

Semi-Automated Approach to Map Clinical Concepts to SNOMED CT Terms by Using Terminology Server

Abdul Mateen RAJPUT ^{a,1}, Karen TRIEP ^a and Olga ENDRICH ^a

^aMedical 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. Terminology Servers provide an interface, which can be used to automate the process of retrieving data. Snowstorm is a terminology server developed by SNOMED International. In this work, the feasibility of using Snowstorm to automate the data retrieval and mapping has been discussed.

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

1. Introduction

SNOMED CT [1] is the most comprehensive clinical terminology that provides a uniform way of describing health conditions and their associated information. It is used by healthcare professionals around the world to exchange standardized electronic health information. It has facilitated the digitization of the healthcare system significantly by improving semantic interoperability. In 2016, it was considered as the best core reference terminology for eHealth deployments in Europe [2].

The implementation of SNOMED CT could help in a wide range of tasks, from data entry to clinical decision support [3]. The tasks can be divided into terminology services and recording services. Terminology services offer functions to data irrespective of their references [3]. Snowstorm is a free Terminology server [4] developed by SNOMED International. It is open source and can be implemented in a short time. It provides an interface where diverse features of SNOMED CT can be explored. Many instances of Snowstorm are operational around the globe [5–7]. For the current research, an instance running on the IHTSDO server [5] was used.

The encoding of clinical concepts to SNOMED CT has been an important topic and a recent review described many different tools and methods applied so far [8]. However, almost all of the papers discussed in the review only share the findings of the processing of free text and not the structured text. To the author's knowledge, this is the first mapping approach where structured text and a terminology server have been used.

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 objective of this work is to encode all the clinical concepts to SNOMED CT and use the encoded data for clinical decision support and data analytics, as the first step microbiological associated concepts were chosen.

Since there were hundreds of concepts, an automated querying approach was devised by using KNIME (Konstanz Information Miner) [9]. 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 [10–13]. 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 concepts regarding microbiological clinical findings, stored in 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 366 microorganisms was chosen initially to test the method. Our setup is a local Windows 10 machine with KNIME version 4.5.1.

2.1. Data preparation

The dataset contains local identifiers and concepts name used in the microbiology laboratory. Most microorganisms' names were in standard English medical language, however, some of them were in German (Table 1), which have to be translated into English first as there is no German version of SNOMED CT available.

2.2. Web service description

The International Edition, MAIN, of SNOMED CT with the web service running at [5] was used. The description of different endpoints and operations is listed on the same Swagger interface page. The Snowstorm offers various features to improve search results e.g. search can be constrained to a particular semantic tag i.e. Organism, disorder, etc. Among other filters, there is also a possibility to search only concepts which are currently “active”. The property “active” separates the concepts, which are intended for current use, from “inactive” concepts [14].

2.3. KNIME workflow

A KNIME workflow was developed to automate the whole process of querying, saving, and parsing the output from the Snowstorm Terminology Server. The workflow reads the concepts list, starts the loop, generates and runs the queries, and saves and parses the retrieved JSON files after ending the loop. In addition, it also filters the output to extract data of interest. Following is the detailed description of each node of the workflow, shown in figure 1:

- **Table Creator:** This node creates an empty data table, which can be populated with text like a spreadsheet table. The dataset of source terms was loaded in this node.

- String Manipulation: This node modifies the text and offers many options to do that. For example, search and replace, capitalize or remove leading and trailing white spaces.
- Table Row to Variable Loop Start: It converts rows coming from the output of the previous node to variables; makes it easier to run through a loop over them. In our case, these were rows containing the names of microorganisms.
- Java Edit Variable (simple): This node provides an interface to write simple java codes.
- Variable to Table Row: This node converts variable names to table rows.
- Get Request: This node uses a Get method to a given server with various configurations, such as time delay between requests made by the Get method etc.
- Loop End: Loop End runs a loop over the total number of rows, which are indexed by “Table Row to Variable Loop Start” and ends the loop.
- JSON Path: This node offers query language for JSON, similar to XPath for XML.
- Ungroup: This node creates a list for each list of collection values.
- Rule-based Row Filter: This node takes a list of user-defined rules and tries to match them to each row.

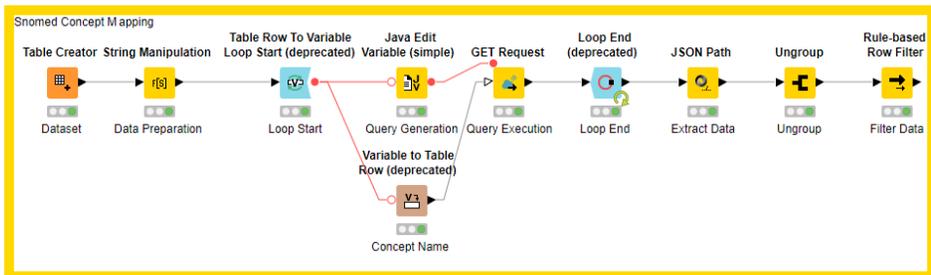


Figure 1. KNIME workflow shows the data processing by different nodes. The workflow starts from left and green sign shows that the workflow has been executed successfully. The deprecated nodes are updated in the latest versions of KNIME but they still functions.

3. Results

Out of 366 clinical concepts, 320 of them were unique, 46 concepts were repetitions of the same concepts with a different local identifier e.g. Hafnia alvei has two local identifiers; hal and halv. The workflow retrieved mapped SNOMED CT data of 303 unique concepts successfully. Since some clinical concepts were duplicated due to a different local identifier, the output has 583 rows with SNOMED CT codes and fully specified names (figure 3). The 63 concepts did not retrieve any SNOMED CT data at all; analyzing them revealed that most of the concepts were in the German language (see Table 1).

Table 1. The list of concepts, which did not retrieve any data from the Snowstorm web service.

List of concepts	Reason for unsuccessful data retrieval
Leptotrichia spp.	spp.
Bacillus cereus/thuringiensis	/
Bacteroides fragilis Gruppe	Gruppe
Coryneforme Bakterien	Bakterien
Enterobacter cloacae Komplex	Komplex
Grampositive Stäbchen	Stäbchen

The successful execution of the query yielded a JSON file parsed in the workflow. Figure 2 shows the JSON file of the query where the concept “*Actinomyces europaeus*”

```

"conceptId": "113414004",
"active": true,
"definitionStatus": "PRIMITIVE",
"moduleId": "900000000000207008",
"effectiveTime": "20020131",
"fsn": {
  "term": "Gleimia europaea (organism)",
  "lang": "en"
}

```

Figure 2. The JSON output of the query executed at Snowstorm Terminology server.

is used. The retrieved files contain various properties of the concept. The Concept ID contains the identifier of the concept. As discussed above, the active property shows whether the concept is intended for current usage by showing “True”. If it is not intended for the current usage then “False” is mentioned next to the active tag [14]. The active property is there because over the period, concepts become obsolete as new concepts take their place. The FSN tag shows the Fully Specified Name of the concept, which is a unique description in SNOMED CT.

The success of retrieval was mixed in the results. Some of the queries did not retrieve any data due to a typographical error or because there was a German word in the source concept. Others retrieved data of exactly one SNOMED CT term, which was identical to the concept.

Few queries retrieved a lot of the terms, which have to be manually filtered. This happens because a term is available in different contexts and has different Semantic tags i.e. Organism, disorder, etc. For such cases, queries can be constrained by adding such Semantic tags so only output containing associated metadata can be retrieved [16].

S column1	S idAndFsnTerms	B active
Actinomyces%20europaeus	113414004 <u>Gleimia europaea (organism) </u>	true
Actinomyces%20funkei	419012004 Actinomyces funkei (organism)	true
Aerococcus%20viridans	78803006 Aerococcus viridans (organism)	true
Bjerkandera%20adusta	415884002 Bjerkandera adusta (organism)	true
Enterococcus%20avium	87875008 Enterococcus avium (organism)	true
Geotrichum%20species	34324005 <u>Genus Geotrichum (organism) </u>	true

Figure 3. The parsed JSON file with the highlighted retrieved FSN terms. The highlighted terms shows that the concept queried (left) was one of the synonyms of the retrieved FSN term.

Figure 3 shows that a term was retrieved even though it was not identical to the concept queried. This occurs because the concept used in the query was one of the synonyms of the FSN term.

In one of the rare examples, it was discovered that not all the terms available at Snowstorm web service are identical to the SNOMED CT browser [17]. For example, figure 4 shows the concept of “Actinomyces georgiae” returns the FSN “Schaalia georgiae (organism)” from the web service but on the SNOMED browser the code has a different FSN i.e. “Actinomyces georgiae (organism)”. However, the code associated with both terms is the same i.e. “113415003”.

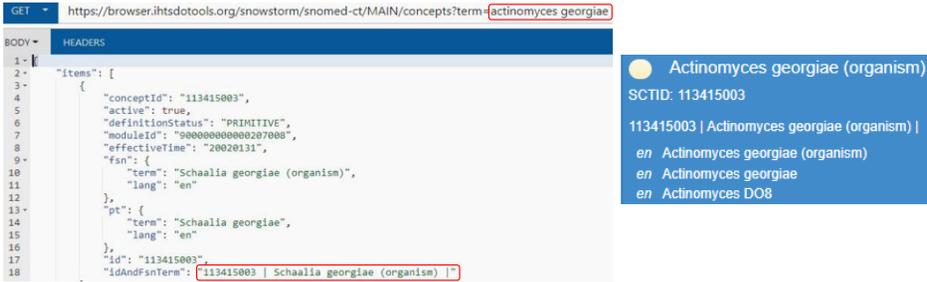


Figure 4. Different FSN terms retrieved from Snowstorm web service and Snomed browser. Left: Data retrieved from the Snowstorm web service. Right: Data retrieved from the SNOMED CT browser.

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. Terminology servers, by providing APIs, offer features to do it with systematic approaches and reduce the manual work. In this work, Microbiological data was retrieved from a Terminology Server (Snowstorm) provided by SNOMED International by using an automated approach with KNIME. Even though the Microbiological dataset was chosen, this method can be applied to any type of clinical dataset.

The quality of data retrieved from Snowstorm was satisfactory as manual validation, by using the SNOMED CT browser, was performed after the SNOMED CT data was retrieved by using this approach. During the work, there were challenges because the language used in the local laboratory was not English, the amount of data retrieved also required manual reviews, and Snowstorm seems to have a different version of SNOMED CT as different data was retrieved from the SNOMED CT browser as compared to Snowstorm web service. Further exploration of this approach could be done by using various features offered by Snowstorm API.

References

[1] SNOMED - 5-Step Briefing. SNOMED, <https://www.snomed.org/snomed-ct/five-step-briefing> (accessed 27 January 2022).

[2] Rossander A, Lindsköld L, Ranerup A, et al. A State-of-the Art Review of SNOMED CT Terminology Binding and Recommendations for Practice and Research. *Methods Inf Med* 2021; **60**: e76–e88.

- [3] 8. SNOMED CT Implementation - SNOMED CT Starter Guide - SNOMED Confluence, <https://confluence.ihtsdotools.org/display/docstart/8.+snomed+ct+implementation> (accessed 27 January 2022).
- [4] IHTSDO/snowstorm. Java, SNOMED International, <https://github.com/IHTSDO/snowstorm> (2022, accessed 19 January 2022).
- [5] Swagger UI, <https://browser.ihtsdotools.org/snowstorm/snomed-ct/swagger-ui.html> (accessed 27 January 2022).
- [6] Swagger UI, <https://snowstorm.msal.gov.ar/swagger-ui.html> (accessed 27 January 2022).
- [7] Swagger UI, <https://snowstorm.test-nictiz.nl/swagger-ui.html> (accessed 27 January 2022).
- [8] Gaudet-Blavignac C, Foufi V, Bjelogrić 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.
- [9] Open for Innovation. KNIME, <https://www.knime.com/open-for-innovation-0> (accessed 25 January 2022).
- [10] Rajput AM. Semi-Automated Method to Generate Simulated Clinical Data from OpenEHR Platform - Think!EHR. *Stud Health Technol Inform* 2020; **275**: 232–233.
- [11] Rajput AM. Semi-Automated Approach to Populate any FHIR Server via POST Method. *Stud Health Technol Inform* 2021; **281**: 1114–1115.
- [12] Rajput AM, Gurulingappa H. Semi-automatic Approach for Ontology Enrichment Using UMLS. *Procedia Computer Science* 2013; **23**: 78–83.
- [13] Rajput AM, Gronke A, Johannis W. Semi-automated approach to validate and enrich LOINC codes by FHIR Server. In: *2021 5th International Conference on Medical and Health Informatics*. New York, NY, USA: Association for Computing Machinery, pp. 304–307.
- [14] active - SNOMED CT Glossary - SNOMED Confluence, <https://confluence.ihtsdotools.org/display/DOCGLOSS/active> (accessed 16 March 2022).
- [15] 3.1.4 Meaning of the Active Field - Release File Specification - SNOMED Confluence, <https://confluence.ihtsdotools.org/display/DOCRELFMT/3.1.4+Meaning+of+the+Active+Field> (accessed 24 January 2022).
- [16] semantic tag - SNOMED CT Glossary - SNOMED Confluence, <https://confluence.ihtsdotools.org/display/DOCGLOSS/semantic+tag> (accessed 27 January 2022).
- [17] SNOMED International Browser, <https://browser.ihtsdotools.org/?> (accessed 27 January 2022).