Facilitating SNOMED-CT Template Creation by Targeting Stopwords

Rashmi Burse, Michela Bertolotto and Gavin McArdle
University College Dublin, Belfield, Dublin 4, Ireland

Keywords: Quality Assurance, Biomedical Ontologies, SNOMED-CT Templates, Lexical Auditing, Semantic Analysis, Biomedical Named Entity Recognition.

Abstract: Quality Assurance (QA) of biomedical ontologies is a major challenge in the health-informatics domain. One of the preliminary ways in which we can maintain the quality of a biomedical ontology is by ensuring consistency in the modelling styles of biomedical concepts. Maintaining consistency in the lexical, structural and ontological modelling of biomedical concepts reduces a concept's susceptibility to errors. SNOMED-CT, which is one of the most widely adopted biomedical ontologies, strives to achieve this consistency by creating templates for logical definitions based on the description of biomedical concept names. The work presented here in based on the observation that the majority of the SNOMED-CT templates contain stopwords (non-medical terms) in their description that indicate a relationship between two medical concepts. We hypothesize that the process of creating SNOMED-CT templates can be automated to a large extent by targeting stopwords. In this work, we present a method that exploits stopwords in concept names to create templates for the structural and logical modelling of lexically and semantically similar biomedical concepts. The results have shown promising potential by extracting a multitude of SNOMED-CT templates, exhibiting more than 200 templates for the stopword *of*. Given the high demand for QA of biomedical ontologies, these results are highly beneficial in automating the existing mechanisms employed in maintaining consistency in the modeling of SNOMED-CT concepts. The presented method can be used as a complementary process to mitigate the manual efforts of SNOMED-CT curators. Furthermore, auditing potentially incomplete definitions of SNOMED-CT concepts using the extracted templates has identified 49-87% inconsistent concepts for the stopwords *of* and *in* in the biomedical ontology.

1 INTRODUCTION

Biomedical ontologies are referenced by Electronic Health Record (EHR) systems to populate clinical information in patient records to achieve semantic interoperability across healthcare organizations (Burse et al., 2021; Duarte et al., 2014; Kim et al., 2020; Willett et al., 2018). Considering the end goal, consistency within a biomedical ontology becomes implicitly critical. However, given the huge volume and variety of biomedical ontologies, maintaining consistency in the logical definitions of clinical concepts (even within the same ontology) is often difficult to achieve. The cause of such inconsistencies can be attributed to the prolonged development process and the involvement of a team of multi-disciplinary experts like healthcare professionals, ontology experts, and computer scientists (Unertl et al., 2012; Ceusters et al., 2004).

SNOMED-CT (IHTSDO, 2002a) is one of the world's most widely adopted biomedical ontologies with more than 300,000 concepts. Despite its high rate of adoption, studies have shown that there is scope for improvement in the quality of SNOMED-CT. The development of efficient auditing techniques for Quality Assurance (QA) of biomedical ontologies, including SNOMED-CT, has been a major challenge in the health-informatics domain (Amith et al., 2018). The auditing techniques can be broadly categorized into structural, lexical, semantic and hybrid approaches (Amith et al., 2018). A major part of lexical auditing techniques involves detecting inconsistencies in the structural modeling of lexically similar concepts. Many SNOMED-CT concepts with similar lexical formats for their Fully Specified Names (FSNs) have different structural modelling styles (Agrawal, 2018; Burse et al., 2022b). A consequence of such an unwarranted variety in the modelling styles of lexically similar concepts is that they are highly prone to missing attribute relationships. Consistency in the lexical and structural

Burse, R., Bertolotto, M. and McArdle, G.
Facilitating SNOMED-CT Template Creation by Targeting Stopwords.
DOI: 10.5220/0011660500003414
In *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2023) - Volume 5: HEALTHINF*, pages 279-286
ISBN: 978-989-758-631-6; ISSN: 2184-4305

modelling styles of biomedical concepts decreases the probability of missing attribute relationships thus ensuring complete logical definitions and superior quality of biomedical ontologies.

To provide excellent quality of healthcare delivery, maintaining consistency in the modeling styles of SNOMED-CT concepts is crucial. SNOMED-CT strives to achieve this by developing modelling templates and description patterns (IHTSDO, 2002b). SNOMED-CT documentation provides a list of description templates to model concepts conforming to a specific lexical format. Currently this is done by domain experts by filling an authoring template (IHTSDO, 2002c). The creation of the templates is undertaken as a part of the SNOMED-CT QI (Quality Improvement) Project (IHTSDO, 2019). We believe the work presented here, that develops SNOMED-CT modeling templates in a semi-automated way, can act as a complementary process to mitigate the manual effort of domain experts.

The method proposed in this paper is based on the observation that the majority of the templates created by SNOMED-CT contain stopwords (non-medical terms) in their description that indicate a relationship between two medical concepts. For instance the preposition *of*, which is usually considered to be a stopword in the processing of blocks of free natural language text, in the concept name *Neoplasm of heart* divulges significant information on the semantic relationship between the medical terms *Neoplasm* and *heart* ensuring the presence of a relationship *finding site - heart* in the concept's logical definition. We hypothesize that the process of creating SNOMED-CT templates can be automated to a large extent by targeting stopwords. The work presented here is an extension of the technique mentioned in (Burse et al., 2022b). To improve on the results obtained in (Burse et al., 2022b), we have extended the method to enhance the quality of the original research by adding lexical information that considers the beginning token of a concept name in template creation. The extended implementation has complemented the previous results by extracting a multitude of new templates, ranging from 6 templates for the stopword *of* in the original implementation to as high as 200+ templates in the extended implementation.

2 BACKGROUND

SNOMED-CT (IHTSDO, 2002a) categorizes biomedical information into 19 major hierarchies. Concepts in SNOMED-CT are represented using a unique identifier (SCTID) and a unique name called Fully Specified Name (FSN). Concepts are linked to each other using two types of relationships

- IS-A relationships, which are hierarchical in nature and represent subsumption relationships between two concepts that belong to the same hierarchy.
- Attribute relationships, which give more information about a concept by linking it to concepts from other hierarchies based on domain and range constraints defined in the logical model of SNOMED-CT.

Furthermore, SNOMED-CT contains two types of concepts

- Fully Defined (FD) concepts (IHTSDO, 2002a), which are sufficiently defined to distinguish them from other concepts.
- Primitive / Partially Defined (PD) concepts, which are not sufficiently defined. There are multiple reasons as to why a concept may be a PD concept (IHTSDO, 2002d). One of the reasons being that the attribute relationships that distinguish the concept from other concepts may not be present in its definition.

The presented method is an extension of the work (Burse et al., 2022b). To summarize, the method presented in (Burse et al., 2022b) is based on two core assumptions:

- FD concepts are more suitable for machine processing because firstly, they are sufficiently differential, i.e., the concept has at least one sufficient definition that distinguishes it from any concepts or expressions that are neither equivalent to, nor sub-types of, the defined concept (IHTSDO, 2002a); secondly, FD concepts are manually inspected by SNOMED-CT authors (IHTSDO, 2002d) to attain the FD status and therefore are more reliable than PD concepts. The higher the number of FD concepts in an ontology, the more reliable the biomedical ontology is (Schulz et al., 2009). Based on this assumption (Burse et al., 2022b) treats FD concepts as a benchmark to create templates and audit PD concepts that are lexically and semantically similar.
- Stopwords, although often typically disregarded in the processing of natural language text, are a rich source of semantic information in the highly constrained lexical structure of concept names in biomedical ontologies and can be exploited to identify missing attribute relationships in biomedical concepts.

Based on these assumptions, the method (Burse et al., 2022b) targets a list of stopwords from the PubMed stopword list[1] (Pubmed, 2022) and groups lexically similar concepts together. These concepts are then semantically analyzed by using an atomic annotator (Burse et al., 2022a). The atomic annotator labels the individual tokens in each FSN with a semantic domain tag (Burse et al., 2022a). Concepts belonging to the same lexical and semantic pattern are grouped together into sample-sets, i.e., a sample-set is a collection of SNOMED-CT concepts exhibiting the same lexical and semantic pattern. The FD sample-sets are processed using an intersection set logic to calculate the attribute relationships common to all concepts within a sample-set. These attributes are then assumed to be mandatory for that sample-set and a mandatory relationship template is created for each stopword and each sample-set within a stopword (e.g. Finding site is a mandatory attribute relationship for Disorder of Body structure (DOB). Finally, these templates are used to audit PD concepts belonging to matching sample-sets.

3 METHOD

The presented method is an extension of the work (Burse et al., 2022b). Let us refer to the method in (Burse et al., 2022b) as the Basic Algorithm (BA). To extend the BA , an additional lexical layer was applied to all sample-sets in order to extract more specific templates by sub-grouping concepts starting with the same lexical token within a sample-set (e.g. the sample-set Disorder of Body structure (DOB) creates a general template whereas Neoplasm of Body structure (DOB-Neoplasm) creates a more specific template). The skewed distribution of FD and PD templates, which was a limitation of the basic algorithm that prevented the auditing of potentially inconsistent concepts, was handled as follows (Let us refer to this work as the Enhanced Algorithm (EA)):

- If a FD template did not exist for a PD concept, the PD concept was audited against a more general template. For example, DID-hypertrichosis (DID stands for Disorder in Disorder) was audited against DID since a FD template did not exist for DID-hypertrichosis.
- Ideally sample-sets were created if more than one concept starting with the same lexical token existed within a semantic pattern but, to handle skewed template distribution, exceptional PD sample-sets with a single concept were created in case a FD template existed for them. For example, a new sample-set was created for a single PD concept *Astrocytoma of retina* (SCTID-255026000) because DOB-astrocytoma was a template in FD concepts.

An interesting observation was an increase in the number of mandatory attributes for general templates in BA after specific templates were created for subgrouped concepts within the semantic pattern in EA. For example, while the mandatory attribute template for DID (DID stands for Disorder in Disorder) included *Associated with* (SCTID - 47429007) in BA, the mandatory attribute template for DID had three attributes (*Associated morphology* (SCTID – 116676008), *Finding site* (SCTID – 363698007), and *Associated with* (SCTID - 47429007)) in EA. While this has identified additional inconsistent PD concepts in the auditing step, the accuracy of these inconsistency detections remains an issue of manual inspection. To assist the domain experts in gauging the accuracy of these suggestions, we have provided a confidence level for each of the identified inconsistency. The confidence level is denoted using three degrees, namely, low, medium, high. The fundamental reasoning behind calculating the confidence levels is that the relevance of the superfluous attributes added to general templates in EA needs to be manually inspected. We need to determine whether the presence of these attributes can be credited to an impartial intersection logic result or the additional attributes are simply common because of the elimination of other concepts, by sub-grouping, that originally belonged to this general category. The confidence level is assigned as follows.

1. If a specific concept is audited against a matching specific template then the confidence level is **high**.
2. If a general concept is audited against a matching general template then:
   (a) If at least one of the suggested attributes is present in the BA template then the confidence level is **high**. The reasoning is that BA templates capture the essence of the stopword in the majority of the cases. For example, the stopword *caused by* translated to mandatory attribute *Causative agent* (SCTID- 246075003). The stopword *in* translated to mandatory attribute *Associated with* (SCTID- 47429007), i.e., how Disorder 1 is associated with Disorder 2 in DID, thus the confidence level is higher for

---

[1] A few non-medical terms, like *following* and *caused*, were added by (Burse et al., 2022b) to the list of PubMed stop-words after observing repeating lexical patterns in SNOMED-CT concepts. The analyzed stopwords also include combinations e.g. *due to*.

this suggestion.

   (b) If none of the suggested attributes is present in the BA template then the confidence level is **medium**, owing to the doubt in superfluous attributes in EA general templates.

3. If a specific concept is audited against a more general template then:
   (a) If at least one of the suggested attributes is present in the BA template then the confidence level is **medium**.
   (b) If none of the suggested attributes is present in the BA template then the confidence level is **low**, e.g. DID-myopathy is validated against DID and the suggested missing attribute is *Associated morphology* (SCTID - 116676008). *Associated morphology* is not a mandatory attribute for DID in BA, thus the confidence level is low for this suggestion.

Table 1 summarizes the additional features added in the EA of the method. The next section discusses the improvements in the results obtained using the EA over BA.

Table 1: Summary of enhancements made in the presented method over BA.

Implementation	Description	Example
Basic Algorithm (BA)	Templates calculated as Variable–stopword–Variable. (3 exceptions DOS-abuse, DOS-overdose, and DOD-sequela)	[DIS]-GEN-of-[BOD] (DOB)[2] for "disorder of body structure"
Enhanced Algorithm (EA)	**Template Creation:** Templates calculated as Constant-stopword-Variable. The first token was scrutinized to create additional sample-sets within a semantic pattern. Skewed distribution of FD and PD templates was handled. **Auditing PD concepts:** A confidence level was added to each identified inconsistency / missing attribute suggestion in PD concepts.	Subgrouping within DOB based on the first token. E.g. Abscess-GEN-of-[BOD] (DOB-Abscess).

4 RESULTS

4.1 Template Generation

After enhancing the BA with EA, the number of extracted templates has increased for 8 out of the 11 stopwords analyzed by the method. The stagnant number of templates for stopwords *during, to,* and *into* can be ascribed to the extremely limited number of FD concepts available for analysis (6,3,3 respectively). Although the count is constant, the template for *into* changed from DITB (DITB stands for Disorder into Body structure) in BA to a more specific DITB-hemorrhage in EA. On the other hand, the templates for *during* (DDP) (DDP stands for Disorder during Procedure) and *to* (DTB) (DTB stands for Disorder to Body structure) did not change. Figure 1. illustrates the statistics with the help of a bar chart.

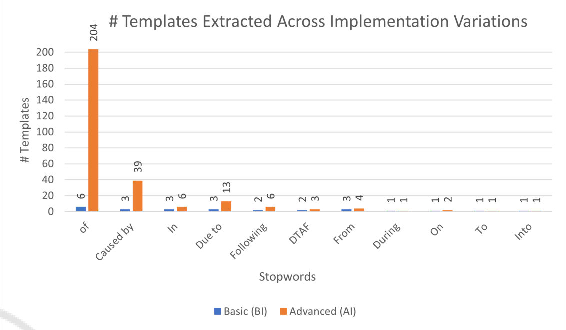

Figure 1: Number of templates extracted in EA and BA implementations of the method.

Furthermore, in EA the number of mandatory attributes detected for each specific sub-grouped template has also increased in the majority of the cases. This is logical, given the increased specificity of the new templates. The maximum number of mandatory attributes per template increased from 2 in BA to 5 in EA. Table 2 lists a few examples of general BA templates and their corresponding specific EA counterparts to demonstrate the increase in the number of mandatory attributes extracted.

It is noteworthy that in the majority of the cases, the original general templates still exist in EA, owing to a number of concepts starting with an individual lexical token that did not have any other concept starting with the same lexical token to pair/sub-group with. For example, 89 SNOMED-CT concepts (such as *Achalasia of esophagus* (SCTID- 45564002) and *Acrosyndactyly of toe* (SCTID – 699049007)) are still included in DOB sample-set despite the generation of 197 additional specific templates under DOB, i.e. DOB-firstLexicalToken. In total, only 3 of the 26 BA templates were not present in the EA results; however they were replaced by more detailed templates. More specifically, DdtO (DdtO stands for Disorder due to Object) was replaced by DdtO-injury, DOnB (DOnB stands for Disorder on Body structure) by DOnB-callosity and DOnB-ulcer, and DITB (DITB stands for Disorder into Body structure) by DITB-

---

[2] Semantic pattern from (Burse et al., 2022b)

Table 2: Comparative evaluation of the number of mandatory attributes extracted by BA and the corresponding EA counterparts,.

BA Template	Semantic pattern Description	BA Mandatory Attributes (SCTID-FSN)	EA Template	EA Mandatory Attributes (SCTID-FSN)
DFP	Disorder Following Procedure (DFP)	255234002-After	DFP-infection	255234002-After 370135005-Pathological process
			DFP-thrombophlebitis	255234002-After 363698007-Finding site 116676008-Associated morphology
DdtO	Disorder due to Object	42752001-Due to 246075003-Causative agent	DdtO-injury	42752001-Due to 246075003-Causative agent 116676008-Associated morphology
DdtD	Disorder due to disorder	42752001-Due to	DdtD- erythrocytosis	363698007-Finding site 42752001-Due to 363713009-Has interpretation 363714003-Interprets
			DdtD-hypermelanosis	116676008-Associated morphology 363698007-Finding site 42752001-Due to
DOB	Disorder of body structure	363698007-Finding site	DOB-agenesis	370135005-Pathological process 363698007-Finding site 246454002-Occurrence 116676008-Associated morphology
			DOB-abrasion	116676008-Associated morphology 363698007-Finding site 42752001-Due to
			DOB-derangement	116676008-Associated morphology 363698007-Finding site 363713009-Has interpretation 363714003-Interprets
			DOB-actinomycosis	363698007-Finding site 246075003-Causative agent 370135005-Pathological process
DCBO	Disorder caused by organism	246075003- Causative agent	DCBO-mycetoma	116676008-Associated morphology 363698007-Finding site 246075003-Causative agent 263502005-Clinical course 370135005-Pathological process

hemorrhage.

4.2 Auditing PD Concepts

Figure 2. illustrates the number of inconsistent PD concepts identified, per stopword, by each of the two implementations. The EA templates have identified a promising percentage of inconsistent PD concepts containing the stopwords *of* (49.2%), *in* (87.5%), and *due to* (72.2%). The identified PD concepts are deemed inconsistent because they do not follow the modelling styles of FD concepts, which is assumed to be the ground truth by our semi-automated method. We recommend a further manual inspection of the highlighted inconsistencies by SNOMED-CT curators. Of note, the absence of inconsistencies in PD concepts for the majority of the stopwords does not reflect poorly on the potential of the EA method. In such cases, either PD concepts were unavailable for auditing or the mandatory attributes were already present in their definitions and hence they were not flagged as inconsistent.

5 DISCUSSION

There are several differences between the templates extracted by the presented EA method vs SNOMED-CT templates created by authors. These are discussed below.

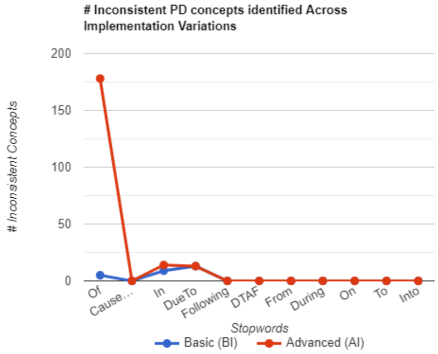

Figure 2: Number of inconsistent PD concepts identified by EA and BA templates.

Process. While the SNOMED-CT templates are manually created to restrict both FD and PD concepts to a certain structural modeling style, our semi-automated method extracts mandatory attributes by assuming FD concepts to be a source of ground truth. Thus, the quality of the extracted templates depends on the quality of the existing FD concepts. There have been cases of outlier detection within FD concepts (Burse et al., 2022b). Outliers are FD concepts that do not conform to the modelling style of the majority of the FD concepts. Such outliers were eliminated manually to avoid degrading the quality of the extracted templates by omitting mandatory attributes exhibited by most of the FD concepts. The extracted templates have identified a promising percentage of inconsistent PD concepts containing the stopwords *of*, *in*, and *due to* (Figure 2).

Level of Detail. Firstly, our templates focus only on identifying mandatory attributes based on the presence of those attributes in FD concepts conforming to a particular lexical and semantic pattern, thus the cardinality in our templates is always assumed to be 1..*, i.e., the mandatory attribute should be present at least once in the logical definition. We do not extract non-mandatory attributes with varying cardinalities or restrict the upper limit of the cardinality of an extracted mandatory attribute. For example, the SNOMED-CT template for Neoplasm of [body structure] has non-mandatory attributes like *occurrence, pathological process, causative agent, clinical course,* and *finding site* with a cardinality of 0..1, whereas our template has only extracted the mandatory attributes *associated morphology* and *finding site*, with a cardinality of 1..*, for DOB-neoplasm. The presented example further highlights one of the discrepancies found between the general rules of SNOMED-CT templates and the specific templates (IHTSDO, 2002b). While general rule 6(a) states that ”of [body structure]” should be removed from the FSN, if the attribute finding site is not present, this rule is clearly violated in many of the specific SNOMED-CT templates including the previous example, Neoplasm of [body structure]. In contrast, the templates extracted by our method abide by all the general rules (IHTSDO, 2002b).

Secondly, our templates do not account for the cardinality within a role-group. In SNOMED-CT, a role-group is an association between a set of attribute value pairs that causes them to be considered together. Role-group information is omitted for two reasons, (a) the aim of our method is to identify mandatory attributes that help make a concept FD and the presence of the attribute in at least one role-group is more important than deciding exactly which role-group. Furthermore, the role-group information would have been relevant if we had extracted other non-mandatory attributes to group them with the mandatory attributes in specific role-groups. Since non-mandatory attributes are not within the scope of our template, the role-group information is extraneous to our method. (b) The role-group number between different concepts can vary, causing the intersection logic to fail. For example, while attribute X falls in role-group 1 in concept A, the same attribute can fall under role-group 2 in concept B, thus rendering the intersection result NULL despite its mandatory nature.

5.1 Analysis of Extracted Templates

While the quantitative analysis of the extracted mandatory attributes is straightforward and can be fully automated, to ensure high standards qualitative analysis of the automatically extracted templates requires manual inspection by a domain expert. While on the one hand, the increased specificity due to subgrouping effectively captures significant details in some templates, e.g., DOnB-callosity has mandatory attributes *finding site, associated morphology,* and very specifically *causative agent* that captures the cause of callosity, i.e., friction. On the other hand, given the complex variety in SNOMED-CT concepts, sub-grouping may not always be conducive, e.g., in cases where the same template brings out necessary details for some concepts but appends unnecessary mandatory attributes for some others. This happens in cases where the mandatory attributes depend on the nature of the disorder in the FSN and just conducting a lexical and semantic analysis is not sufficient. For example, when *Myopathy in osteomalacia* (SCTID-240092003) is audited against the template DID, the suggested attribute *Associated morphology* seems to be accurate because *Osteomalacia* (SCTID- 4598005) has the attribute *Associated morphology*. But when *Myopathy in acromegaly* (SCTID-240089002) is audited against DID, neither *myopathy* nor *acromegaly* have the attribute *Associated morphology*, which makes it necessary to examine the suggested mandatory attribute for concepts in DID-myopathy sample-set.

However, on examination of the previous example, the irrelevance of the attribute suggestion, *Associated morphology*, for *Myopathy in acromegaly* (SCTID-240089002) can either be attributed to the different nature of the atomic disorders in the FSN or simply be backtracked to the individual disorders and their definition status. For example, while *Osteomalacia* (SCTID- 4598005) is a FD concept, *Acromegaly* (SCTID- 74107003) is a PD concept. Backtracking to the definition status of the atomic disorders in the FSN, the slight irrelevance in the suggested mandatory attributes for certain concepts may not reflect poorly on the quality of the extracted templates but rather be due to the fact that the individual PD concepts forming the FSN need to be scrutinized instead. If that is the case then our templates audit not only concepts falling under the sample-set but also trace back to individual disorders forming the semantic pattern.

Another interesting observation was that in most of the cases the attributes extracted for the specific sub-grouped templates in EA always include the attributes of the more general template, denoting that specificity already accounts for inheritance while defining a concept logically. To summarize, while the quantitative analysis is straightforward, the qualitative analysis of the extracted mandatory attributes would require manual inspection.

5.2 Limitations

One of the limitations of the method is the analysis of cases where PD concepts already have all the mandatory attributes in their definitions and are still assigned the PD status. Determining the reason for the PD status of such concepts is out of the scope of our semi-automated method and perhaps can be attributed to one of the reasons mentioned in (IHTSDO, 2002d). A manual inspection might be necessary to find the root cause of such incomplete definitions and what attributes need to be added in order to fully define them (IHTSDO, 2002d). Alternatively, another reasoning could be that these concepts are assigned the PD status in error, even after the presence of all mandatory attributes as per their FD counterparts, and should be assigned FD status instead. In the latter case, our templates will tremendously benefit in increasing the number of FD concepts in SNOMED-CT and thereby increasing its rate of adoption (Schulz et al., 2009).

Furthermore, although the EA method has solved the problem of the skewed distribution of FD and PD templates, this issue can only be handled if the most general sample-set exists in the FD template. For example, PD DID-myopathy concepts could be audited against DID template. However, if DID (the most general template) were absent there would not be any template to audit the skewed PD DID-myopathy concepts. A possible solution to this problem would be manually completing the definition of at least one of the PD defined concepts to extract a FD template and then using that template to audit the remaining PD concepts belonging to the matching sample-set (Burse et al., 2022b).

Finally, given the semi-automatic nature of the method, the level of in-depth analysis (as available in manually curated SNOMED-CT templates) may not always be feasible. However, extraction of the basic templates is currently the best option to standardize thousands of existing SNOMED-CT concepts into predefined templates. The basic templates extracted automatically can then be refined into sophisticated templates after manual inspection by domain experts. Although the level of templates extracted by this semi-automated method will not be as detailed as the ones that were manually created by the domain experts, this is a good start to identify repeating patterns and ascertain if there are any inconsistencies in the way these concepts are currently modeled. Complete automation of QA of biomedical ontologies will continue to be a challenge in the health-informatics domain. We believe that our contribution will aid in a complementary way to ease the manual efforts of SNOMED-CT curators.

6 CONCLUSION & FUTURE WORK

In this work, we presented an improved method to extract mandatory attribute relationship templates for SNOMED-CT concepts by considering FD concepts as a source of ground truth. The method has extracted a multitude of new templates over the basic implementation (Burse et al., 2022b). The auditing results that identified inconsistent PD concepts have shown promising potential to highlight inconsistencies in the modeling styles of lexically similar concepts. An interesting direction for future work would be resolving the atomic annotator bottleneck (Burse et al., 2022b; Burse et al., 2022a) in order to increase the coverage of SNOMED-CT concepts being analyzed. Indeed the atomic annotator restricts the number of SNOMED-CT concepts being processed based on the length of their FSNs and limits the number of atomic dictionaries created due to its semi-automatic nature. We believe the results would further improve after resolving this bottleneck.

REFERENCES

Agrawal, A. (2018). Evaluating lexical similarity and modeling discrepancies in the procedure hierarchy of snomed ct. *BMC Medical Informatics and Decision Making*, 18.

Amith, M. T., He, Z., Bian, J., Lossio-Ventura, J. A., and Tao, C. (2018). Assessing the practice of biomedical ontology evaluation: Gaps and opportunities. *Journal of biomedical informatics*, 80:1–13.

Burse, R., Bertolotto, M., and Mcardle, G. (2022a). A novel atomic annotator for quality assurance of biomedical ontologies. In *HEALTHINF*.

Burse, R., Bertolotto, M., O'Sullivan, D. M., and Mcardle, G. (2021). Semantic interoperability: the future of healthcare.

Burse, R., Mcardle, G., and Bertolotto, M. (2022b). Targeting stopwords for quality assurance of snomed-ct. *International journal of medical informatics*, 167:104870.

Ceusters, W., Smith, B., Kumar, A., and Dhaen, C. (2004). Mistakes in medical ontologies: where do they come from and how can they be detected? *Studies in health technology and informatics*, 102:145–63.

Duarte, J., Castro, S., Santos, M. F., Abelha, A., and Machado, J. (2014). Improving quality of electronic health records with snomed. In *CENTERIS 2014*.

IHTSDO (2002a). Ihtsdo snomed international confluence. https://confluence.ihtsdotools.org/. last accessed 26/10/2021.

IHTSDO (2002b). Sct modelling templates and description patterns. https://confluence.ihtsdotools.org/display/SCTEMPLATES/SCT+Modeling+Templates+and+description+patterns. last accessed 25/09/2022.

IHTSDO (2002c). Sct template specification. https://confluence.ihtsdotools.org/display/SCTEMPLATES/Template+specification. last accessed 25/09/2022.

IHTSDO (2002d). What does it mean if a concept is fully-defined or primitive and how do i tell the difference. https://ihtsdo.fresh desk.com/support/solutions/articles/4000050378-what-does-it-mean-if-a-concept-is-fully-defined-or-primitive-and-how-do-i-t. last accessed 15/12/2021.

IHTSDO (2019). Snomed ct quality improvement project. https://confluence.ihtsdotools.org/download/attachments/87042649/201936_Cathy_Richardson.pdf?version=1&modificationDate=1573141746000&api=v2. last accessed 25/09/2022.

Kim, J., Macieira, T. G. R., Meyer, S. L., Ansell, M., Bjarnadottir, R. I., Smith, M. B., Citty, S. W., Schentrup, D. M., Nealis, R. M., and Keenan, G. M. (2020). Towards implementing snomed ct in nursing practice: A scoping review. *International journal of medical informatics*, 134:104035.

Pubmed (2022). Pubmed stopwords. https://pubmed.ncbi.nlm.nih.gov/help/#help-stopwords. last accessed 30/09/2022.

Schulz, S., Suntisrivaraporn, B., Baader, F., and Boeker, M. (2009). Snomed reaching its adolescence: Ontologists' and logicians' health check. *International journal of medical informatics*, 78 Suppl 1:S86–94.

Unertl, K. M., Novak, L. L., Gadd, C. S., and Lorenzi, N. M. (2012). The science behind health information technology implementation: Understanding failures and building on successes. In *AMIA*.

Willett, D. L., Kannan, V., Chu, L., Buchanan, J. R., Velasco, F. T., Clark, J., Fish, J. S., Ortuzar, A. R., Youngblood, J. E., Bhat, D., and Basit, M. A. (2018). Snomed ct concept hierarchies for sharing definitions of clinical conditions using electronic health record data. *Applied Clinical Informatics*, 9:667 – 682.