Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220549

Implementing SNOMED CT in Open Software Solutions to Enhance the Findability of COVID-19 Questionnaires

Carina Nina VORISEK[a,1], Eduardo Alonso ESSENWANGER [a], Sophie Anne Ines KLOPFENSTEIN[a,b], Julian SASS[a], Jörg HENKE[c], Carsten Oliver SCHMIDT[c], Sylvia THUN[a], on behalf of the NFDI4Health Task Force COVID-19
[a]Berlin Institute of Health at Charité – Universitätsmedizin Berlin
[b]Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Berlin
[c]University Medicine of Greifswald

Abstract. SNOMED CT fosters interoperability in healthcare and research. This use case implemented SNOMED CT for browsing COVID-19 questionnaires in the open-software solutions OPAL/MICA. We implemented a test server requiring files in a given YAML format for implementation of taxonomies with only two levels of hierarchy. Within this format, neither the implementation of SNOMED CT hierarchies and post-coordination nor the use of release files were possible. To solve this, Python scripts were written to integrate the required SNOMED CT concepts (Fully Specified Name, FSN and SNOMED CT Identifier, SCTID) into the YAML format (YAML Mode). Mappings of SNOMED CT to data items of the questionnaires had to be provided as Excel files for implementation into Opal/MICA and further Python scripts were established within the Excel Mode. Finally, a total of eight questionnaires containing 1.178 data items were successfully mapped to SNOMED CT and implemented in OPAL/MICA. This use case showed that implementing SNOMED CT for browsing COVID-19 questionnaires is feasible despite software solutions not supporting SNOMED CT. However, limitations of not being able to implement SNOMED CT release files and its provided hierarchy and post-coordination still have to be overcome.

Keywords. SNOMED CT, Semantic Interoperability, Standardization, COVID-19

1. Introduction

The NFDI4Health Task Force COVID-19, an interdisciplinary German network project within the National Research Data Infrastructure for Personal Health Data (NFDI4Health) initiative is aiming for a better overview on public health, epidemiological and clinical studies targeting the current pandemic [1]. Therefore, three interlinked platforms were developed containing: 1) a Central Search Hub for study level information, 2) a COVID-19 study hub SEEK platform, linking COVID-19 studies with their metadata, documents (assets) and other information, 3) and Opal and MICA, containing items from selected COVID-19 survey instruments [2]. OPAL/MICA are open software solutions built for managing and harmonizing epidemiological data [3].

Using a common terminology can improve interoperability and secondary use of data within data infrastructures. Currently, SNOMED CT is the most comprehensive clinical healthcare terminology worldwide providing more than 350.000 concepts. Its organization SNOMED International comprises 41 member countries [4]. The three components "concepts", "descriptions" and "relationships" enable SNOMED CT's features such as compositional grammar, expression constraint queries and post-coordination offering possibilities to combine SNOMED CT concepts [5]. In MICA, a simple semantic structure to classify study items is available, the Maelstrom taxonomy, which is composed of 18 domains and 135 subdomains [6].

To provide a faceted search of COVID-19 questionnaires, we aimed to implement SNOMED CT in addition to the already used Maelstrom taxonomy into MICA/OPAL, to further improve the findability of COVID-19 questionnaires. This study evaluates whether SNOMED CT can be implemented into open software solutions, which were initially not focused on the support of SNOMED CT.

2. Methods

As a first step, we mapped SNOMED CT concepts to data items of COVID-19 questionnaires already stored in OPAL/MICA as described previously [7]. Next, a test server based on OPAL/MICA was implemented for our use case. As OPAL/MICA requires files in a given YAML format for implementation of taxonomies and allows only two levels of hierarchy, using SNOMED CT release files for import was not possible. Therefore, a Python script was written to integrate the required SNOMED CT concepts (Fully Specified Name, FSN and SNOMED CT Identifier, SCTID) into the YAML format (YAML Mode). In addition, OPAL/MICA require specific Excel file formats for implementing questionnaires and their mappings to standards which is targeted by the Excel mode of the Python Script. Mapping files, Excel formats and the Python script are provided on GitHub [8].

3. Results

3.1. Mapping to SNOMED CT

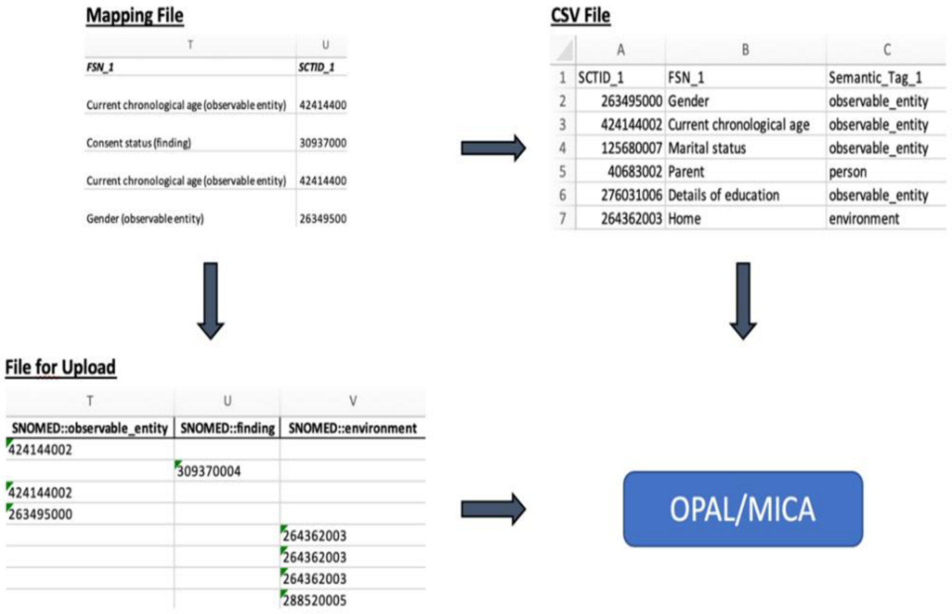
A total of eight questionnaires containing 1.178 data items were mapped to SNOMED CT and then stored in OPAL/MICA. OPAL/MICA require specific Excel file formats allowing only one SNOMED CT concept mapped to one data item. The mapping limitations are presented in detail in **table 1**.

Table 1. Limitations of Mapping SNOMED CT concepts to Data Items of the Questionnaires

Data Item	Does the patient suffer from a cardiovascular disease? History of heart revascularization
Current mapping in OPAL/MICA	Disorder of cardiovascular system (disorder)
Possible mapping using post-coordination	|Past history of procedure (situation)|:|Associated procedure (attribute)|=|Heart revascularization (procedure)|

3.2. Python Script - Excel mode:

The Excel mode (**Figure 1)** read the */Files* folder and looped over all its Excel files. Every file contained “FSN_X” and “SCTID_X” columns (X = non-negative value). The script filtered out the semantic tags listed on the “FSN_X” column and created a “SNOMED::X” column with “X” depending on the semantic tag for all semantic tags. The prefix SNOMED is defined in the taxonomy definition given to OPAL/MICA that is the precondition before any mappings can be uploaded completely. The values of the “SCTID_X” column was mapped to the corresponding “SNOMED::X” column. Finally, the script created Excel files in the required format in the */Output* folder.



Mapping File

T	U
FSN_1	*SCTID_1*
Current chronological age (observable entity)	42414400
Consent status (finding)	30937000
Current chronological age (observable entity)	42414400
Gender (observable entity)	26349500

CSV File

	A	B	C
1	SCTID_1	FSN_1	Semantic_Tag_1
2	263495000	Gender	observable_entity
3	424144002	Current chronological age	observable_entity
4	125680007	Marital status	observable_entity
5	40683002	Parent	person
6	276031006	Details of education	observable_entity
7	264362003	Home	environment

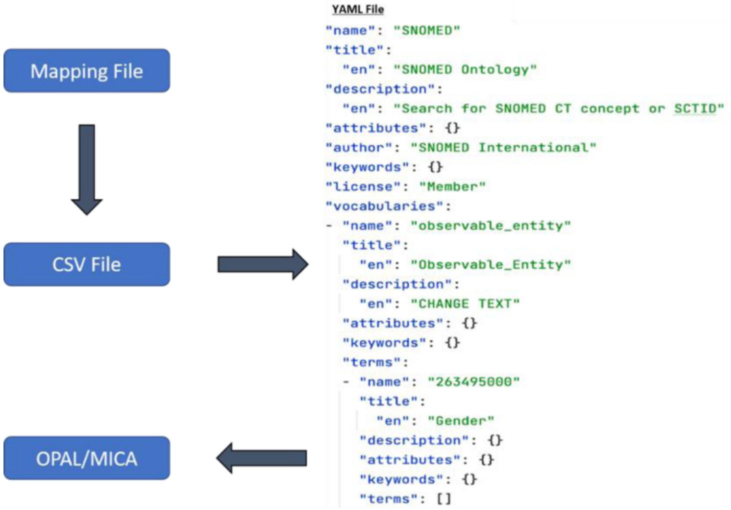
File for Upload

T	U	V
SNOMED::observable_entity	SNOMED::finding	SNOMED::environment
424144002		
	309370004	
424144002		
263495000		
		264362003
		264362003
		264362003
		288520005

Figure 1. Shows the Workflow of the Excel Mode in the Python Script.

3.3. Python Script - YAML mode:

This mode created the YAML file (**Figure 2**) in a given format required by OPAL/MICA. Excel files located in the */Files* folder were read and merged into one CSV file with the following headers “SCTID_X”; “FSN_X”; “Semantic_Tag_X”. The YAML file was created together with the csv2yaml.py file and finally saved in the */Output* folder.

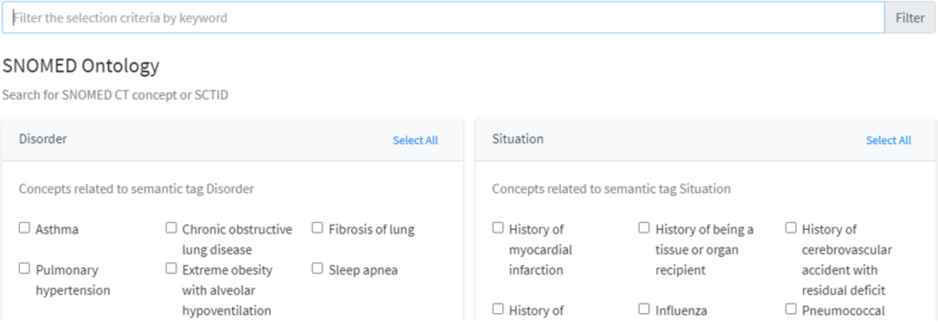


```
"name": "SNOMED"
"title":
  "en": "SNOMED Ontology"
"description":
  "en": "Search for SNOMED CT concept or SCTID"
"attributes": {}
"author": "SNOMED International"
"keywords": {}
"license": "Member"
"vocabularies":
- "name": "observable_entity"
  "title":
    "en": "Observable_Entity"
  "description":
    "en": "CHANGE TEXT"
  "attributes": {}
  "keywords": {}
  "terms":
  - "name": "263495000"
    "title":
      "en": "Gender"
    "description": {}
    "attributes": {}
    "keywords": {}
    "terms": []
```

Figure 2. Shows the Workflow of the YAML Mode in the Python Script.

3.4. Display of Results in OPAL/MICA

User experience was improved by categorizing the SNOMED CT concepts by their semantic tag utilizing the two level hierarchy option in OPAL/MICA. SCTIDs were provided in the background, therefore browsing was possible using the FSN as well as the SCTID. Implementation results in OPAL/MICA are provided in **figure 3**.



SNOMED Ontology

Search for SNOMED CT concept or SCTID

Disorder

Concepts related to semantic tag Disorder

☐ Asthma ☐ Chronic obstructive lung disease ☐ Fibrosis of lung
☐ Pulmonary hypertension ☐ Extreme obesity with alveolar hypoventilation ☐ Sleep apnea

Situation

Concepts related to semantic tag Situation

☐ History of myocardial infarction ☐ History of being a tissue or organ recipient ☐ History of cerebrovascular accident with residual deficit
☐ History of ☐ Influenza ☐ Pneumococcal

Figure 3. SNOMED CT displayed in OPAL/MICA.

4. Discussion and Conclusion

To our knowledge, faceted search using SNOMED CT in open software solutions has not been implemented nor tested. This use case showed that implementing SNOMED CT for browsing questionnaires related to COVID-19 studies is feasible despite software solutions initially not designed to support SNOMED CT. However, we faced several limitations using OPAL/MICA: We needed to adhere to specific formatting requirements allowing only one SNOMED CT concept per data item consequently inhibiting

SNOMED CT's grammar and post-coordination. Due to the required YAML format when importing new taxonomies into OPAL/MICA, the use of SNOMED CT's release files was not possible. Converting between these different file formats manually fostered slow working processes, enormous workload and most likely human error.

Therefore, Python Scripts were written to transfer the Excel files established during the mapping process into Excel files implementing the annotated questionnaires (Excel Mode) and SNOMED CT as taxonomy into OPAL/MICA (YAML Mode). With the help of our Python Scripts, we created an automated workflow from Mapping to Implementation. To further integrate SNOMED CT features such as querying, grammar and post-coordination into OPAL/MICA, the software applications need to be adapted and terminology servers need to be established.

This study demonstrates the implementation of SNOMED CT in open software solutions often used in academic institutions, smaller healthcare facilities as well as low- and middle-income countries with limited financial opportunities. Therefore, our novel approach using terminologies in open-source environment can enhance the success of their implementation. In addition, we aim to extend our use-case beyond COVID-19 studies and hence elevate the overall findability of study items from clinical, epidemiological and public health studies.

5. Acknowledgements

This work was done as part of the NFDI4Health Task Force COVID-19 (www.nfdi4health.de). We gratefully acknowledge the financial support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project Number 451265285, PI 345/17-1/SCHM 2744/9-1.

References

[1] Task Force COVID-19 - NFDI4Health. https://www.nfdi4health.de/de/task-force-covid-19/ (accessed August 11, 2021).

[2] Schmidt CO, Darms J, Shutsko A, et al. Facilitating Study and Item Level Browsing for Clinical and Epidemiological COVID-19 Studies, Stud. Health Technol. Inform. 281 (2021) 794–798.

[3] Doiron D, Marcon Y, Fortier I, et al. Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination., Int. J. Epidemiol. 46 (2017) 1372–1378.

[4] SNOMED - Members. https://www.snomed.org/our-customers/members (accessed January 18, 2022).

[5] SNOMED CT Starter Guide - SNOMED CT Starter Guide - SNOMED Confluence. https://confluence.ihtsdotools.org/display/DOCSTART/SNOMED+CT+Starter+Guide?preview=/28742871/47677485/doc_StarterGuide_Current-en-US_INT_20170728.pdf (accessed January 18, 2022).

[6] Bergeron J, Doiron D, Marcon Y, et al. Fortier, Fostering population-based cohort data discovery: The Maelstrom Research cataloguing toolkit, PLoS One. 13 (2018) e0200926.

[7] Vorisek CN, Klopfenstein SI, Sass J, et al. Evaluating Suitability of SNOMED CT in Structured Searches for COVID-19 Studies, Stud. Health Technol. Inform. 281 (2021) 88–92.

[8] BIH-CEI/SNOMEDCT_OPAL. https://github.com/BIH-CEI/SNOMEDCT_OPAL (accessed January 18, 2022).