Validation and simplification of expression constraints

V. Giménez, J.A. Maldonado, D. Boscá, D. Moner | VERATECH FOR HEALTH

SNOMED International
SNOMED CT EXPO 2020
Virtual Conference October 8-9

Abstract

We present some advanced functionalities, which are integrated in our Expression Constraint execution engine, in order to validate and optimize them, such as MRCM validation, pre- and post-execution simplification and visualization and validation of results.

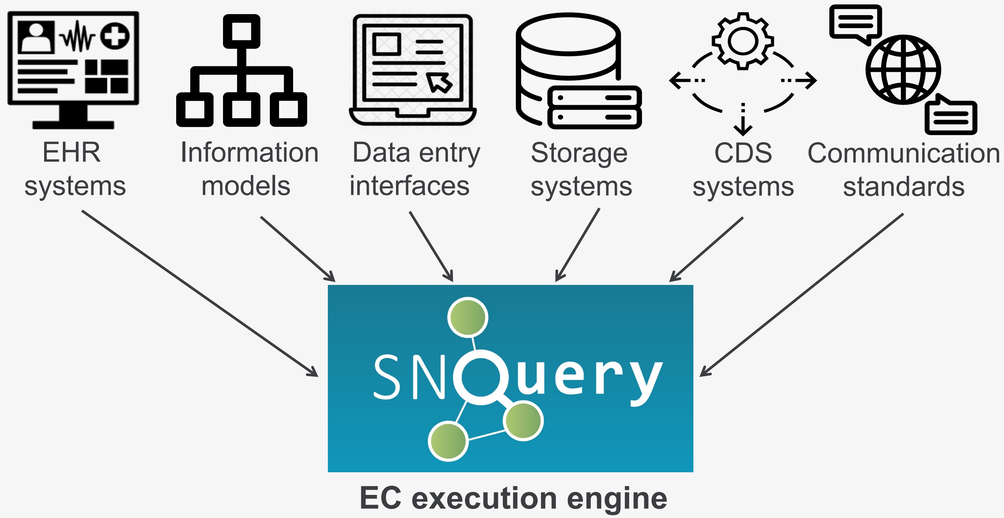
Introduction

SNOMED CT Expression Constraint Language (ECL) [1] is a language developed by **SNOMED International** for the definition of SNOMED CT **Expression Constraints** (ECs). ECs are executable expressions that define bounded sets of clinical meanings intensionally.

The intensional definition of sets of SNOMED CT concepts (henceforth, subsets) is important in several uses cases. For example, subsets are used in **terminology binding** between a clinical information model artifact and a terminology artifact, for instance to define the set of valid values of codified data.

Additionally, ECL is used in the definition of the **SNOMED CT Machine Readable Concept Model** (MRCM).

In order to support the ECL it is required an **EC execution engine** that should query an SNOMED CT database. Potential users of such engine include:



EC execution engine

In order to use SNQuery EC engine, we have implemented a web platform, which can be freely accessed:

https://snquery.veratech.es

Validation and simplification of expression constraints

V. Giménez, J.A. Maldonado, D. Boscá, D. Moner | VERATECH FOR HEALTH

METHODS

SNOMED International
SNOMED CT EXPO 2020
Virtual Conference October 8-9

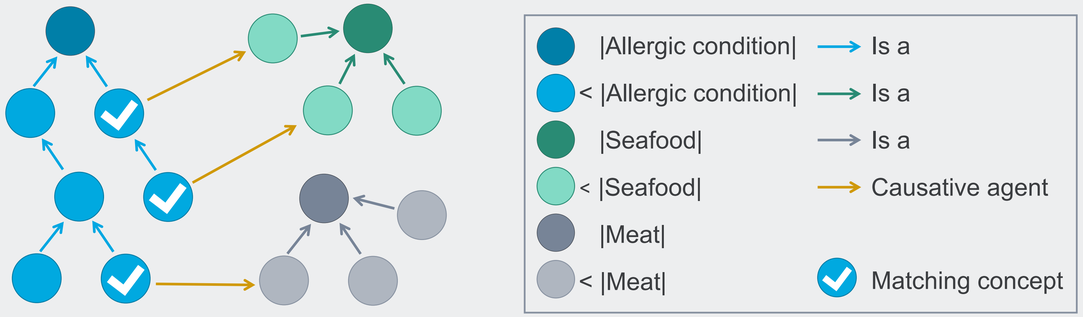
Preliminary

Expression Constraint Language (ECL)

ECL enables the **intensional definition of sets of clinical meanings** by defining **ECs**. To define a simple EC, a **constraint operator** is applied to a **focus concept** in order to **navigate** over SNOMED CT hierarchies. Example: "< 473011001 |Allergic condition|" (descendants of |Allergic condition|). **Refinements** allow **filtering matching clinical meanings** by adding conditions on constrained focus concepts in the form of **attribute refinements**. Example:

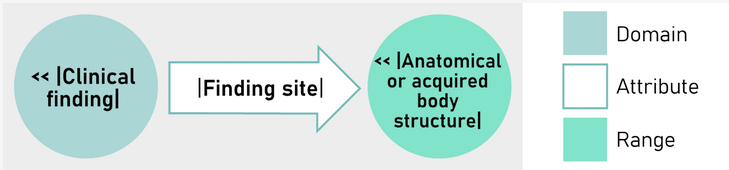
```
< 473011001 |Allergic condition (disorder)|:
246075003 |Causative agent (attribute)| = < 44027008 |Seafood (substance)| OR
246075003 |Causative agent (attribute)| = < 28647000 |Meat (substance)|
```

Allergic conditions that have as causative agent any type of seafood or meat



Machine Readable Concept Model (MRCM)

SNOMED CT is governed by the Concept Model, which is a **set of rules** that define the **domain** and **range** of **attribute** relationships. MRCM is defined by using the **ECL**. For instance, the domain of the attribute |Finding site| is defined by the EC "<< 404684003 |Clinical finding|" and its range is defined by "<<442083009 |Anatomical or acquired body structure|".



Example of a MRCM triplet (domain-attribute-range)

Graph databases

Graph databases bring a number of **potential advantages** over traditional database systems, such as relational. They emphasize **connectedness** of data that fits the **polyhierarchical** and **ontological** nature of SNOMED CT. Graph databases also provide a **great flexibility** to **add new nodes** and **relationships** to the graph without affecting existing queries [2]. The more robust representation of SNOMED CT facilitates the execution of **complex queries** that go beyond subsumption queries (i.e. the descendants of a given concept) as required by ECL.

Validation and simplification of expression constraints

V. Giménez, J.A. Maldonado, D. Boscá, D. Moner | VERATECH FOR HEALTH

SNOMED International
SNOMED CT EXPO 2020
Virtual Conference October 8-9

Results

Storage of the SNOMED CT database

Since SNOMED CT is by definition a **directed** and **acyclic graph**, the use of a **graph database** seems to be a logical choice. We have used **Neo4j** to store SNOMED CT by means of the **Property Graph Data Model** [3].

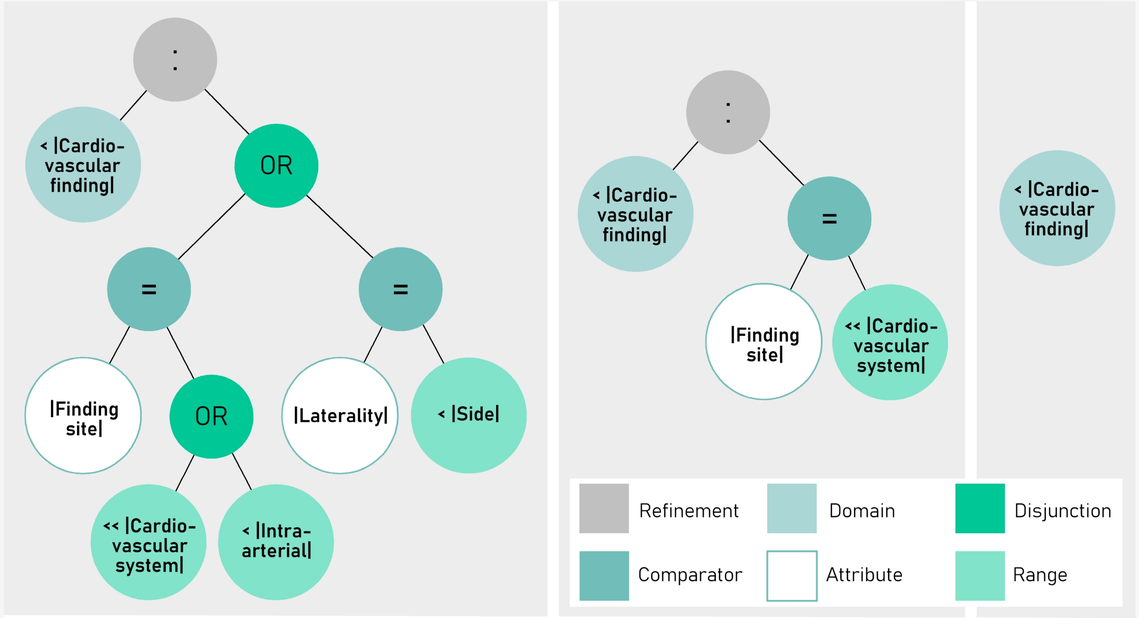
EC validation and execution

For the validation process, it is necessary to **check** the triplets **domain-attribute-range** of the EC against the **rules** defined in the **MRCM** and the **logical definition of the focus concept**. For the execution of ECs, the engine translates the ECL into Neo4j **Cypher Query Language** before **querying** the SNOMED CT database.

EC simplification

The main idea is to **reduce the complexity and execution time** of ECs. We have applied different techniques in order to simplify ECs:

- **Pre-execution**: MRCM-, logic definition- and subsumption-based for the removal of superfluous refinements.
- **Post-execution**: mining of the answer subset in order to narrow down the focus concept and/or removal of the whole refinement.



Example: syntax tree representations of the original EC (left) and the ECs obtained after applying both the MRCM- (center) and the logic definition-based simplification (right) methods

Subsets visualization

After executing an EC, the answer subset is presented in **sortable tabular form**. Optionally, we show the **circle packing** [4] of the subset, where each circle represents a root concept of a sub-hierarchy. These concepts are members of the answer subset or they have at least one descendant in it. Only intermediate concepts are shown.

Validation and simplification of expression constraints

V. Giménez, J.A. Maldonado, D. Boscá, D. Moner | VERATECH FOR HEALTH

SNOMED International
SNOMED CT EXPO 2020
Virtual Conference October 8-9

Discussion

Practical application:

Consistency validation of refsets and terminology

The data shown in circle packing is useful for reviewing and **validating refsets** and the **terminology** itself. Refsets can have several **potential problems**, such as the existence of **homographs** that may cause the inclusion of **wrong concepts**. The terminology might have concepts improperly classified. As an example on validating the terminology itself, given the subset of those disorders of the body system that are located in the blood vessel structure:

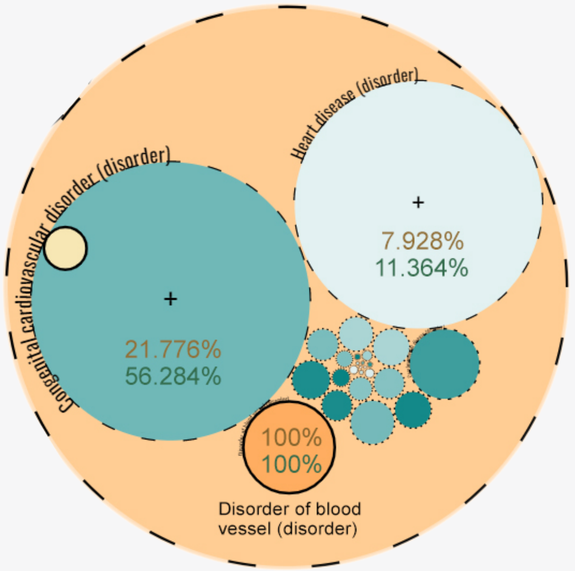
```
< 362965005 |Disorder of body system (disorder)|:
363698007 |Finding site (attribute)| = << 59820001 |Blood vessel structure|
```

whose simplification (post-execution) is:

```
<< 27550009 |Disorder of blood vessel (disorder)|
```

we can asseverate that every disorder of the body system concept having as finding site a type of blood vessel structure is descendant of 27550009 |Disorder of blood vessel (disorder)|.

We show the circle packing representation of the previous example:



Note **100% of the concepts of the subset are descendants of |Disorder of blood vessel| (recall = 1),** 7.93% are descendants of |Heart disease| and 21.78% of |Congenital cardiovascular disorder| (since SNOMED CT is a poly-hierarchical terminology, a concept can have multiple parents). **100% of the descendants of |Disorder of blood vessel| are included in the subset (precision = 1)**, 11.36% of the descendants of |Heart disease| are included in the subset and 56.28% of the descendants of |Congenital cardiovascular disorder| are included in the subset.

References

[1] *SNOMED CT ECL Specification and Guide*. SNOMED International, 2020.

[2] I. Robinson, J. Webber and E. Eifrem, *Graph Databases. New opportunities for connected data*. 2015.

[3] R. Angles, “The Property Graph Model,” Proceedings of the 12th Alberto Mendelzon International Workshop on Foundations of Data Management (CEUR Workshop Proceedings), 2018.

[4] C. R. Collins and K. Stephenson, “A circle packing algorithm,” Computational Geometry: Theory and Applications, vol. 25, no. 3, 2003, doi: 10.1016/S0925-7721(02)00099-8.