Proceedings of MultilingualBIO 2020, pages 15–20
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Localising the Clinical Terminology SNOMED CT by Semi-automated Creation of a German Interface Vocabulary

Stefan Schulz[1,2], Larissa Hammer, David Hashemian-Nik, Markus Kreuzthaler [1]
[1]Institute for Medical Informatics, Statistics and Documentation,
Medical University of Graz, Auenbruggerplatz 2/V, 8036 Graz, Austria,
[2]Averbis GmbH, Salzstraße 15, 79098 Freiburg i.Br., Germany
{stefan.schulz, markus.kreuzthaler}@medunigraz.at

Abstract

Medical language exhibits large variations regarding users, institutions, and language registers. With large parts of clinical information only documented in free text, NLP plays an important role in unlocking potentially re-usable and interoperable meaning from medical records in a multitude of natural languages. This study highlights the role of interface vocabularies. It describes the architectural principles and the evolution of a German interface vocabulary, which is under development by combining machine translation with human annotation and rule-based term generation, yielding a resource with 7.7 million raw entries, each of which linked to the reference terminology SNOMED CT, an international standard with about 350 thousand concepts. The purpose is to offer a high coverage of German medical jargon, in order to optimise terminology grounding of clinical texts by NLP systems. The core resource is a manually maintained table of English-to-German word and chunk translations, supported by a set of language generation rules. We describe a workflow consisting in the enrichment and modification of this table by human and machine efforts, together with top-down and bottom-up methods for terminology population. A term generator generates the final vocabulary by creating one-to-many German variants per SNOMED CT English description. Filtering against a large collection of domain terminologies and corpora drastically reduces the size of the vocabulary in favour of terms that can reasonably be expected to match clinical text passages within a text-mining pipeline. An evaluation was performed by a comparison between the current version of the German interface vocabulary and the English description table of the SNOMED CT International release. An exact term matching was performed with a small parallel corpus constituted by text snippets from different clinical documents. With overall low retrieval parameters (with F-values around 30%), the performance of the German language scenario reaches 80 – 90% of the English one. Interestingly, annotations are slightly better with machine-translated (German – English) texts, using the International SNOMED CT resource only.

Keywords: clinical language, under-resourced languages, technical term generation

1. Introduction

Clinical documentation addresses the needs of health professionals to communicate, collect, and share information for joint decision making, to summarize heterogeneous data, and to customize them to provide optimal support to different use cases.

Electronic health records (EHRs), besides their primary purpose of data presentation and visualisation, bear the potential of large data analysis. It has turned out that structured data do not optimally meet clinicians' documentation and communication requirements, which explains their preference of free text and a general tendency of bias regarding structured (and especially coded) clinical data.

Clinical information ecosystems, their support by computers, and particularly the role clinical language plays therein are far from being ideal. Yet modern clinical care, biomedical research and the translation of the latter into clinical care require ontological and terminological standards in order to make clinical information and data reliable, precise and interoperable.

The need for health data interoperability and exchange is addressed by a multitude of terminology and classification systems, which categorize and define technical terms and their meaning (Schulz et al. 2019; Bodenreider et al., 2018). A certain tragedy lies not only in the fact that these systems interoperate with each other only in exceptional cases and their contents are barely mappable, but also that, despite their commitment to language and concept representation, they are far from representing the jargon that clinicians use in their daily practice. Yet there are some reasons to be optimistic, given the increasing acceptance of large, well-curated terminology systems like SNOMED CT (Millar 2016) and LOINC, used by impressive applications like OHDSI, demonstrating the potential of universal terminologies to integrate and compare data extracts from a variety of clinical information sources (Hripcsak et al., 2018).

Clinical language is largely different from the standard language, including the language used in medical literature. Text is produced in a hurry; often entered directly by clinicians, partly by dictation (with subsequent transcription), increasingly by using speech recognition. Often, no documentation standards are used.

In all these cases, parsimony of expression dominates, to the extent that ambiguous expressions, as long as they are short enough, are preferred, assuming the reader has the context to disambiguate them. Abbreviations and acronyms abound, so that many clinical texts appear overly cryptic even to specialists from other disciplines, let alone to patients. Clinical language is furthermore characterized by incomplete sentences, by lack of grammatical correctness and by a wild mixture of hybrid technical terms that blend the host language with fragments of English, Latin and Greek vocabularies.

The vocabulary mismatch between the clinical jargon and the controlled language of medical terminology systems is immense. For instance, the SNOMED CT concept label "Primary malignant neoplasm of lung" (the eighth most common cause of death worldwide) is unlikely to be literally found in any text written by a doctor. Even in scientific texts (which are of better editorial quality), such artificial terms are highly uncommon. There is no single occurrence of the above term in 27 million MEDLINE records (opposed to about 150,000 hits for the synonym

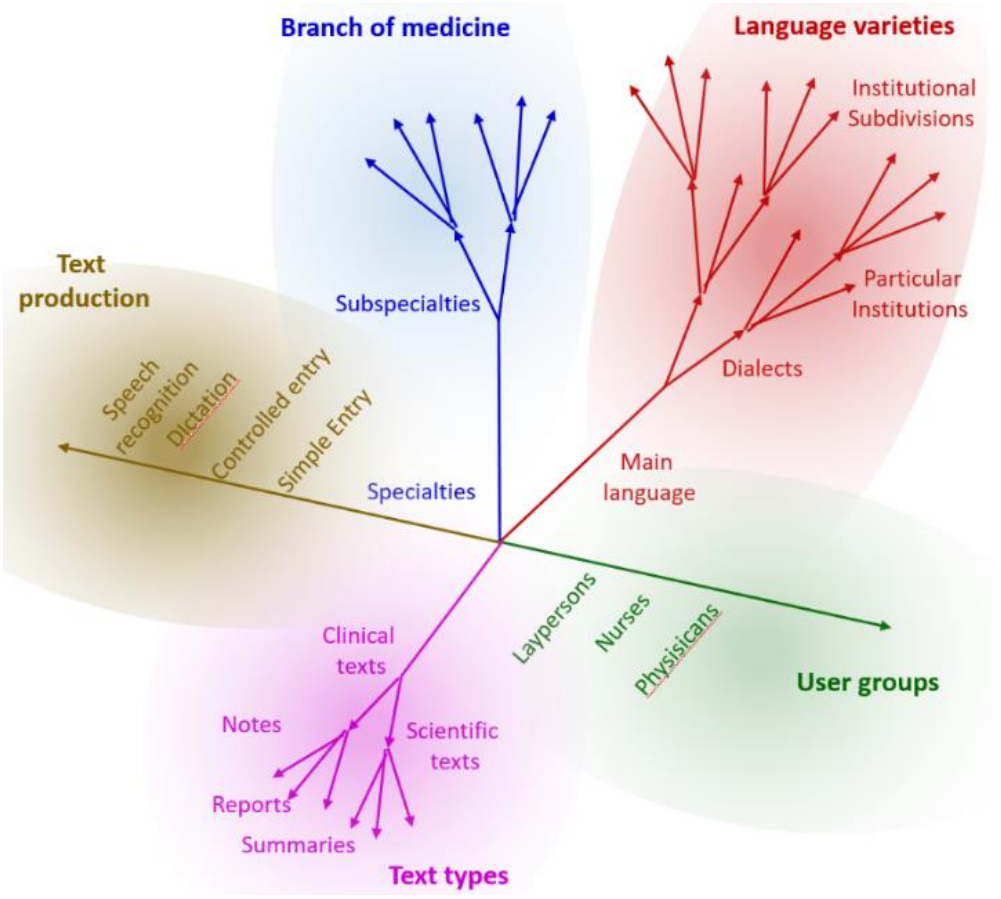

Fig. 1 – Determinants of medical interface terms

"lung cancer"). It is no wonder that terminology implementation studies have shown that standardized terms are often inadequate for clinical use (Højen et al., 2014).

This gap can only be filled by a bottom-up, community-driven non-prescriptive terminology building approach (Schulz et al, 2017). Interface vocabularies, also known as (user) interface terminologies have been proposed (Kalra et al, 2016) as mediators between the real-world clinical language in a given setting (local, national, and user-specific) and international terminology standards like SNOMED CT.

The content of interface vocabularies depends on a series of factors; from language groups (e.g. German, French) to dialects (German spoken in Austria, French spoken in Canada, etc.) to user groups (physicians, nurses, laypersons) to institutions (department A of hospital X, clinic B inside health centre Y) to document types and document sections. The choice of terms also depends on the way the text is produced. Fig. 1 shows how medical language registers are shaped along several axes.

Ideally, an interface vocabulary maps every lexeme into this space, so that most terms become unambiguous according to the context in which they are used. If this context is not known, or not specified in the dictionary, lexical ambiguity becomes a main source of errors in natural language processing.

A manual creation of highly fragmented and specialised interface vocabularies prohibits itself. Instead, automated means should support interface vocabulary creation and management.

There are several use cases for clinical interface vocabularies, some of which are directly related to NLP systems, particularly systems for semantic search within term lists of whole documents, and information extraction. However, collections of interface terms also are important as parts of mono and bilingual dictionaries for specific technical or scientific domains.

Equally important is their use as source for value sets for structured data entry within data acquisition forms, where the terms should be close to the users' language preferences.

2. Materials and Methods

2.1 Source Terminology

SNOMED CT (Millar, 2016) is an ontology-based terminology owned and maintained by the standards development organisation SNOMED International. SNOMED CT is intended to provide the vocabulary needed to represent electronic health records. The current international release has about 350,000 active representational units, called SNOMED concepts. They represent language-independent meanings and are rooted in a formal framework based on the description logics OWL-EL. In this sense, SNOMED CT is very advanced compared to other terminologies. E.g., the concept

Thyroiditis is defined as logically equivalent to a *Disorder* with *Inflammatory morphology* that is located at some *Structure of the thyroid gland*. For several natural languages, SNOMED concepts are linked to one or more technical terms via a so-called description table. E.g., the international English version includes about 950,000 such terms, divided into fully specified terms, i.e. self-explaining, often synthetic labels like the one discussed in the previous section, and synonyms, which are closer to the clinical language in use and correspond to what we have introduced as "interface terms" in section 1. SNOMED CT terms range from single words ("appendectomy", "aspirin") to complex phrases and even sentences ("Computed tomography of neck, chest, abdomen, and pelvis without contrast", "Product containing only sulfamethoxazole and trimethoprim in parenteral dose form").

Besides English, Spanish is the only language for which an official SNOMED CT version, maintained by SNOMED International, exists. Danish and Swedish versions have been locally created, however with only one localised term per concept. For other languages (French, Dutch), partial localization efforts are ongoing. However, for many important languages (German, Italian, Russian, Japanese, Chinese) no SNOMED CT language resources exist, let alone for the multitude of smaller languages, despite their importance in clinical documentation and communication.

2.2 Resources for term harvesting and scoring

Several domain-specific, German language clinical corpora with clinical discharge summaries have been collected at the authors' institution, thanks to several projects on clinical NLP, with authorisation from the institutional ethics committee. In particular, a corpus with about 30,000 cardiology summaries, one with about 5,000 melanoma-related summaries, and one with about 2,000 colorectal cancer summaries were harvested. Another source of clinical language was a database with about 1.7 million unique clinical problem list entries. In addition, the official Austrian drug dictionary was used as a source. For scoring and filtering the machine-generated German interface vocabulary, a collection of 17 German medical terminology systems was exploited, together with a dump from the leading German Medical Journal, the German Wikipedia, filtered by domain, and several drug repositories with drug names, ingredients, and additional drug-related information.

2.3 General Method

A more detailed description of the workflow can be found elsewhere (Hashemian Nik et al., 2019). The main idea of our approach is the combination of machine translation with human translation and validation, as well as a generative process that assembles translations of complete SNOMED CT terms out of their – often highly repetitive – single word or short chunk translations. Briefly, the terminology building process can be described as follows:

Pre-processing

1. Definition of the source terminology (in our case, English textual descriptions (terms) linked to SNOMED CT codes);
2. Identification of terms that are identical across existing translations (e.g. Latin names of organisms);
3. Rule-based chunking of terms into single tokens, noun phrases and prepositional phrases;
4. Sorting chunks and words by decreasing frequency;
5. Submission to neural Web-based translation engines (Google translate, DeepL).

These steps have to be repeated for new terms that come with each semi-annual updates of SNOMED CT.

Specification and implementation

6. Specification of grammar-specific annotations, e.g. POS, gender, number and case for nouns, case for prepositions.
7. Implementation of term building routines, e.g. for adjective / noun inflection and single-word composition, using Python scripts.

Manual curation

8. Manual checking of chunk translation results;
9. Adding new synonyms and spelling variants
10. Adding short forms (acronyms, abbreviations);
11. Identifying ambiguous source terms and adding context-aware translations of longer phrases;

Term creation and manual validation

12. Execution of term assembly routines;
13. Manual assessment of results for formal, stylistic, and content correctness; accordingly repeating former steps, particularly 6, 7, 10, 11.

Bottom-up enhancement by corpora

14. Creation of n-gram lists ("real-world chunks") from clinical corpora, according to the rules developed in 3;
15. Manual mapping of real-world chunks to chunk translation table, iteration of steps 12 and 13.

Validation of progress

16. Validation against benchmarks; blind checking of results against fully machine-translated terms;
17. Manual validation of concept annotations within an NLP pipeline that uses the terminology on real clinical texts.

Enhancement and filtering

18. Exclusion of short (length < 4 characters) acronyms, unless embedded in context (e.g., "CT" is excluded, "CT scan" is preserved).
19. Selection of resources (corpora, dictionaries, databases) to be used as sources of truth for filtering and enhancement;
20. Semi-automated addition of brand names using national drug databases;
21. Creation of rules to harvest spelling variants from external sources (e.g. "ce" vs. "ze", hyphenation vs. spaces or single-word compounds);
22. Defining scoring metrics based on token and n-gram occurrences in external sources;
23. Manual collection of negative examples to constitute patterns for term candidate rejection.

Terminology profiling

24. Using scores and other parameters to filter the terminology according to different usage profiles, e.g. for text mining or value set creation.

2.4 Benchmarking

The interface terminology is periodically checked against a benchmark that was built on top of the results of a multilingual manual SNOMED CT annotation experiment (Miñarro-Gimémez et al., 2018, 2019) for which a small (average 3650 words), but highly diverse corpus had been built, composed by text snippets from clinical documents in six European languages (English, Swedish, French, Dutch, German and Finnish), out of which a parallel corpus was created by medical translators. Texts were annotated with SNOMED CT codes by terminology experts. These texts and related code assignments had never been used in the interface vocabulary building process.

For interface vocabulary benchmarking we re-used the German and English portions of this parallel corpus, together with the SNOMED CT codes attached. For the SNOMED CT representation, two reference standards were used:

- Reference standard R1: annotations using the English, French, Swedish and Dutch versions of SNOMED CT on the respective parallel texts performed by nine terminologists, totalling 2,090 different SNOMED CT codes;
- Reference standard R2: annotation using the English (International) version of SNOMED CT on the English portion of the corpus, performed by two annotators, totalling 1075 different codes (reference standard 2).

These reference standards were used to compare the following scenarios by using a very simple term mapper:

a. SNOMED codes retrieved by matching terms of our German interface vocabulary with the German portion of the corpus;
b. SNOMED codes retrieved by matching English terms of the International SNOMED CT description table[1] with the German portion of the corpus, machine-translated into English by using the freely available Google translator;
c. SNOMED codes retrieved by matching English terms of the International SNOMED CT description table[1] with the English portion of the corpus.

The concept mapper is based on exact match between one or more decapitalised tokens, iterating over the vocabulary reversely ordered by string length. For each match of a lexicon entry the corresponding string is removed from the corpus and the SNOMED CT code(s) assigned to it is (are) stored. The resulting code sets are compared to the set of codes in R1 and R2; and precision, recall and F-measures are calculated.

3. Results

The work started in 2014 with limited resources (one part-time terminologist and one to three medical students working on average 8 hours per week). Since then, it has been subject to constant optimization and quality improvement.

The current size of the terminology is about 7.7 million records, each record consisting of the SNOMED identifier, an interface term ID, the English source term and the automatically generated German interface term. Table 1 shows an example of eight German interface terms, automatically created out of two English SNOMED terms. All eight translations are correct in content and understandable, but only those in bold are grammatically correct and likely to be found in clinical documents. The interface terms were generated out of 125 thousand German word / chunk translations from about 100,000 English words / chunks.

An analysis of the current quality of the interface terminology by blinded human assessment terminology stated equivalence regarding content correctness when comparing a random interface term with the (only) term that resulted from the machine translation system DeepL (Hashemian Nik et al., 2019). However, the results show deficits regarding grammar, spelling, and style issues of the current state of the interface vocabulary. The same study revealed that a case insensitive, spelling variation tolerant match between an ideal translation suggested by a domain expert (not knowing the generated results) occurred with half of the machine-generated interface terms.

The combinatory explosion observed especially with long SNOMED term translations, many of which are not ideal and some of them not even understandable makes filtering and profiling necessary.

Code	English	German
53701004	Sebaceous gland activity	Glandula sebacea Tätigkeit
		Glandula sebacea Aktivität
		Talgdrüsentätigkeit
		Talgdrüsenaktivität
	Sebaceous gland secretion	Glandula sebacea Absonderung
		Glandula sebacea Sekretion
		Talgdrüsenabsonderung
		Talgdrüsensekretion

Table 1: Example of a SNOMED CT code, two English terms and eight generated German terms

We started with three profiles, viz. (i) one for text mining, limited to terms with a maximum of six tokens; another one (ii) in which only terms that literally matched the resources (cf. subsection 1.2) were preserved; and a third one (iii) which allowed more flexibility regarding plausibility checking, and in which up to 50 synonyms above a quality threshold were accepted.
Whereas (i) yielded 506 thousand interface terms (6.5% of the raw list), (ii) yielded only 89 thousand (1.2%), and (iii) 387 thousand. The corresponding coverage of SNOMED CT codes was 39% for (i), 17% for (ii) and 29% for (iii).

[1] which includes canonical and interface terms

The rationale for producing different profiles is explained by the use cases to be served by the interface vocabulary. For text mining purposes, exact or moderately fuzzy matches of terms with more than six tokens are very unlikely. On the other hand, implausible terms (because of combinations), which hardly ever match are harmless.

In cases where interface terms are created for human use (e.g. supporting picklists or auto-completion functionality for data entry), well-formedness, comprehensibility and currency are crucial. However, by using a strict filter, many of the synthetically created labels, like the above-discussed "primary malignant neoplasm of the lung" would be thrown out, because they do not occur in medical documents and not even in other terminology sources. The benchmark results are given in Table 2 and Table 3.

Experiment R1	Precision	Recall	F_1
a. German texts	0.49	0.16	0.28
b. German texts, machine translated	0.48	0.16	0.28
c. English texts	0.50	0.19	0.31

Table 2: Retrieval performance using reference standard R1 (pooled annotations by nine terminology experts, performed with English, Swedish, Dutch, and French SNOMED CT translations, performed on the respective language portion of the ASSESS-CT parallel corpus).

Experiment R2	Precision	Recall	F_1
a. German texts	0.36	0.23	0.29
b. German texts, machine translated	0.37	0.25	0.30
c. English texts	0.41	0.31	0.35

Table 3: Retrieval performance using reference standard R2 (pooled annotations by two terminology experts, performed with the SNOMED CT version on the English language portion of the ASSESS-CT parallel corpus).

4. Discussion and Outlook

We have outlined a complex heuristics that generates German interface terms for SNOMED CT concepts. In a previous work, we had demonstrated that its quality was roughly comparable to fully machine-generated terms. The advantage of our approach however was the high term productivity compared with machine translation, especially the assembly of term variants that are rare but useful, especially for data entry and text mining. A natural next step would be to exploit neural term harvesting approaches for additional terminology enrichment. Word embeddings might help retrieve new synonyms, but they also will require large amounts of training resources, which are difficult to acquire in a clinical context, let alone sharable among researchers.

Another strand of future work is the increased incorporation of acronyms and other short forms into the resource. So far, we have re-used existing acronym lists and have manually expanded acronyms from our clinical sources, but the ambiguity of two and three character acronyms is high. This is the reason why single acronyms with four characters and are suppressed in our pipeline, whereas longer terms containing them are released, e.g. "DM type 1" where sense disambiguation can be expected from the local context.

The benchmarking results provide interesting insights in the problems around terminology grounding of clinical texts, the peculiarities of a huge terminology like SNOMED CT, about the current quality of the German interface terminology and finally about the raison d'être of terminology translations in general.

It must be emphasised that the results given in tables 2 and 3 were not the result of text analysis in a NLP pipeline (for which better results should be expected, but of an overly simple term matching algorithm). The problem of finding the right SNOMED CT code for a passage of clinical text – even by terminology experts – was described in depth by Miñarro-Gimémez et al. (2018, 2019), who reported an astonishingly low inter-annotator agreement of about 40% (Krippendorff's α). That a team of nine annotators had come up with more than double the numbers of codes for the same content (in four languages), compared to a pair of coders (for English only) sheds light on the high degree of personal discretion involved. Of course, this meant that for many chunks of clinical meaning there were many annotations with semantically closely related codes, which explains the overall low recall, especially in the R1 scenario. The expert annotation task had also privileged SNOMED pre-co-ordinations, e.g. for "Fracture of the neck of the femur", which did not match expressions in the text like "The neck of the left femur was broken". Our term matching text might have matched the single codes "neck", "femur", "left", and "broken". However, this phenomenon is expected in all scenarios. Another characteristic of the corpus, which explains low performance values, is the frequency of acronyms and other short forms, e.g. the roman numbers "I" to "XII" for the cranial nerves.

Coming back to the primary purpose of this benchmarking, viz. the comparison of the German interface vocabulary created by the authors with the nearly one-million English term list that comes with the International SNOMED CT release (and which includes many close-to-user terms), the figures are remarkable insofar the performance of the German language scenario reaches 80 to 90% of the performance of the English one.

Finally, the figures on the alternative strategy, viz. machine-translating non-English clinical texts to English with Google Translate and checking against the original English SNOMED CT term list, could be a starting point for a radical re-thinking of multilingual text processing. Is it still worthwhile developing multilingual resources if neural machine translation (even not trained with specific clinical text) yields increasingly better results? Concentrating human efforts on improving the already very rich inventory of tools and resources for English could then be a better idea than creating and maintaining language resources for a multitude of different languages with insufficient financial and human resources.

Current versions of the resource can be downloaded from http://user.medunigraz.at/stefan.schulz/mugit/

5. Acknowledgements

This work was partly supported by the EU projects SEMCARE – 7th FP grant 611388; ASSESS-CT - H2020-PHC-2014-15, grant 643818.

6. Bibliographical References

Bodenreider, O., Cornet, R., Vreeman, D.J. (2018). Recent Developments in Clinical Terminologies - SNOMED CT, LOINC, and RxNorm. *Yearb Med Inform*, 27(1):129-139.

Hashemian Nik, D., Kasáč, Z., Goda, Z., Semlitsch, A., Schulz, S. (2019). Building an Experimental German User Interface Terminology Linked to SNOMED CT. *Stud Health Technol Inform*, 264:153-157.

Højen, A.R., Elberg, P.B., Andersen, S.K. (2014). SNOMED CT adoption in Denmark - why is it so hard? *Stud Health Technol Inform.* 205:226-230.

Hripcsak, G., Levine, M.E., Shang, N., Ryan, P.B. (2018). Effect of vocabulary mapping for conditions on phenotype cohorts. *JAMIA*, 25(12):1618-1625.

Kalra, D., Schulz, S., Karlsson, D., Vander Stichele, R., Cornet, R., Rosenbeck Gøeg, K., Cangioli, G., Chronaki, C., Thiel, R., Thun, S., Stroetmann, V. (2016). *Assessing SNOMED CT for Large Scale eHealth Deployments in the EU. ASSESS CT Recommendations*. http://assess-ct.eu/final-brochure.html.

Millar J.(2016). The Need for a Global Language - SNOMED CT Introduction. *Stud Health Technol Inform*, 225:683-685.

Miñarro-Giménez, J.A., Cornet, R., Jaulent, M.C., Dewenter, H., Thun, S., Gøeg, K.R., Karlsson, D., Schulz, S. (2019) Quantitative analysis of manual annotation of clinical text samples. *Int J Med Inform.*:123:37-48

Miñarro-Giménez, J.A., Martínez-Costa, C., Karlsson, D., Schulz, S., Gøeg, K.R. (2018). Qualitative analysis of manual annotations of clinical text with SNOMED CT. *PLoS One*. Dec 27:3(12)

Schulz, S., Daumke, P., Romacker, M., López-García, P. (2019). Representing oncology in datasets: Standard or custom biomedical terminology? *Informatics in Medicine Unlocked*, 15:100186.

Schulz, S., Rodrigues, J.M., Rector, A., Chute, C.G. (2017). Interface Terminologies, Reference Terminologies and Aggregation Terminologies: A Strategy for Better Integration. *Stud Health Technol Inform*, 245:940-944.