DELIVERABLE

Project Acronym: **ASSESS CT**

Grant Agreement number: **643818**

Project Title: **Assessing SNOMED CT for Large Scale eHealth Deployments in the EU**

D2.2 – Use of terminologies for representing structured and unstructured clinical content – interim report

Authors:

Stefan Schulz	Medical University of Graz (Austria)
Jose Antonio Miñarro-Giménez	Medical University of Graz (Austria)
Daniel Karlsson	University of Linköping
Kirstine Rosenbeck Gøeg	University of Ålborg
Kornél Markó	Averbis GmbH

Project co-funded by the European Commission within **H2020-PHC-2014-2015/H2020_PHC-2014-single-stage**		
Dissemination Level		
PU	Public	
PP	Restricted to other programme participants (including the Commission Services	
RE	Restricted to a group specified by the consortium (including the Commission Services	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Revision History, Status, Abstract, Keywords, Statement of Originality

Revision History

Revision	Date	Author	Organisation	Description
1	25 Feb	Stefan	MUG	First Draft
2	27 Feb	Stefan	MUG	Final Draft
3	2016-02-28	Daniel	LIU	Review + Task 2.5 additions
4	2016-02-28	Kirstine Rosenbeck Gøeg	AAL	Review + additions + re-organisation
5	2016-02-29	Daniel Karlsson	LIU	Pre-pre-final
6	2016-02-29	Stefan Schulz	MUG	Pre-Final
7	2016-03-08	Dipak Kalra	Eurorec	Final Review
8	2016-03-08	Empirica	Empirica	Editorial revision
9	2016-03-10	Stefan Schulz	MUG	Final version

Date of delivery	Contractual:	29.02.2016	Actual:	10.03.2016
Status	final ☒ /draft ☐			

Abstract (for dissemination)	In a variety of experiments using terminology settings (SNOMED CT against alternative, hybrid terminology), aspects of terminology use have been studied, and results have been collected and analysed. Most of this is provided by the interim deliverable D2.1. Deliverable 2.2 describes the methodology of applying Natural Language Processing (NLP) techniques for automatic annotation of clinical texts, a qualitative analysis of annotation disagreement, and a comparative study of the results from the terminology binding (structured) and free-text annotation (unstructured) experiments. The analyses of this deliverable are still in early phases of development and a richer analysis will be provided in the final deliverable (D2.4).
Keywords	SNOMED CT, UMLS, Terminology coverage and quality, Electronic Health Records, Semantic Interoperability, eHealth in Europe

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of Content

1 Executive Summary

Workpackage 2 of ASSESS CT targets empirical evidence for the fitness for purpose of SNOMED CT, compared to other terminology scenarios. In a variety of experiments using terminology settings reflecting the ASSESS CT scenarios, aspects of terminology use have been studied, and results have been collected and analysed. Results related to terminology coverage and inter-annotator agreement have been reported in the ASSESS CT interim deliverable D2.1. This deliverable focuses on preliminary results from an experiment applying Natural Language Processing (NLP) techniques for automatic annotation of clinical free-text documents, a qualitative analysis of annotation disagreement, and a comparative study of the results from the terminology binding (structured) and free-text annotation (unstructured) experiments.

For the NLP experiment, as a basis for gathering evidence, a SNOMED CT-only setting was compared to a hybrid terminology, based on a subset of the Unified Medical Language System (UMLS) meta-thesaurus. Results from this experiment are preliminary.

Both the annotations made in the free-text annotation and the terminology binding study have been studied qualitatively to identify categories of terminology use and specifically types of disagreement. Similar types of disagreement occur in both free-text and binding experiments, most likely due to features of the terminologies used.

The analyses of this deliverable are still in early phases of development and a richer analysis will be provided in the final deliverable.

2 Introduction

2.1 About this document

This document is an interim deliverable (D2.2) of the ASSESS CT project. As such, it reflects ongoing work and preliminary results as of February 2016. Entitled "Use of terminologies for representing structured and unstructured clinical content", it is based on content of deliverable D2.1 "Multilingual and multidisciplinary study of terminology coverage and quality". As both topics are closely intertwined, D2.2 should be read as an extension of D2.1, but for reporting purposes both documents need to be separate which implies redundancies in the text. Text passages that repeat or summarize content from D2.1 are therefore picked out in *italics*. In comparison with D2.1, this document reports more incomplete results and should be seen as a snapshot of work in progress.

First, we will present preliminary findings of a Natural Language Processing (NLP) experiment. Here we broaden the view on representation, because we show an alternative to manual representation including an overview of the NLP algorithm and its performance.

Secondly, we will present a first comparative analysis of all the representation experiments (free text, terminology binding and NLP) highlighting differences in concept coverage, term coverage and inter-annotator agreement.

Thirdly, we will analyse several cases of inter-annotator disagreements. The type of disagreements indicate what measures could be taken to improve consistency of selected concepts. The relative ease or difficulty to improve consistency is an important factor when choosing one terminology over another for the representation of clinical content.

This document will be completed as the final deliverable 2.4, with a complete record of project results and a more in-depth discussion. The final deliverable will be submitted with extended annexes, including manuscripts to be submitted to scientific journals.

2.2 Goals, and objectives of ASSESS CT

The goal of ASSESS CT is to improve semantic interoperability of eHealth services in Europe by investigating the fitness of the international clinical terminology SNOMED CT, the world's most comprehensive multilingual healthcare terminology, as a potential standard for EU-wide medical documentation. SNOMED CT claims to provide codes for comprehensively representing the content of health records.

As health care systems are organized nationally, the EU has not taken any steps so far towards the adoption of a standardized health terminology, and up to now, SNOMED CT has been introduced in only part of the EU member states. However, as the mobility of EU citizens is increasing and national boundaries are loosen for a more internationalized market for health care services, the question of interoperability of health care data gains importance at a European level. The ASSESS CT consortium is addressing this challenge by investigating the current use of SNOMED CT, analysing reasons for adoption / non-adoption, and identifying success factors, strengths and weaknesses related to SNOMED CT and to alternative terminologies.

The adoption of SNOMED CT is scrutinized against two alternative scenarios, viz. (i) to abstain from actions at the EU level, and (ii) to devise an EU-wide semantic interoperability framework alternative without SNOMED CT. These scenarios were addressed in WP2 through three different terminology settings: SNOMED CT only (SCT_ONLY), a UMLS-derived alternative terminology set (UMLS_EXT), and a German only terminology setting (LOCAL) corresponding to a scenario where each country maintains their own terminology without or with minimal EU level coordination.

2.3 ASSESS CT Workpackage 2

The ASSESS CT Workpackage 2 is conducting comparative studies, all of which attempt the following two questions:

- *How well does SNOMED CT address selected use cases, compared to an alternative setting, which uses a mix of existing terminologies without SNOMED CT, adapted to the needs of EU member states?*
- *How well does SNOMED CT address selected use cases, compared to the current state of affairs, i.e. sticking to the terminologies already in used across EU member states?*

All use cases are committed to the overall goal of semantic interoperability, assuming its positive impact on patient safety and health service cost-effectiveness. ASSESS CT assumes that this requires standardization of meaning at international level. As a result, semantic artefacts are required to introduce language-independent meaningful units (concepts) in a precision and granularity sufficient for clinical documentation purposes across clinical disciplines and specialties. These concepts should ultimately be unambiguous by means of formal or textual definitions, as well as due to fully specified names. Availability of term variants and synonyms is another desideratum. Workpackage 2 addresses three use cases:

- *Use of SNOMED CT vs. other terminologies for manually annotating clinical texts in different languages. This is mainly justified by the fact that natural language documents contain the terms clinicians use in their daily practice. The more easily these terms can be linked to concepts in a terminology, the higher is its quality. This depends on two aspects, viz. (i) the granularity of content provided by the concepts in the terminology (concept coverage) and (ii) the wealth of clinically relevant synonyms or entry terms in the terminology (term coverage). Another quality criterion is inter-annotator agreement: the more the annotation results coincide between two or more annotators, the more precisely defined and/or self-explaining is the terminology. This is seen as a second quality criterion.*
- *Use of SNOMED CT vs. other terminologies for providing textual values for structured data entry forms. Despite the predominance of text, structured data entry is increasingly important in clinical documentation, especially for clinical research, quality control, disease registries, and billing. The structuring of clinical information is provided by binding the meaning of the data elements of information models to external terminologies and by constraining value sets for coded data elements.*
- *Use of SNOMED CT vs. other terminologies for machine annotation of clinical text in different languages. The main rationale is the fact that natural language continues being the main carrier of clinical information. The ongoing adoption of Electronic Health Record (EHR) systems, is substituting paper charts by computer-based charts, but often with no change of content structure, which continues highly compacted text, often with idiosyncratic and error-laden terms and passages. Natural Language Processing (NLP) has developed powerful tools and techniques to analyse human language and map its content to controlled terminologies. This use case uses an off-the-shelves text processing pipeline tailored to several languages.*

All use cases provide indicators for SNOMED CT's technical fitness for use. As technical fitness for use is a prerequisite for clinical fitness for use, and samples of clinical data are used for the studies, clinical fitness for use can be indirectly assessed. The evidence created by the studies proposed in WP2 is assumed to disseminate knowledge about the current state of SNOMED CT, in order to inform policy dialogues and strategic planning processes that are necessary to set the course for EU-wide clinical reference terminologies.

3 Natural language processing as a tool for annotation of clinical text

3.1 The importance of free text documentation in electronic health records

Free text documentation is arguably the most faithful expression of the findings, ideas and intentions of a documenting clinician, not coerced into a predefined clinical model or template, or to predefined value lists. Assuming that there will be continued development of clinical models and value lists to capture clinical documentation to a greater extent, insights derived from free text documentation allow us to anticipate the future direction of clinical modelling, and therefore the future expectations of clinical terminology systems.

Natural language processing (NLP) technology is improving in accuracy; and it will therefore be increasingly used to detect important clinical facts from historic or current free text documentation that can be mapped to care pathways, decision support systems and reporting systems. It is therefore important that the terminology that is used for structured and coded data capture is also capable of serving these NLP needs.

3.2 Corpora for NLP experiment

The corpora used were the same as for the manual free text annotation task.

For both the manual and machine annotation tasks, a multilingual corpus was necessary. To this end, clinical texts in six languages (Dutch, English, French, Finnish, German, and Swedish) were collected by the consortium partners. The acquisition of corpora was done in a way supposed to approximate representativeness in terms of clinical domains, document sections, and document types. Finally, 60 document snippets (400 – 600 characters), 10 for each language were selected. Each snippet was translated into all other languages by professional translators and then reviewed and corrected by Workpackage members. The output, a parallel corpus consisting of 60 text snippets per language, was tokenized in order to generate the input for the manual and machine annotation experiments. For the experiments described in this document, only the English, French, Dutch, and Swedish texts were used, as no SNOMED CT translations are available for German and Finnish.

3.3 Terminologies for NLP experiment

The terminologies used were the same as for the manual free text annotation task.

In order to respond to the overall requirements of ASSESS CT, viz. comparing SNOMED CT to alternative terminologies, the following two custom terminology settings were used. All of them were filtered by selected UMLS Semantic groups[1]. These groups constitute pairwise disjoint divisions of all concepts in the UMLS Metathesaurus. Via SNOMED CT – UMLS mappings, the same semantic groups are also used to partition SNOMED CT.

- *SNOMED CT, international version (English) August 2015, Swedish version, as well as Dutch and French fragments provided by the Belgian government where the terminology is currently being localised, were included. Only concepts from selected UMLS semantic groups were used, in order to exclude terminology content that is outside of the clinical domain proper. We also excluded the SNOMED CT "Situation"*

[1] ANAT = Anatomy, CHEM = Chemicals & Drugs, CONC = Concepts & Ideas, DEVI = Devices, DISO = Disorders, GENE = Genes & Molecular Sequences, LIVB = Living Beings, OBJC = Objects, PROC = Procedures

hierarchy, which provides pre-coordinated concepts to express context like negation, certainty, time etc. The reason for this is that we consider this as belonging to information models, not terminologies, with the "Situation" hierarchy constituting an information model inside SNOMED CT. This terminology setting is named ***SCT_ONLY****.* In order to make SNOMED CT available also for German texts, a semi-automated method was used to generated German language terms out of English SNOMED CT descriptions. This method was based on a resource developed in the EU project SEMCARE.

- *In order to address alternative settings, we created an alternative hybrid terminology, including terminologies already in use. Starting point was the UMLS Metathesaurus, a repository of over 160 biomedical terminologies in different languages, with all of their content linked to unique identifiers (CUIs). Criteria for inclusion in the ASSESS CT alternative setting were sources that are actively updated*[2]*. From these, the following sources were excluded: (i) sources the use of which makes only sense in an U.S. context, such as U.S: drugs, (ii) sources in languages other than English, Dutch, French, German, Swedish, Finnish, (iii) SNOMED versions and Read code versions, (iv) sources out of scope regarding our data (nursing, dentistry).*
  Additional localised terminologies were added, for which the English version was part of the UMLS selection, e.g. MeSH, ATC, ICD in several languages.
  This terminology setting is termed ***UMLS_EXT***. It is huge for English with about two million concepts and four million terms. In comparison, the Swedish subset is small with 32,000 concepts, mostly represented by one term only. The Dutch and French subsets of UML_EXT have about 175,000 concepts, with 284,000 terms (French) and 342,000 (Dutch). The German subset has about 85,000 concepts and 228,000 terms.
- A set of terminologies for a strictly local setting was compiled for German, which included terminologies only available for German without mappings to UMLS or other terminologies, e.g. a German procedure terminology. This terminology setting is termed **LOCAL**. It has about 119,000 concepts and 201,000 terms.

3.4 Natural language processing pipeline

A natural language processing (NLP) pipeline is built to extract specific information units from unstructured texts. For ASSESS CT, the task is to automatically identify terminology codes by mapping clinical texts to the target terminology settings (SCT_ONLY, UMLS_EXT, LOCAL). The NLP pipeline consists of a number of components such as sentence detector, tokenizer, stemmer, morpho-semantic analyser, part-of-speech-tagger, chunkers and parsers. Each of these components can be individually adapted to different use cases to process text in different languages. These text analysis components, so called Analysis Engines (AEs), stick together in the Apache UIMA[3] (Unstructured Information Management Architecture) framework building an overall solution for different use cases. UIMA provides all necessary methods to create custom annotators and to put them together as an aggregated analysis engine (AAE) consisting of a sequence of multiple AEs. Each AE takes the output (annotations) of previous AEs as input and generates new annotations, building up structured information.

For ASSESS CT, the following components are used:

Sentence Detector: Due to the ambiguity of punctuation, the problem of sentence boundary recognition is known to be a non-trivial task. For instance, a full stop might appear in arbitrary contexts such as inside abbreviations or as a decimal point, resulting in different meanings.

Tokenizer: Sequences of characters are split into individual tokens, which do not only describe single words but also punctuation marks such as periods and commas. As in the

[2] http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/active_release.html

[3] http://uima.apache.org/

case of the sentence recognition problem, words itself may contain numbers ("B2B"), additional characters ("Dr."), or dashes ("e-mail"), making tokenization a non-trivial task. The ASSESS-CT pipeline uses a rule-based approach, which can be applied to different languages in different domains. The Tokenizer takes the sentences annotated by the sentence detector as input.

Stemmer: A stemmer reduces each token to its word stem, e.g. “mobiles” and “mobility” are reduced to the stem “mobil”. Reducing the number of inflections by mapping terms to their stem improves the concept mapping process and reduces the number of features in the case of statistical methods. The stemmer is rule based and is available for a number of languages.

Morpho-semantic analyser: Many European languages, such as German or Swedish, are fusional languages, characterised by complex forms of composition, derivation and inflection. This makes morphological analysis an inevitable subtask for processing text in such languages.

Concept Mapper: With the annotations generated by the previous AEs (linguistic pre-processing), the concept mapper is able to create mappings between concepts of a terminology and free-text phrases in various morphological layers such as word-, stem- and subword-level. The component contains a number of features such as concept disambiguation and filtering to improve the mapping process.

Nevertheless, the background terminology essentially determines the quality of the mapping, as evidenced by the following text examples that correspond to the concepts “Metastasis to lung” (285604008 in SNOMED CT) or “Lung metastasis” (C0153676 in the UMLS), in English and German:

- *… Lungenmetastasen …*
- *… Lungenfiliae …*
- *… pulmonale Metastasierung…*
- *… pulmonary metastases …*
- *… pulmonary relapse of a metastasis …*
- *… metastases in the lung …*

etc.

Expansive paramediastinal tumor shadow visible on x - rays and CT . An external bronchoscopy established a mucous membrane infiltration at segment level by a poorly differentiated adenocarcinoma . Mediastinoscopy shows the tracheobronchial and bifurcation lymph nodes are clear of tumors . No risk - enhancing restrictions with regard to pulmonary function . Surgical treatment is therefore indicated .

Figure 1: Example of concept annotations

Figure 1 exemplarily shows concept annotations, i.e. sequences of words that map to terminological entries in a reference system (here: UMLS).

3.5 Running the NLP experiment and data analysis

In this task, the terminological resources described above are used to generate concept annotations on the experimental corpora by using a NLP pipeline. In contrast to the manual annotation experiment, there are no human annotators in the loop.

In a first run, the simplest configuration for the NLP pipeline is used to generate baseline results. In subsequent iterations, different configuration settings are tested to improve the performance of the system, but without modifying the underlying terminologies.

For the baseline evaluation, the concept mapper maps words and sequences of words by applying a very simple linguistic pre-processing step only (i.e. stemming, that is applied by

default in many information retrieval settings). Therefore, annotations are made if and only if the word stems of the words in the texts correspond to the word stems of terminological entries. Future runs will also incorporate:

- Stopwords: Articles, preposition etc. will be ignored during the concept mapping procedure
- Decompounding: Concept mapping will be performed on parts of words (by processing texts and terminological entries in the same way)
- Nouns & Adjectives: Mapping will be only performed on nouns, adjectives, and combinations of them
- Acronym detection for avoiding linguistic normalization (e.g. “AIDS” vs. “to aid”)
- Combinations between these criteria

The manually created corpus annotations serve as a reference standard, i.e. their annotations are considered as being correct. The evaluation pipeline then produces a list with annotations correctly found by the text mining pipeline (true positive= TP), annotations not found by the text mining pipeline (false negative= FN) and annotations incorrectly found by the text mining pipeline (false positive = FP).

As overall result, the following standard metrics are calculated based on all documents:

- Recall $= \frac{TP}{(TP+FN)}$
- Precision $= \frac{TP}{TP+FP}$
- F-score $= 2 * \frac{Precision*Recall}{Precision+Recall}$

Since there are two manually generated annotation sets per language, the NLP evaluation pipeline runs on both sets, and the results are averaged afterwards. Figure 2 shows the average values on all scenarios and languages considered.

3.6 Preliminary results

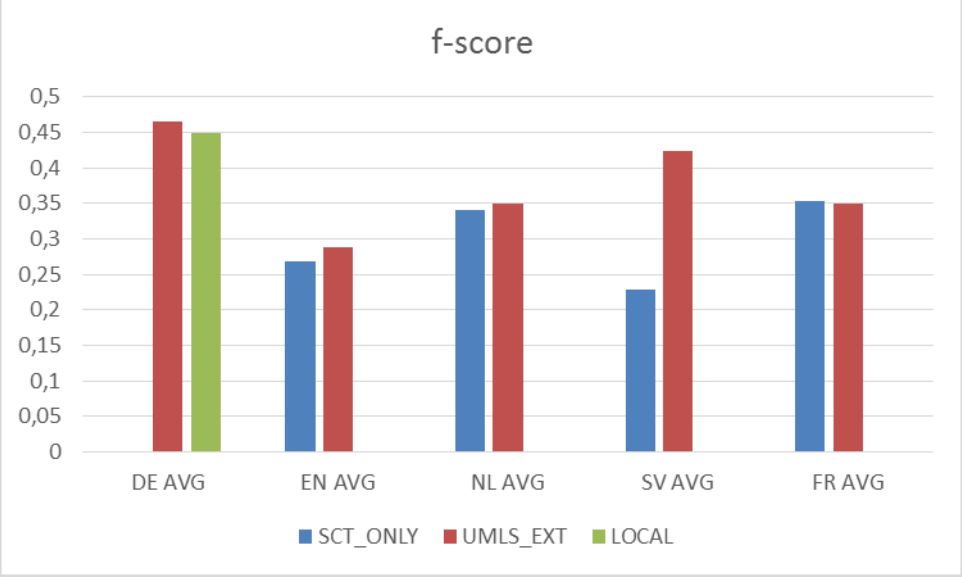

Figure 2: Average f-score values for the simple baseline configuration

For the present – and keeping in mind that naïve mapping approach is used here in order to generate figures for the baseline condition only – UMLS_EXT and SCT_ONLY seem to perform without significant differences. An exception is the Swedish language setting, where

UMLS_EXT performs better than SCT_ONLY. Explanations for this difference may include that only one term per SNOMED CT concept has been translated serving the purpose of a Fully Specified Term in Swedish, i.e. a term that is comparatively unambiguous but often more complex and less likely to be used in everyday language. This is, however, also true for French and Dutch, and therefore needs to be further analysed together with additional inspections, including different NLP configurations, which will shed more light on the quality of the automatically generated annotations and experimental settings.

In addition, for the 20 documents annotated twice, a gold standard is being created by consortium members with a medical background who decide on a "best mapping" in each case of inter-annotator disagreement.

4 Qualitative assessment of inter-annotator disagreement

The type of inter-annotator disagreements indicate what measures could be taken to improve consistency of selected concepts. The relative ease or difficulty to improve consistency is an important factor when choosing one terminology over another for the representation of clinical content.

4.1 Inter-annotator disagreement in free-text annotation experiment

Disagreement has been discussed at several work package meetings to give insight into the kinds of disagreement present in the data. Across all settings and languages the rate of inter-annotator disagreement is considerable which, when used in practice, would yield unsatisfactory retrieval results leading to lack of interoperability. Here are three disagreement cases examined qualitatively.

- "Nitroglycerin retard" – according to the guidelines, when annotating drugs preference should be given to substance and not product codes. On the other hand, preference should be given to pre-coordinated codes. In this case, one annotator chose the product code 346465006 |Modified release glyceryl trinitrate (product)|, whereas the other one chose the substance code 387404004 |Nitroglycerin (substance)|. (There is no SNOMED CT concept for 'modified release', only 'modified release tablet' which would be too specific here)
- "mucous membranes in mouth, pharynx, and on the larynx": Here one coder combined 33044003 |Structure of mucous membrane of oropharynx (body structure)| with 71248005 |Structure of mucous membrane of larynx (body structure)|, the other one chose only 113277000 |Oral mucous membrane structure (body structure)|
- "Fragmental fractures of the two upper vertebrae of the cervical spine". Here, one annotator combined
  13321001 |Fracture, comminuted (morphologic abnormality)|, 125606003 |Fracture of cervical spine (disorder)|, 14806007 |Bone structure of atlas (body structure)|, and 39976000 |Bone structure of axis (body structure)|, whereas the other one combined 112624007 |Fragmentation (morphologic abnormality)|, 207984009 |Fracture of second cervical vertebra (disorder)|, and 207983003 |Fracture of first cervical vertebra (disorder)|

More disagreement cases will be collected and discussed in the final report. In addition, a case study will be performed aiming at identifying semantic equivalence and/or proximity by exploiting the axiomatic structure underlying pre-coordinated SNOMED-CT concepts.

4.2 Inter-annotator disagreement in terminology binding experiment

In this part, the qualitative results are also preliminary and incomplete, and their analysis and interpretation is ongoing. In particular, the alternative terminology use is in an earlier phase of analysis compared to SNOMED CT use. A complete analysis will be provided in the final deliverable.

4.2.1 Methods

To analyse the annotations, we evaluated each element in the information model by looking at the set of annotation plus the comments provided by the annotators for each setting. We

coded the type of disagreement and the significance of it for each information model snippet. One example of such a coding is: "**Systematic hierarchy mismatch** (1 coder in 10/10 codes)". After each such coding step, we explained how we got to this conclusion.

For example (one coder in 10/10 codes): One coder systematically coded with Clinical findings rather than Observable entities, for example 301283003|Finding of rate of respiration (finding)| vs. 86290005|Respiratory rate| or 301113001|Finding of heart rate (finding) vs. 364075005 | Heart rate (observable entity) |. The coder that maps to findings adds in a comment:

"*I have mapped this to findings, because I assume this list is used as entries for the result of evaluating each of these areas. As you see Glasgow coma scale can't be mapped using findings - but the Observable hierarchy maps perfectly to these entries. ... Mapping all entries in this template using Observables gives me a full map of 100% using pre-coordinated concepts. So, in order for me to determine which of the proposals to go for I would need to understand whether these entries represent the heading/label of each entry or they represent the evaluation results. If they represent the tables, then I would use the observables in all cases, but if they represent the evaluation results, then I would use the findings...*"

This example tells that the coder wondered whether it was the left hand side or the right hand side representation of the content that should be represented

We have not yet finished this analysis. However, in the final deliverable we will be able to show the type and significance of coding differences.

4.2.2 Preliminary results

Preliminary findings suggest that common annotation differences were:

Overlap vs. disjunction of attribute-value pair semantics: In attribute-value pairs among the clinical model elements, annotators differed in how much of the attribute semantics would be redundantly represented in the value code, i.e. to which extent the value should be interpretable outside the context of the information model. For example, for the information model attribute "Headache location", the value set includes various sites where a headache may be experienced such as "parietal left" or "occipital". Some annotators used codes for the anatomical locations, for example "left parietal region", whereas other annotators used codes including headache, for example "headache experienced in left parietal region".

Hierarchy uncertainties: SNOMED CT has two hierarchies, Substance and Pharmaceutical / biologic product, which are closely related. Unless guidelines are provided explaining the use of the two hierarchies, the two often offer equally well suitable representations, leading to lack of agreement on which one to choose. However, using SNOMED CT defining attributes, transformation between the two forms is close to trivial. Furthermore, SNOMED CT implementation projects would typically provide guidelines that would restrict the use the two hierarchies. As the example in the method section suggest, the two hierarchies' observable entity and clinical findings were also used interchangeably.

Difference in level of concept model consistency in information model patterns: Some groupings in the SNOMED CT setting have been made with awareness of concept model patterns i.e. the attribute have to be associated with the value set in a way that follows SNOMED CT concept model rules. These disagreements exist mostly because not all annotators mastered these rules. Secondly, some information model patterns could be represented using two equally good terminology patterns, which causes disagreement even for experienced coders.

Insufficient medical or medical informatics knowledge: Some model elements have been misinterpreted due to lack of medical knowledge or misrepresented due to lack of medical informatics knowledge.

Generally, agreement seemed to relate to coverage. When there is a lack of coverage, the annotators have to be more creative which will likely increase disagreement.

5 Comparative analysis

As the results of the NLP experiment are still preliminary, the comparative analysis is not conclusive. For example, we do not yet know what content coverage can be obtained with NLP techniques. Consequently, this chapter is mostly just a comparison of the free text annotation experiment and the terminology binding experiment.

5.1 Concept coverage in English experiments

In the free text annotation experiments, the coders did not agree on what a full match, inferred match, partial match or none-match is. Looking at the concept coverage scores for the English part of the experiment: using full, inferred and partial match as indication of concept coverage resulted in a concept coverage of 94% for the UMLS_EXT and 92% for SNOMED CT; whereas concept coverage is 88% for the UMLS_EXT setting and 86% for SNOMED CT using only full and inferred match as indicators.

The terminology binding experiment resulted in a significantly better coverage of content using SNOMED CT than the set of four chosen terminologies (ATC, ICD-10, LOINC and MeSH) which were used to represent the alternative setting. However, indications of coverage i.e. full, inferred, partial or none were assigned very differently by annotators. Thus, we conclude that the annotators experienced, in general, higher content coverage of SNOMED CT than the alternative terminologies, but they did not agree on what a full match, inferred match, partial match or none-match is.

Using full, inferred and partial match as indication of content coverage resulted in a content coverage of 80 % for the combination of ATC, ICD-10, LOINC and MeSH, and 94% for SNOMED CT. When using only full and inferred match as indication, content coverage is 51 % for the combination of ATC, ICD-10, LOINC and MeSH while that for SNOMED CT is 79 %.

Across experiments, coverage is very high for all settings. The most notable difference is that the combination of ATC, ICD-10, LOINC and MeSH seems to perform worse in the terminology binding experiment than UMLS_EXT does in the free text annotation experiment. At this point in time, we cannot explain the difference. Maybe it is due to the significantly larger size of UMLS_EXT (approx. 2,000,000 concepts) compared to the size of the hybrid terminology (approx. 100,000 concepts) or the relatively better performance of SNOMED CT in terminology binding use cases than in free text annotation use cases. Another confounder may be that the free text annotation experiment used a customized browser which rendered the two terminology settings SCT_ONLY and UMLS_EXT in the same way; whereas in the terminology binding experiment, various external browsers with possibly different features were used.

Another smaller difference is that the content coverage of SNOMED CT in the terminology binding experiment is higher than in the free text annotation experiment. However, this might be due to the relative ease of representing structured content compared with unstructured content, or the differences in annotator expertise level. The terminology binding experiment also used the original browsers, whose usability is optimised to the terminology they represent.

6 Discussion and outlook

In three experiments, existing terminologies in two settings have been used to annotate pieces of information in either structured form, as in clinical information models, or unstructured form, as in clinical free text.

6.1.1 Qualitative and comparative analysis

Using terminologies inside clinical information models and using terminologies for annotating free text documents are two quite distinct terminology use cases. Clinical information models provide a structure and context for terminology use and the target of binding is formally defined whereas in free text annotation target of annotation is dependent of the identification of semantic units (chunks) by the annotator, adding variability to terminology use.

It can be argued that using standardized terminologies inside clinical information models is the primary use case for those terminologies, i.e. the standardized terminologies are built for being used inside certain structures. Codes from typical terminologies do need some structure to provide context to allow meanings to be sufficiently disambiguated, at a minimum for spatio-temporal specification, i.e. what or who the information is about and when it holds/held true. The general question of demarcation between representing meaning in the clinical information models verses that in the terminology, while has been the topic of projects such as HL7 TermInfo and SemanticHealthNet, is far from resolved. This is also reflected, in particular, in the results of the terminology binding experiment.

Similar differences were seen when studying disagreement in the use of terminologies in the free text annotation and the terminology binding experiments. In both experiments, there were disagreements on the selection of hierarchies within SNOMED CT, even though there were guidelines to minimize such variability. When using groupings, i.e. multiple codes, to express meaning, annotators in both studies often showed a level of creativity beyond what can be safely interpreted. Here, it is important that annotators have sufficient knowledge of the terminologies used and specifically their limitations.

The participants in the free text experiment all had medical background but some lacked medical informatics training, whereas in the terminology binding, all had medical informatics experience but not medical background. The combined results from both experiments indicate that both medical and medical informatics knowledge is needed to optimize the use of standardized terminologies.

6.2 Limitations

The analysis is still ongoing. Hence, results and conclusions based on results may change.

6.3 Additional results to be expected

The analysis of NLP results will be developed further for deliverable D2.4; same holds for the comparative analysis between experiment results. Forthcoming deliverables will have a stronger focus on interpretation of results in terms of recommendations.