|  |  |
| --- | --- |
|  | P:\EuroPartners-General\EU-Program\FP7-ICT\Running projects\EURECA\Logo\EURECA-Logo-Final.jpg |

ICT-2011-288048

EURECA

Enabling information re-Use by linking clinical Research and CAre

IP

Contract Nr: 288048

Deliverable: 4.3 & 9.3 Initial proposal for the mapping formalism and mappings to EHR and CT models

Due date of deliverable: (01-31-2014)

Actual submission date: (MM-DD-YYYY)

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

Responsible WP: UPM - PHILIPS

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) |  |

# DOCUMENT INFO

## Author

|  |  |  |
| --- | --- | --- |
| Author | Company | E-mail |
| Sergio Paraíso Medina | UPM | sparaiso@infomed.dia.fi.upm.es |
| David Perez del Rey | UPM | dperez@infomed.dia.fi.upm.es |
| Raúl Alonso Calvo | UPM | ralonso@infomed.dia.fi.upm.es |
| Santiago Aso Lete | UPM | saso@infomed.dia.fi.upm.es |
| Laura Hollink | VUA | l.hollink@vu.nl |
| Pajolma Rupi | XEROX | pajolma.rupi@xrce.xerox.com |
| Ahmed Ibrahim | Philips | ahmed.a.ibrahim@philips.com |
| Anouar Laarif | GBG | Anouar.Laarif@germanbreastgroup.de |
| Ηaridimos Kondylakis | FORTH | kondylak@ics.forth.gr |
| Sheng Yu | UOxF | sheng.yu@oncology.ox.ac.uk |
| Scott Marshall | Maastro | m.scott.marshall@maastro.nl |
| Cyril Krykwinski | IJB | [cyril.krykwinski@bordet.be](mailto:cyril.krykwinski@bordet.be) |
| Holger Stenzhorn | UdS | holger.stenzhorn@uniklinikum-saarland.de |
| Kerstin Rohm | Fraunhofer | kerstin.rohm@ibmt.fraunhofer.de |

## Documents history

|  |  |  |
| --- | --- | --- |
| Document version # | Date | Change |
| V0.1 | 26.11.2013 | Starting version, template |
| V0.2 | 26.11.2013 | Definition of ToC |
| V0.3 | 12.04.2014 | First complete draft |
| V0.4 |  | Integrated version (send to WP members) |
| V0.5 |  | Updated version (send PCP) |
| V0.6 |  | Updated version (send to project internal reviewers) |
| Sign off |  | Signed off version (for approval to PMT members) |
| V1.0 |  | Approved Version to be submitted to EU |
|  |  |  |

## Document data

|  |  |
| --- | --- |
| Keywords |  |
| Editor Address data | Name: Sergio Paraíso Medina  Partner: UPM  Address: Facultad de Informática  Universidad Politécnica de Madrid  Campus de Montegancedo, s/n  28660 Boadilla del Monte, Madrid, Spain  Phone: +34 91 336 74 45  E-mail: sparaiso@infomed.dia.fi.upm.es |
| Delivery date |  |

## Distribution list

|  |  |  |
| --- | --- | --- |
| Date | Issue | E-mailer |
|  |  |  |

Table of Contents

[0 DOCUMENT INFO 2](#_Toc385364206)

[0.1 Author 2](#_Toc385364207)

[0.2 Documents history 2](#_Toc385364208)

[0.3 Document data 2](#_Toc385364209)

[0.4 Distribution list 3](#_Toc385364210)

[1 Introduction (UPM) 7](#_Toc385364211)

[2 Mapping formalisms (UPM) 8](#_Toc385364212)

[2.1 Core dataset annotation (PHILIPS) 8](#_Toc385364213)

[2.1.1 Label to Core dataset (PHILIPS) 8](#_Toc385364214)

[2.1.2 Free text to CDM (Xerox) 12](#_Toc385364215)

[2.1.2.1 Introduction 12](#_Toc385364216)

[2.1.2.2 First mapping proposal 12](#_Toc385364217)

[2.1.2.3 Future work 17](#_Toc385364218)

[2.1.3 Terminology linking service (VUA) 18](#_Toc385364219)

[2.1.3.1 The TLS as described in Deliverable 4.2 18](#_Toc385364220)

[2.1.3.2 Functional description of classes and methods 18](#_Toc385364221)

[2.1.3.3 User documentation 20](#_Toc385364222)

[2.2 Core dataset to EURECA CDM (UPM) 22](#_Toc385364223)

[2.2.1 EURECA guidelines 22](#_Toc385364224)

[2.2.2 Data Push Service 23](#_Toc385364225)

[2.2.2.1 ETL tools 23](#_Toc385364226)

[2.2.2.2 Normalization pipeline 24](#_Toc385364227)

[2.2.3 Query normalization 26](#_Toc385364228)

[3 Data source mappings to EURECA (clinical partners) 29](#_Toc385364229)

[3.1 Institut Jules Bordet Data 29](#_Toc385364230)

[3.1.1 Data source types of information 29](#_Toc385364231)

[3.1.2 Core dataset annotations 29](#_Toc385364232)

[3.1.3 HL7 v3 message template selection 29](#_Toc385364233)

[3.1.4 HL7 v3 message generation 32](#_Toc385364234)

[3.2 MAASTRO Clinic Data 32](#_Toc385364235)

[3.2.1 Data source types of information 32](#_Toc385364236)

[3.2.2 Core dataset annotations 32](#_Toc385364237)

[3.2.3 HL7 v3 message template selection 34](#_Toc385364238)

[3.2.4 HL7 v3 message generation 34](#_Toc385364239)

[3.3 University of Oxford data 34](#_Toc385364240)

[3.3.1 Data source types of information 34](#_Toc385364241)

[3.3.2 Core dataset annotations 35](#_Toc385364242)

[3.3.3 HL7 v3 message template selection 36](#_Toc385364243)

[3.3.4 HL7 v3 message generation 36](#_Toc385364244)

[3.4 Saarland University Hospital data 36](#_Toc385364245)

[3.4.1 Data source types of information 36](#_Toc385364246)

[3.4.2 Core dataset annotations 38](#_Toc385364247)

[3.4.3 HL7 v3 message template selection 39](#_Toc385364248)

[3.4.4 HL7 v3 message generation 40](#_Toc385364249)

[3.5 GBG data 42](#_Toc385364250)

[3.5.1 Data source types of information 42](#_Toc385364251)

[3.5.2 Core dataset annotations 42](#_Toc385364252)

[3.5.2.1 Request the possible SNOMED code for a term 42](#_Toc385364253)

[3.5.2.2 SNOMED CT Code selection 43](#_Toc385364254)

[3.5.3 HL7 v3 message template selection 43](#_Toc385364255)

[3.5.4 HL7 v3 message generation 43](#_Toc385364256)

[3.6 Connectors to other models / tools 44](#_Toc385364257)

[3.6.1 TranSMART (PHILIPS) 44](#_Toc385364258)

[3.6.1.1 Representing Dataset Explorer tree 44](#_Toc385364259)

[3.6.1.2 ETL process 44](#_Toc385364260)

[3.6.1.3 Dataset Explorer screenshots 46](#_Toc385364261)

[3.6.1.4 Tools 47](#_Toc385364262)

[3.6.2 Observational Medical Outcomes Partnership - OMOP (UPM) 47](#_Toc385364263)

[3.6.3 INDIVO X (FORTH) 50](#_Toc385364264)

[3.6.4 ObTiMA (Fraunhofer IBMT) 52](#_Toc385364265)

[3.6.4.1 Workflow 52](#_Toc385364266)

[3.6.4.2 Future work 53](#_Toc385364267)

[4 CONCLUSIONS 54](#_Toc385364268)

[5 REFERENCES 55](#_Toc385364269)

[6 ANNEX A: OMOP data mapping 56](#_Toc385364270)

**TABLE OF FIGURES**

[Figure 1: BioPortal XML result of an annotation 9](#_Toc383704590)

[Figure 2: Example of output file with input text and annotated concepts 10](#_Toc383704591)

[Figure 3: SPARQL query for retrieving the semantic type(s) of a concept 11](#_Toc383704592)

[Figure 4: Example of the triple-based representation 13](file:///C:\Users\Sergio\Dropbox\proyecto%20eureca\deliverables\D4.3-9.3\D4.3-9.3-Initial%20proposal%20for%20the%20mapping%20formalism_0.3.docx#_Toc383704593)

[Figure 5: Main classes in the triple-based representation of eligibility criteria 13](#_Toc383704594)

[Figure 6: Triple-based representation of a criteria item example defining the gender 14](file:///C:\Users\Sergio\Dropbox\proyecto%20eureca\deliverables\D4.3-9.3\D4.3-9.3-Initial%20proposal%20for%20the%20mapping%20formalism_0.3.docx#_Toc383704595)

[Figure 7: Triple-based representation of a criteria item example defining the age 14](file:///C:\Users\Sergio\Dropbox\proyecto%20eureca\deliverables\D4.3-9.3\D4.3-9.3-Initial%20proposal%20for%20the%20mapping%20formalism_0.3.docx#_Toc383704596)

[Figure 8: Example of mapping NLP output to HL7 Patient class 15](file:///C:\Users\Sergio\Dropbox\proyecto%20eureca\deliverables\D4.3-9.3\D4.3-9.3-Initial%20proposal%20for%20the%20mapping%20formalism_0.3.docx#_Toc383704597)

[Figure 9: Triple-based representation of a criteria item example defining a Laboratory\_or\_Test\_Result criteria type 15](file:///C:\Users\Sergio\Dropbox\proyecto%20eureca\deliverables\D4.3-9.3\D4.3-9.3-Initial%20proposal%20for%20the%20mapping%20formalism_0.3.docx#_Toc383704598)

[Figure 10: Example of mapping NLP output to HL7 Observation class 16](#_Toc383704599)

[Figure 11: Triple-based representation of a criteria item example defining a Treatment criteria type 16](file:///C:\Users\Sergio\Dropbox\proyecto%20eureca\deliverables\D4.3-9.3\D4.3-9.3-Initial%20proposal%20for%20the%20mapping%20formalism_0.3.docx#_Toc383704600)

[Figure 12: Example of mapping NLP output to HL7 Procedure class 16](#_Toc383704601)

[Figure 13: Triple-based representation of a criteria item example defining a Treatment criteria type 17](file:///C:\Users\Sergio\Dropbox\proyecto%20eureca\deliverables\D4.3-9.3\D4.3-9.3-Initial%20proposal%20for%20the%20mapping%20formalism_0.3.docx#_Toc383704602)

[Figure 14: Example of mapping NLP output to HL7 Substance Administration class 17](#_Toc383704603)

[Figure 15: UML class diagram of the TLS. 19](#_Toc383704604)

[Figure 16: Part of the BioPortal data model 20](#_Toc383704605)

[Figure 17: Deployment example 22](#_Toc383704606)

[Figure 18: ETL process using Kettle 24](#_Toc383704607)

[Figure 19: General ETL process 25](#_Toc383704608)

[Figure 20: Normalization of a concept 26](#_Toc383704609)

[Figure 21: Query Normalization process 27](#_Toc383704610)

[Figure 22: Observation query template 28](#_Toc383704611)

[Figure 23: Complete CRF “F1 Registration” of the SIOP/GPOH trial in a paper based form 37](#_Toc383704612)

[Figure 24: CRF “F1 Registration” of the SIOP/GPOH trial provided ObTiMA 38](#_Toc383704613)

[Figure 25: ETL process CDM into tranSMART 44](#_Toc383704614)

[Figure 26: Concept tree of the Dataset Explorer 46](#_Toc383704615)

[Figure 27: Comparison of two subsets in Data Explorer 47](#_Toc383704616)

[Figure 28: The entity-relationship diagram (ERD) of the CDM data tables and relationships between them 48](#_Toc383704617)

[Figure 29: The entity-relationship diagram (ERD) of the Vocabulary tables and relationships between them 49](#_Toc383704618)

[Figure 27. Example mappings of IndivoX to the EURECA CDM 51](#_Toc383704619)

[Figure 28 Example query for retrieving an Ultrasonography from EURECA DWH 52](#_Toc383704620)

# Introduction (UPM)

This report focuses on describing how clinical data sources from different partners in EURECA consortium have been mapped and loaded into EURECA semantic interoperability layer and Common data model (CDM). We explain the work performed by the different clinical partners in association with technical partners, to achieve semantic integration of original datasets into the EURECA platform.

Methods described here aim to solve two main types of heterogeneities, (i) terms from original data sources should be mapped to Core dataset, and (ii) the structure of data sources should also be linked to the EURECA CDM. In addition, un-structured data has to be structured through Natural Language Processing (NLP) methods to be loaded into the EURECA CDM.

Starting from the EURECA clinical datasets made available by the clinical sites, sets of concepts, relevant for the project were studied in deliverable 4.2. Core dataset, defined deliverable 4.2, was based on existing domain ontologies and selected to maximize the coverage of the semantics of data.

Some of shared data sources already use standard terminologies for representing concepts, terminologies such as NCIt or ICD-10. To translate (or map) such coded data, the Terminology linking service was applied. The Terminology linking service was introduced in deliverable 4.2 and exploits mappings defined in BioPortal [1].

For other sources that are not actually coded, data have been previously annotated and reviewed by clinical and technical partners. Tools such as the Xerox Concept Annotator, MetaMap, or BioPortal LOOM can be applied to provide automatic annotations. Mappings presented in this document have been also validated manually by clinical experts.

Finally, mappings to connect the EURECA platform to other data models from related projects such as tranSMART (i2b2), IndivoX PHR, OMOP and Obtima are also described in this document.

# Mapping formalisms (UPM)

This section presents the foundation and processed defined to map data from clinical partners to the EURECA platform. Section 2.1 presents annotations of terms, both in structured and un-structured data, to EURECA Core d

ataset. Section 2.2 describes mappings between Core dataset concepts and the EURECA CDM, depicting how clinical sources have to be transformed to be loaded in the common infrastructure.

## Core dataset annotation (PHILIPS)

In Deliverable 4.2 [1] we focused on the definition of the core dataset that sufficiently covers the semantics of our clinical domains of interest, which is breast cancer. We presented an analysis of the semantics of the clinical eligibility criteria based on the medical ontologies SNOMED-CT, medDRA and LOINC. These ontologies were chosen due to their wide adaptation in the clinical research domain. We compared eligibility criteria in the domain of breast cancer with others trials in the domains of lung cancer, sarcoma and nephroblastoma. We concluded that (1) a relatively small group of concepts occurred in a large number of clinical trials, and (2) the effort of adding new trials is low since the additional sets of concepts that need to be mapped to relevant data are small. We also evaluated the concepts that appeared in the clinical patient datasets provided by our clinical partners, and evaluated the coverage of SNOMED-CT, medDRA and LOINC for each dataset. Our results showed that SNOMED-CT has the highest coverage rate, but it lacks coverage of specific concepts related to radiotherapy and genes. Currently, there is no alternative for the radiotherapy domain, but results in Deliverable 4.2 showed that there were gene related concepts that were covered by the HUGO Gene Nomenclature Committee (HGNC). The results in Deliverable 4.2 also indicated that the majority of concepts extracted from the clinical datasets can be annotated using the Core dataset, in order to form a semantic unified data vocabulary representation.

## Label to Core dataset (PHILIPS)

We used BioPortal to annotate the medical concepts stored in the patient datasets provided by our clinical partners. Bioportal provides an API which can be implemented in different scripting and programming languages. We created a python script for annotating the datasets with the following input parameters:

1. The concepts that were extracted from the datasets;
2. A list of the selected ontologies (identifiers), e.g. if one wants to annotate “Neoplasm of breast” with SNOMED-CT, medDRA and LOINC, the input would be [“Neoplasm of breast”, {‘46896, ‘42280’ ‘47637’}].

The output is the annotated concepts from the selected ontologies. The API supports JSON and XML as output. Figure 1 shows an example of a BioPortal result.

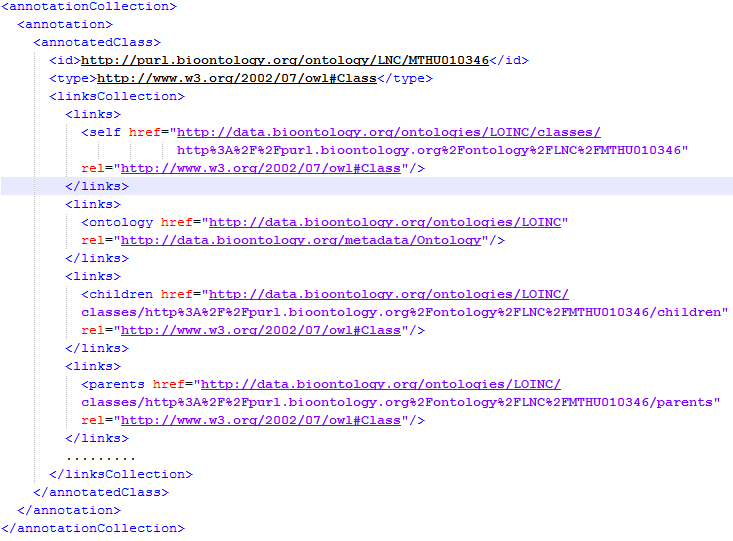


Figure 1: BioPortal XML result of an annotation

The next step is transforming the XML data and removing irrelevant attributes. This will allow our clinical partners to validate the results much faster since they don’t have to inspect irrelevant information. Figure 2 shows a simplified example of an entry (dates and identifiers are removed) of the transformed XML data. The figure shows the original input text and the associated annotations. The ‘localConceptId’ is comprised of the ontology code and the concept code.



Figure 2: Example of output file with input text and annotated concepts

To determine whether a match is correct or not is not a trivial task. Figure 2 shows that there are multiple SNOMED-CT matches for the input ‘oxazepam’. There are different ways to verify which match is valid. One approach is to inspect the different branches in which the concepts reside, and propagate up the tree to the root level concept. In our example we choose the concept for which the root level node would be 'Substance' in the SNOMED-CT hierarchy, since ‘oxazepam’ is a medication. Another approach is to use the UMLS semantic types of the annotated concepts, as they can provide additional information about the semantics. The semantic types are not shown in Figure 1 because the new Bioportal API[[1]](#footnote-1) does not include the semantic types in its output anymore (semantic types were included in the deprecated API). However, the new API provides a SPARQL interface to query ontologies using the SPARQL standard[[2]](#footnote-2). We created a function that retrieved the semantic type(s) of an annotated concept (Figure 3). The function requires a Bioportal API key (API\_KEY), an ontology code ($ontologyCode) and a concept code ($conceptCode).

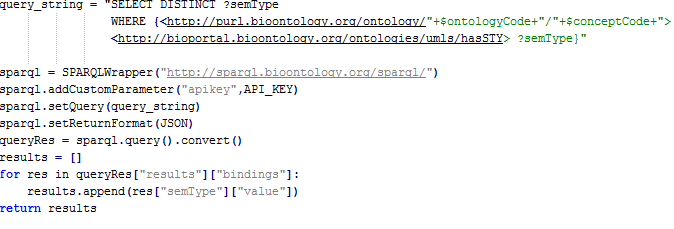


Figure 3: SPARQL query for retrieving the semantic type(s) of a concept

The last step is creating the HL7 messages with the transformed XML results (Figure 2). We use Mirth Connect to create the HL7 messages. This is a semi-automatic process since we cannot immediately map the XML data to the HL7 templates. Creating a HL7 message for a Substance Administration is straightforward. For every entry in the XML file we check whether a SNOMED-CT concept is a child of the root concept ‘Substance’ (LOINC and medDRA concepts are discarded). If that is the case, we extract the other attributes (e.g. identifiers and dates), and create the template. The difficult part is choosing the right template if the XML data contains disease, therapeutic, or diagnostic related concepts. For these kinds of data, we involve our clinical partners and discuss which template is suited for which concept. Again, the semantic types can offer extra insight for choosing the right template.

TODO Ahmed

## Free text to CDM (Xerox)

Exchanging clinical data is a process whose sensitivity is highly attributed to the data format in the respective exchanging entities. The free text availability of a significant part of this data gives rise to the need of structuring it which eases the process of exchanging it. This task is addressed in WP3: *Information extraction from free text* where we use Natural Language Processing (NLP) expertise to process the free text and annotate the mentions of medical concepts (deliverable D3.1 [1]) and relations between them (deliverable D3.2 [2]). The NLP output consists of a triple-based representation. The triple-based format is not among the supported formats by the EURECA Data Push Service and as a result a mapping process between the current triple-based format and one of the supported formats is needed to be performed. The mapping between these two formats gives the possibility to store the data into EURECA CDM and on a later phase, it can be retrieved by using the available data access services.

The rest of this section describes an overview of the manual mapping work performed so far, using the recommended format: HL7v3 messaging (the supported formats are detailed in D3.3, chapter 2 [3]).

The first subsection introduces the context of the mapping work between the NLP output and HL7v3 messages. The second subsection describes the exploration of the possibilities of mapping our NLP triple-based representation to attributes of RIM classes of HL7v3 messages. The third subsection is focused on discussing the main future directions of the mapping.

#### Introduction

The work performed for developing the NLP tools is based on publicly available data, among them being the clinical trials data, and more precisely the clinical trial eligibility criteria (CTEC) section. This data was downloaded from clinicaltrials.gov website, being restricted to breast cancer domain.

CTEC comprises clinical features which serve as factors for qualifying or disqualifying a patient from participating to a clinical trial. Even though the main intention of HL7v3 messages consists of using them for exchanging clinical patient data, having no access to such data for the time being led us to use the CTEC data instead. The advantage of using it is that in both cases we deal with clinical features which evaluate the health status of the patients.

#### First mapping proposal

The triple-based representation of CTEC produced by the NLP tool is described in details in deliverable D3.2, chapter 3 [2]. The following example (Figure 4) defines the representation of the criteria item: “*No evidence of ventricular ectopics greater than 4/min on EKG*”.

ec1 **hasText** "*No evidence of ventricular ectopics greater than 4/min on EKG*"

ec1 **excludes** f1

f1 **isA** Laboratory\_or\_Test\_Result

f1 **hasObject** x1

x1 **hasTerm** "ventricular ectopics"

x1 **hasCUI** "C0151636"

f1 **hasValue** x2

x2 **hasMin** 4

x2 **hasUnit** "/min"

f1 **hasProcedure** x3

x3 **isA** Diagnostic\_Procedure

x3 **hasTerm** "EKG"

x3 **hasCUI** "C1623258"

Figure 4: Example of the triple-based representation

The eligibility criterion (ec1) is an instance of the main class[[3]](#footnote-3): Eligibility\_Criterion shown in Figure 5.

|  |  |  |
| --- | --- | --- |
| Class | Property | Property value type |
| EC (Eligibility\_Criterion) | requires | Fact |
| excludes | Fact |
| hasText | String |
| Fact | isA | Criterion Fact Class |
| UMLS\_Concept\_Value | hasTerm | String |
|  | hasCUI | UMLS Concept Unique Identifier |
| Quantitative\_Value | equals | Number |
|  | hasMin | Number |
|  | hasMax | Number |
|  | hasUnit | String |

Figure 5: Main classes in the triple-based representation of eligibility criteria

We have based our work on the Unified Medical Language System (UMLS) and we use it as the coding system in HL7 messages. In this context, the concept unique identifiers (CUIs) are used as the codes for the terms which are identified by the NLP tools. The criteria types defined in the conceptual schema of CTEC are taken into consideration and they provide useful information for obtaining a contextualization of the criteria items before mapping them to the HL7 message structure.

The following examples illustrate how the triple-based representations can be projected into HL7 messages. There is one example taken for each of the four RIM classes: patient, observation, procedure and substance administration.

**Patient**

The CTEC representation criteria types which can be mapped to patient entity attributes are: Age (mapping to *birthtime* patient entity attribute) and Gender (mapping to *administrativeGenderCode* patient entity attribute).

An example taken for each of these attributes illustrates the mapping process in details.

The first example specifies the gender of the patient to be enrolled in the clinical trial: *“Female”*. Its corresponding triple-based representation is shown in

ec1 **hasText** “Female”

ec1 **includes** f1

f1 **isA** Gender

f1 **hasValue** c1

c1 **hasTerm** “Female”

c1 **hasCUI** “C0015780”

Figure 6: Triple-based representation of a criteria item example defining the gender

The two main triples that will be used for the mapping are: “f1 **isA** Gender” which defines under which attribute the information will be stored and “c1 **hasCUI** “C0015780”” which defines the code of *administrativeGenderCode* attribute.

The second example specifies the age of the patient to be enrolled in the clinical trial:

*“Age ≥ 18 years”*

ec2 **hasText** “Age ≥ 18 years”

ec2 **includes** f2

f2 **isA** Age

f2 **hasValue** v2

v2 **hasMin** “18"

Figure 7: Triple-based representation of a criteria item example defining the age

The two triples that will be used in this case are: “f2 **isA** Age” which defines under which attribute the information will be stored and “v2 **hasMin** “18” which will be translated into the birthdate and will give the value to the *birthtime* attribute.

***Note****: Due to the fact that the clinical trials contain non-precise values of the age attribute, meanwhile the clinical patient data contain a precise value of it (or the birthdate) we have supposed having a precise value for the example above (i.e. we have supposed having the criteria item: “Age=18 years” instead of: “Age ≥ 18 years”)*

The corresponding mapping of these two criteria items to the HL7 messaging standard is the following:



Figure 8: Example of mapping NLP output to HL7 Patient class

The other attributes: id and name of the patient are missing due to the nature of the data we are working on (i.e. such information is not present in the clinical trials data).

**Observation Class**

Most of the criteria types defined in the CTEC representation (Diagnosis, Laboratory\_or\_Test\_Result etc.) fall under the Observation RIM class. The following example illustrates the mapping between the content of *Laboratory\_or\_Test\_Result* criteria type and the corresponding attributes of the Observation class. The criteria item taken as an example is: *“No evidence of ventricular ectopics greater than 4/min on EKG”.* The triple-based representation of this example is the following:

ec1 **hasText** "*No evidence of ventricular ectopics greater than 4/min on EKG*"

ec1 **excludes** f1

f1 **isA** Laboratory\_or\_Test\_Result

f1 **hasObject** x1

x1 **hasTerm** "ventricular ectopics"

x1 **hasCUI** "C0151636"

f1 **hasValue** x2

x2 **hasMin** 4

x2 **hasUnit** "/min"

f1 **hasProcedure** x3

x3 **isA** Diagnostic\_Procedure

x3 **hasTerm** "EKG"

x3 **hasCUI** "C1623258"

*HL7 Observation*

*HL7 ObservationRange*

*HL7 Methodology*

Figure 9: Triple-based representation of a criteria item example defining a Laboratory\_or\_Test\_Result criteria type

The criteria item above represents a test result which is excluded. In terms of HL7 messages it defines an Observation class which has an observation range and a specific methodology used for doing the observation.

The HL7 message subpart representing this criteria item is the following:



Figure 10: Example of mapping NLP output to HL7 Observation class

We have made use of the negation indicator (“*negationInd*” attribute of the Observation class) for representing the fact that the observation is negated. The date of the Diagnosis is not specified in this criteria item example and as a result the “*effectiveTime*” attribute of the Observation class is not populated with the respective information.

**Procedure Class**

The criteria item example taken for the Procedure class requires the eligible patient to have undergone a specific surgical intervention. The criteria item is: “*Patient must have undergone lumpectomy*” and its respective triple-based representation is shown in Figure 11:

e2  **hasText**  "Patient must have undergone lumpectomy"

e2  **includes**  e1

e1  **isA**  Treatment

e1  **hasCUI**  "C0851238"

e1  **hasTerm**  "lumpectomy"

Figure 11: Triple-based representation of a criteria item example defining a Treatment criteria type

HL7 mapping proposal for this example is the following one:

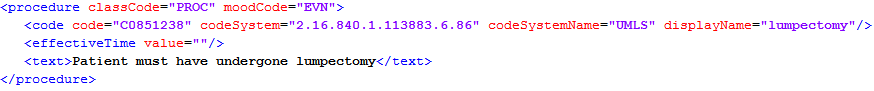


Figure 12: Example of mapping NLP output to HL7 Procedure class

The HL7 message contains the information about the surgical intervention activity called “lumpectomy” including its corresponding UMLS concept unique identifier.

**Substance Administration Class**

The last example taken is the example which represents a treatment criteria type in the context of the triple-based representation and a substance administration class in the context of HL7 messaging standa rd.

The example: “*Received treatment with bevacizumab*” has the following triple-based representation:

e3  **hasText**  "Received treatment with bevacizumab"

e3  **includes**  e1

e2  **hasCUI**  "C0796392"

e2  **hasTerm**  "bevacizumab"

e1  **hasDrug**  e2

e1  **hasCUI**  "C0087111|C1522326|C1533734|C1705169|C3161471"

e1 **isA** Treatment

e1  **hasTerm**  "treatment"

Figure 13: Triple-based representation of a criteria item example defining a Treatment criteria type

Its corresponding HL7 mapping proposal consists of the following representation:

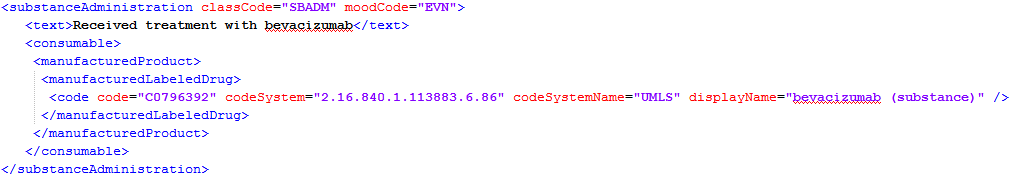


Figure 14: Example of mapping NLP output to HL7 Substance Administration class

#### Future work

Using the CTEC data for elaborating the mapping between the NLP tool output and the HL7 messaging standard has the drawback of not having all the important/mandatory attributes to be mapped due to the nature of this data. For example, the date of the diagnosis is not present in CTEC, which leads to the fact that the respective attribute value in the HL7 message will be missing. As a result, one of the most important future directions will be the adaptation of this mapping to clinical patient data.

The examples shown in the previous subsection give a first intuition of how to automatically map the CTEC data to HL7 messages and this will be the second important future direction. The usage of UMLS semantic information will be a helpful indicator in this perspective.

As stated in D3.2, section 3.2 [2], complex facts (a fact being an instance of one of the classes defined in the conceptual schema for CTEC, figure 11, deliverable D3.2 [2]) that involve conjunction, disjunction, temporal information and/or negation of subfacts are not handled by the current relation extraction tool. The support of these complex facts by the future version of the tool will be followed by analysing and performing new ways of obtaining the corresponding mapping to HL7 messages.

Having a precise list of attributes considered as mandatory for the exchange of HL7 messages will give another useful contribution to the focus of the mapping process.

## Terminology linking service (VUA)

The objective of the terminology linking service is to translate concepts of different terminologies and the EURECA Core dataset. The arguments will be the concept, original terminology and the target terminology, returning the concept according to target terminology.

The Terminology Linking Service (TLS) is delivered as a software package written in Java. It consists of:

* A library of four Java classes
* An example program
* Javadoc pages

#### The TLS as described in Deliverable 4.2

As described in D4.2 the TLS is a wrapper around the BioPortal SPARQL endpoints that provides functionality that is expected to be used in the project, in the form of 8 convenience methods providing access to the chosen terminologies SNOMED CT, ICD 10, ICD 9, CTCAE, MeSH, MedDRA, NCIT, LOINC and the links between them.

Convenience methods:

1. Given a label, find the concept URL in a given terminology
2. Given a label, find the concept URLs in all EURECA terminologies
3. Given a regular expression find URLs in a given terminology of concepts with labels that match the expression.
4. Given a regular expression find URLs in all EURECA terminologies of concepts with labels that match the expression.
5. Given a concept URL, find URLs of broader concepts
6. Given a concept URL, find URLs of narrower concepts
7. Given a concept URL, find the URL of the linked concept(s) in a given terminology
8. Given a concept URL, find the URL of the linked concepts in all EURECA terminologies.

#### Functional description of classes and methods

Figure 15 shows a class diagram of the TLS, displaying the four classes – TerminologyService, TerminologyLinkingService, BioPortalRequest and BioPortalOntologyAcronyms – and their methods. The eight convenience methods that form the main functionality of the TLS are in blue italic font. Below we will give a description of each class.

The **BioPortalRequest class** is used by all other classes. It is sole purpose is to connect to the BioPortal API to send a SPARQL query and retrieve the results. We have decided not to use any of the known RDF stores such as Jena or Sesame in our implementation, to make sure we do not depend on BioPortal’s continuing support for these libaries. Instead, we make a direct HTTP request to the bioportal server. A BioPortalRequest object is constructed using BioPortalRequest(String service, String apikey), where the service is the URI of the BioPortal SPARQL endpoint (either <http://sparql.bioontology.org/mappings/sparql> or <http://sparql.bioontology.org/sparql>) and the apikey is a key we requested for the EURECA project. The code for this class was based on example code by NCBO published on <https://github.com/ncbo/sparql-code-examples/>.



Figure 15: UML class diagram of the TLS.

The **TerminologyService class** is the implementation of convenience methods 1 to 6 that query the concepts, labels and hierarchical relations in the selected EURECA terminologies. The convenience methods formulate a corresponding SPARQL query given a concept label, a regular expression or concept URI. If the acronym of a EURECA terminology is provided, the SPARQL query is constructed so that only this terminology is queried; if no terminology is provided, all 8 selected EURECA terminologies are queried.

BioPortal provides serveral options for the formats in which the results are returned: “text/plain”, “application/json”, “application/rdfxml”, “text/csv” or “text/tab-separated-values”. In the TerminologyService class the format is by default set to “text/tab-separated-values”. In this setting, the results are returned as a string with tabs separating the variables (rows) and newline characters separating the solutions (columns). This setting can be changed using the setAcceptFormat(String acceptFormat) method, or requested using the getAcceptFormat() method.

The EURECA project apikey and the right SPARQL service for this type of queries (<http://sparql.bioontology.org/sparql>) are hardcoded in this class. However, if necessary they can be set with the setSparqlService(String sparqlService) and setApikey(String apikey) methods, or requested with the getSparqlService() and getApikey() methods.

The **TerminologyLinkingService class** contains the 2 convenience methods that deal with mappings between concepts in different terminologies. The format of the results, api key and SPARQL service can be set in the same way as in the TerminologyService class. The SPARQL service is set to <http://sparql.bioontology.org/mappings/sparql> by default here, which is the right service for queries regarding mappings.

In the TerminologyLinkingService, we restrict the results to links with concepts in a given terminology (convenience method 7) or with any of the eight EURECA terminologies (convenience method 8). Technically, we restrict to a VirtualOntology. A VirtualOntology groups several (possibly old) versions of the same Ontology in BioPortal. For example, the VirtualOntology of MeSH has four Ontology versions. In this way, we return links with concepts in all of MeSH’ ontology versions and we don’t burden the user with selecting a version.

The **BioPortalOntologyAcronyms class** is used by the TerminologyLinkingService class. It queries BioPortal for all VirtualOntologies and their acronyms. This allows users of convenience methods 7 and 8 to use the acronym of the terminology, rather than the URI or ID, which are used by BioPortal to identify VirtualOntologies.

To clarify this point, Figure 16 illustrates part of the BioPortal schema that is related to ontology versions and mappings. For brevity, only the relations with the target terminology are shown; the schema to denote the source terminology is the same.



Figure 16: Part of the BioPortal data model

Technically, each time a TerminologyLinkingService object is created, we create a new BioPortalOntologyAcronyms object in which we store the identifiers and acronyms in a a twodimensional array. The method getVirtualOntologyID(String acronym) is used in the TerminologyLinkingService class to retieve a an ID given an acronym of a terminology.

In **all four classes**, the runtime of all SPARQL queries is printed to standard out. When the user provides a terminology acronym that is not in the list of selected EURECA terminologies, a warning is printed to standard out.

#### User documentation

Standard Javadoc pages are provided for all classes. In an example java program called TLSExample.java we illustrate and comment on the use of the 8 convenience methods and show how to set and get parameters in the four classes discussed above.

The 8 convenience methods above can, for example, be called using the following java code snippets:

1. ts.getConcepts("Alanine aminotransferase increased", "CTCAE")
2. ts.getConcepts("Alanine aminotransferase increased")
3. ts.getMatchingConcepts("^Fever malign", "MDR")
4. ts.getMatchingConcepts("^Grade 1 Pulmonary fibros.s")
5. ts.getBroaderConcepts("http://purl.bioontology.org/ontology/MSH/D011981")
6. ts.getNarrowerConcepts("http://purl.bioontology.org/ontology/MSH/D011967")
7. tls.getMappedConcepts("http://purl.bioontology.org/ontology/MSH/D008305", "MDR")
8. tls.getMappedConcepts("http://purl.bioontology.org/ontology/MSH/D008305")

The possible input values for the arguments of the methods are discussed as comments in TLSExample.java. Also, we note that users should be careful when using convenience methods 3 and 4. Depending on the regular expression, these queries can take up to several minutes. In general, regular expressions that provide the first characters of the label, e.g. by using the start of line character ^ as in "^Fever malign", give faster results than expressions that do not specify where the given characters can be found in the label, as in "ever maligna". This is presumably due to the properties of the BioPortal index.

## Core dataset to EURECA CDM (UPM)

## EURECA guidelines

Standards and technologies used at each clinical partner are frequently institution specific. Therefore, such differences must be solved with a practical solution that requires the minimal amount of effort and knowledge about the infrastructure of the EURECA platform.

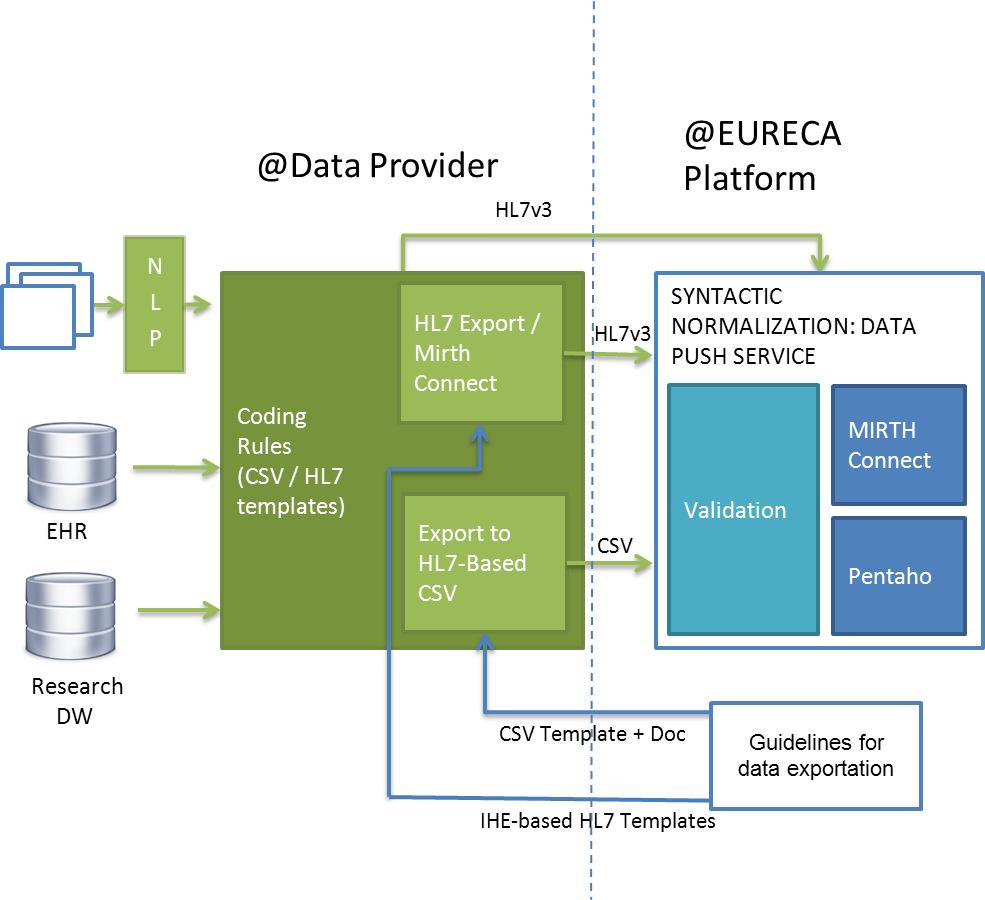


Figure 17: Deployment example

As Figure 17 shows, the EURECA platform will provide a set of technical guidelines for achieving data exportation from hospitals and research institutions. The goal of these guidelines is to help to create a minimal set of standard exchange messages that covers data from a hospital required to integrate their data into EURECA.

The data export process at each institution can be divided in two main tasks:

1. Terms in clinical datasets have to be annotated using standard vocabularies. For the annotation step, a tool to facilitate this task has been developed, Concept2HL7[[4]](#footnote-4). Conept2HL7 is a Core dataset browser developed by the UPM, binding the normalized version of Cored Dataset concepts to the HL7 RIM-based CDM. The Core dataset, as commented before, mainly contains SNOMED CT, LOINC and HGNC concepts.
2. Development of channels to generate HL7 messages. Guidelines are provided with a group of templates for HL7 v3 CDA messages. These templates have been built following IHE recommendations. They cover information that clinical partners plan to export to EURECA.

Then, these HL7 v3 messages templates could be adapted to facilitate the creation of exportation channels. Once the HL7 template is selected, it is just required to map the information from the clinical partner dataset to fields marked in templates as parameters. For message generation, other tools such as Mirth Connect are being used in the project.

The main goal of the documentation guidelines is to describe to IT technicians from institutions, how to model information extracted from their data sources and into the EURECA CDM. I.e. how to represent different types of information from their information systems using most suitable HL7 v3 template from those provided. Besides, some recommendations are added for the mapping between their data sources to the HL7 v3 message fields.

## Data Push Service

#### ETL tools

The ETL process (Extract, Transform and Load) consists on extracting information from data sources and inserting this information into the CDM. To achieve this goal, we have mainly used two open source software: Mirth Connect and Pentaho Data Integration (Kettle). Both solutions are used by the Data Push Service to insert information into de CDM.

In the Mirth Connect case, it is important to note that this software is specifically designed for HL7 messages. Mirth Connect uses “channels” that act like a pipe. In this pipe the input type and the output type are specified, and then, the information is processed inside of the pipe transforming the input data into the correct output data type.

The Data Push Service also uses a channel to automatically process the generated HL7 messages. HL7 messages must be generated according to CDA standard (Clinical Document Architecture). In order to facilitate the creation of messages and guarantee that all the information in the message can be correctly inserted into the CDM, templates for specific HL7 messages have been created.

General ETL tools can also be used to load data into de EURECA CDM, such as Pentaho (Kettle). In this tool, the ETL process is built as a data flow using an intuitive user interface (UI). The graphical design created by the drag and drop option of Kettle, is stored by this software as an xml document. Like in Mirth Connect case, this program can be used to extract the information from different sources and insert it into the CDM.

In the next Figure, an example of transformation using Pentaho Kettle can be observed:

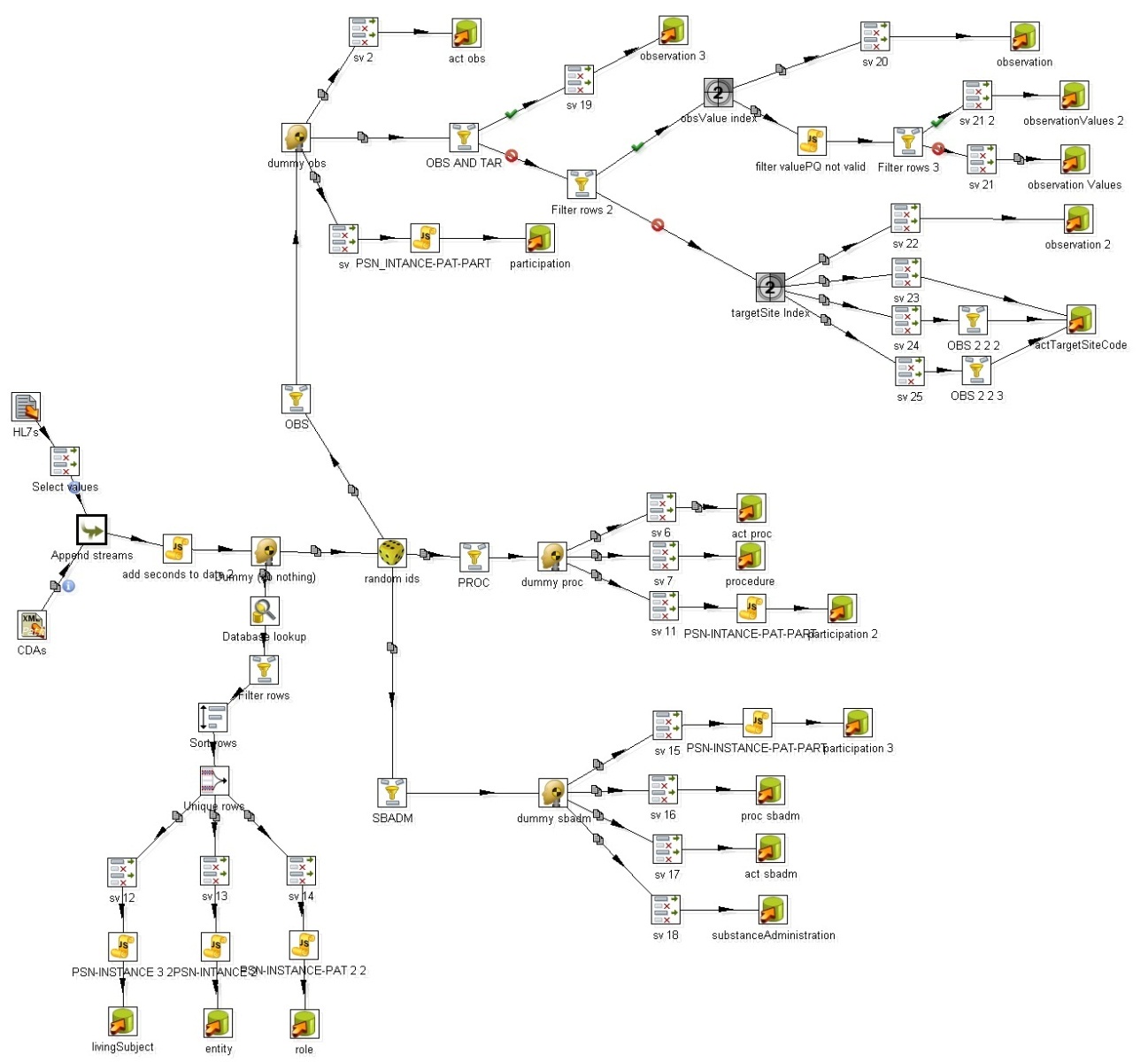


Figure 18: ETL process using Kettle

More information about Kettle can be found in deliverable D.1.3 of EURECA.

Once the information has been inserted into the CDM using the Data Push Service (with Pentaho or Mirth Connect), this information can be queried through the Data Acess Service using SPARQL query language.

#### Normalization pipeline

Once data is loaded into the EURECA CDM, even following HL7 standards, there exist different forms of representing the same information. The normalization pipeline process aims to homogenize such representation of information in the CDM. For this purpose, Core dataset concepts present in data are transformed into a normalized form. This transformation is mainly based on the central vocabulary of the Core ataset, that is, the SNOMED-CT terminology. Concepts from other terminologies in Core dataset (LOINC, HGNC) are not transformed.

First, the normalization pipeline process translates concepts encoded using terminologies that are not included in the Core dataset, e.g. ICD, into equivalent concepts from SNOMED-CT terminology. This step is performed by the Terminology Linking Service, which is explained in section 2.1.3 of this deliverable.

Once all concepts stored into the CDM are coded using the Core dataset, the normalization process transforms SNOMED-CT concepts into its normal form. The most informative version of an SNOMED-CT concept is when represented in normal form. This form also allows retrieval of pre and post-coordinated SNOMED CT expressions.

The first step is to check if concepts stored in the “code” attribute of the “Act” entity of the HL7-based CDM are normalized. If the concept is represented using the normal form, no transformation is required. On the other hand, when a concept needs to be normalized, the Core dataset service returns the set of concepts and relationships that compounds its normal form. In this case, new information has to be frequently stored in other attributes – e.g. method code or target site code – or even in other classes of the model – e.g. entity, role and participation. Besides that new attributes and/or entities are generated and associated to the Act instance, the original code is always maintained and stored in the “codeOrig” attribute defined in the CDM. The new concept obtained from normalization is stored in attribute “code”, both belonging to RIM “Act” table.

In figure 20, the normalization of the SNOMED concept ‘Trocar Biopsy’ is depicted as an example. Trocar Biopsy is represented with SNOMED CT code SCT\_5337005. Its normalization returns two codes, “SCT\_129314006 | Biopsy” and “SCT\_118418003 | Trocar”, which are stored in “Act method code” table and “entity table” respectively – following the information retrieved from the terminology binding. Thus, in this example, due to normalization new relationships have to be created to store the data.



Figure 20: Normalization of a concept

It should also be mentioned that the Normalization pipeline process keeps the semantics of the data. Therefore, the normalized CDM is a superset of the original version of than the data loaded into the CDM.

## Query normalization

The Query Normalization Services aims to facilitate the generation of queries to the CDM. This process starts with (i) a Core dataset concept or (ii) a Core dataset concept plus a CDM context. It returns a query template (Figure 21) that contains the SPARQL query necessary to retrieve the minimum information related to the given concept. The template additionally contains a set of optional filters and attributes that can be applied / retrieved about the Core dataset concept on CDM. Finally, it is possible to execute the desired query on the CDM.

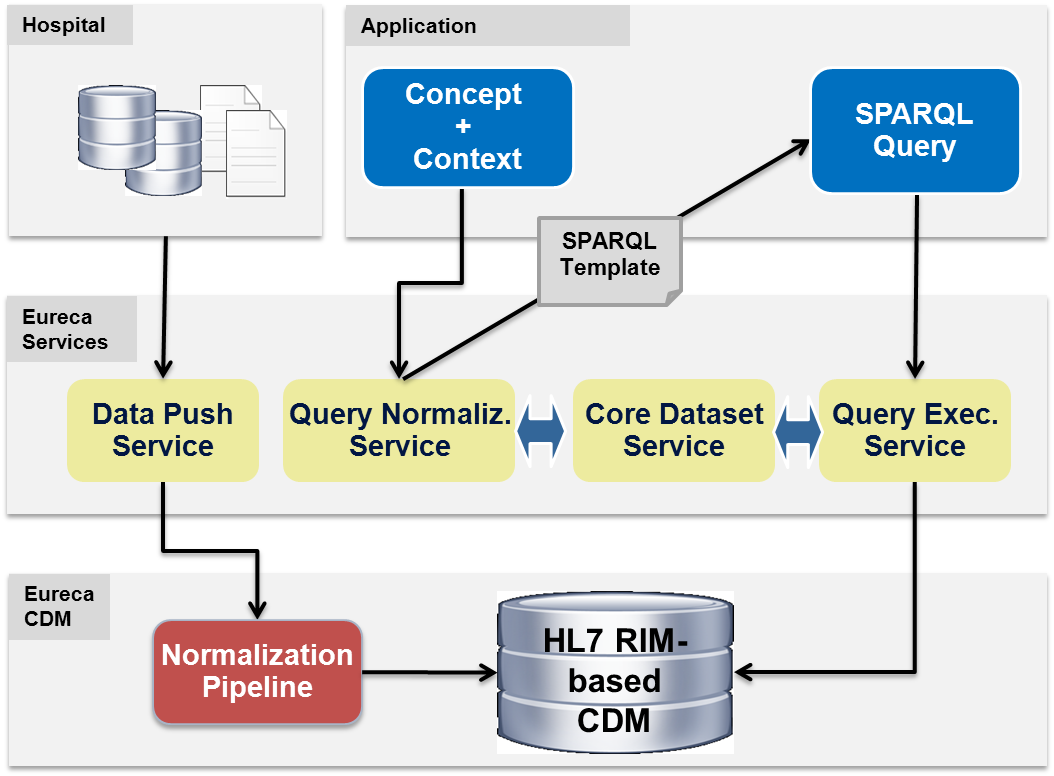


Figure 21: Query Normalization process

In order to retrieve the corresponding template for a given concept, this process uses Core dataset information inferring semantic knowledge of the medical vocabularies on CDM. Therefore, the SNOMED normalization method and the Terminology Binding of the Core dataset are used.

The SNOMED Normal Form is used to translate the given Core dataset concept into the normalized form of the expression. Then, the Terminology binding is applied to map the normalized expression with the CDM. Once this information is obtained, it is possible to select the correct query template for retrieving information on CDM. For executing this task, a Query Template Library (QTL) has been created. QTL is defined by a set of query templates based on CDM domain elements, covering all possibilities of representing Core dataset information on the CDM. Thus, five templates have been defined for that purpose:

* Observation: for querying data of RIM observations related to a patient, entities, targets and values.
* Procedure: for querying data of RIM procedures related to a patient, entities, targets and values.
* Substance Administration: for querying data of RIM substances administration related to a patient, entities and doses.
* Entity/Devices/Products/Genes: for querying data of RIM entities related to a patient.
* Demographic information: for querying data of RIM living subject information, that is, information related to a patient.

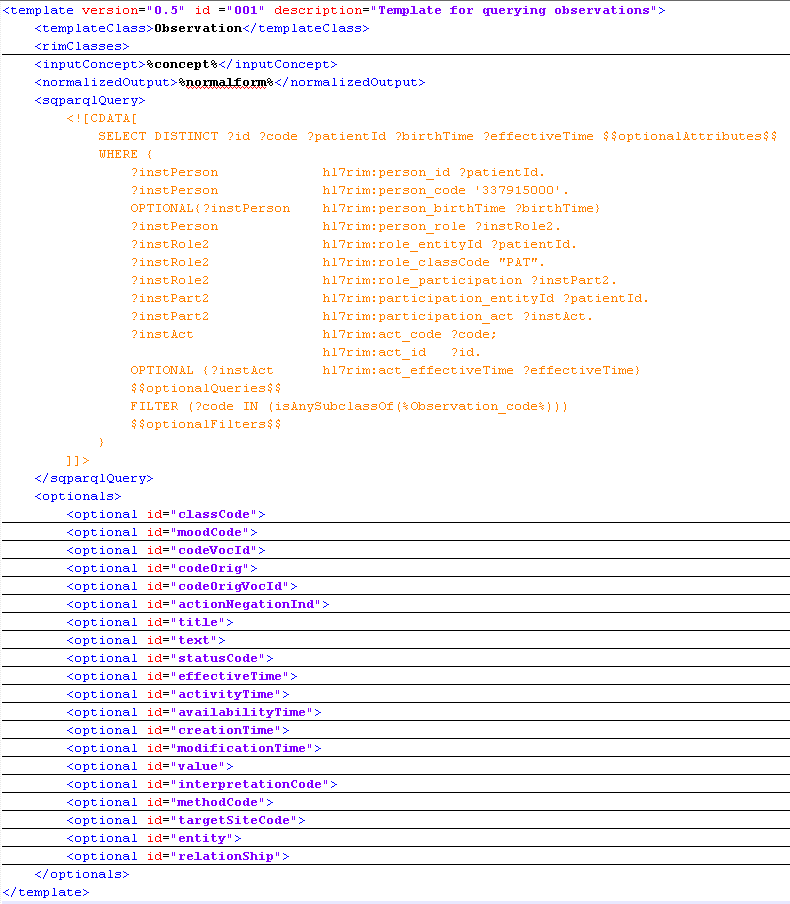


Figure 22: Observation query template

Figure 22 shows an example of the query template for observations. Once the correct query template is obtained, and replacing the required parameters, a query to retrieve data from the normalized EURECA CDM can be built without knowledge about the model. The Query Template Library, will be further explained in deliverable 4.4 “Initial prototype of the semantic interoperability framework”.

# Data source mappings to EURECA (clinical partners)

## Institut Jules Bordet Data

## Data source types of information

The *Institut Jules Bordet*, as part of the clinical partners within EURECA, provides a set of clinical data that are meant to be stored in the CDM of the project platform, so as to be exploited for the different services implemented during the project development.

The clinical data that are used to develop and validate the EURECA tools, and which are used to define the core dataset are coming from the following sources:

* EHR data (Anatomical pathology, Consult reports, Laboratory, Chemotherapy prescription, Multidisciplinary Team (MDT) reports)
* PACS
* TOP trial
* Cancer registry

More information on these datasets could be found in the previous deliverable D9.1[[5]](#footnote-5).

Until now (January 2014), 898 patients already gave their informed consent so that IJB can use and provide their data for research according to legal and ethical requirements, and the amount of patients keeps on increasing.

## Core dataset annotations

## HL7 v3 message template selection

HL7 messages templates that are already implemented at IJB are adapted to fit to the information contained in each dataset. But even if HL7v3 messages do not take charge of coding and annotating concepts enclosed into the unstructured/textual data all services of the project are working on, such CDA files encapsulate part or the totality of text entries.

*Table 1* gives an overview of all the document types that contain some textual/unstructured data and that should be aimed to be used to help extracting relevant information to develop the EURECA scenarios. Some of them are already encoded in HL7v3 CDA formats. The others should obviously ideally be encoded in such a format as well, but contain mainly raw text data up to now.

Document types are coded with both LOINC codes and internal Bordet OID (Object Identifiers) when LOINC terminological terms do not express well enough the concept in question. OIDs make it possible to create customised concepts and allow defining unique identifiers, even between several clinical institutions.

At IJB, internal customised OIDs all start with the same following root OID: 2.25.299518904337880959076241620201932965147, followed by the ones that are specific for each concepts.

|  |  |  |  |
| --- | --- | --- | --- |
| **Document type (in French)** | **Term.** | **Code** | **Terminology description** |
| Note de consultation | LOINC | 11488-4 | Consult note |
| Note d'admission | LOINC | 34744-3 | Nurse Admission evaluation note |
| Annotation administrative | OID | 1.7A | - |
| Annotation Médicale | OID | 1.7M | - |
| Annotation de cytaphérèse | OID | 1.7C | - |
| Résumé d'hospitalisation | LOINC | 34105-7 | Hospital Discharge summary |
| Draft rapport d'hospitalisation | LOINC | 34130-5 | Hospital Progress note |
| Soins Infirm. Consultations | LOINC | 11536-0 | Nurse notes |
| Résumé de radiothérapie | LOINC | 34831-8 | Radiation oncology consult note |
| Protocole opératoire | LOINC | 11504-8 | Provider-unspecified Operation note |
| Lettre de chirurgie | OID | 1.29 | - |
| Lettre de depistage | OID | 1.28 | - |
| Lettre Accueil | OID | 1.30 | - |
| Lettre | OID | 1.27 | - |
| Endoscopie | LOINC | 18751-8 | Endoscopy study |
| Résumé CMO | OID | 1.2 | - |
| Journal USI | OID | 1.7.1 | - |
| Journal urgence | OID | 1.7.2 | - |
| Echocardiographie | LOINC | 11522-0 | Cardiac echo |
| Echographie gynécologique | LOINC | 34778-1 | Obstetrics and Gynecology Note |
| Protocole d'Anatomo-Pathologie | OID | 1.410 | - |
| Protocole de Radiologie | LOINC | 11528-7 | Radiology |
| Scanner-RMN | OID | 1.611 | - |
| Médecine nucléaire | OID | 1.620 | - |
| Rapport de prise en charge en hôpital de jour | LOINC | 34106-5 | Physician Hospital Discharge summary |
| Annotation -Urgences | LOINC | 11488-4 | Consult note |
| Bilan urodynamique | OID | 1.50 | - |
| Bilan d'évaluation gériatrique | OID | 1.60 | - |
| Rapport ICSO | OID | 1.770 | - |
| Protocole d'imagerie - DEXA | LOINC | 38269-7 | Study report Skeletal system DXA |
| Dossier nutrition | LOINC | 34800-3 | Nutrition and dietetics Consult note |
| Note d'urgence | LOINC | 15507-7 | Provider-unspecified ED Progress note |
| Protocole d'échographie vésico-prostatique | OID | 1.616 | - |
| Note de psychologue | LOINC | 34792-2 | Psychology Note |

Table 1: Internal EHR document type codifications

Part of these OIDs should be mapped to standard terminologies/ontologies (like SNOMED, NCI or other UMLS terminology) when possible, but ask long-term efforts and involvement from both healthcare IT and medical staff.

We defined a list of possibly interesting sections for the whole set of document types that have been described hereinbefore. This list can be seen on *Table 2* and coded with both LOINC and OIDs as well. These section codes define the basis for the generation of HL7v3 documents on our site.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sections (French)** | **Term.** | **Code** | **Terminology description** |
| Motif d’admission | LOINC | 46241-6 | Hospital admission diagnosis Narrative - Reported |
| Affection actuelle | LOINC | 11450-4 | Problem list - Reported |
| Antécédents | LOINC | 11348-0 | History of past illness Narrative |
| Mode de vie | LOINC | 29762-2 | Social history Narrative |
| Examen physique | LOINC | 29545-1 | Physical findings Narrative |
| Examens complémentaires | OID | 5.530.6 | - |
| Evolution en cours d'hospitalisation | LOINC | 8648-8 | Hospital course Narrative |
| Traitement en cours d'hospitalisation | LOINC | 42346-7 | Medications on admission (narrative) |
| Traitement de sortie | LOINC | 10183-2 | Hospital discharge medications Narrative |
| Conclusion(s) | LOINC | 11450-4 | Problem list - Reported |
| Disposition(s) | LOINC | 18776-5 | Plan of treatment (narrative) |
| Antécédents néoplasiques | OID | 5.530.12 | - |
| Antécédents médicaux | OID | 5.530.13 | - |
| Antécédents chirurgicaux | LOINC | 47519-4 | History of Procedures Document |
| Antécédents gynéco-obstétricaux | OID | 5.530.15 | - |
| Anamnèse | LOINC | 11348-0 | History of past illness Narrative |
| Examen clinique | LOINC | 59776-5 | Procedure findings Narrative |
| Diagnostic | LOINC | 30954-2 | Relevant diagnostic tests/laboratory data Narrative |
| Evolution | OID | 5.530.19 | - |
| Histoire de l'affection néoplasique | OID | 5.530.20 | - |
| Détails de la radiothérapie | OID | 5.530.21 | - |
| Traitement à domicile | LOINC | 42346-7 | Medications on admission (narrative) |
| Motif de la consultation | LOINC | 29299-5 | Reason for visit Narrative |
| Examen sénologique | OID | 5.530.24 | - |
| Examen gynécologique | OID | 5.530.25 | - |
| Attitude | OID | 5.530.26 | - |
| Examens réalisés dans le cadre du bilan | OID | 5.530.27 | - |
| Interrogatoire systématique | LOINC | 10210-3 | Physical findings of General status Narrative |
| Antécédents familiaux | LOINC | 10157-6 | History of family member diseases Narrative |
| Echographie endovaginale | OID | 5.530.30 | - |
| Intervention chirurgicale | OID | 5.530.31 | - |
| Traitement | OID | 5.530.32 | - |
| Suites opératoires | OID | 5.530.33 | - |
| Anatomopathologie | OID | 5.530.34 | - |
| Radiothérapie | OID | 5.530.35 | - |
| Examens réalisés | LOINC | 46240-8 | History of hospitalizations + History of outpatient visits Narrative |
| Conclusions et dispositions | OID | 5.530.37 | - |
| Antécédents gynécologiques | OID | 5.530.38 | - |
| Protocole technique | OID | 5.530.39 | - |

Table 2: CDA section types

For a future development a subset of these sections should be selected and adapted for each type of document, as our CDA still mostly are XML-like files containing some text entries with full raw text or semi-structured text. In addition, more well-adapted templates, such as the ones provided by UPM, should be more appropriate to be able to process semantically and correctly the data.

## HL7 v3 message generation

HL7v3 CDA messages are generated by an internal script, built on the information that has been shortly described hereinbefore. There are no formal templates used for that purpose yet.

## MAASTRO Clinic Data

## Data source types of information

MAASTRO Clinic contributes patient data to EURECA that was collected during radiotherapy treatment for cancer. Although there are multiple sources of data within the MAASTRO IT environment, the bulk of data that is available in a structured form is extracted from MAASTRO’s own EMR (Electronic Medical Record) system, called EMD (Electronische Medische Dossier). Some data in the EMR is also (Dutch) free text that results when a clinician writes notes and observations into the EMR. Several other data sources exist in the clinic but the initial priority was given to data which is of critical importance during the evaluation of eligibility for MAASTRO trials such as TNM status.

## Core dataset annotations

MAASTRO has supplied to the EURECA consortium patient data in the form of XML files. Subsequently these data files were converted to HL7 messages according to the EURECA specifications. Specifically three datasets were shared via the EURECA Central Data Protection (CDP) authority:

* 174\_LungCancerPatient.xml
* 479\_EURECA\_BREAST.xml
* 500\_EURECA\_BREAST.xml

In these files 4.626 lung cancer, 3.372 and 1.087 breast cancer patients were included respectively.

The data elements included in the above files were both in the form of structured data and “buried” in free text. The structured data elements were all already coded concepts, either using the NCI Thesaurus (NCIt) or a custom ontology maintained by MAASTRO. The free text data elements were processed by another partner (Philips) to extract concepts. The extracted concepts were reviewed by MAASTRO and examples are given in the table below:

|  |  |
| --- | --- |
| **Concept** | **Coded/Free text** |
| Breast cancer | Coded (NCIt) |
| Lung cancer | Coded (NCIt) |
| Age of the patient at pathology | Coded (custom) |
| Date of diagnosis | Coded (custom) |
| T stage | Coded (NCIt) |
| N stage | Coded (NCIt) |
| M stage | Coded (NCIt) |
| Gender | Coded (NCIt) |
| Clinical trial participation | Coded (custom) |
| Contact date | Coded (custom) |
| Weight | Coded (custom) |
| WHO/ECOG performance status | Coded (NCIt) |
| **Various comorbidities and procedures** | **Extracted from medical history free text** |
| *COPD* |  |
| *Hypercholesterolemia* |  |
| *Osteoarthritis* |  |
| *Hypertension* |  |
| *Asthma* |  |
| *Pulmonary embolism* |  |
| *Cardiac dysrhythmia* |  |
| *Hyperthyroidism* |  |
| *Hysterectomy* |  |
| *Osteoporosis* |  |
| *Tonsillectomy* |  |
| *Cholecystectomy* |  |
| *Glaucoma* |  |
| *Fibromyalgia* |  |
| *Hypothyroidism* |  |
| *Rheumatism* |  |
| *Caesarean section* |  |
| *Polyneuropathy* |  |
| *Jaundice* |  |
| *Psoriasis* |  |
| *Hernia* |  |
| *Atrial fibrillation* |  |
| *Cataract* |  |
| *Epilepsy* |  |
| *Thrombosis* |  |
| *Myocardial infarction* |  |
| *Cardiac arrest* |  |
| *Breast reduction* |  |
| *Migraine* |  |
| *Thyroid Surgery* |  |
| *Salpingitis* |  |
| *Artificial cardiac pacemaker* |  |
| *Cervical cancer* |  |
| *Stomach hemorrhage* |  |
| *Abortion* |  |
| *Oophorectomy* |  |
| *Macular degeneration* |  |
| *Gastric perforation* |  |
| *Tachycardia* |  |
| *Carpal tunnel syndrome* |  |
| *Neurofibromatosis* |  |
| *Pancreatitis* |  |
| *Venous varices* |  |
| *Mitral valve regurgitation* |  |
| Various medications and comorbidities (not listed) | Extracted from medication free text |

Table 3: Maastro Core dataset annotations

## HL7 v3 message template selection

The generic *Observation* template was used to generate HL7 messages that represent tumour status (TNM status and tumour staging), with one message per patient containing cancer type and TNM status. We also used a modification of the template for Antibiotic Treatment or substanceAdministration (SBADM) for the medications extracted from free text by Philips. For the medical oncological history (concepts extracted by Philips), we used two different templates: 1) “Disease, Syndrome and Finding” for Diagnosis and 2) “Care activity procedure” for medical procedures.

## HL7 v3 message generation

MIRTH Connect has been used to process the XML and to generate HL7 v3 messages from the contents using the chosen templates. UPM was able to import TNM status with NCI Thesaurus codes. The medicines, diagnoses, and procedures extracted by Philips from MAASTRO’s free text were used to produce HL7 messages with SNOMED codes. In cases where the data was structured and typed, this was a mapping process with some extra logic. The exception to the MIRTH Connect channel approach was with data that was extracted from free text, such as medical oncology history. In that case, the “Root” type of the extracted concept had to be checked in order to choose the appropriate message template, i.e. Diagnosis vs. Procedure. For that purpose, we used the getRootConcept() Web Service from UPM.

## University of Oxford data

## Data source types of information

As a clinical partner, university of Oxford provides a number of different types of clinical data that can be exported to the EURECA platform. These data sources contain clinical data that have been collected for a number of studies. These clinical data mainly include information that has been manually extracted from the following clinical systems:

* Electronic patient record system (main hospital EPR system)
* Pathology reports
* Clinical notes
* Radiology and imaging system

However, the data available for the exportation to EURECA platform are manually curated and stored in local research database instances. Therefore data need to be exported from the local research database software and transformed into HL7 messages.

## Core dataset annotations

There are mainly two main datasets that can be exported from the local research database:

* Breast cancer patient clinical outcome dataset
* Sarcoma patient pathology information

These datasets originated from different research projects that require the accumulation of clinical information about the patients. Patient information were manually looked up on various hospital information systems and entered into the research database software.

In order to export the data in a meaningful way so that the clinical information can be processed in a semantically correct manner, clinical data need to be annotated with a standard clinical terminology. In the following table, annotations of the data fields are made to map different types of clinical information to SNOMED-CT concepts.

|  |  |
| --- | --- |
| **Data item** | **SNOMED-CT concept** |
| Age | Age (Code 102518004) |
| Menopause | Menopause finding (Code 276477006) |
| TSize | Tumor size (Code 263605001) |
| Nstatus | Generic lymph node tumour invasion status stage (Code 258309004) |
| Nnumb | Number of lymph nodes involved by malignant neoplasm (Code 443527007) |
| Nsampled | Lymph node tissue sample (Code 309078004) |
| ERElisa | Enzyme-linked immunosorbent assay (Code 76978006) |
| EGFR | Assay technique (Code 272392009) |
| Grade1 | Histological grading systems (Code 277457005) |
| ReGrade | Histological grading systems (Code 277457005) |
| Histology | Neoplasm (morphologic abnormality) (Code 108369006) |
| RT | Radiation oncology AND/OR radiotherapy (Code 108290001) |
| Chemo | Chemotherapy (procedure) (Code 367336001) |
| Side | Laterality (attribute) (Code 272741003) |
| Operation | Excision of breast tissue (Code 69031006) |
| Axilla | Neoplasm of axilla (disorder) (Code 126639006) |
| eRFSsteep | Surviving free of recurrence of neoplastic disease (Code 445150007) |
| tRFSsteep | Duration of recurrence-free survival (Code:445397003) |
| eRFSsteep10y | *Requires Post-coordination* |
| tRFSsteep10y | *Requires Post-coordination* |
| eDRFSsteep | *Requires Post-coordination* |
| tDRFSsteep | *Requires Post-coordination* |
| eDRFSsteep10y | *Requires Post-coordination* |
| tDRFSsteep10y | *Requires Post-coordination* |

Table 4: Oxford Core dataset annotations

## HL7 v3 message template selection

Due to the nature of the clinical information in the research database, HL7 message templates are selected accordingly to fit the purpose of exporting clinical data in a standardised way. Among the HL7 message templates that have been provided in the mapping guidelines, the *Genetic Finding* template is used extensively to represent data that were obtained from genetic test results. The generic *Observation* template is also an important candidate to generate HL7 messages that represent the data about patient diagnostic information.

## HL7 v3 message generation

Currently the HL7 XML files are generated by a script that reads in the exported datasets and creates HL7 messages based on the given template. The placeholders in the templates will be filled by the value of the mapped data field from the research datasets. When more data become available, mapping channels need to be created by using MirthConnect to automate the HL7 message generation.

## Saarland University Hospital data

The Saarland University Hospital will provide clinical data stemming from several different sources. Clinical trial data will be provided by the ontology-based clinical trial management system ObTiMA (see the relevant documentation at http://obtima.org and <http://p-medicine.eu/downloads/deliverables> for more detailed information on this). Further, clinical data will be provided by their hospital information system that is based on i.s.h.med (an extension of SAP R3 for clinical environments) together with laboratory data retrieved form the laboratory management system M/Lab.

The department for Pediatric Oncology and Hematology at the Saarland University Hospital develops and deploys ObTiMA as clinical trial management system for the SIOP 93-01/GPOH and SIOP 2001/GPOH trials. In regard to EURECA, the deliverables D2.2 and D3.3 describe the intended plan and design of how to implement mechanisms to extract and transform clinical trial data from ObTiMA into the HL7 CDA format which is compatible with the EURECA CDM. The following sections below describe in detail the steps required to generate HL7 v3 messages based on the provided set of predefined, specific HL7 CDA templates out of the available data sets in ObTiMA. The focus here at is set on the export of data collected in the above mentioned clinical trial.

## Data source types of information

The SIOP 93-01/GPOH clinical trial was carried out between April 1994 and December 2001 with 962 patients with Wilms' tumour were enrolled [1] and the SIOP 2001/GPOH clinical as its follow up running until September 2011. The relevant patient parameters/data were captured by specific case report forms (CRFs) , such as a CRF for registering a patient in the trial or a CRF to describe the pre-operative chemotherapy of a patient.

To clarify the approach we employed, we exemplarily describe each step performed to export t the questions asked on the CRF “F1 Registration”. Figure 23 shows the paper based form of this particular CRF (which is used to collect the data before the deployment of ObTiMA).

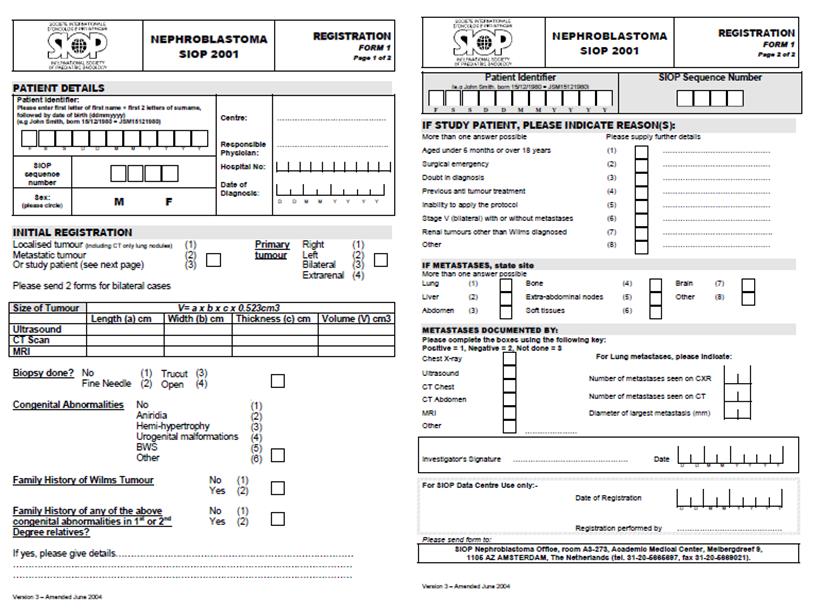


Figure 23: Complete CRF “F1 Registration” of the SIOP/GPOH trial in a paper based form

As first step, the paper-based form of this CRF was taken and used as blueprint to create the same in an associated electronic form within ObTiMA’s clinical trial builder. The patient-related data of the mentioned trials were then entered at the clinical trial sites into ObTiMA based on this CRF for newly enrolled patients. For existing patient data, which had been stored in an external database predating the development of ObTiMA, a tool was developed to import this data into the electronic CRFs created in ObTiMA.

Figure 24 shows a screenshot taken from ObTiMA where the structured format of the section “size of tumour” of the CRF “F1 Registration” is depicted. On this image, one can see that e.g. that the entry field for tumour size is typed to hold a numerical value with centimetre as is measurement unit i whereby e.g. the location of the primary tumour can be selected via radio buttons from a set of four answer possibilities defined during the initial set-up of this CRF in ObTiMA’s trial definition phase.

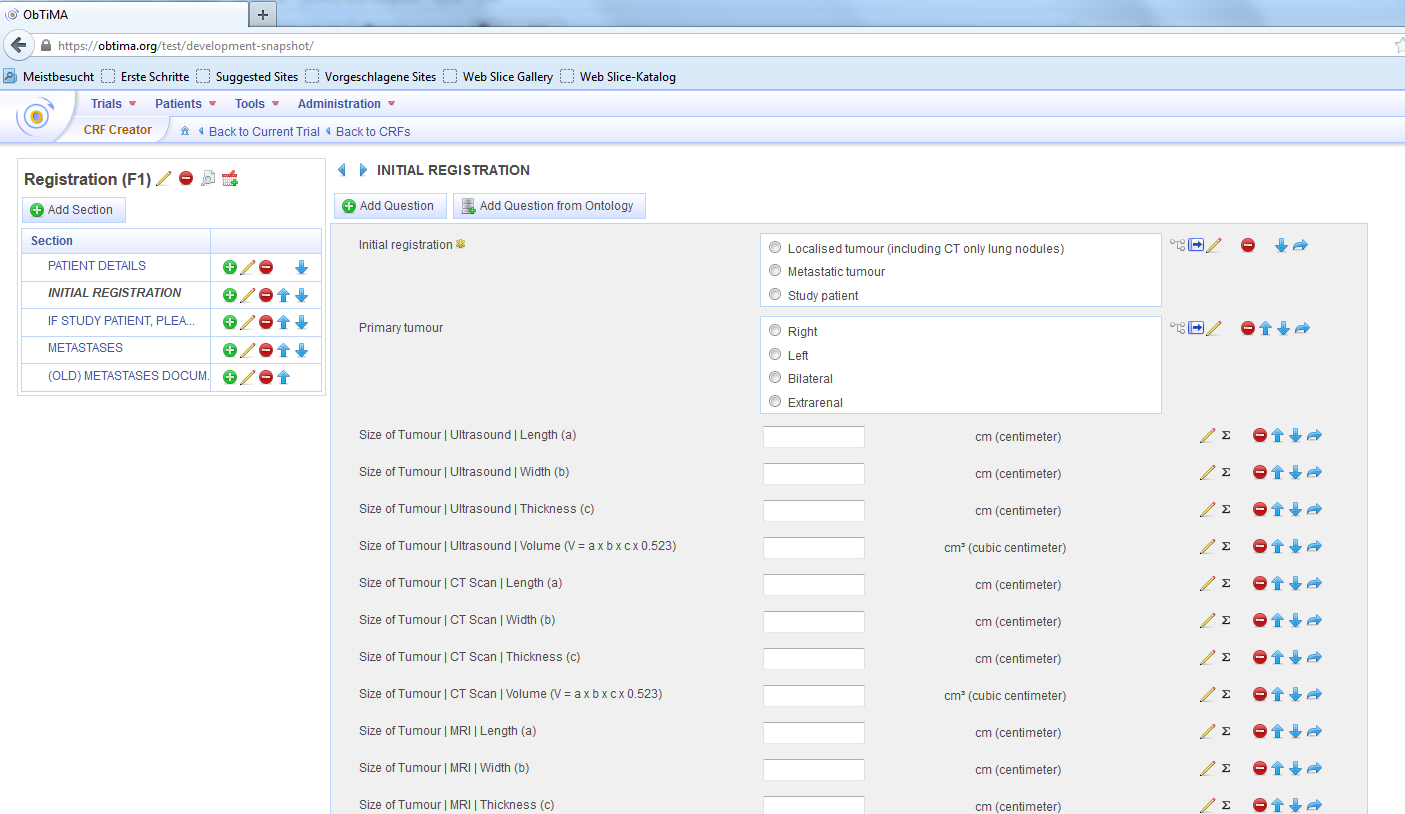


Figure 24: CRF “F1 Registration” of the SIOP/GPOH trial provided ObTiMA

## Core dataset annotations

The extraction and translation of the single CRF questions from ObTiMA into the needed EURECA CDM format was carried out through a particular mapping process which required the manual set-up of a specific mapping file in ObTiMA.

In order to perform this task, we checked all questions and the respective answer possibilities (in the case of multiple choice questions) found on the CRF “F1 Registration” with the help of the web application Concept2HL7[[6]](#footnote-6) provided by UPM.

For most of the questions and answer possibilities, a clear and obvious classification with a suitable item from the CDM was possible, like it is shown in the following table:

|  |  |  |  |
| --- | --- | --- | --- |
| **CRF Question** | **CRF Answer** | **CDM Item Code (SNOMED-CT)** | **CDM Item Label**  **(SNOMED-CT)** |
| Congenital Abnormalities |  | 66091009 | Congenital disease (disorder) |
|  | No | 373067005 | No (qualifier value) |
|  | Aniridia | 69278003 | Congenital aniridia (disorder) |
|  | Hemi-hypertrophy | 56007004 | Congenital hemihypertrophy |
|  | Urogenital malformations | 287085006 | Genitourinary congenital anomalies |
|  | BWS | 81780002 | Beckwith-Wiedemann syndrome |
|  | Other | 74964007 | Other (qualifier value) |

Table 5: Example for EURECA Core dataset annotations for the CRF question “Congenital Abnormalities”

Problems arose for the classification of some complex coordinated expressions, like “Aged under 6 years or over 18 years” or “Renal tumours other than Wilms diagnosed”. No direct mapping of those expressions to single concepts is feasible, neither to SNOMED-CT nor any other from the set of ontologies/terminologies included within the CDM. But, as it is not mandatory at this stage to fully map all questions and answer possibilities to single CDM items, we currently disregard those items in the mapping file but will return to solving this issue at a later stage during the project.

## HL7 v3 message template selection

The SIOP 93-01/GPOH and SIOP 2001/GPOH clinical trial data will be exported in form of HL7 v3 messages from ObTiMA. This process requires a message generation that is based on translating the SIOP data into corresponding, specific HL7 CDA templates. Each of the HL7 CDA templates consists of several specific sections with every section representing a particular EURECA CDM item. Therefore we selected for every question/answer block on the CRF “F1 Registration” in ObTiMA a suitable HL7 CDA section from the given set of predefined CDA templates.

For some question/answer items, like “Patient”, corresponding HL7 CDA sections exist and thus a translation into a CDM item does not need to be performed. But for most other items, like “Congenital Abnormalities”, a suitable item needs to be found within the CDM with the help of the HL72Concept tool before selecting a HL7 CDA suitable section.

The following examples illustrate this procedure:

* CRF question “Patient Details”:

<recordTarget>

<patientRole>

<id extension="%%PATIENT\_ID%"/>

<patient>

<administrativeGenderCode code="%% Male(248153007) or Female(248152002))%%" codeSystem="2.16.840.1.113883.5.1"/>

<birthTime  
 value="%%BIRTHDATE(YYYYMMDDHHMMSS)%%"/>

</patient>

</patientRole>

</recordTarget>

* CRF question “Congenital Abnormalities” leads to concept “Observation”:

The corresponding answer possibilities have to be represented through an observation section:

<entry>

        <**observation classCode="OBS"** moodCode="EVN">

               <templateId root=""/>

                     <id root="%%ID\_OF\_THE\_OBSERVATION%%"/>

                       <code ="%%code respected Congenital Abnormality%%"

displayName="%%SNOMED CT literal of respected Congenital Abnormality %%"/>

                                codeSystem="2.16.840.1.113883.6.96"

                                codeSystemName="SNOMED CT"/>

                        <statusCode code="COMPLETED"/>

         </observation>

  </entry>

* CRF question “Biopsy done” leads to concept “Procedure”:

The procedure (biopsy) and its applied method can be described through a Procedure section

<entry>

   <**procedure classCode="PROC"** moodCode="EVN">

<code ="86273004"

codeSystem="2.16.840.1.113883.6.96" c

codeSystemName="SNOMED CT"

displayName="**Biopsy (procedure)**"/>

<effectiveTime value="20130219"/>

<methodCode

code="%%code of biopsy method%%" codeSystem="2.16.840.1.113883.6.96"

codeSystemName="SNOMED CT"

displayName="%% SNOMED CT literal of biopsy method%%"/>

</procedure>

</entry>

The summarizing and integration of all HL7 CDA sections composed on the basis of the SIOP data form the specific, completeHL7 v3 message.

In parallel with the discovered issue to map complex expressions onto fitting SNOMED-CT concepts discovered, it is hard to find suitable HL7 CDA sections for some complex expressions. Also in parallel to the same issue with SNOMED-CT concepts, those expressions are disregarded for the first exemplary implementation but will be tackled at a later stage.

## HL7 v3 message generation

All filled-in questions of the CRFs of the given SIOP trials will be transferred to the EURECA CDM through HL7 v3 messages.

As pointed out above, this message generation is carried out through the use of the fitting predefined HL7 CDA templates combined with the extracted and translated SIOP specific data. This translation to HL7 CDA is performed based on the manually created,corresponding mapping file which includes the CDM annotations.

As also mentioned before, some HL7 CDA sections like “patient”, “author” or “custodian” can be filled-in directlywhereas entries for the HL7 CDA sections “observation”, “diagnosis”, “measurement” or “procedures” require a translation into the proper CDM items first. This is realised through the generated mapping file.

The following examples show generated sections:

* CRF question “Patient”:

<recordTarget>

<patientRole>

<id extension="123458abc"/>

<patient>

<name/>

<administrativeGenderCode code="248152002" codeSystem="2.16.840.1.113883.5.1"/>

<birthTime value="19980109000000"/>

</patient>

</patientRole>

</recordTarget>

* CRF question “Congenital Abnormalities” with given answer “no” leads to the following observation section:

<observation classCode="OBS" moodCode="EVN" negationInd=”true”>

              <templateId root=""/>

              <id root="364839024"/>

              <code code="66091009"

                     displayName="Congenital disease (disorder)"

                     codeSystem="2.16.840.1.113883.6.96"

                     codeSystemName="SNOMED CT"/>

              <statusCode code="COMPLETED"/>

  </observation>

* CRF question “Biopsy done?” with answer “fine needle” leads to the following procedure section:

<procedure classCode="PROC" moodCode="EVN">

<code ="86273004"

codeSystem="2.16.840.1.113883.6.96"

codeSystemName="SNOMED CT"

displayName="Biopsy (procedure)"/>

<effectiveTime value="20130219"/>

<text>Biopsy</text>

<methodCode code="48635004"

codeSystem="2.16.840.1.113883.6.96"

codeSystemName="SNOMED CT"

displayName="Fine needle biopsy"/>

</procedure>

To conclude, it can be said that most questions and answer possibilities on the CRF “F1 Registration” can be readily transferred into HL7 v3 messages based on the translation into CDM items and HL7 CDA templates. Therefore we believe that most of the SIOP data collected through the other existing CRFs of the trials can be transferred applying the same process into the EURECA environment. There still exist some issues in regard to complex expressions but those issues are expected to be solvable in the near future by applying some further analysis into e.g. the possibility of post-coordinating SNOMED-CT concepts to form complex concepts.

## GBG data

The GBG Forschungs GmbH is going to deliver data of the following trials:

- TBP (Metastatic study)

- GAIN (Adjuvant study)

- GeparQuattro (NeoAdjuvant study)

These trials are completed, their data were analysed and the results were published.

## Data source types of information

The data is available in the csv format and is about:

- Baseline data

- Adverse Events during Therapy

The data does not contain free text fields, the following terms present a part of the column headers of the csv file:

* + Thromboembolic events
  + Eye disorders
  + Anaemia, grade 3-4
  + Neutropenia, grade 3-4
  + nausea
  + vomiting
  + mucositis
  + constipation

The values in corresponding cells are:

1 = Adverse event was detected by the patient

0 = Adverse event was not detected by the patient

## Core dataset annotations

The annotation of the data in SNOMED CT is a semi-automatic process. As already mentioned, the data does not contain free text field, the annotation will be done for the columns names in the csv file.

#### Request the possible SNOMED code for a term

For this part we use the BioPortal REST-Services (search service) to request a possible SNOMED CT code for a specific term, the response is then parsed to extract the SNOMED information. At least the result list is displayed for the selection to the clinical expert.

#### SNOMED CT Code selection

The Clinical expert at the GBG identifies the SNOMED CT concept from the result list, the selection will be saved to the database.

The stored information will be used as Template for selection and for the HL7 message generation.

## HL7 v3 message template selection

The template selection for a specific term is done manually by using one of the defined templates from UPM at the EURECA wiki:

(<http://atlas.ics.forth.gr/EURECA/wiki/index.php/A_list_of_available_HL7_v3_templates_to_export_data>) or by using the tool from UPM concept2HL7.

The selected template is also stored in the database and can be used for the HL7 message generation.

## HL7 v3 message generation

The collected data in the preview steps will be exported as HL7 v2 CSV from the database.

## Connectors to other models / tools

## TranSMART (PHILIPS)

We have developed tools that extract data, which resides in a EURECA infrastructure, into the knowledge management platform tranSMART[[7]](#footnote-7). As a result, tranSMART users can use the Dataset Explorer feature of tranSMART to query extracted data (in aggregate form).

#### Representing Dataset Explorer tree

Dataset Explorer is an i2b2[[8]](#footnote-8)-based tool that represents concepts in tree form and lets one compare two sets of groups based on one or more points of comparison. We use a subset of the classes in the HL7 RIM as the basic structure (i.e. the root level nodes) in the Dataset Explorer tree since the concepts of the Core dataset are translated to those classes in the CDM. We use SNOMED-CT and LOINC to represent the child nodes concepts. We do not provide the full range of predecessor concepts but only use the concepts that are extracted from the data warehouses in combination with the root concepts, e.g. if we extract the concept *Surgical procedure* we create the path *Procedure//Surgical procedure* instead of *Procedure//Procedure by method//Surgical procedure.* A full range of predecessors increases the complexity of the Dataset Explorer significantly.

#### ETL process

The first step (Figure 25) is identifying the concepts that reside in the EURECA data warehouse. These concepts are extracted from the HL7 RIM ‘Act’ table. We also extract additional information regarding demographics (age and gender), observation values (e.g. tumour size in mm) and target site codes (e.g. location of the tumour). We built a data flow with Pentaho Kettle to extract the data from the EURECA data warehouse. These intermediate results are placed in text files such that they can be processed further. After we extract the data, we transform it such that it can be imported by tranSMART.

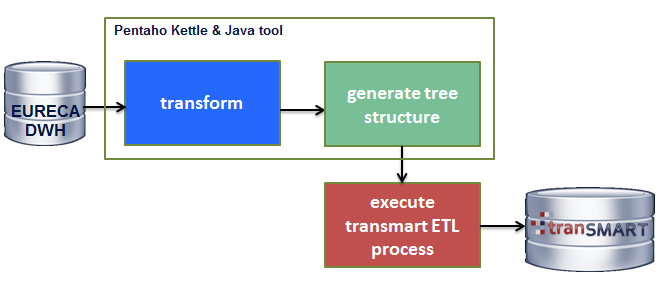


Figure 25: ETL process CDM into tranSMART

We built a Java tool that takes the intermediate results as input and generates the tree structure in a tranSMART compatible format. TranSMART has its own ETL process for importing data into their data warehouse. A disadvantage of the tranSMART ETL process is that it does not fully support generating dynamic trees. We illustrate this via the following example: Suppose we have the concepts A, B and C in our data set, and we know that A is a parent of B, and B is a parent of C. The tranSMART ETL process does not allow us to create the path *A/B/C* dynamically but only *A/B*, *B/C* or *A/C.* In our case, we need to inspect the data and create the tree based on the existing data. An alternative is not to inspect the data and load all the SNOMED-CT concepts into the Dataset Explorer. SNOMED-CT has over 300.000 concepts and it is therefore not feasible to have a full list with all the SNOMED-CT concepts, as this would significantly increase the complexity and usability of the Dataset Explorer. Also, it does not make any sense to store concepts that do not appear in the data into the tree (queries will return empty results).

The next step is passing the data, which has been transformed, to the tranSMART ETL process. We refer the reader to [3] for more information about this process.

We use the method *getRootConcept* from the *CoreDataSetService* to generate the tree structure. The method retrieves the root name of a given concept, e.g. *getRootConcept* (“surgical procedure”) returns ‘Procedure’ as root concept, which in turn will result in the path *Procedure//surgical procedure* in the Dataset Explorer tree. Finally, the data is loaded into tranSMART. After the data has been loaded, tranSMART users can use the Dataset Explorer to query extracted data, in aggregate form (Figure 23 and Figure 24).

#### Dataset Explorer screenshots

Figure 26 shows a screenshot of the concept tree of the Dataset Explorer. The number between brackets is the number of subject involved (e.g. 48 subjects are diagnosed with ‘infiltrating duct carcinoma’ and for 50 subjects the ‘age’ is present). Also, the labels ‘abc’ stand for textual values and the labels ‘123’ for integer values.

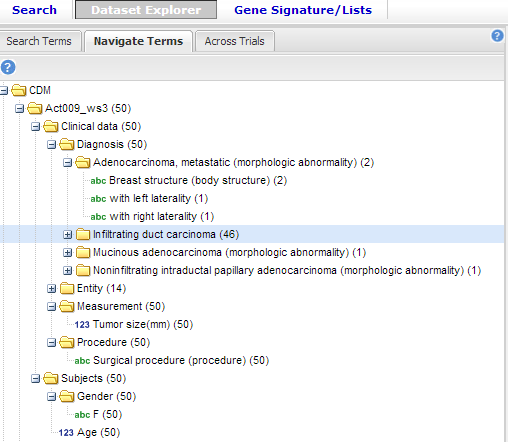


Figure 26: Concept tree of the Dataset Explorer

Figure 27 shows a comparison between two subsets. Users can drag and drop concepts from the left component into the right component. The subsets share the same criteria ‘Infiltrating duct carcinoma’ but different tumour size (< 20mm and >= 20mm) and age (age >= 50 and age < 50).

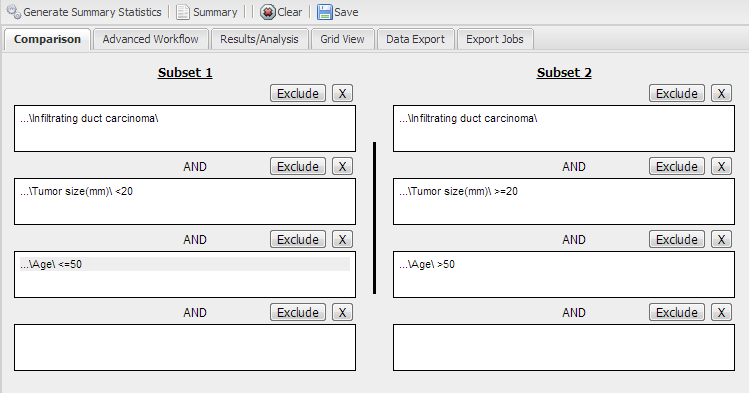


Figure 27: Comparison of two subsets in Data Explorer

#### Tools

We built a data flow to transform the data and a Java tool that generates the tree structure (Figure 23). An advantage of Pentaho Kettle is that we could extract and transform data, which resides in a EURECA data warehouse, relatively easy. When generating the tree, we were required to inspect individual concepts, merge concepts, call a web service and merge the final results. A disadvantage of Kettle is that a lot of case distinctions (e.g. inspecting individual concepts and basing different decisions on them, and merging records based on different criteria) increase the complexity of the data flow and introduce overhead in the development time. Also, implementing web services, and caching the responses require more effort in Kettle, which also result in overhead in the development time.

## Observational Medical Outcomes Partnership - OMOP (UPM)

Observational Medical Outcomes Partnership (OMOP) created a framework for observational research, which can accommodate observational data of different types (both claims and EHRs) from sources around the world. This framework consists of two components: the CDM and the Standard Vocabulary.

The purpose of the CDM is to standardize the format and content of the observational data, so standardized applications, tools and methods can be applied to them.

The CDM needs to support the conduct of research to identify and evaluate associations between interventions (drug exposure, procedures, healthcare policy changes etc.) and outcomes caused by these interventions (condition occurrences, procedures, drug exposure etc.). Outcomes can be efficacious (benefit) or adverse (risk). Often times, specific cohorts (e.g., myocardial infarction, acute liver failure) may be defined for treatments or outcomes, using clinical events (diagnoses, observations, procedures, etc.) that occur in predefined temporal relationships to each other. The CDM, combined with a method for standardizing its content (via the Vocabulary), will ensure that research methods can be systematically applied to produce meaningfully comparable results.

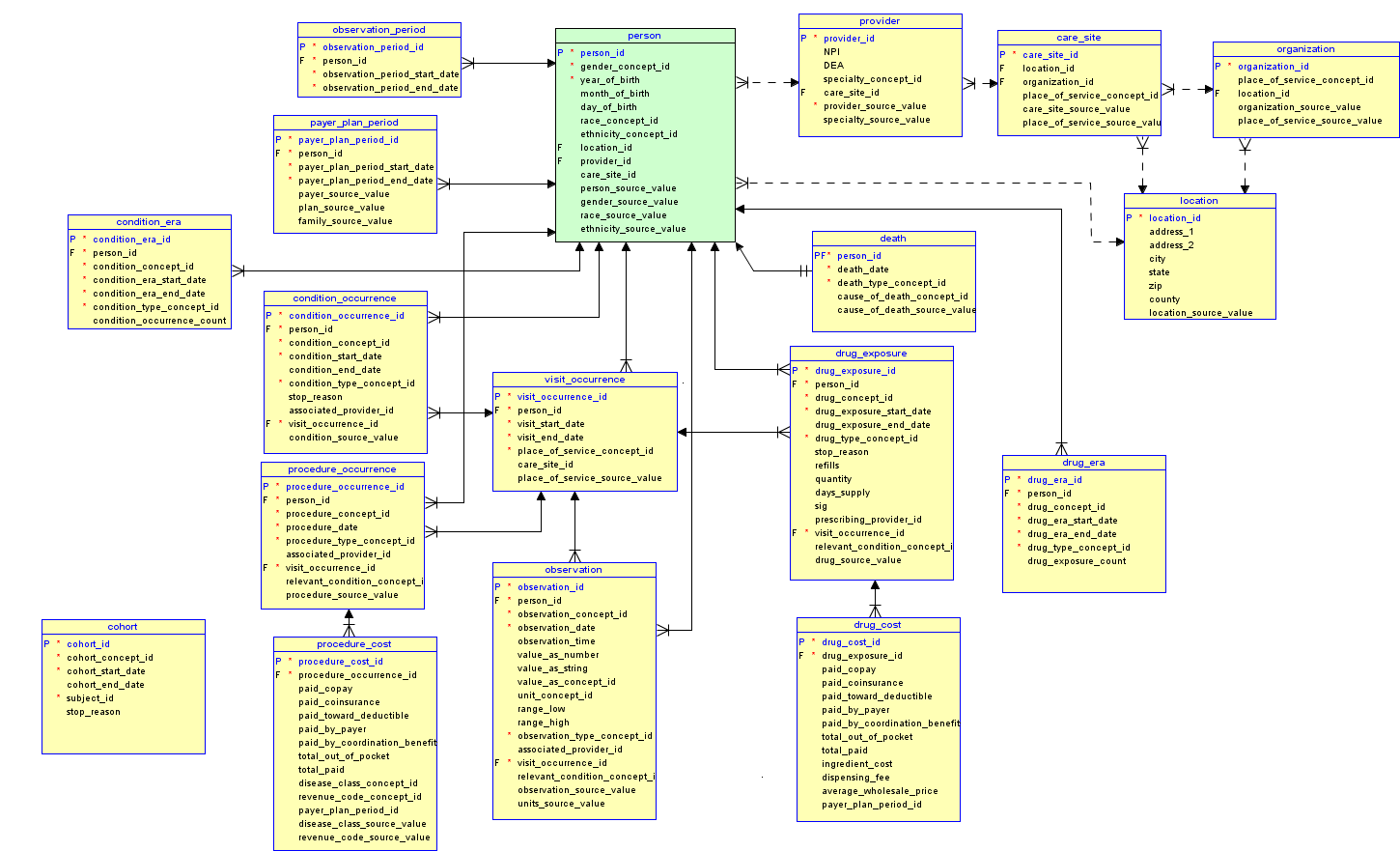


Figure 28: The entity-relationship diagram (ERD) of the CDM data tables and relationships between them

The Standard Vocabulary is a tool that enables transparent and consistent content across disparate observational databases, and serves to support in conducting efficient observational research.

The Standard Vocabulary contains all of the code sets, terminologies, vocabularies, nomenclatures, lexicons, thesauri, ontologies, taxonomies, classifications, abstractions, and other such data.

All content, such as drugs or conditions, are referred to by concepts. The Standard Vocabulary is needed to understand and make use of these concepts. The Standard Vocabulary also provides additional class concepts, relationships and ancestry relationships between concepts and a source to map concepts that is needed to convert non-standard vocabularies to Standard Vocabularies during the ETL process of external data.

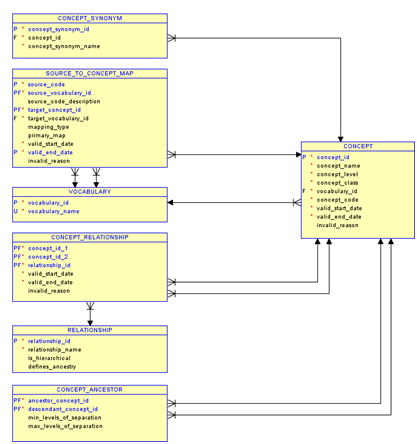


Figure 29: The entity-relationship diagram (ERD) of the Vocabulary tables and relationships between them

Some examples of SNOMED-CT and LOINC concepts stored in OMOP Standard Vocabulary are provided. They follow this structure:

CONCEPT\_ID","CONCEPT\_NAME","CONCEPT\_LEVEL","CONCEPT\_CLASS","VOC*ABULARY\_ID","CONCEPT\_CODE","VALID\_START\_DATE","VALID\_END\_DATE","INVALID\_REASON"|*

* Gender: *442986,"Female",2,"Clinical finding",1,"248152002",1970-01-01,2099-12-31*
* Procedure: *4046027,"Care of patient with infectious disease",1,"Procedure",1,"133895001",1970-01-01,2099-12-31*
* Observation: *3050116,"Catalase [Enzymatic activity/mass] in Red Blood Cells",1,"LOINC Code",6,"49221-5",2007-05-21,2099-12-31*

**Data Mapping**

The creation of a database in OMOP format requires mapping and diverse transformations of the data to meet the specifications of the CDM, and the content has to conform to the Standard Vocabularies (vocabulary mapping).

A detailed discussion about the EURECA HL7-RIM based model and specific aspects of mapping and ETL conversions of data to the standard OMOP CDM is provided. This includes a description of major tables of the CDM schema and special data handling (specific aspects of mapping and converting data) required for each table. In each section, the tables and their mapping are individually reviewed along with any source specific rules and exceptions.

## INDIVO X (FORTH)

IndivoX is a personal health platform that enables individuals to own and manage a complete, secure, digital copy of their health and wellness information. For more information on data exchange between IndivoX and EURECA CDM see deliverable 3.3 “Service for uniform access to clinical trial data and other external sources” . Here we will only focus on the initial mapping of Information in IndivoX to EURECA CDM. IndivoX stores information concerning the following categories

* Problems
* Allergies
* Medications
* Laboratory results
* Procedures

Bellow we shown the mappings of these basic categories to EURECA CDM sub-models

|  |  |
| --- | --- |
| IndivoX | EURECA CDM |
| Problems | Observations |
| Allergys | Observations |
| Medications | Substance Administration |
| Laboratory results | Observations |
| Procedures | Procedure |

Table 6: Mapping of IndivoX information categories to EURECA-CDM sub-models

However, besides mapping to the general categories of the CDM more fine-grained in needed to enable bidirectional information exchange between IndivoX and EURECA DWH.

To achieve that we exploit a set of terminologies from the Semantic Core dataset to annotate and semantically enrich these types of information. IndivoX uses SNOMED CT for problem coding, HL7 V3 for immunizations and LOINC for coding laboratory values. To be able to export the data stored in IndivoX to the EURECA DWH the proper translations had to be implemented. For the time being we present bellow an initial list of such mappings in Figure 27 and then an example SPARQL query in Figure 28. As we proceed towards the implementation of the PHR demonstrator to be presented in the second review this list will be revised and enhanced.

**Renal Symptoms**

ACT\_title: Kidney disease

ACT\_code: 90708001

ACT\_effectiveTime: 2008-11-13

ActTargetSiteCode\_title: Kidney structure

ActTargetSiteCode\_code: 64033007

**Follow up of known malformation**

ACT\_title: Follow-up examination normal

ACT\_code: 185388001

ACT\_effectiveTime: 1996-08-23

ActTargetSiteCode\_title: Developmental abnormality

ActTargetSiteCode\_code: 21390004

**Antenatal Diagnosis**

ACT\_title: Prenatal examination and care of mother

ACT\_code: 18114009

ACT\_effectiveTime: 1996-08-23

ActMethodCode\_title: Examination - action

ActMethodCode\_code: 302199004

ActTargetSiteCode\_title: Female genital structure

ActTargetSiteCode\_code: 53065001

ActTargetSiteCode\_title: Neoplasm

ActTargetSiteCode\_code: 108369006

**Sonography**

ACT\_title: Ultrasonogaphy

ACT\_code: 16310003

ACT\_effectiveTime: 1996-08-23

ActMethodCode\_title: Ultrasound imaging - action

ActMethodCode\_code: 278292003

ActTargetSiteCode\_title: Nephroblastoma

ActTargetSiteCode\_code: 25081006

**Abdominal CT Scan**

ACT\_title: CT of abdomen

ACT\_code: 169070004

ACT\_effectiveTime: 1997-08-23

ActMethodCode\_title: Computed tomography imaging - action

ActMethodCode\_code: 312251004

ActTargetSiteCode\_title: Abdominal structure

ActTargetSiteCode\_code: 113345001

ActTargetSiteCode\_title: Nephroblastoma

ActTargetSiteCode\_code: 25081006

**Abdominal MRI**

ACT\_title: MRI of abdomen

ACT\_code: 241621009

ACT\_effectiveTime: 1997-08-25

ActMethodCode\_title: Magnetic resonance imaging - action

ActMethodCode\_code: 312250003

ActTargetSiteCode\_title: Abdominal structure

ActTargetSiteCode\_code: 113345001

ActTargetSiteCode\_title: Nephroblastoma

ActTargetSiteCode\_code: 25081006

Figure 27. Example mappings of IndivoX to the EURECA CDM

Moreover, an example SPARQL query for retrieving an “Ultrasonography” is the following

SELECT DISTINCT ?id ?code ?entityId ?birthTime ?effectiveTime

WHERE

{

?instObs hl7rim:observation\_code "16310003";

hl7rim:observation\_methodCode "278292003";

hl7rim:observation\_targetSiteCode "25081006";

hl7rim:observation\_code ?code;

hl7rim:observation\_id ?id;

hl7rim:observation\_effectiveTime ?effectiveTime;

hl7rim:observation\_participation ?obs\_part;

hl7rim:observation\_targetSiteCode ?tarSiteCode;

hl7rim:observation\_targetSiteCodeTitle ?tarSiteTitle;

hl7rim:observation\_targetSiteCodeVocId ?tarSiteVocId;

hl7rim:observation\_methodCode ?methodCode;

hl7rim:observation\_methodCodeTitle ?methodTitle;

hl7rim:observation\_methodCodeVocId ?methodVocId;

hl7rim:observation\_valueST ?valueST;

hl7rim:observation\_valuePQ ?valuePQ;

hl7rim:observation\_units ?units;

hl7rim:observation\_refRangeMin ?rangeMin;

hl7rim:observation\_refRangeMax ?rangeMax;

hl7rim:observation\_title ?title;

hl7rim:observation\_codeVocId ?codeVocId.

?obs\_part hl7rim:participation\_entityId ?entityId.

?livSubj hl7rim:livingSubject\_id ?entityId;

hl7rim:livingSubject\_birthTime ?birthTime.

}

Figure 28 Example query for retrieving an Ultrasonography from EURECA DWH

## ObTiMA (Fraunhofer IBMT)

As already described in chapter 3.4 - Saarland University Hospital, we will provide mechanisms in order to export trial data from ObTiMA into the EURECA environment.

#### Workflow

The workflow of the ObTiMA export interface can be roughly abstracted into the following steps:

1. Create CRF template in ObTiMA

The patient related data of a trail are captured through specific CRFs. The trial builder of ObTiMA enables the CRF generation by several predefined question types like "input number" "input date" or questions with multiple predefined answers, so that most of these filled in parameters are available in a structured form.

1. Create CDA template(s) and mapping file (Core dataset annotations) for CRF

As described in the chapter 3.4.1 the CDA templates and the corresponding mapping has to be built manually

1. Upload CDA template(s) and mapping file in ObTiMA
2. Store CRF template + CDA template(s) + mapping file in CRF Repository of ObTiMA for reuse in later trials
3. Fill out CRF
4. Generate HL7v3 message based on CDA template
5. Export HL7v3 message to EURECA environment.

#### Future work

The manual generation of the CDA templates within a corresponding mapping file is a very cumbersome process. As ObTiMA enables the creation of ontology based CRFs, in the future we will evaluate to integrate the EURECA Core dataset into ObTiMA in order to map the entries of the CRFs automatically to the Core dataset concepts without manual steps. Furthermore, we will explore to generate the CDA template automatically from the mapping results.

# CONCLUSIONS

This report has shown the work performed by clinical partners in collaboration technical partner for integrating shared clinical data sources into the EURECA platform. This process includes the automatic annotation of data, which initially does not contain codification of concepts, using concepts from EURECA Core dataset, i.e. SNOMED-CT, LOINC and HGNC. Coded data using other terminologies has been mapped to the EURECA Core dataset with the Terminology Linking Service (TLS). The TLS is a component of the EURECA semantic layer that provides access to links between biomedical terminologies used within EURECA. Annotations obtained in this deliverable are an update of mappings presented in the previous deliverable D4.2 for some data sources such as GBG and IJB data sets.

Most clinical data sources have implemented channels for generating HL7 v3 messages based on IHE profiles. Processes developed here have achieved the integration of clinical data represented using Core dataset concepts (mostly SNOMED-CT) and using the CDM schema, based in HL7 v3 Reference Information Model. Semantics of annotations and data loaded have been validated by experts of clinical side.

Data populated in EURECA platform has been transformed using a normalization pipeline that homogenizes data representation and enhances the query capabilities on stored data. Thus, thanks to this data normalization within the CDM, different data sources can be queried by EURECA application layer through a unique and homogeneous endpoint.

Besides how to load current available clinical data sources into the EURECA CDM, this deliverable has explained how data can be exported from EURECA to other tools. Connectors to OMOP data model, IndivoX PHR, Obtima CTMS and tranSMART have been implemented and described in this document. Since the semantics of clinical partners data sources has been integrated into EURECA CDM, and coded using Core dataset, these connectors to external systems can be used on any of these data sources. Although the process is not completely automatic, this work represent a further step to facilitate the use, not only to applications developed within the EURECA project, but also to other widely adopted frameworks used in biomedical research.

# REFERENCES

1. Patricia L. Whetzel, Natalya Fridman Noy, Nigam H. Shah, Paul R. Alexander, Csongor Nyulas, Tania Tudorache, Mark A. Musen: BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Research 39(Web-Server-Issue): 541-545 (2011)
2. S. Aït-Mokhtar, B. De Bruijn, C. Hagège and P. Rupi, “EURECA Deliverable 3.2: Initial prototype for relation identification between concepts,” 2013.
3. S. Aït-Mokhtar, C. Hagège and P. Rupi, “EURECA Deliverable 3.1: Initial prototype for concept extraction out of EHR free text”, 2013.
4. Deliverable 3.3, “Service for uniform access to clinical trial data and other external sources”
5. Deliverable 4.2, “Initial proposal for the core datasets”.

# ANNEX A: OMOP data mapping

The creation of a database in OMOP format requires mapping and diverse transformations of the data to meet the specifications of the CDM, and the content has to conform to the Standard Vocabularies (vocabulary mapping).

A detailed discussion about the EURECA HL7-RIM based model and specific aspects of mapping and ETL conversions of data to the standard OMOP CDM is provided. This includes a description of major tables of the CDM schema and special data handling (specific aspects of mapping and converting data) required for each table. In each section, the tables and their mapping are individually reviewed along with any source specific rules and exceptions.

**Table Name: PERSON**

The field mapping is performed as follows:

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| PERSON\_ID | Person.id | If Person.id is not an integer value, then generate a new id.  Example: id=fictitious1  parseInt(id.substring(10,id.length)) | Foreign key inherited from ‘Entity’ and ‘LivingSubject’ tables. |
| GENDER\_CONCEPT\_ID | LivingSubject.administrativeGenderCode | System generated  New instance for ‘Concept’ table  concept\_id=442986(female)/ 442985(male)  vocabulary\_id=1 (SNOMED)  concept\_code=LivingSubject.administrativeGenderCode  if concept\_code=‘248152002’  concept\_name=’female’  if concept\_code=‘248153007’  concept\_name=’male’  concept\_class=’Clinical Finding’  concept\_level=2  valid\_start\_date= 1970-01-01  valid\_end\_date=2099-12-31 | It is not necessary to develop a mapping from SNOMED to HL7 administrative sex\*, because Gender belongs to Clinical Finding class in SNOMED, which is included in OMOP Standard Vocabulary.  \*HL7 administrative sex is included in OMOP Standard Vocabulary. |
| YEAR\_OF\_BIRTH | LivingSubject.birthTime | select datepart(yyyy, LivingSubject.birthTime) |  |
| MONTH\_OF\_BIRTH | LivingSubject.birthTime | select datepart(mm, LivingSubject.birthTime) |  |
| DAY\_OF\_BIRTH | LivingSubject.birthTime | select datepart(dd, LivingSubject.birthTime) |  |
| RACE\_CONCEPT\_ID | Person.raceCode | System Generated  New instance for ‘Concept’ table  vocabulary\_id=1 (SNOMED)  concept\_code=person.raceCode  concept\_class=’Social Context’  concept\_level=0  valid\_start\_date= 1970-01-01  valid\_end\_date=2099-12-31 | A Mapping to CDC Race vocabulary could be developed.  This fields would fill an instance for ‘source\_to\_concept\_map’ table:  source\_code=Person.raceCode  source\_vocabulary\_id=1  target\_vocabulary\_id =13  concept\_level=1  concept\_class=’Race’  valid\_start\_date= 1970-01-01  valid\_end\_date=2099-12-31 |
| ETHNICITY\_CONCEPT\_ID |  |  | Person.ethnicGroupCode in HL7-RIM.  Otherwise, use concepts from the OMOP standard vocabulary: vocabulary\_id=44 (Ethnicity) |
| LOCATION\_ID | Entity.id | Entity.id when Entity.classCode=’PLC’ | Entity could be a place. But we would have to add Place table from HL7-RIM. |
| PROVIDER\_ID | Entity.id | Entity.id when Entity.classCode=’PROV’/’SBADM’ |  |
| CARE\_SITE\_ID | Entity.id | Entity.id when Entity.classCode=’ORG’ |  |
| PERSON\_SOURCE\_VALUE | Entity.code |  | Reference to the code in the original vocabulary.  (SNOMED-CT code for Homo Sapiens: 337915000) |
| GENDER\_SOURCE\_VALUE | LivingSubject.administrativeGenderCode |  |  |
| RACE\_SOURCE\_VALUE | Person.raceCode |  |  |
| ETHNICITY\_SOURCE\_VALUE |  |  | Not Available (any mapping is made for the Ethnicity) |

**Table Name: DRUG\_EXPOSURE**

The field mapping is performed as follows:

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| DRUG\_EXPOSURE\_ID | Exposure.id (Act.id) | Act.classCode=’EXPOS’ or Act.classCode=’SBADM’  If Exposure.id is not an integer value, then generate a new id. | This entity will be a mixture between Exposure and SubstanceAdministration tables. Both inherit their key from Act entity. |
| PERSON\_ID | Person.id |  |  |
| DRUG\_CONCEPT\_ID | Entity.code  Entity.classCode  Entity.name  (Sometimes the information of the drug is stored in the act table when classCode=’SBADM’) | System Generated  New instance for ‘Concept’ table:  vocabulary\_id=1 (SNOMED)  concept\_code=Entity.code  concept\_class=’Substance’  concept\_level=0  concept\_name=Entity.name  valid\_start\_date= 1970-01-01  valid\_end\_date=2099-12-31 | Drug could be an instance of Material (Subclass of Entity in HL7-RIM model).    A Mapping to RxNorm vocabulary could be developed.  This fields would fill an instance for ‘source\_to\_concept\_map’ table:  source\_code=Entity.code  source\_vocabulary\_id=1  target\_vocabulary\_id =8  concept\_level=0/1/2 |
| DRUG\_EXPOSURE\_START\_DATE | Act.activityTime |  |  |
| DRUG\_EXPOSURE\_END\_DATE | Act.activityTime |  |  |
| DRUG\_TYPE\_CONCEPT\_ID |  |  | Use concepts from  OMOP Drug Exposure Type vocabulary.  vocabulary\_id=36  concept\_level=1 |
| STOP\_REASON |  |  | Not Available |
| REFILLS | SubstanceAdministration.doseCheckQuantity |  |  |
| QUANTITY | SubstanceAdministration.doseQuantity |  |  |
| DAYS\_SUPPLY |  | Days\_supply= select datepart(dd,end\_date) - select datepart(dd, start\_date)  If Days\_supply <0  Days\_supply=Days\_supply+(31 or 30) | Drug Prescription |
| SIG | SubstanceAdministration.routeCode |  | The directions on the drug prescription |
| PRESCRIBING\_PROVIDER\_ID | Person.id |  |  |
| VISIT\_OCCURRENCE\_ID |  | System generated key for Visit\_Occurrence entity. | Participation (Visit) involves an Entity (Person) and a role (Patient). |
| RELEVANT\_CONDITION\_CONCEPT\_ID | Entity.code  Entity.name  Entity.creationTime | System generated  New instance for ‘Concept’ table  vocabulary\_id=1 (SNOMED)  concept\_code=Entity.code  concept\_class=’Clinical Finding’  concept\_level=1/2/3  concept\_name=Entity.name  valid\_start\_date= Entity.creationTime  valid\_end\_date=2099-12-31 |  |
| DRUG\_SOURCE\_VALUE | Entity.codeOrig or Entity.code |  |  |

**Table Name: CONDITION\_OCCURRENCE**

The field mapping is performed as follows:

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| CONDITION\_OCCURRENCE\_ID |  | System Generated |  |
| PERSON\_ID | Person.id |  |  |
| CONDITION\_CONCEPT\_ID | Entity.code  Entity.name  Entity.creationTime | System Generated  New instance for ‘Concept’ table  vocabulary\_id=1 (SNOMED)  concept\_code=Entity.code  concept\_class=’Clinical Finding’  concept\_level=1/2/3  concept\_name=Entity.name  valid\_start\_date= Entity.creationTime  valid\_end\_date=2099-12-31 |  |
| CONDITION\_START\_DATE |  |  | Not Available. |
| CONDITION\_END\_DATE |  |  | Not Available. |
| CONDITION\_TYPE\_CONCEPT\_ID |  |  | Use concepts from OMOP Condition Ocurrence Type vocabulary.  vocabulary\_id=37  concept\_level=1 |
| STOP\_REASON |  |  | Not Available. |
| ASSOCIATED\_PROVIDER\_ID | Person.id |  |  |
| VISIT\_OCCURRENCE\_ID |  | System Generated |  |
| CONDITION\_SOURCE\_VALUE | Entity.codeOrig |  |  |

**Table Name: VISIT\_OCCURRENCE**

The field mapping is performed as follows:

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| VISIT\_OCCURRENCE\_ID | Participation.typeCode  Participation.entityId  Participation.roleId  Entity.classCode  Role.classCode | System Generated  Role that matches with Participation.roleId should have the value ‘PAT’ for its classCode attribute.  Entity.classCode=’PSN’ (Entity.id=Participation.entityId)  Participation.typeCode=’PART’ |  |
| PERSON\_ID | Participation.entityId |  |  |
| VISIT\_START\_DATE | Participation.timeStart |  |  |
| VISIT\_END\_DATE | Participation.timeEnd |  |  |
| PLACE\_OF\_SERVICE\_CONCEPT\_ID | Entity.code  Entity.classCode  Entity.name | Entity.classCode should be ’PLC’  New instance for ‘Concept’ table  vocabulary\_id=1 (SNOMED)  concept\_code=Entity.code  concept\_class=’Environment or geographical location’  concept\_name=Entity.name  concept\_level=0  valid\_start\_date= 1970-01-01  valid\_end\_date=2099-12-31 | A Mapping to CMS Place of Service vocabulary could be developed.  This fields would fill an instance for ‘source\_to\_concept\_map’ table:  source\_code=Entity.code  source\_vocabulary\_id=1  target\_vocabulary\_id =14 |
| CARE\_SITE\_ID | Entity.id | Entity.id when Entity.classCode=’ORG’ |  |
| PLACE\_OF\_SERVICE\_SOURCE\_VALUE | Entity.code or Entity. codeOrig |  | Use the code if a mapping to CMS Place of Service vocabulary is done, or the original code if it is not. |

**Table Name: PROCEDURE\_OCCURRENCE**

The field mapping is performed as follows:

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| PROCEDURE\_OCCURRENCE\_ID | Procedures.id | If Procedures.id is not an integer value, then generate a new id. |  |
| PERSON\_ID | Person.id |  |  |
| PROCEDURE\_CONCEPT\_ID | Act.code  Act.name  Act.creationTime  Act.classCode | System Generated  Act.classCode should be ’PROC’  New instance for ‘Concept’ table  vocabulary\_id=1 (SNOMED)  concept\_code=Act.code  concept\_class=’Procedure’  concept\_name=Act.title  concept\_level=1/2/3  valid\_start\_date= Act.creationTime  valid\_end\_date=2099-12-31 |  |
| PROCEDURE\_DATE | Act.effectiveTime |  |  |
| PROCEDURE\_TYPE\_CONCEPT\_ID |  |  | Use concepts from OMOP Procedure Occurrence Type vocabulary.  vocabulary\_id=38  concept\_level=1 |
| ASSOCIATED\_PROVIDER\_ID | Person.id |  |  |
| VISIT\_OCCURRENCE\_ID |  | System Generated |  |
| RELEVANT\_CONDITION\_CONCEPT\_ID |  |  | Not Available. |
| PROCEDURE\_SOURCE\_VALUE | Act.codeOrig or Procedures.id |  |  |

**Table Name: OBSERVATION**

The field mapping is performed as follows:

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| OBSERVATION\_ID | Observation.id | If Observation.id is not an integer value, then generate a new id. |  |
| PERSON\_ID | Participation.entityId | Where Participation.actId matches with this Observation.id && the entity with that id has entity.classCode== 'PSN' |  |
| OBSERVATION\_CONCEPT\_ID | Act.code  Act.codeVocId  Act.classCode | System Generated  Act.classCode should be’OBS’  New instance for ‘Concept’ table  vocabulary\_id=Act.codeVocId  if (Act.codeVocId==2.16.840.1.113883.6.96) vocabulary\_id=1 (SNOMED)  if (Act.codeVocId==2.16.840.1.113883.6.1) vocabulary\_id=6 (LOINC)  concept\_code=Act.code  concept\_class=’Clinical Finding’(SNOMED) / ’LOINC Code’(LOINC)  concept\_name=Act.title  concept\_level=1/2/3(SNOMED) / 1(LOINC)  valid\_start\_date= Act.creationTime  valid\_end\_date=2099-12-31 |  |
| OBSERVATION\_DATE | Act.effectiveTime |  |  |
| OBSERVATION\_TIME | Act.activityTime |  |  |
| VALUE\_AS\_NUMBER | ActObservationValues.valuePQ |  |  |
| VALUE\_AS\_STRING | ActObservationValues.valueST |  |  |
| VALUE\_AS\_CONCEPT\_ID |  | System generated | Add ActObservationValues.code to the model (LOINC/SNOMED codes) |
| UNIT\_CONCEPT\_ID | ActObservationValues.units | System generated | Concepts from UCUM vocabulary  vocabulary\_id=11  concept\_name= ActObservationValues.units  Normalize the units with this vocabulary. |
| RANGE\_LOW | ActObservationValues.referenceRangeMin |  |  |
| RANGE\_HIGH | ActObservationValues.referenceRangeMax |  |  |
| OBSERVATION\_TYPE\_CONCEPT\_ID |  |  | Use concepts from OMOP Observation Type vocabulary.  vocabulary\_id=39  concept\_level=1 |
| ASSOCIATED\_PROVIDER\_ID | Person.id |  |  |
| VISIT\_OCCURRENCE\_ID |  | Select the visit whose entity is included in a participation with the actId of this observation |  |
| RELEVANT\_CONDITION\_CONCEPT\_ID |  |  | Not Available. |
| OBSERVATION\_SOURCE\_VALUE | Act.codeOrig or Observation.id |  |  |
| UNITS\_SOURCE\_VALUE | ActObservationValues.units |  |  |

**Table Name: OBSERVATION\_PERIOD**

The field mapping is as follows:

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| OBSERVATION\_PERIOD\_ID |  | System generated |  |
| PERSON\_ID | Person.id |  |  |
| OBSERVATION\_PERIOD\_START\_DATE | Participation.roleId  Participation.classCode  Entity.id  Participation.entityId  Act.id (Observation.id)  Participation.actId  Entity.classCode  Act.classCode  Participation.actId  Participation.entityId  Act.creationDate | Role that matches with Participation.roleId should have the value ‘PAT’ for his classCode attribute.  Entity.classCode=’PSN’ (Entity.id=Participation.entityId)  Act.classCode=’OBS’  (Act.id=Participation.actId)  Check all the Observations (acts\*) for this Person (entity\*) and select the lower date (Act.creationDate).  \*The relation between entities and acts is stored in Participation table. | Date of the first Observation instance of the same Person. |
| OBSERVATION\_PERIOD\_END\_DATE | Participation.roleId  Participation.classCode  Entity.id  Participation.entityId  Act.id (Observation.id)  Participation.actId  Entity.classCode  Act.classCode  Participation.actId  Participation.entityId  Act.creationDate | Role that matches with Participation.roleId should have the value ‘PAT’ for his classCode attribute.  Entity.classCode=’PSN’ (Entity.id=Participation.entityId)  Act.classCode=’OBS’  (Act.id=Participation.actId)  Check all the Observations (acts\*) for this Person (entity\*) and select the higher date (Act.creationDate).  \*The relation between entities and acts is stored in Participation table. | Date of the last Observation instance of the same Person. |

**Table Name: DRUG\_COST**

The field mapping is as follows:

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| DRUG\_COST\_ID |  | System generated | Cost of a Drug Exposure |
| DRUG\_EXPOSURE\_ID | Act.id (Exposure.id) | Act.classCode should have ’EXPOS’ value | Foreign Key from Act |
| PAID\_COPAY |  |  | Not Available |
| PAID\_COINSURANCE |  |  | Not Available |
| PAID\_TOWARD\_DEDUCTIBLE |  |  | Not Available |
| PAID\_BY\_PAYER |  |  | Not Available |
| PAID\_BY\_COORDINATION\_BENEFITS |  |  | Not Available |
| TOTAL\_OUT\_OF\_POCKET |  |  | Not Available |
| TOTAL\_PAID |  |  | Not Available |
| INGREDIENT\_COST |  |  | Not Available |
| DISPENSING\_FEE |  |  | Not Available |
| AVERAGE\_WHOLESALE\_PRICE |  |  | Not Available |
| PAYER\_PLAN\_PERIOD\_ID |  |  | Not Available |

**Table Name: PROCEDURE\_COST**

The field mapping is as follows:

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| PROCEDURE\_COST\_ID |  | System generated |  |
| PROCEDURE\_OCCURRENCE\_ID | Procedures.id |  |  |
| PAID\_COPAY |  |  | Not Available |
| PAID\_COINSURANCE |  |  | Not Available |
| PAID\_TOWARD\_DEDUCTIBLE |  |  | Not Available |
| PAID\_BY\_PAYER |  |  | Not Available |
| PAID\_BY\_COORDINATION\_BENEFITS |  |  | Not Available |
| TOTAL\_OUT\_OF\_POCKET |  |  | Not Available |
| TOTAL\_PAID |  |  | Not Available |
| DISEASE\_CLASS\_CONCEPT\_ID |  |  |  |
| REVENUE\_CODE\_CONCEPT\_ID |  |  | Not Available |
| PAYER\_PLAN\_PERIOD\_ID |  |  | Not Available |
| DISEASE\_CLASS\_SOURCE\_VALUE |  |  |  |
| REVENUE\_CODE\_SOURCE\_VALUE |  |  | Not Available |

**Table Name: PAYER\_PLAN\_PERIOD**

The field mapping is as follows:

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| PAYER\_PLAN\_PERIOD\_ID |  | System generated |  |
| PERSON\_ID | Person.id |  |  |
| PAYER\_PLAN\_PERIOD\_START\_DATE |  |  | Not Available |
| PAYER\_PLAN\_PERIOD\_END\_DATE |  |  | Not Available |
| PAYER\_SOURCE\_VALUE |  |  | Not Available |
| PLAN\_SOURCE\_VALUE |  |  | Not Available |
| FAMILY\_SOURCE\_VALUE |  |  | Not Available |

**Table Name: PROVIDER**

The field mapping is as follows:

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| PROVIDER\_ID | Person.id | If Person.id is not an integer value, then generate a new id. |  |
| NPI |  |  | Not Available |
| DEA |  |  | Not Available |
| SPECIALTY\_CONCEPT\_ID |  |  | Not Available. |
| CARE\_SITE\_ID | Entity.id | Entity.id when Entity.classCode=’ORG’ |  |
| PROVIDER\_SOURCE\_VALUE | Entity.code |  |  |
| SPECIALTY\_SOURCE\_VALUE |  |  | Not Available. |

**Table Name: LOCATION**

The field mapping is as follows:

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| LOCATION\_ID | Entity.id | Entity.id when Entity.classCode=’PLC’  If Entity.id is not an integer value, then generate a new id. |  |
| ADDRESS\_1 |  |  | place.addr could be added from HL7-RIM model |
| ADDRESS\_2 |  |  | place.addr could be added from HL7-RIM model |
| CITY |  |  | Not Available |
| STATE |  |  | Not Available |
| ZIP |  |  | Not Available |
| COUNTY |  |  | Not Available |
| LOCATION\_SOURCE\_VALUE | Entity.code |  |  |

**Table Name: ORGANIZATION**

The field mapping is as follows:

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| ORGANIZATION\_ID | Entity.id | Entity.id when Entity.classCode=’ORG’  If Entity.id is not an integer value, then generate a new id. |  |
| PLACE\_OF\_SERVICE\_CONCEPT\_ID | Entity.code  Entity.classCode  Entity.name | New instance for ‘Concept’ table  vocabulary\_id=1 (SNOMED)  concept\_code=Entity.code  concept\_class=’Environment or geographical location’  concept\_name=Entity.name  concept\_level=0  valid\_start\_date= 1970-01-01  valid\_end\_date=2099-12-31 | A Mapping to CMS Place of Service vocabulary could be developed.  This fields would fill an instance for ‘source\_to\_concept\_map’ table:  source\_code=Entity.code  source\_vocabulary\_id=1  target\_vocabulary\_id =14  concept\_level=1 |
| LOCATION\_ID | Entity.id | Entity.id when Entity.classCode=’PLC’ |  |
| ORGANIZATION\_SOURCE\_VALUE | Entity.codeOrig or Entity.code |  |  |
| PLACE\_OF\_SERVICE\_SOURCE\_VALUE | Entity.codeOrig or Entity.code |  |  |

**Table Name: CARE\_SITE**

The field mapping is as follows:

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| CARE\_SITE\_ID | Entity.id | Entity.id when Entity.classCode=’ORG’  If Entity.id is not an integer value, then generate a new id. |  |
| LOCATION\_ID | Entity.id | Entity.id when Entity.classCode=’PLC’ |  |
| ORGANIZATION\_ID | Entity.id | Entity.id when Entity.classCode=’ORG’ |  |
| PLACE\_OF\_SERVICE\_CONCEPT\_ID | Entity.code  Entity.classCode  Entity.name | System generated  Entity.classCode=’PLC’  New instance for ‘Concept’ table  vocabulary\_id=1 (SNOMED)  concept\_code=Entity.code  concept\_class=’Environment or geographical location’  concept\_name=Entity.name  concept\_level=0  valid\_start\_date= 1970-01-01  valid\_end\_date=2099-12-31 | A Mapping to CMS Place of Service vocabulary could be developed.  This fields would fill an instance for ‘source\_to\_concept\_map’ table:  source\_code=Entity.code  source\_vocabulary\_id=1  target\_vocabulary\_id =14  concept\_level=1 |
| CARE\_SITE\_SOURCE\_VALUE | Entity.codeOrig or Entity.code |  |  |
| PLACE\_OF\_SERVICE\_SOURCE\_VALUE | Entity.codeOrig or Entity.code |  |  |

**Table Name: DEATH**

The field mapping is as follows:

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| PERSON\_ID | Person.id |  |  |
| DEATH\_DATE | LivingSubject.deceasedTime |  |  |
| DEATH\_TYPE\_CONCEPT\_ID |  |  | There is not any death information in HL7-RIM based models.  Use concepts from OMOP Death Type vocabulary.  vocabulary\_id=45  concept\_level=1 |
| CAUSE\_OF\_DEATH\_CONCEPT\_ID |  |  | Not Available |
| CAUSE\_OF\_DEATH\_SOURCE\_VALUE |  |  | Not Available |

**Table Name: COHORT**

The field mapping is as follows:

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| COHORT\_ID |  |  | Not Available. |
| COHORT\_CONCEPT\_ID |  |  | Not Available. |
| COHORT\_START\_DATE |  |  | Not Available. |
| COHORT\_END\_DATE |  |  | Not Available. |
| SUBJECT\_ID |  |  | Not Available. |
| STOP\_REASON |  |  | Not Available. |

**Added tables**

New tables completing OMOP CDM in order to store the biggest amount of information as possible extracted from the HL7 based model.

Added table: GENE\_RELATIONSHIP

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| Gene\_Relationship\_id |  | System generated |  |
| Gene\_id | Entity.id | Entity.classCode=’MAT’ |  |
| Observation\_id | Observation.id | Participation.entityId=Gene\_id  Participation.actCode=Observation\_id  Participation.roleId  Role.entityId=Gene\_id  Role.classCode should be ’SPEC’ |  |

Added table: GENE

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| Gene\_id | Entity.id | Entity.classCode should be ’MAT’ |  |
| Gene\_Concept\_id | Entity.code  Entity.codeVocId  Entity.name | concept\_code=Entity.code  if (Entity.codeVocId== 2.16.840.1.113883.6.281)  vocabulary =HGNC (needs to be introduced in OMOP Standard Vocabulary)  concept\_level=0  concept\_name=Entity.name  valid\_start\_date= 1970-01-01  valid\_end\_date=2099-12-31 | There is not any vocabulary in OMOP standard related to genes in order to make a mapping from HGNC. |

Added table: ACT\_RELATIONSHIP

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| idA | ActRelationship.idA |  | Observation/Drug\_Exposure/Procedure\_Occurrence instance |
| idB | ActRelationship.idB |  | Observation/Drug\_Exposure/Procedure\_Occurrence instance |
| typeA | ActRelationship.idA  Act.classCode | If (Act.classCode==’PROC’) typeA=’Procedure\_Occurrence’  If(Act.classCode==’OBS’)  typeA=’Observation’  if(Act.classCode==’EXPOS’ || ‘SBADM’)  typeA=’Drug\_Exposure’ |  |
| typeB | ActRelationship.idB  Act.classCode | If (Act.classCode==’PROC’) typeB=’Procedure\_Occurrence’  If(Act.classCode==’OBS’)  typeB=’Observation’  if(Act.classCode==’EXPOS’ || ‘SBADM’)  typeB=’Drug\_Exposure’ |  |
| typeCode | ActRelationship.typeCode |  |  |

Added table: TARGET\_SITE

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| Act\_Id | ActTargetSiteCode.id |  | Primary foreign key with the id of the Observation/Procedure\_Occurrence it belongs to. |
| Act\_class | Act.classCode |  | Text field indication the table this act belongs to. |
| Target\_Site\_Concept\_Id | ActTargetSiteCode.code  ActTargetSiteCode.codeVocId  ActTargetSiteCode.title | System generated  New instance for ‘Concept’ table  vocabulary\_id=1 (SNOMED)  concept\_code= ActTargetSiteCode.code  concept\_class=’Attribute’/’Qualifier value’/’Body structure’  concept\_name=Act.title  concept\_level=0  valid\_start\_date= 1970-01-01  valid\_end\_date=2099-12-31 |  |

Added table: INTERPRETATION\_CODE

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| Act\_Id | ActObservationInterpretationCode.id |  | Primary foreign key with the id of the Observation it belongs to. |
| Intepretation\_Concept\_Id | ActObservationInterpretationCode.code | System generated  New instance for ‘Concept’ table  vocabulary\_id=1 (SNOMED)  concept\_code=ActObservationInterpretationCode.code  concept\_class=’ Qualifier value’  if(ActObservationInterpretationCode.code== 62482003)  concept\_name=Low  if(ActObservationInterpretationCode.code== 85696000)  concept\_name=Negative  if(ActObservationInterpretationCode.code== 394424008)  concept\_name=Positive  if(ActObservationInterpretationCode.code== 261665006)  concept\_name=Unknown  concept\_level=0  valid\_start\_date= 1970-01-01  valid\_end\_date=2099-12-31 |  |

**Source Independent Data Mapping**

Description of the mapping process for Drug and Condition Era’s.

**Table Name: DRUG\_ERA**

All Drug Eras are recorded in the DRUG\_ERA table based on the following field mapping:

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| DRUG\_ERA\_ID |  | System generated. |  |
| PERSON\_ID | Person.id |  |  |
| DRUG\_CONCEPT\_ID | Entity.code  Entity.name  Entity.classCode | System Generated  New instance for ‘Concept’ table:  vocabulary\_id=1 (SNOMED)  concept\_code=Entity.code  concept\_class=’Substance’  concept\_level=0  concept\_name=Entity.name  valid\_start\_date= 1970-01-01  valid\_end\_date=2099-12-31 | Drug could be an instance of Material (Subclass of Entity in HL7-RIM model).    A Mapping to RxNorm vocabulary could be developed.  This fields would fill an instance for ‘source\_to\_concept\_map’ table:  source\_code=Entity.code  source\_vocabulary\_id=1  target\_vocabulary\_id =8  concept\_level=0/1/2 |
| DRUG\_ERA\_START\_DATE | Participation.roleId  Participation.classCode  Entity.id  Participation.entityId  Act.id (Exposure.id)  Participation.actId  Entity.classCode  Act.classCode  Participation.actId  Participation.entityId  Act.creationDate | Role that matches with Participation.roleId should have the value ‘PAT’ for his classCode attribute.  Entity.classCode=’PSN’ (Entity.id=Participation.entityId)  Act.classCode=’EXPOS’  (Act.id=Participation.actId)  Check all the Exposures (acts\*) for this Person (entity\*) and select the lower date (Act.creationDate).  \*The relation between entities and acts is stored in Participation table. |  |
| DRUG\_ERA\_END\_DATE | Participation.roleId    Participation.classCode  Entity.id  Participation.entityId  Act.id (Exposure.id)  Participation.actId  Entity.classCode  Act.classCode  Participation.actId  Participation.entityId  Act.creationDate | Role that matches with Participation.roleId should have the value ‘PAT’ for his classCode attribute.  Entity.classCode=’PSN’ (Entity.id=Participation.entityId)  Act.classCode=’EXPOS’  (Act.id=Participation.actId)  Check all the Exposures (acts\*) for this Person (entity\*) and select the higher date (Act.creationDate).  \*The relation between entities and acts is stored in Participation table. |  |
| DRUG\_TYPE\_CONCEPT\_ID |  |  | Use concepts from  OMOP Drug Exposure Type vocabulary.  vocabulary\_id=36  concept\_level=1 |
| DRUG\_EXPOSURE\_COUNT |  |  | Count the number of Drug\_Exposures related with this Person.id and this Drug\_Concept\_id. |

**Table Name: CONDITION\_ERA**

Condition Era table is constructed through an aggregation of individual Condition Occurrences recorded in the CONDITION\_OCCURRENCE table.

All Condition Eras are recorded in the CONDITION\_ERA table based on the following field mapping:

| **Destination Field** | **Source Field** | **Applied Rule** | **Comment** |
| --- | --- | --- | --- |
| CONDITION\_ERA\_ID |  | System generated |  |
| PERSON\_ID | Person.id |  |  |
| CONDITION\_CONCEPT\_ID | Entity.code  Entity.name  Entity.creationTime | System generated  New instance for ‘Concept’ table  vocabulary\_id=1 (SNOMED)  concept\_code=Entity.code  concept\_class=’Clinical Finding’  concept\_level=1/2/3  concept\_name=Entity.name  valid\_start\_date= Entity.creationTime  valid\_end\_date=2099-12-31 |  |
| CONDITION\_ERA\_START\_DATE |  |  | Not Available. |
| CONDITION\_ERA\_END\_DATE |  |  | Not Available. |
| CONDITION\_TYPE\_CONCEPT\_ID |  |  | Use concepts from OMOP Condition Ocurrence Type vocabulary.  vocabulary\_id=37  concept\_level=1 |
| CONDITION\_OCCURRENCE\_COUNT |  |  | Count the number of Condition\_Ocurrence related with this Person.id and this Condition\_Concept\_id. |

1. http://data.bioontology.org/documentation [↑](#footnote-ref-1)
2. http://sparql.bioontology.org/ [↑](#footnote-ref-2)
3. For an explicit list of criterion fact classes and their respective properties, please refer to section 3 of deliverable D3.2 [2]. [↑](#footnote-ref-3)
4. http://kandel.dia.fi.upm.es:8078 [↑](#footnote-ref-4)
5. EURECA, C. Krykwinski, D. Van Vyve, N. Graf, A. Dekker, K. Mehta, F. Buffa, N. Forgó, M. Goralczyk, B. Lodzig, H. Kondylakis, L. Koumakis, R. Siebes, A. Kaptein, and K. De Schepper, “Delivrable: D9.1 Report on the development environment and on the available test data,” 2013. [↑](#footnote-ref-5)
6. http://kandel.dia.fi.upm.es:8078 [↑](#footnote-ref-6)
7. <http://transmartfoundation.org/> [↑](#footnote-ref-7)
8. https://www.i2b2.org/ [↑](#footnote-ref-8)