECA platform increase by the inclusion of new datasets, which are rectal and head/neck cancer domains and an extension of the existing breast cancer dataset, this document shows exemplary specific analyses and evaluations in order to detect if the Core Dataset covers the concepts of the new datasets sufficiently. The evaluation results and the consequences to the Core Dataset are also discussed.

1.1 Structure of the Deliverable

This deliverable is structured as follows:

Chapter 2 summarises the technical characteristics and structure of the Core Dataset. The necessary features to enable an extension of the Core Dataset are also shown.

Chapter 3 describes specific analyses in order to evaluate if the Core Dataset covers the concepts of the new datasets in the head/neck and rectal cancer domains sufficiently. These analyses are based on the eligibility criteria in publicly available datasets.

In chapter 4, the concepts of the Core Dataset are validated against the clinical datasets of two clinical partners:
First, the datasets from the MAASTRO clinic, which belong to the domain of either head/neck, or rectal cancer. Second, the dataset from the University of Oxford, which belongs to cancer follow up including adverse events and drugs, as an amendment of the previously evaluated datasets of deliverable 4.2.

2 The EURECA Core Dataset

The Core Dataset is an essential part of the EURECA Semantic Interoperability Platform. The platform homogenizes the different clinical data sources and enables a uniform access to these data by the EURECA end user applications (see Figure 2).

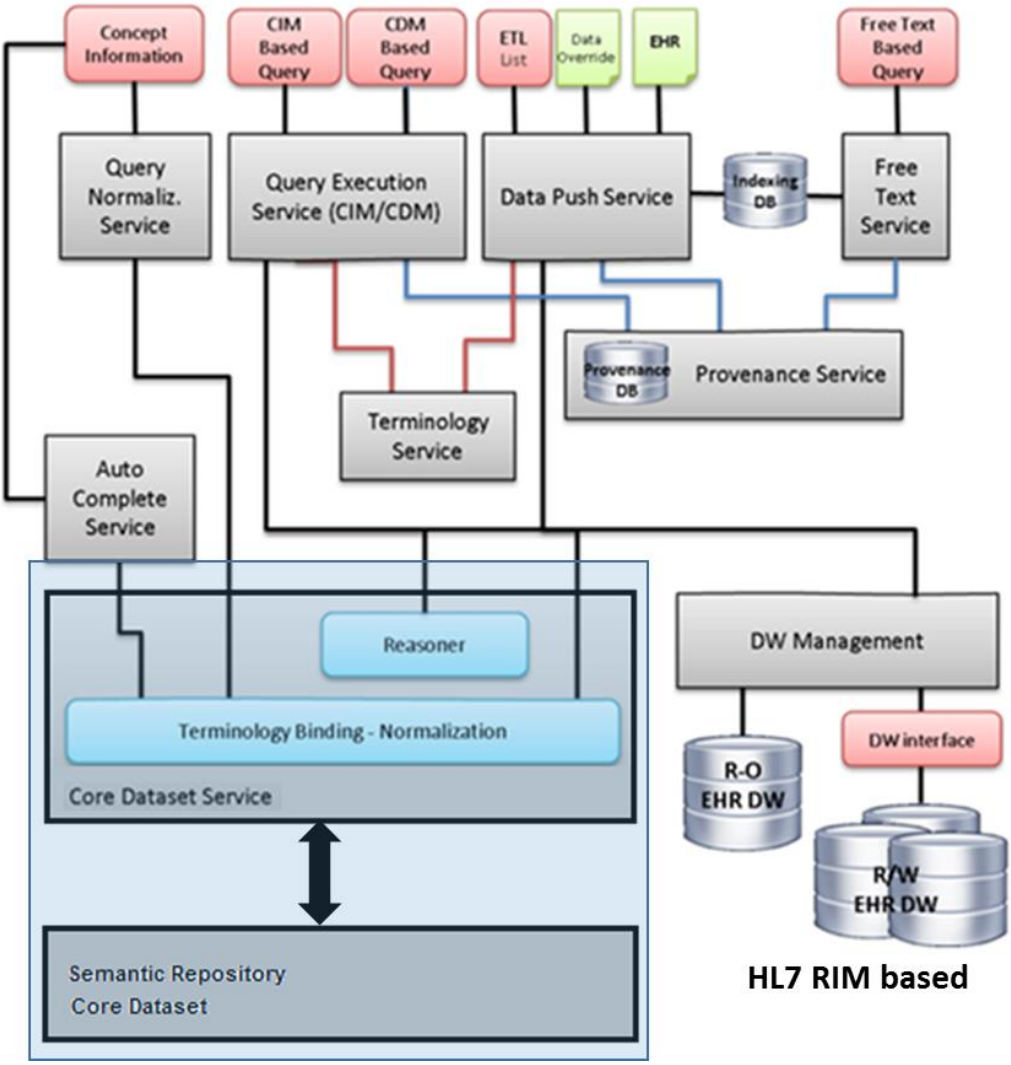


Figure 2: Semantic Interoperability Platform

This is primarily realised through a data description by a consistent data model - the EURECA CIM. The CIM consists of the CDM, the Core Dataset and the relations between them.
The CDM enables a structured storage of the clinical data following a common schema (HL7 RIM based data repository) and provides mechanisms to link the medical vocabulary (see Figure 1).
The Core Dataset contains the medical vocabulary, relationships between the terminology concepts, and specific services in order to infer semantic knowledge to the different components of the EURECA platform.

The current chapter describes the structure and characteristics of the Core Dataset. The reasons for which a Core Dataset extension and the implemented technologies are required are also described.

Beside the technical details of this chapter, the focus of chapter 3 and 4 will be on the evaluation of the new datasets and its resulting consequences for the Core Dataset.

2.1 Structure of the Core Dataset

As mentioned before, the Core Dataset contains the medical vocabulary, the relationships between the terminology concepts and specific services in order to infer semantic knowledge for the different components of the EURECA platform. The following section describes in detail the structure and features of the Core Dataset.

2.1.1 Core Dataset Vocabularies

The Core Dataset stores and represents the medical datasets of the EURECA specific cancer domains. Deliverable 4.2 showed that a sub-set of the following standard vocabularies represents the required clinical datasets optimally:

- SNOMED-CT
- LOINC
- HGNC

SNOMED-CT covers most of the required clinical concepts. *LOINC* covers specific laboratory results and *HGNC* represents the unique and meaningful names of every known human gene. Details of the Core Dataset definition can be found in deliverable 4.2. [1]

2.1.1.1 Pre- and post-coordinated methods

SNOMED-CT provides specific functionalities that enable the management of pre- and post-coordinated concepts and thereby allow a more efficient administration of the huge amount of SNOMED-CT terms. In particular, the post-coordinated methods offer the possibility to “build” specific medical concepts, which are not directly covered by the Core Dataset concepts. This functionality will be exemplified in chapter 4 using the new medical datasets of the medical partners.
Through ***pre-coordinated methods***, it is possible to picture the “is_a” relationship of a specific concept like shown for the antibiotic Anthracycline in Figure 3.

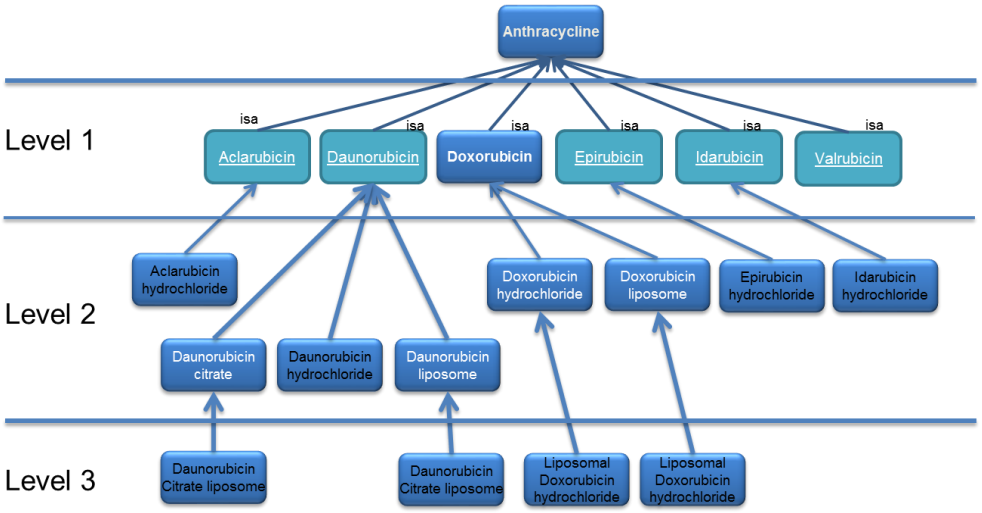


Figure 3: Hierarchical tree structure of the antibiotic Anthracycline in SNOMED-CT (pre-coordinated structure

Post-coordinated methods provide the possibility to create a specific clinical term with two SNOMED-CT codes such as shown in Figure 4 for "Histologically-confirmed breast cancer".

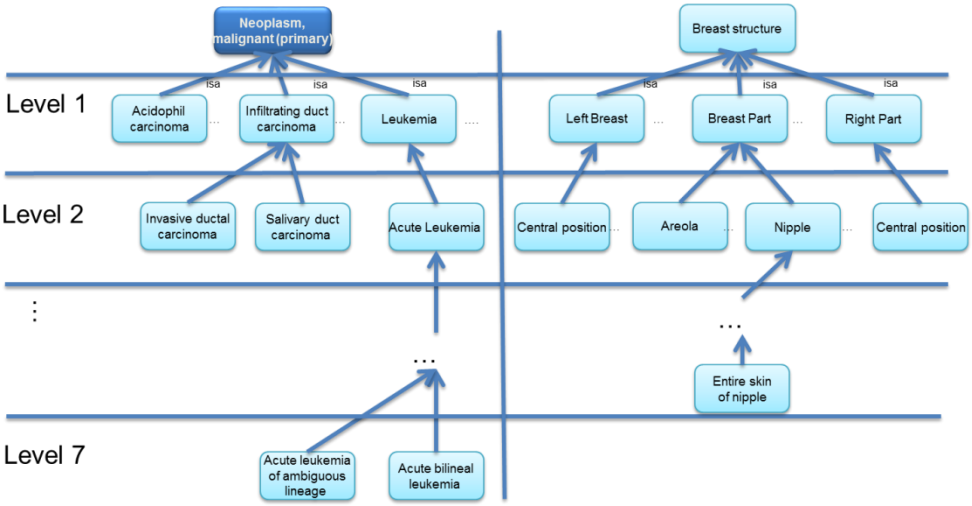


Figure 4: Post-coordination concept for "histologically confirmed breast cancer"

2.1.2 Semantic Repository of the Core Dataset

The specific semantic requirements, in particular the representation of the medical vocabulary and the relationships between the medical concepts, are technically realised

through a semantic repository – a so-called Sesame Server[4]. The Sesame Server provides a framework for managing and querying RDF data as terminologies that are stored as OWL[5] resources.

Therefore, a EURECA specific OWL file has been created to represent the medical concepts in triples. The OWL file represents the Core Dataset and comprises of SNOMED-CT, LOINC, and HGNC concepts and their linkage to the CDM. [2]
The following example shows a fragment of this OWL file (SNOMED-CT concept 245849007 "Post-surgical breast structure" and its linkage to the CDM).

```
<owl:Class rdf:about="SCT_312285003">
        <rdfs:label xml:lang="en">
                Post-surgical breast structure (morphologic abnormality)
        </rdfs:label>
        <skos:altLabel xml:lang="en">Post-surgical breast structure</skos:altLabel>
        <prv:containedBy rdf:resource="http://test.org#Procedure_targetSiteCode"/>
        <prv:containedBy rdf:resource="http://test.org#Observation_code"/>
        <prv:containedBy rdf:resource="http://test.org#Observation_targetSiteCode"/>
        <rdfs:subClassOf rdf:resource="SCT_245849007"/>
</owl:Class>
```

2.1.3 The Core Dataset Services

The Core Dataset Services enable the interaction of the Core Dataset, available through the semantic repository, with the components of the EURECA platform. The available services are the following:

- The Auto-Complete Service
- The Reasoner Service
- Terminology Binding Service

Figure 5 visualises the functionalities of the Core Dataset Services. These functionalities are described in detail in the following sections.

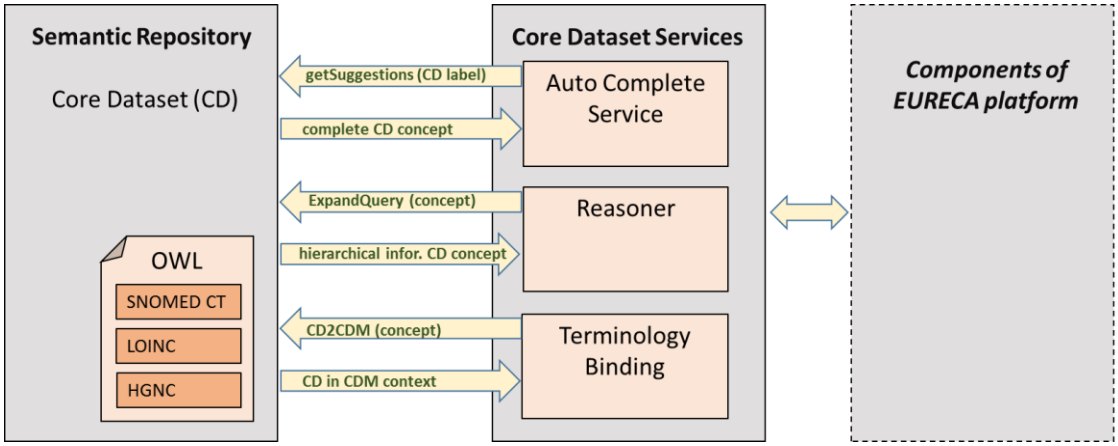


Figure 5: Core Dataset Services

[4] http://rdf4j.org/
[5] Web Ontology Language, http://www.w3.org/TR/owl-semantics/

2.1.3.1 The Auto Complete Service

The Auto Complete Service enables retrieval of detailed information about a specific Core Dataset Concept. This is realised through the getSuggestions (parameter: Core Dataset label) function. The function returns for a specific label, for example “Neoplasm”, an object containing title, code, and the CDM context, such as “concept: Neoplatic disease | 55342001, isDiagnosis”. [3]

2.1.3.2 Reasoner

The Reasoner enables getting hierarchical knowledge out of a specific SNOMED-CT Core Dataset concept. It is possible to retrieve all “is_a” relationships of a specific concept, using the pre-coordinated structure of SNOMED-CT (see Figure 3).

2.1.3.3 Terminology Binding

This component is used for the linkage of the EURECA CDM and the Core Dataset concepts. The CD2CDM function of this component enables the extraction of the necessary information from the Core Dataset in order to map a specific concept to the CDM.
The following example shows part of the response of the CD2CDM function for the concept “106221001 (Genetic finding)”. [2]

```
<ax21:code xsi:type="ax21:classifiedConcept">
   <ax21:attribute>code</ax21:attribute>
   <ax21:code>106221001</ax21:code>
   <ax21:label>Genetic finding (finding)</ax21:label>
   <ax21:table>Observation</ax21:table>
</ax21:code>
```

2.1.4 Management tools of the Core Dataset

The following tools are used to maintain the mentioned EURECA OWL file:

- Protégé[6]: Free and open-source ontology editor and framework for building intelligent systems.
- Knoodl[7]: Distributed Information Management System (DIMS). Contains tools for creating, managing, analysing and visualizing RDF/OWL descriptions.
- NeOn Toolkit[8]: Open source multi-platform ontology engineering environment. Includes management, editing and visualization functionalities for ontologies, OWL/RDF support, several import mechanisms and support for reuse and management of networked ontologies.

Protégé is used for searching inconsistencies and the generation of specific OWL extensions. Knoodl and NeOn are used for collaborative ontology edition and management of the OWL file.

[6] http://protege.stanford.edu/
[7] http://knoodl.com/ui/home.html
[8] http://neon-toolkit.org/wiki/Main_Page

2.2 Reasons for a Core Dataset Extension

The Core Dataset needs to be periodically re-worked and updated. The reasons are described here, whereby the following sections 2.3 and 2.4 are focusing on the technology in order to process such an extension without much effort.

- **New versions of standard vocabularies:** Vocabularies are constantly updated by their contributors. This updates imply that a given concept might change, become deprecated or is added (for more details see section 2.4.1).
- **New medical domains:** Before the datasets of a new medical domain can be represented by the EURECA platform, it has to be examined whether the concepts of the Core Dataset are able to semantically represent these datasets. Examples of this process are shown in chapter 3 and 4 of this document. The result of these analyses can require a Core Dataset Extension (for more details see section 2.4.2)

2.3 Technologies for a Core Dataset Extension

The extension of the Core Dataset's datasets, require an update of the OWL file that stores the Core Dataset specific concepts and its mappings to the CDM. This process is primarily performed by two specific applications:

- The **Script Generator**: An EURECA specific script is used to update (or initially create) the OWL file. It is based on a script for generating an OWL file of the SNOMED-CT release provided by IHTSDO[9]. The original script was extended for the inclusion of LOINC and HGNC terminologies and to enable the storage of the mappings of the Core Dataset concepts to the CDM.

- The **TermBindingLoader:** This application is a Java program that analyses and determines the mapping formalisms between the Core Dataset terminologies and the HL7 RIM based CDM.

The whole OWL update requires six (6) specific steps, but in some cases it is sufficient to update the linkage of the Core Dataset concepts to the CDM (steps 3-6). The steps of a complete Core Dataset update can be summarised as follows:

1. Gather the different release files of the vocabularies (e.g. new SNOMED CT release from IHTSDO).
2. Execute the *Script Generator* within these release files in order to produce a new version of the OWL file that contains the updates of respective vocabularies.
3. Evaluate, whether the terminology binding rules of the *TermBindingLoader* have a mapping from the vocabulary concepts to the CDM objects and update those rules where required.
4. When the binding rules are completed (step 3), the *TermBindingLoader* can generate a set of files which contain the mapping information between the concepts of the vocabularies and the CDM.

[9] The International Health Terminology Standards Development Organisation http://www.ihtsdo.org/

5. The generated files of step 4 have to be used as an input for the *Script Generator* and the script itself has to be executed again in order to update the mappings to the CDM.
6. After a second execution of the *Script Generator*, the OWL resource includes the required mappings to the CDM and can be deployed at the Sesame server.

2.4 The Core Dataset extension

The following section describes, based on concrete examples, when the described technologies of the previous section needs to be executed in order to extend the Core Dataset.

2.4.1 New versions of standard vocabularies

The standard vocabularies represents the required concepts of the Core Dataset. The contributors update these vocabularies constantly, implying that a given concept might change, can become deprecated, or is added.
The SNOMED-CT versioning is managed by the SNOMED-CT International Request Submission (SIRS[10]), which is maintained and distributed by the IHTSDO.
LOINC has a similar versioning system[11] (also maintained by the IHTSDO), as well as HGNC does[12].
The contributors provide these new versions by one or more release files. These release files initiate a complete Core Dataset update, as described in step 1-6 above.

2.4.2 New medical domains

The Core Dataset tailored to the EURECA specific clinical domains, which are breast cancer, lung cancer, sarcoma and nephroblastoma. If datasets of new medical domains have to be stored and used by the EURECA platform, it has to be examined whether the concepts of the Core Dataset are able to represent semantically these datasets (see chapter 3 and 4).
Otherwise, the Core Dataset has to be extended. If an extension is required, it is usually sufficient to update the binding rules of the identified not covered concepts to the CDM (see section 2.4.2.1), but sometimes it is required to add an entire new vocabulary to the Core Dataset (see section 2.4.2.2).

2.4.2.1 Terminology binding extension and modification

As mentioned before, the concepts of complex vocabularies such as SNOMED-CT have to be linked to the HL7 RIM based CDM.
Therefore, each Core Dataset concept has to have a mapping to a specific object of the CDM. An unknown terminology binding of a concept results in an unknown storage place in the CDM (see figure 6).

[10] https://sirs.nlm.nih.gov/
[11] https://loinc.org/submissions
[12] http://www.genenames.org/cgi-bin/request

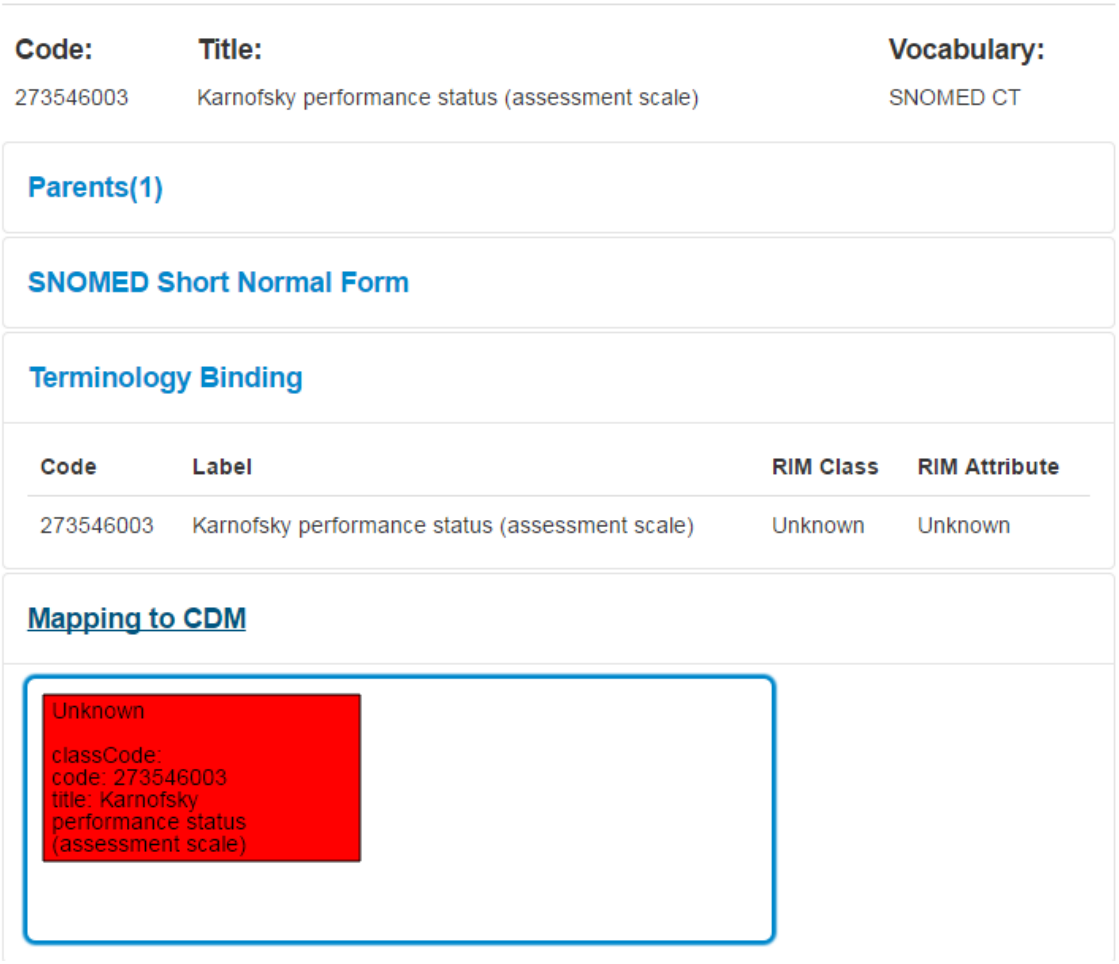


Figure 7: Example of unknown mapping

The proposed solution is to find another convenient concept (e.g. with the specific SNOMED-CT methods, see section 2.1.1.1).
If this workaround cannot be achieved, a new mapping rule (Core Dataset concept to CDM) has to be added - meaning the terminology binding rules of the TermBindingLoader have to be modified and the OWL file has to be updated (step 3-6 of the mentioned workflow above must be re-executed). After that, the OWL resource is updated with the new relationships and mappings and the new concept is completely integrated in the Core Dataset.

2.4.2.2 Add entire vocabulary to the Core Dataset

The inclusion of datasets of new medical domains could require a completely new standard vocabulary to represent the required concepts with the Core Dataset.

As a first step, the concepts of the new vocabulary have to be analysed considering how to integrate them into the specific Core Dataset structure. The OWL file provides specific labels to enable the inclusion of the concepts of new vocabularies.
The following OWL file fragment explains these characteristics in detail:

- owl:Class: Core Dataset concept
- rdfs:label: label of the concept

- skos:altLabel: alternative labels for the concept
- prv:containedBy: mapping of the concept in the CDM
- rdfs:subClassOf: direct parents of the concept
- owl:Restriction: restriction block in owl
- owl:onProperty: relationships of the concepts (e.g. SCT_116676008 is the associated morphology relationship)
- owl:someValuesFrom: value associated to the relationship (e.g. the associated morphology relationship, SCT_78319006 is the Nonvenomous insect bite (morphologic abnormality))

```
<owl:Class rdf:about="SCT_15034009">
  <rdfs:label xml:lang="en">Nonvenomous insect bite of breast with infection (disorder)</rdfs:label>
                                                                                    Synonym
  <skos:altLabel xml:lang="en">Insect bite, nonvenomous, of breast, infected</skos:altLabel>
  <skos:altLabel xml:lang="en">Nonvenomous insect bite of breast with infection</skos:altLabel>

                                                                                    CDM Binding
  <prv:containedBy rdf:resource="http://www.gib.fi.upm.es/hl7rim-common-data-model/#Observation_code"/>

                                                                                    hierarchical relationship
  <rdfs:subClassOf rdf:resource="SCT_211105005"/>
  <rdfs:subClassOf rdf:resource="SCT_56892000"/>
  <rdfs:subClassOf rdf:resource="SCT_23004002"/>
  <owl:equivalentClass><owl:Class>
  <owl:intersectionOf rdf:parseType="Collection">
    <owl:Class rdf:about="SCT_211105005"/>
    <owl:Class rdf:about="SCT_56892000"/>
    <owl:Class rdf:about="SCT_23004002"/>
    <owl:Restriction>
      <owl:onProperty rdf:resource="RoleGroup"/>
      <owl:someValuesFrom>
        <owl:Class>
        <owl:intersectionOf rdf:parseType="Collection">
          <owl:Restriction>
            <owl:onProperty rdf:resource="SCT_116676008"/>
            <owl:someValuesFrom rdf:resource="SCT_78319006"/>
          </owl:Restriction>
          <owl:Restriction>
            <owl:onProperty rdf:resource="SCT_363698007"/>
            <owl:someValuesFrom rdf:resource="SCT_82038008"/>
          </owl:Restriction>
        </owl:intersectionOf>
        </owl:Class>
      </owl:someValuesFrom>
    </owl:Restriction>
    <owl:Restriction>
      <owl:onProperty rdf:resource="RoleGroup"/>
      <owl:someValuesFrom>
        <owl:Restriction>
          <owl:onProperty rdf:resource="SCT_370135005"/>
          <owl:someValuesFrom rdf:resource="SCT_441862004"/>
        </owl:Restriction>
      </owl:someValuesFrom>
    </owl:Restriction>
  </owl:intersectionOf>
  </owl:Class></owl:equivalentClass>
</owl:Class>
```

Steps 1-6 need to be executed in order to rebuild the OWL file. The Script Generator has to generate the new OWL file and the concepts of the new vocabulary must be analysed in order to update the terminology binding to the CDM (see previous section). Chapter 2 summarised the technical characteristics of the current Core Dataset. It described the structure, which enables semantic storage of medical data by the standard

vocabularies SNOMED-CT, LOINC and HGNC. Furthermore, the available features and required steps of a Core Dataset extension have been discussed.

The clinical partners focused out that the Core Dataset should cover a wider range of medical datasets. As the inclusion of datasets in the domain of head/neck and rectal cancer are foreseen, the following chapter describes the required evaluation of publicly available datasets and their technical consequences for the Core Dataset.

3 Evaluation of datasets for the EURECA Core Dataset extension

As chapter 2 summarises the technical characteristics of the Core Dataset that enable a Core Dataset extension, this chapter describes the required preceding dataset evaluation in order to detect whether the Core Dataset covers new medical domains sufficiently.

As the clinical partners decided that the inclusion of datasets in the cancer domains head/neck and rectal cancer will increase the medical benefit of the EURECA infrastructure, the chapter will further describe the evaluation of these datasets and the resulting consequences for the Core Dataset. These analyses were carried out through the eligibility criteria in these domains in public available datasets[13].

These results need to be validated through “real” datasets of the clinical partner. These clinical validations are described using the MAASTRO datasets in chapter 4.

3.1 Identification of datasets for extension by public available clinical trial data

In Deliverable 4.2, we analysed eligibility criteria of clinical trial data in the domains of breast cancer, lung cancer, sarcoma and nephroblastoma. In this section, we extend the analysis by also representing the semantics of clinical trials in the domain of head/neck cancer, and rectal cancer (Table 1). As in Deliverable 4.2, we use the BioPortal annotator to annotate the semantics of the criteria with the medical ontologies: SNOMED-CT and LOINC.

Clinical domain	***Number of trials***
Head and neck cancer	931
Rectal cancer	198

Table 1: Number of clinical trials’ eligibility criteria extracted from ClinicalTrial.gov

3.1.1 Evaluation of the results

Table 2 shows the averages of the number of ontology concepts per trial in the investigated domains and for two ontologies.

	SNOMED-CT	***LOINC***
Head and neck cancer	89	69
Rectal cancer	86	65

Table 2: Average number of concepts per trial

We evaluated the occurrences of concepts across different trials, as shown in Figure 8. The curves, shown in figures (a-d), indicate that a large ratio of the criteria is similar. These findings are quite similar to the findings in Deliverable 4.2. There we noticed that there was also a high re-occurrence of concepts in the domains of breast cancer and lung cancer.

[13] https://clinicaltrials.gov/

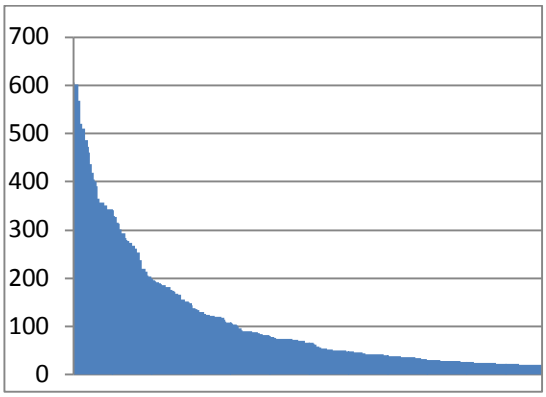


Figure 7a LOINC: count of unique concepts for head and neck cancer

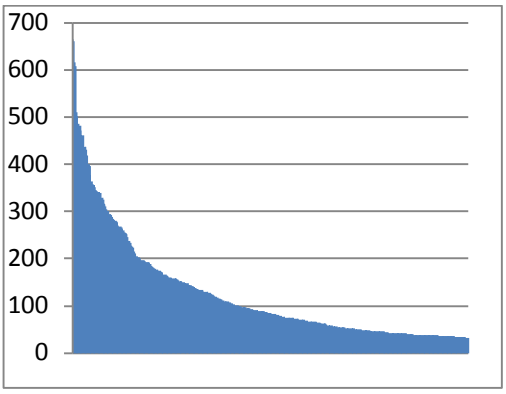


Figure 7b SNOMED-CT: count of unique concepts for head and neck cancer

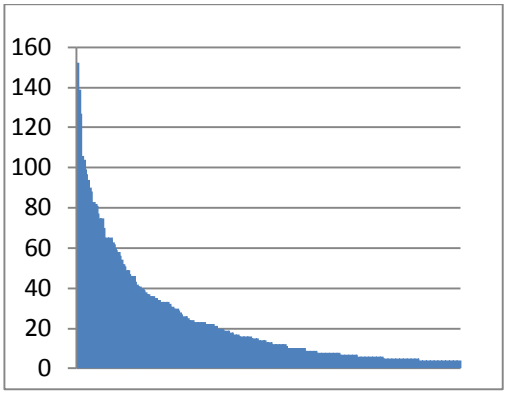


Figure 7c LOINC: count of unique concepts for rectal cancer

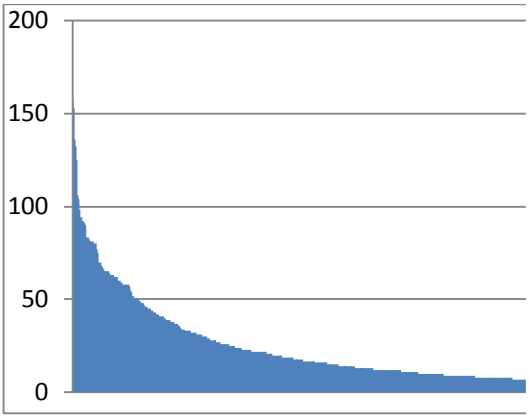
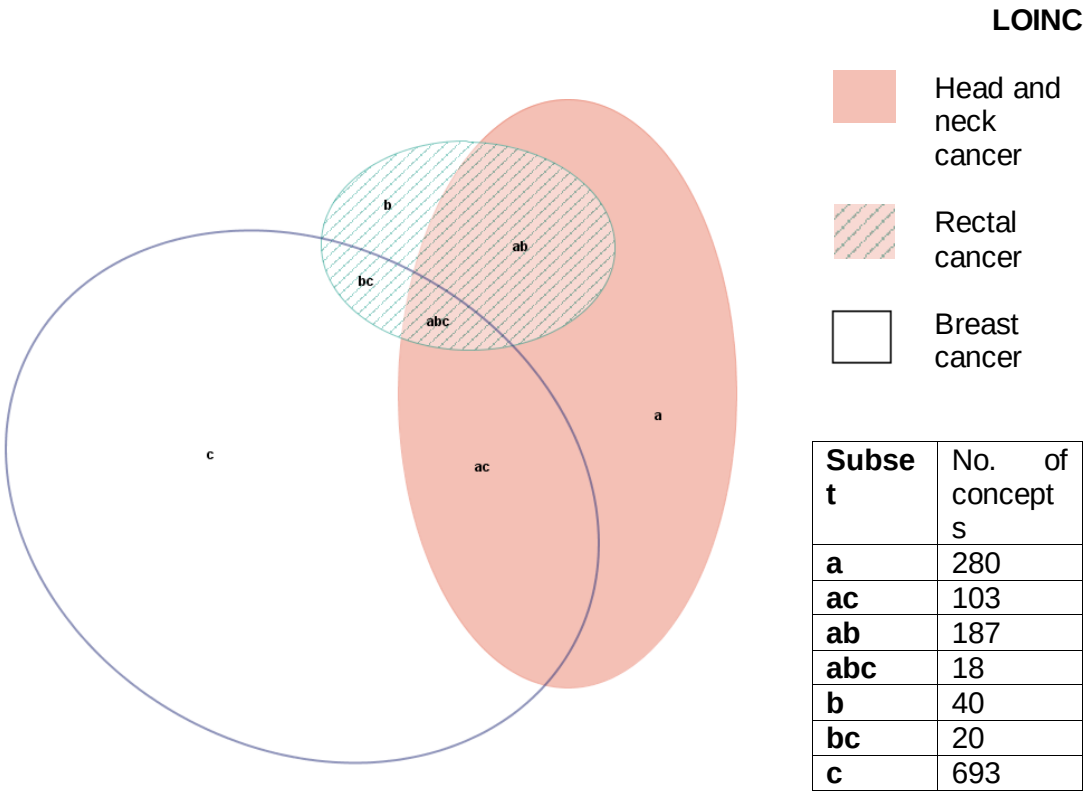


Figure 7d SNOMED-CT: count of unique concepts for rectal cancer

Figure 8: The reoccurrences of concepts across trials. The number of head and neck cancer, and rectal cancer (y-axis) trials that include the top 500 most frequently occurring concepts (x-axis) out of SNOMED-CT and LOINC.

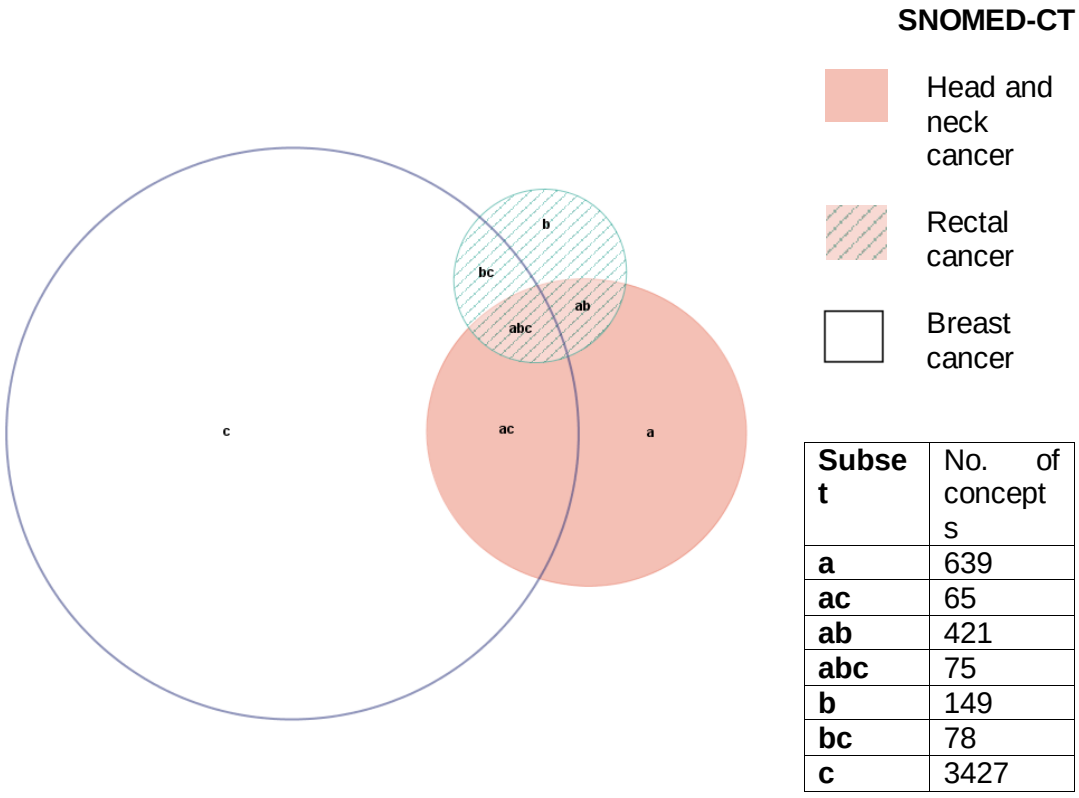
3.1.1.1 Subset of Concepts

Figure 9 and Figure 10 compare the sets of the concepts for the different domains and two ontologies. We also included the breast cancer analysis results presented in deliverable 4.2, as this is currently the largest domain within our semantic solution. As opposed to the results presented in deliverable 4.2, which included high overlaps among the eligibility criteria of the breast cancer and lung cancer domains, the figures below show that there is very little overlap between the head and neck cancer domains, and the rectal cancer and breast cancer domains.



Subset	No. of concepts
a	280
ac	103
ab	187
abc	18
b	40
bc	20
c	693

Figure 9: Sets of LOINC concepts for head and neck cancer, rectal cancer and breast cancer



Subset	No. of concepts
a	639
ac	65
ab	421
abc	75
b	149
bc	78
c	3427

Figure 10: Sets of SNOMED-CT concepts for head and neck cancer, rectal cancer and breast cancer

3.1.2 Consequences for the EURECA Core Dataset

This section discusses the evaluations of the previous sections, in order to determine the consequences to the Core Dataset. We restricted these analyses to SNOMED-CT occurrences, because it is the main Core Dataset vocabulary.

The identified SNOMED-CT concepts from clinicaltrial.gov for head/neck and rectal cancer trials have been annotated automatically (4619 for head/neck and 2237 for rectal cancer). These concepts have been ranked against the total number of occurrences. The first one hundred concepts have been examined by manually checking their presence in the Core Dataset and their classification to HL7 RIM-based CDM classes.

Approximately 50% of the concepts (55% head/neck, 49% rectal) have been directly classified and mapped to classes of the CDM. The rest of the concepts could not be classified directly. They can be divided into three different types:

1. Qualifier value concept
2. Concept not included in the Core Dataset
3. Social concept

1. Refinements provide the possibility to describe the SNOMED-CT concepts precisely by attribute-value pairs. For example, the concept “fracture of femur” can be refined through the attribute “associated morphology” value “compression fracture” pair.

71620000|fracture of femur|:116676008|associated morphology|=21947006|compression fracture

This attribute-value combination is important in order to classify the complete semantic meaning of a specific concept. However, some specific value concepts are used for applying qualifying characteristics to a concept (e.g. “left”, “severe”). They are called “***Qualifier value***”.
Qualifier values do not have a complete meaning by themselves and they are not represented in the SNOMED-CT hierarchy (see section 2.1.1.1). For example, “Acute bronchitis’” is decomposed into the SNOMEC-CT pre-coordinated concepts “Bronchitis (disorder)” and the qualifier attribute-value pair “Course-Acute”, where ‘acute’ is the qualifier value. Therefore, qualifier values do not have a direct mapping to the CDM. However, they are mapped to the CDM as an additional information to the main SNOMED-CT concept, which they are “qualifying”. The term ‘Left breast’ for example, can be decomposed as “Breast structure” with laterality ‘”left”. Both concepts “breast structure” and ‘”left” are linked to “targetSite” attribute of the CDM.

Qualifier values represent approximately 80% of the mentioned unclassified concepts. This high number result because the annotation was performed by using an automatic tool on a free-text data source. This tool annotates all components of the post-coordinated concepts as single words (as qualifier values).

2. Some of the evaluated concepts are currently ***not included*** in the Core Dataset (approximately 13% of the mentioned unclassified concepts). Most of these concepts belong to previous SNOMED-CT releases, and are marked as deprecated in the SNOMED-CT version that is used in the Core Dataset (see section 2.4.1).

3. ***Social concepts*** are used to represent social conditions and circumstances significant to health care. They represent 7% of the cases of unclassified concepts. In the current state of the Core Dataset, this set of concepts is not mapped to any attribute of the CDM. Usually, it is sufficient to store this kind of information by specific CDM attributes (e.g. race code, marital status).

The analyses of this chapter showed that the Core Dataset covers most annotated concepts from the analysed eligibility criteria that are necessary to describe the datasets in the cancer domains head/neck and rectal sufficiently.

At the first sight, the high number of detected SNOMED-CT qualifier values could prove the greatest challenge for the inclusion of the new cancer domains, because they can lead to an incomplete semantic meaning. However, a more detailed analysis showed, that most of those concepts can be also pictured by regular post-coordinated SNOMED-CT concepts in order to avoid these problems.

Some concepts, in particular most of the social concepts, can be sufficiently represented using specific CDM attributes. Therefore, these concepts do not require an explicit concept in the EURECA Core Dataset.

Finally, to summarise the above observations, a) it can be concluded that the OWL file needs to be updated by a few specific SNOMED-CT concepts and b) the terminology binding rules need to be updated as well in order to sufficiently represent the new cancer domains with the Core Dataset. These steps are described in section 2.3 ”Technologies for a Core Dataset Extension” in detail.

Our evaluation results have been validated by “real” clinical datasets of MAASTRO. This process is described in the following chapter.

4 Evaluation of Core Dataset extensions by clinical partners

Chapter 2.2 showed why the medical concepts, which are described through the Core Dataset, are changing continuously. This constant change requires the medical datasets of the clinicians to be evaluated frequently to ensure that the datasets of the clinicians can be represented semantically using the functionalities of the EURECA platform.

During the project activities, the clinical partners concluded that datasets in the head/neck and rectal cancer domains would improve the medical impact of the EURECA platform. Therefore, public available datasets of these domains were evaluated in order to identify the corresponding required Core Dataset extensions (see previous chapter). The first part of this chapter validates these evaluation results with “real” datasets of the clinical partner Maastricht Radiation Oncology (MAASTRO).

Additionally, the University of Oxford (UOXF) determined that the Core Dataset should also represent follow up, adverse event and drug data of their breast cancer datasets. While the general breast cancer dataset was already validated in the deliverable 4.2 [1], the second part of this chapter validates whether the expanded breast cancer datasets could be covered with the Core Dataset.

4.1 Evaluation of MAASTRO Datasets

This section analyses whether the evaluated Core Dataset (see previous chapter) covers the datasets of the new selected cancer domains sufficiently. We describe if the Core Dataset concepts cover the concepts contained in the ontologies and umbrella protocols that MAASTRO employs for rectal cancer and head/neck cancer.

The result of the coverage of the MAASTRO datasets is listed in Table 3 and Table 4, where for each concept it is indicated whether it is covered by the standard ontologies of the Core Dataset. As can be seen the coverage especially from SNOMED-CT and LOINC is good. The not covered concepts are marked red and some not applicable concepts are marked blue.

It should be mentioned that some medical vocabularies do not require a specific annotation with the Core Dataset’s vocabularies, as they can be directly stored into specific fields of the HL7 RIM based CDM (e.g. “Hospital”).

Sometimes, a specific clinical term cannot be directly represented through the Core Dataset’s concepts. It is required to put some terms together in a post-coordinated way (see section 2.1.1.1). As shown in the comment fields of Table 3 and Table 4, post-coordination has to be used from time to time to express specific concepts. It is also possible to create a post-coordinated concept in both SNOMED-CT and LOINC (for example by composing terms for “tumor regression” and “score”).

Concept	SNOMED-CT	LOINC	Comment
Rectal Cancer	X		363351006 Malignant tumor of rectum

Curative intent	X	X	These are the intents of the treatments and must be added with the act of the treatment. [108290001] Radiation oncology AND/OR radiotherapy + [373808002] Curative - procedure intent || [363676003] Palliative intent
Hospital	X	X	Can be stored in CDM. No special vocabulary required.
Patient Number		X	Can be stored in CDM. No special vocabulary required.
Study number			Can be stored in CDM. No special vocabulary required.
Age at Diagnosis	X	X	Can be stored in CDM. No special vocabulary required. Effective time of the diagnosis - Birth date
Age at start of radiotherapy (first fraction)			Can be stored in CDM. No special vocabulary required. Effective time of the radiotherapy - birth date
Gender (Male/female)	X	X	Can be stored in CDM. No special vocabulary required.
Ethnicity	X	X	Can be stored in CDM. No special vocabulary required. RaceCode
Height	X	X	50373000 Body height measure
Weight before start of treatment	X	X	27113001 Body weight at a given effective time before treatment
Body Mass Index (BMI)	X	X	60621009 Body mass index
ECOG (WHO) performance scale before start of treatment	X	X	424122007 ECOG performance status finding. At a given effective time before treatment
Respiratory comorbidity	X	X	
Cardiovascular comor bidity	X	X	
Diabetes	X	X	[73211009] Diabetes mellitus
Neoplastic comorbidity	X		
Multidisciplinary (MDT) management	X		
Medication	X	X	432102000 Administration of substance
Prescription	X	X	Not applicable. When using SNOMED-CT to summarise information about a particular type of medication (e.g. use of a local anesthetic during a procedure), a SNOMED-CT expression that includes information about the nature of the substance administered MAY be used. However, this form SHOULD NOT be used for communicating about the prescription, supply, or personal administration of medication.
Pre-existing QoL general challenges			"Record the worst grade of general complaints according the EORTC QLQ-C30 and EQ-DL5, which occurred within 4 weeks before the date of histology". We need to store the information about the grades of the complaints at a given time.

Pre-existing QoL rectal challenges			"Record the worst grade of rectal complaints according the EORTC QLQ-C29, which occurred within 4 weeks before the date of histology". We need to store the information about the grades of the complaints at a given time.
Blood & Serum test	X		166932001 Measurement of level of drug in blood (example, there are also more general concepts)
Staging imaging			The stage of cancer may be set by a CT or a MR. So this information need to be altogether with the diagnosis of X cancer, the stage, the CT/MR imaging that stage this, etc. [450436003] Positron emission tomography with computed tomography, [113091000] Magnetic resonance imaging, [82918005] Positron emission tomography, [363023007] Computerized axial tomography of site
Staging System	X		
Clinical T stage	X		[78873005] T category
Clinical N stage	X		[277206009] N category
Clinical M stage	X		[277208005] M category
Clinical extramesorectal lymph-nodes			
Mesorectal Fascia (MRF)	X		[26077002] Disorder of fascia + [245460006] Mesorectum
Metastasis location	X		[128462008] Secondary malignant neoplastic disease. Target Site Code
Stage Grouping	X		
Tumor Location to the lower limit (not the central location)			
Tumor length before treatment	X		[399375005] Tumor size finding with a given effectiveTime before the effectiveTime of the treatment
Methods of tumor investigation	X		Method codes from the diagnosis of the tumor
Date of Histology			Can be stored in CDM. No special vocabulary required. Effective time
Histology	X	X	[443961001] Malignant adenomatous neoplasm, [72495009] Mucinous adenocarcinoma, [255046005] Neuroendocrine tumor
Grade of Histology	X	X	[12619005] GX grade, [54102005] G1 grade, [1663004] G2 grade, [61026006] G3 grade, [258245003] G4 grade
Tumor Markers	X	X	[250724005] Tumor marker measurement
Tumor Markers – specimen	X		[445405002] Specimen obtained by surgical procedure [258415003] Biopsy sample
Diagnostic CT	X		The [303678006] CT of regions concept from the diagnostic HL7v3 template

Diagnostic PET	X		The [82918005] Positron emission tomography concept from the diagnostic HL7v3 template
Diagnostic MR	X		The [113091000] Magnetic resonance imaging concept from the diagnostic HL7v3 template
Treatment pathways	X		[367336001] Chemotherapy [108290001] Radiation oncology AND/OR radiotherapy. And check the effectiveTimes of the acts, to know if it is concurrent or sequential
Rectal cancer treatment characteristics/ Pre-surgery Chemotherapy treatment characteristics			
Chemotherapy administered	X	X	
Date of start chemotherapy	X		Can be stored in CDM. No special vocabulary required. Effective time start
Date of end chemotherapy	X	X	Can be stored in CDM. No special vocabulary required. Effective time end
Chemotherapy regimen (specify chemotherapy drugs)	X	X	
Chemotherapy cycles prescribed			[399042005] Chemotherapy cycle. StatusCode
Chemotherapy cycles delivered			[399042005] Chemotherapy cycle. StatusCode
Preoperative Radiotherapy on primary tumor (T) treatment characteristics/ Preoperative Radiotherapy on lymph-nodes (N) treatment characteristics/ Radiotherapy treatment characteristics on primary tumor (T+N)/ Radiotherapy on metastasis (M+) treatment characteristics/ Radiotherapy treatment characteristics on M+/ yStaging (in case of preoperative treatment)/ Surgery treatment characteristics			
Surgery performed	X		[387713003] Surgical procedure. effectiveTime
Date of surgery		X	Can be stored in CDM. No special vocabulary required. Effective time
Type of local surgery	X		86743009 Local excision (procedure) 398740003 Colostomy (procedure) 16564004 Hartmann operation, rectal resection (procedure) 174312005 Reversal of Hartmann's procedure (procedure)
Type of regional surgery	X		[58347006] Excision of lymph node 66459002 Unilateral (qualifier value) 51440002 Right and left (qualifier value)
Surgery technique	X		86481000 Laparotomy (procedure) 73632009 Laparoscopy (procedure) 448156002 Excision of rectum by transanal approach

			39046009 Electron microscopy transmission technique, complete (procedure)
Surgery toxicities	X		[428794004] Fistula, 40733004 Infectious disease (disorder), 48422000 Open wound with complication (disorder), etc.
Stoma	X		[225577002] Stoma finding
Type of stoma	X		concept chosen from the "[225577002] Stoma finding" hierarchy
Date of stoma placement			effectivetime of the stoma placement
Date of stoma removal			effectivetime of the stoma removal
Metastasectomy	X		[703467000] Excision of secondary malignant neoplasm
Metastasis location	X		[[128462008] Secondary malignant neoplastic disease. target site lung, liver, brain, bone.
Pathological TNM (y)pStaging			
Staging System	X		
Pathological T stage – (y)pT	X		
Pathological N stage – (y)pN	X		
Pathological M stage – (y)pM	X		
Stage Grouping	X		
Circumferential Resection Margin (CRM)	X		[396271002] Surgical circumferential margin finding
Intestinal margin			
pCR (SNOMED = Comp	X		[399056007] Complete therapeutic response
TRG scoring system	X		
TRG score	X	X	[396445003] tumor regression finding?
Residual tumor (R)	X		[17964000] Residual tumor stage
Death	X		DeceasedInd from the livingSubject, [419099009] Dead
Date of Death	X	X	

Table 3: Coverage by EURECA ontologies of concepts in rectal cancer umbrella protocol

Concept	SNOMED-CT	LOINC	Comment
Primary tumor site	X		TargetSiteCode of the tumor
Second primary tumor, Third, Fourth			TargetSiteCodes of the tumor
local recurrence of previously resected tumor			
Microlaryngoscopy	X		[173035009] Microlaryngoscopy
Panendoscopy	X		
CT neck	X		[169068008] CT of neck
MRI neck (vessels)	X		[241665001] MRI of neck vessels | [72221006] Magnetic resonance imaging of neck
US-FNAC Fine needle aspiration citology			

FDG PET-CT	X		[450436003] Positron emission tomography with computed tomography || [443560005] Positron emission tomography of fluorodeoxyglucose metabolism of brain with computed tomography
PET-CT with hypoxia tracer			[450436003] Positron emission tomography with computed tomography
WHO score	X		[373802001] WHO performance status finding
ACE-27	X		[446363004] Adult comorbidity evaluation-27 score
Weight	X	X	
Length	X	X	
Hb	X		
Several neck surgery specific terms (e.g. Supraomohoidale halsklierdissectie) missing			
Type of chemotherapy		X	
Clinical T stage	X		[399504009] cT category
Additional T (e.g. larynx, a, b)			
Resectability			
primary tumor volume	X		[228791009] Gross tumor volume
SUVmax primary tumor volume			statistical information
Mean SUV value inside primary tumor volume			statistical information
Standard deviation of SUV value inside primary tumor volume			statistical information
Clinical N stage	X		[399534004] cN category
Number of pathological lymph nodes			
Histology of pathological lymph node	X		[423284006] Squamous cell carcinoma of skin of neck. Or undifferentiated = no mapping needed
Grade of pathological lymph node	X		[371494008] pN category
P16	X	X	[391147004] HPV - Human papillomavirus test positive [115326008] Human papillomavirus, type 16
Various toxicity grades (e.g. Xerostomia)	X		[87715008] Aptyalism
local recurrence			
regional recurrence			
curative radiotherapy	X	X	[108290001] Radiation oncology AND/OR radiotherapy. [373808002] Curative - procedure intent
palliative radiotherapy	X		[108290001] Radiation oncology AND/OR radiotherapy. [363676003] Palliative intent
Surgery	X	X	
Chemotherapy	X	X	
distant metastases	X	X	[55440008] M1 category or [128462008] Secondary malignant neoplastic disease

Table 4: Coverage by EURECA ontologies of concepts in head/neck cancer umbrella protocol

In particular Table 3 and 4 show that most of the datasets of MAASTRO's umbrella protocols for rectal cancer and head/neck cancer can be represented through the Core Dataset. Only a few concepts (marked in red) cannot be represented. The blue marked concepts are not applicable (see comment "Prescription") or only needed for statistical reasons.

Some of the concepts that have not been mapped are complex clinical terms which involve not a set of concepts (like in SNOMED-CT post-coordination), but additional temporal information between different "acts" in the CDM. In those cases, the base "acts" can be stored in the CDM and the related specific concepts could be determined by querying the CDM (e.g. regional recurrence implies two "acts" of a diagnosis of carcinoma in the CDM for the same patient, and additionally both diagnoses have to have a "target site" in the same region).

4.2 Evaluation of UOXF's Datasets

The following section is an amendment to deliverable 4.2, in the evaluation of follow-up, adverse event and drug treatment datasets in the breast cancer domain.

The extended Oxford dataset was derived from a clinical study that followed up on patients after a cancer drug treatment. The evaluation task aims to identify whether the Core Dataset is adequate for new datasets from other clinical studies. In this example, the study was conducted with 26 breast cancer patients with follow-ups that record a number of measurements that the investigators wish to observe.

The following table shows the mapping of current clinical variables to SNOMED-CT and HGNC, which are also part of the Core Dataset. Mappings created often use post-coordination of different concepts to be able to fully describe clinical variables. This post-coordination of concepts belonging to the Core Dataset is shown in the column "Core Dataset mappings" including the relations among concepts using RIM attributes from the CDM.

Clinical variable of UOXF	Core Dataset mappings
Comorbidity	
Trial Number	Patient ID
Sign_symptom (symptom short description)	Concepts from 'finding' SNOMED branch. e.g: ActObservationValues: - [38341003] Hypertensive disorder - [449619004] Swelling of upper arm,targetSite = [24028007] Right - [125667009] Contusion
Sign_symptom reported (symptom short description)	N/A To be stored in Act.text (mapping not needed)
start date	N/A To be stored in Act.effectiveTimeStart (mapping not needed)
unknown start date (yes or no)	

Continuous (continuous or intermittent)	Continuous (qualifier value) : 255238004, Intermittent (qualifier value) : 7087005
Ongoing at Start of Treatment (yes or no)	N/A To be stored in Act.statusCode=active/completed (mapping not needed)
grade	Histological grading systems (staging scale) : 277457005
Clinical info	
Age	N/A To be stored in LivingSubject.birthTime
Menopausal Status	Menopause finding (finding) : 276477006 Values: - [76498008] Postmenopausal state - [22636003] Premenopausal state
Primary Tumour Site	Primary tumor site (observable entity) : 399687005 Children of the following concept: [52530000] Body region structure
Primary Diagnosis	Cancer diagnosis based on primary site histological evidence (finding) : 373800009 Children of the following concept: [254837009] Malignant tumor of breast
Primary Diagnosis Date	N/A To be stored in Act.effectiveTimeStart (mapping not needed)
ER Status	Post coordination: Molecular genetic test: 405825005 Entity (HGNC concepts): [ESR1] estrogen receptor 1 ActMethodCode: - [117617002] Immunohistochemistry procedure - [426329006] Fluorescence in situ hybridization ActObservationInterpretationCode: - [POS] Positive - [NEG] Negative
ER Score	Post coordination: Numerical ActObservationValue
PR Status	Molecular genetic test: 405825005 Entity (HGNC concepts): [PGR] progesterone receptor ActMethodCode: - [117617002] Immunohistochemistry procedure - [426329006] Fluorescence in situ hybridization ActObservationInterpretationCode: - [POS] Positive - [NEG] Negative

PR Score	Numerical ActObservationValue from previous column
Her2 Status	Post coordination: Molecular genetic test: 405825005 Entity (HGNC concepts): [ERBB2] v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2 ActObservationInterpretationCode: - [POS] Positive - [NEG] Negative
Her2 Score (ICH)	Post coordination: Molecular genetic test: 405825005 Entity (HGNC concepts): [ERBB2] v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2 ActMethodCode: - [117617002] Immunohistochemistry procedure Numerical ActObservationValue
Her2 Score (Fish)	Post coordination: Molecular genetic test: 405825005 Entity (HGNC concepts): [ERBB2] v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2 ActMethodCode: - [426329006] Fluorescence in situ hybridization Numerical ActObservationValue
Stage at Entry into Trial	Tumor-node-metastasis (TNM) breast tumor staging (tumor staging) : 254326001 ActObservationValues: Children of the following concept: [385356007] Tumor stage finding
Date Breast assessed	Examination of breast (procedure) : 46662001
Time Point Breast Assessed	N/A To be stored in Act.effectiveTimeStart (mapping not needed) of previous variable
Site of mass	Tumor site (observable entity) : 371480007 ActTargetSiteCode: Children of the following concept: [52530000] Body region structure
Size of mass	Tumor size (observable entity) : 263605001 With numerical ActObservationValue
Number of Axillary Nodes	Number of regional lymph nodes involved by malignant neoplasm : 444384007 ActTargetSiteCode: [68171009] Axillary lymph node structure
Status of Axillary Nodes	Numerical ActObservationValue. If 'Status of Axillary Nodes' is 'not metastasized', value would be 0.
Number of Supraclavicular Nodes	Number of regional lymph nodes involved by malignant neoplasm : 444384007 ActTargetSiteCode: [245265003] Supraclavicular lymph node group

Status of Supraclavicular Nodes	Numerical ActObservationValue. If 'Status of Supraclavicular Nodes' is 'not metastasized', value would be 0.
Drug	
BSA	Body surface area (observable entity) : 301898006
Drug name (in this case only the drug used in the trial)	Values of this variable should be mapped using descendants of 'substance' or 'product' branches.
drug date	Act.effectiveTimeStart
drug dose	SubstanceAdministration.doseCheckQuantity (expected quantity)
drug amount mg	SubstanceAdministration.doseQuantity (real quantity) SubstanceAdministration.doseQuantityUnits = 'mg'
Toxicity	
adverse event	Adverse drug event resulting from treatment of disorder (disorder) : 406644009 Values of this variable should be mapped using descendants of 'finding' branches to represent the specific reaction
start date	Act.effectiveTimeStart
stop date	Act.effectiveTimeEnd
Continuous	Continuous (qualifier value) : 255238004, Intermittent (qualifier value) : 7087005
relation to study drug	
Adverse event grade	Value for the Adverse Event observation. Common Toxicity Criteria for Adverse Events: 1 - Mild (255604002) 2 - Moderate (6736007) 3 - Severe (24484000) 4 - Life-threatening (442452003) 5 - Death (399166001) Children of [272141005] Severities.
serious adverse event?	Depends on the 'Adverse event grade'.

Table 5: Mapping of the clinical variables to Core Dataset concepts

As shown in Table 5, the dataset is similar to the previous breast cancer dataset in that they share the same patient assessments such as ER test, HER2 test, tumor evaluation and TNM staging. However, there are other clinical outcomes that are followed up in relation to the drug treatment.

In particular, there are some symptoms after the treatment being followed in later patient assessment. The context of the symptom is associated with breast cancer but it has not been very easy to cover this field with the core dataset. For example for the field such as ‘Primary diagnosis’ and ‘Sign symptom’, due to a wide, variable outcome of the

patient, it may be difficult to evaluate the coverage of SNOMED-CT concepts in this context. However, we completed an exercise that attempts to map all available values of the symptoms and we find SNOMED-CT very satisfactory in covering these sign symptoms described by patients. A common pattern is that most of the values can be composed using a disorder or symptom concept from SNOMED-CT with a concept that describes the location using the attribute 'TargetSite'. The following table lists the concepts that are mapped.

Sign symptom	Mapping	Alternative Mapping
Discomfort in Right Breast	[279084009] Chest discomfort TargetSite: [24028007] Right	[247347003] Discomfort TargetSite: [73056007] Right breast structure
dull pain in right breast	[3368006] Dull chest pain TargetSite: [24028007] Right	[83644001] Dull pain TargetSite: [73056007] Right breast structure
Uncomfortable right breast site	[279084009] Chest discomfort	[247347003] Discomfort TargetSite: [73056007] Right breast structure
dull pain breast	[3368006] Dull chest pain	[83644001] Dull pain [76752008] Breast structure
Sleeplessness in night	[193462001] Insomnia	
Dullache pain at disease site	[83644001] Dull pain TargetSite: [363698007] Finding site	[83644001] Dull pain TargetSite: [263591006] Site of neoplasm UNAPPROVED ATTRIBUTE
Dullache in left breast	[3368006] Dull chest pain TargetSite: [7771000] Left	[83644001] Dull pain TargetSite: [80248007] Left breast structure
Discomfort on Breast	[279084009] Chest discomfort	
Bruise at site of cancer due to core biopsy	[9911007] Core needle biopsy EntryRelationship [classCode=TRIG]: [125667009] Contusion TargetSite: [263591006] Site of neoplasm UNAPPROVED ATTRIBUTE	
Dullache at lump left breast	[83644001] Dull pain TargetSite: [80248007] Left breast structure	[83644001] Dull pain EntryRelationship[classCode=SUBJ]: [89164003] Breast lump TargetSite: [7771000] Left
Dull Pain in Right hip joint	[83644001] Dull pain TargetSite: [362908009] Right hip joint structure	
Bruise on right arm	[125667009] Contusion TargetSite: [368209003] Right upper arm structure	
Left breast lump with overlying skin erythema	[89164003] Breast lump TargetSite: [7771000] Left EntryRelationship[classCode=CONCURRENT]: [444827008] Erythema of skin	
Lower back pain	[279039007] Low back pain	

Swollen Right Breast	[300885006] Swelling of breast TargetSite: [24028007] Right	
Swollen Right Arm	[449619004] Swelling of upper arm TargetSite = [24028007] Right	
Pain Right Arm	[287046004] Pain in right arm	
Hypertension	[38341003] Hypertensive disorder	
Pain Right upper Arm	[22253000] Pain TargetSite: [368209003] Right upper arm structure	
Pain at breast lump	[22253000] Pain TargetSite: [76752008] Breast structure	[22253000] Pain EntryRelation-ship[classCode=SUBJ]: [89164003] Breast lump
Dull pain & lump Right Breast	[83644001] Dull pain + [89164003] Breast lump TargetSite: [24028007] Right	
Lump in Left breast	[89164003] Breast lump TargetSite: [7771000] Left	
Redness Left Breast	[247441003] Erythema TargetSite: [80248007] Left breast structure	
Pain in right breast	[53430007] Pain of breast TargetSite: [24028007] Right	[22253000] Pain TargetSite: [73056007] Right breast structure
Pain in breast lump	[22253000] Pain TargetSite: [76752008] Breast structure	[22253000] Pain EntryRelation-ship[classCode=SUBJ]: [89164003] Breast lump
Adverse events		
Headache	[25064002] Headache	
Tiredness	[84229001] Fatigue	
Vomiting	[422400008] Vomiting	
Heartburne	[16331000] Heartburn	
Nausea	[422587007] Nausea	
Bleeding Picc line Site	[392020005] Peripherally inserted central catheter care EntryRelationship[classCode=TRIG]: [131148009] Bleeding TargetSite: [260738001] Site of insertion UN-APPROVED ATTRIBUTE	
Allergic reaction	[419076005] Allergic reaction	
Dental pain? infec-tion	[27355003] Toothache EntryRelationship[classCode=RSON]: [427898007] Infection of tooth	
Bruise after fall	[125667009] Contusion EntryRelationship: [1912002] Fall	
Impaired wound healing	[271618001] Impaired wound healing	
Tonsilltis	[90176007] Tonsillitis	

Sore throat	[162397003] Pain in throat	
Itchiness on legs	[418290006] Itching TargetSite: [30021000] Lower leg structure	
Hypertension	[38341003] Hypertensive disorder	
Soreness in mouth	[162011005] Sore mouth	
Flu-like symptoms	[95891005] Influenza-like illness	
Bleeding from Gums	[86276007] Bleeding gums	
Pain in breast tumour area	[22253000] Pain TargetSite: [76752008] Breast structure	[22253000] Pain EntryRelationship[classCode=SUBJ]: [126926005] Neoplasm of breast
Mild Sorethroat	[162397003] Pain in throat ObservationValue (severity): [255604002] Mild	
Pain in Tumour Area	[22253000] Pain TargetSite: [263591006] Site of neoplasm UNAPPROVED ATTRIBUTE	[22253000] Pain EntryRelationship[classCode=SUBJ]: [55342001] Neoplastic disease
Cramps in right leg	[55300003] Cramp TargetSite: [32696007] Structure of right lower leg	

Table 6: Mapping of the symptom values to Core Dataset concepts

The standard of clinical information such as age and site of tumor are well-covered by SNOMED-CT. One particular comment about the coverage of the Core Dataset is that the main clinical terminologies such as SNOMED-CT have a good coverage in general to describe the clinical variable being observed, but it is difficult to achieve high specificity because the study that has been conducted is focused on investigating specialised cases of breast cancer. It requires very specific terms to describe the clinical data. For example, all the patients with any 'symptom' that are associated to the treatment should be reported. However, there may not be a concept so specific to describe these conditions.

Chapter 4 shows that the Core Dataset covers nearly all concepts of new clinical datasets. As such, it can be concluded that the required new clinical datasets can be sufficiently represented by the EURECA platform.

5 Conclusion

The Core Dataset is an essential part of the EURECA Semantic Interoperability Platform with its ability to semantically describe the required medical datasets. Specific analyses were executed to determine the Core Dataset in order to describe datasets of the EURECA specific breast, lung, sarcoma and nephroblastoma cancer domains sufficiently. Deliverable 4.2 concluded that a sub-set of standard vocabularies of SNOMED-CT, LOINC and HGNC represents appropriately these clinical datasets. [1]
Based on these results, the Core Dataset was implemented. The content of the Core Dataset undergoes continuous changes due to new standard vocabularies or new cancer domains, which need to be represented.

Chapter 2 summarised that the Core Dataset provides specific technologies in order to execute those extensions without much effort.

The clinical partners decided that the medical impact of the EURECA platform would improve by including new datasets, which required evaluation of the Core Dataset coverage of datasets in the rectal and head/neck cancer domains and an extended breast cancer dataset. Chapter 3 showed, based on the evaluation of public available datasets, that only a few specific SNOMED-CT concepts and terminology binding rules from the Core Dataset to the CDM need to be updated to accommodate datasets of the new cancer domains sufficiently. Chapter 4 approved this assumption by validating “real” clinical datasets of MAASTRO and UOXF.

Ultimately, this deliverable concludes that the Core Dataset covers nearly all concepts of required new clinical datasets and that these datasets can be represented by the technologies of the EURECA Semantic Interoperability Platform.

6 Acronyms

CDM	EURECA Common Data Model
CIM	EURECA Common Information Model
HL7	Health Level 7 (message standard)
HL7 RIM	HL7 Reference Information Model
HGNC	HUGO (Human Genome Organisation) Gene Nomenclature Committee
IHTSDO	International Health Terminology Standards Development Organisation
LOINC	Logical Observation Identifiers Names and Codes
OWL	Ontology Web Language
RDF	Resource Description Framework
SIRS	International Request Submission
SNOMED	Systematized Nomenclature of Human and Veterinary Medicine
SPARQL	Simple Protocol and RDF Query Language

7 References

[1] Deliverable 4.2 “Initial proposal for the core datasets”, 13th March 2014
[2] Deliverable 4.4 “Initial prototype of the semantic interoperability platform”, 27th October 2014
[3] Deliverable 4.3/9.3 “Initial proposal for the mapping formalism and mappings to EHR and CT models”, 12th April 2014