Semi-Automated Approach to Retrieve SNOMED CT Hierarchy of Clinical Terms by Using Terminology Server

Abdul Mateen RAJPUT [a,1], Karen TRIEP [a] and Olga ENDRICH [a]

[a]*Medical Directorate, Medizincontrolling, Inselspital, University Hospital Bern, Insel Gruppe, Bern, Switzerland*

Abstract. SNOMED CT has an enormous number of clinical concepts and mapping to SNOMED CT is considered as the foundation to achieve semantic interoperability in healthcare. Manual mapping is time-consuming and error-prone thus making this crucial step challenging. In addition, hierarchy retrieval of clinical concepts increases the challenges for the user. Terminology Servers provide an interface, which can be used to automate the process of retrieving data. In this work, it is shown that Snowstorm can significantly improve the efficiency of retrieval process if used with semi-automated workflows.

Keywords. SNOMED CT, Snowstorm, Terminology Server, Semantic Interoperability, KNIME, RESTful API Methods

1. Introduction

SNOMED CT [1], a clinical terminology that provides a consistent method for describing health conditions and related information, is widely used by healthcare professionals globally to share standardized electronic health data. Its implementation has greatly helped in digitizing the healthcare system by increasing semantic interoperability.

SNOMED CT can assist in a variety of tasks, from data entry to clinical decision support [2], which can be divided into terminology services and recording services. Snowstorm, a free and open-source terminology server [3] developed by SNOMED International , allows for quick implementation and provides a way to access various features of SNOMED CT. One important feature is Expression Constraint Language (ECL). ECL is a declarative language, which can be used to construct logical expressions to find subset of data.

With regard to interoperability the use of SNOMED CT to encode clinical concepts is important and various tools and techniques have been proposed in literature [4]. In addition to encoding, the classification of the concepts can also be used to group similar concepts for analytical and presentation purposes. For example, subsuming gram-negative bacteria for analyzing the incidence of related disorders. For subsumption the hierarchical data of microorganisms must be made available. This work aims at analyzing

---

[1] Corresponding Author: Abdul-Mateen Rajput, Insel Data Science Center, Inselspital, University Hospital Bern, Insel Gruppe, Bern, Switzerland; E-mail: Abdul.Mateen@insel.ch

the feasibility of SNOMED CT and offered tools to retrieve hierarchical data of microorganisms in order to employ the concepts for clinical decision support and data analysis in future.

Since the approach required querying a terminology server endpoint repeatedly, an automated querying approach was devised by using KNIME (Konstanz Information Miner) [5] . KNIME is an open-source and free data analytics platform that provides a graphical user interface to develop workflows. It has been used for various automated tasks [6–8]. In the current work, we used it to query the web service of Snowstorm by using an automated method.

2. Methods

The University Hospital of Bern, Inselspital, provided the list of microbes associated with clinical findings, stored in the Clinical Data Warehouse (Insel Data Platform). There were more than one thousand concepts covering microorganisms, microbial test methods, and microbial resistant tests. The corpus of 30 microorganisms was chosen initially to test the method.

2.1. Dataset preparation

The dataset contains local identifiers and microorganisms names used in the microbiology laboratory. Only microorganisms were chosen whose names were in standard English medical language excluding some with German names.

The microorganisms were mapped to SNOMED CT identifiers to:

1) Avoid getting more results as one microorganism name could be associated with more than one identifier. Figure 1 shows an example of "Helicobacter Pylori" with two associated terms, even when only "Descendants or Self of" expression of Organism is chosen.

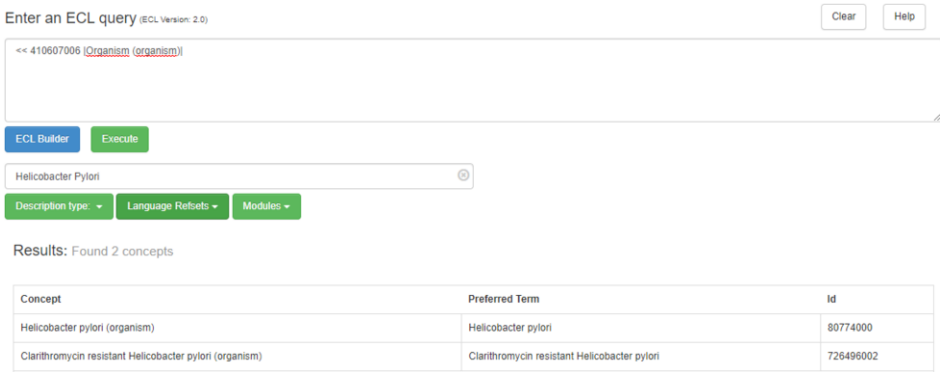2) Get the automated results from the web service since ECL only works with identifiers and not with text terms.

Figure 1: Retrieving all descendants of Organism and filtering organisms with concept "Helicobacter Pylori"

2.2. Web service description and mapping

The 30-June-2022 edition of SNOMED CT with the web service running at [9] was used to query the concepts. The description of different operations is listed on the Swagger interface page. Snowstorm offers various features to improve search results e.g. constraining search to a particular semantic tag i.e. Organism, Disorder, etc.

The following is an example of a query that constrains the results to only terms that are children of concept "Organism" and has Helicobacter Pylori in their names:

https://snowstorm.ihtsdotools.org/fhir/ValueSet/$expand?url=http://snomed.info/s ct?fhir_vs=ecl/<410607006&filter=Helicobacter%20Pylori

Similarly, the microorganisms were searched by using the query and SNOMED CT identifiers linked with the names were retrieved.

2.3. Manual validation and hierarchy retrieval

After executing the above query, whose results are identical to SNOMED CT browser results as shown in figure 1, the selection of only one correct concept was carried out manually. The concept 80774000 |Helicobacter pylori (organism) was chosen from the example query above.

After retrieving the local concepts mapped to SNOMED CT identifiers and manual mapping to correct terms, the final step was to retrieve the hierarchy of those microorganisms. This was achieved by using the ECL expression of "Ancestor of" in order to extract the top-level hierarchy.

A KNIME workflow, similar to the one described in [10], was developed to automate web service querying, saving, and parsing the output from Snowstorm terminology server. The workflow reads the microorganisms list, starts the loop, generates and runs the queries, stores the output, and parses the retrieved JSON files after ending the loop. In addition, it also filters the output to extract the data of interest.

3. Results

Four hundred twenty nine SNOMED CT terms were retrieved from the dataset of thirty microorganisms. Figure 2A shows the SNOMED CT browser results of ECL expression "Ancestor of 80774000 | Helicobacter Pylori" and the same results retrieved from web service (Figure 2B). The advantage of the web service is that the process of retrieving data can be automated in contrast to manual work using a graphical user interface.
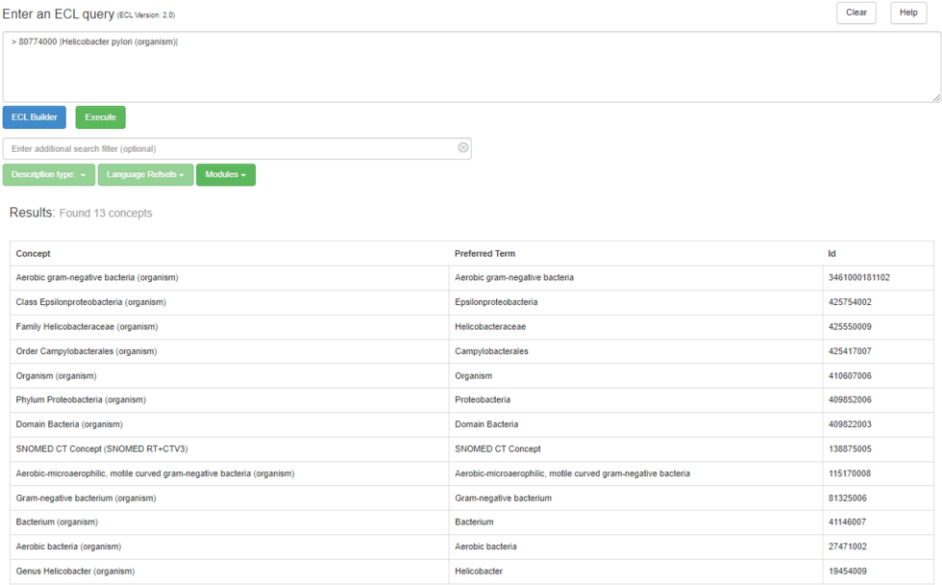
Enter an ECL query (ECL Version: 2.0)   Clear   Help

> 80774000 |Helicobacter pylori (organism)|

[ECL Builder] [Execute]

Enter additional search filter (optional)   ⊗

[Description type: ▼] [Language Refsets ▼] [Modules ▼]

Results: Found 13 concepts

Concept	Preferred Term	Id
Aerobic gram-negative bacteria (organism)	Aerobic gram-negative bacteria	3461000181102
Class Epsilonproteobacteria (organism)	Epsilonproteobacteria	425754002
Family Helicobacteraceae (organism)	Helicobacteraceae	425550009
Order Campylobacterales (organism)	Campylobacterales	425417007
Organism (organism)	Organism	410607006
Phylum Proteobacteria (organism)	Proteobacteria	409852006
Domain Bacteria (organism)	Domain Bacteria	409822003
SNOMED CT Concept (SNOMED RT+CTV3)	SNOMED CT Concept	138875005
Aerobic-microaerophilic, motile curved gram-negative bacteria (organism)	Aerobic-microaerophilic, motile curved gram-negative bacteria	115170008
Gram-negative bacterium (organism)	Gram-negative bacterium	81325006
Bacterium (organism)	Bacterium	41146007
Aerobic bacteria (organism)	Aerobic bacteria	27471002
Genus Helicobacter (organism)	Helicobacter	19454009

Figure 2A: Search results of "Ancestor of Helicobacter Pylori" via graphical user interface of SNOMED CT Browser

S code	S codes	S displays
80774000	3461000181102	Aerobic gram-negative bacteria
80774000	425754002	Epsilonproteobacteria
80774000	425550009	Helicobacteraceae
80774000	425417007	Campylobacterales
80774000	410607006	Organism
80774000	409852006	Proteobacteria
80774000	409822003	Domain Bacteria
80774000	138875005	SNOMED CT Concept
80774000	115170008	Aerobic-microaerophilic, motile curved gram-negative bacteria
80774000	81325006	Gram-negative bacterium
80774000	41146007	Bacterium
80774000	27471002	Aerobic bacteria
80774000	19454009	Helicobacter

Figure 2B: Search results of "Ancestor of Helicobacter Pylori" via SNOWSTORM Terminology Server Web service

The web service query used to retrieve above-mentioned result (figure 2B) is the following:

https://snowstorm.ihtsdotools.org/fhir/ValueSet/$expand?url=http://snomed.info/s ct?fhir_vs=ecl/>80774000

In contrast to GUI results from SNOMED CT browser, only preferred terms are available in the Snowstorm FHIR server. Figure 3 shows the raw output JSON file retrieved after executing the above-mentioned query.

```
{
  "resourceType" : "ValueSet",
  "url" : "http://snomed.info/sct?fhir_vs=ecl/>726496002",
  "expansion" : {
    "total" : 16,
    "offset" : 0,
    "contains" : [ {
      "system" : "http://snomed.info/sct",
      "code" : "3461000181102",
      "display" : "Aerobic gram-negative bacteria"
    }, {
      "system" : "http://snomed.info/sct",
      "code" : "726493005",
      "display" : "Clarithromycin resistant bacteria"
    }, {
      "system" : "http://snomed.info/sct",
      "code" : "425754002",
      "display" : "Epsilonproteobacteria"
    }, {
      "system" : "http://snomed.info/sct",
      "code" : "425550009",
      "display" : "Helicobacteraceae"
```

Figure 3: Raw JSON Output from the SNOWSTORM FHIR Server

4. Discussion and Conclusion

Mapping local concepts to SNOMED CT is a crucial step to improve semantic interoperability but it is complex and time-consuming. Furthermore, requiring hierarchical data makes it even more challenging. Terminology servers, by providing APIs, offer features to reduce the manual work by systematic approaches. In current work, using an automated approach, microbiological data was retrieved from Snowstorm terminology server. Even though the microbiological dataset was chosen, this method can be applied to any type of clinical dataset.

The results retrieved from this approach were identical to what one can get from the SNOMED CT browser but with far less time and efforts thus making this approach much more efficient. During the work, several challenges had to be faced e.g. manually selecting the correct SNOMED CT code associated with microorganism names, and the amount of data that had to be reviewed manually. In future, further exploration of this approach could be done by using various features of Snowstorm API.

References

[1] SNOMED - 5-Step Briefing. SNOMED, https://www.snomed.org/snomed-ct/five-step-briefing (accessed 27 January 2022).
[2] 8. SNOMED CT Implementation - SNOMED CT Starter Guide - SNOMED Confluence, https://confluence.ihtsdotools.org/display/docstart/8.+snomed+ct+implementation (accessed 27 January 2022).
[3] IHTSDO/snowstorm. Java, SNOMED International, https://github.com/IHTSDO/snowstorm (2022, accessed 19 January 2022).
[4] Gaudet-Blavignac C, Foufi V, Bjelogrlic M, et al. Use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for Processing Free Text in Health Care: Systematic Scoping Review. J Med Internet Res 2021; 23: e24594.
[5] Open for Innovation. KNIME, https://www.knime.com/open-for-innovation-0 (2023, accessed 17 March 2023).
[6] Rajput AM. Semi-Automated Approach to Populate any FHIR Server via POST Method. Stud Health Technol Inform 2021; 281: 1114–1115.
[7] Rajput AM, Gurulingappa H. Semi-automatic Approach for Ontology Enrichment Using UMLS. Procedia Computer Science 2013; 23: 78–83.
[8] Rajput AM. Semi-Automated Method to Generate Simulated Clinical Data from OpenEHR Platform - Think!EHR. Stud Health Technol Inform 2020; 275: 232–233.

[9] Swagger UI, https://browser.ihtsdotools.org/snowstorm/snomed-ct/swagger-ui.html (accessed 27 January 2022).

[10] Rajput AM, Triep K, Endrich O. Semi-Automated Approach to Map Clinical Concepts to SNOMED CT Terms by Using Terminology Server. Stud Health Technol Inform 2022; 293: 67–72.