Extracting Subontologies from SNOMED CT*

Warren Del-Pinto[1], Renate A. Schmidt[1], and Yongsheng Gao[2]

[1] Department of Computer Science, University of Manchester, UK
{warren.del-pinto,renate.schmidt}@manchester.ac.uk
[2] SNOMED International, London, UK
yga@snomed.org

1 Introduction

Computing smaller extracts of a larger ontology has been recognised as important for enabling tasks such as ontology creation, review, updating, debugging, navigation, sharing and integration [6, 2, 5]. In addition, reasoning tasks such as querying and classification take less time to execute over a smaller extract than over the original ontology. As the most comprehensive clinical healthcare terminology in the world, SNOMED CT is by necessity a large ontology, containing over 350,000 concepts and a large amount of content is contained in various extensions. As a result, the benefits provided by computing smaller extracts are even more pronounced in this setting. Additionally, the ability to extract and extend content focused on specialist domains can facilitate the navigation and utilisation of specific content within a large terminology that are directly relevant to specialist domains for clinicians and healthcare systems.

Often, reference sets (refsets) [3] are computed or curated by experts to list a subset of concepts that are relevant to a given clinical specialty, such as the General Dentistry Diagnostic refset. However, such lists are not sufficient in applications that rely upon the semantics of the source ontology, where an extract in the form of a standalone ontology is needed.

Modularisation approaches [2, 5] produce such extracts by computing subsets of the stated axioms in the source ontology, such that all entailments with respect to the included concepts are preserved. The computed modules are useful in that they capture the semantics within a domain of interest and can be used in place of the original, larger ontology. However, in practice modules are often large and contain a significant amount of unnecessary information that is not required to capture the modelling of the specified concepts in the domain of interest.

2 Subontology Extraction

We have developed new software to compute concise extracts of SNOMED CT that are semantically complete with respect to a set of input concepts, called

* Thanks to members of our Steering and Working Groups: Rory Davidson, Jim Case, Monica Harry, Kai Kewley and Ghadah Alghamdi. The work was funded by UK EPSRC IAA, the University of Manchester and SNOMED International.

focus concepts. The two main criteria for these extracts, called *subontologies*, are: (i) Focus concepts must be defined equivalently in the subontology and the source ontology. (ii) The transitive closure (with respect to subsumption) between concepts occurring in the extract must be equal in the subontology and the source ontology, up to the signature of concepts in the subontology. The subontology extraction approach automatically identifies additional *supporting concepts* that are required to satisfy condition (i) and includes these in the extracted subontology.

Subontology extraction differs from modularisation approaches in the separation between focus and supporting concepts; while modularisation approaches extract a subset of the original axioms in an ontology, subontology extraction produces equivalent definitions for focus concepts in a compact abstract form, the authoring form (long canonical form in [7]), while supporting concepts are only fully defined if necessary. The hierarchy between concepts in the subontology is then completed by using the classification over the source ontology (SNOMED CT) to identify missing inclusions and add these automatically.

The subontology extraction approach developed in this work supports the language features required by the latest versions of SNOMED CT, including language extensions such as GCI axioms, reflexive roles, transitive roles, role chain axioms and data types, effectively, the description logic ELH^{++} [1].

3 Implementation, Evaluation and Applications

The prototype was implemented in Java, making use of the OWL API and the DL reasoner ELK [4] for classification. A prototype of the tool is available at https://github.com/IHTSDO/snomed-subontology-extraction.

A set of experiments were performed to evaluate the performance of the algorithm in practice. Since the aim is to produce concise extracts, the size of the extracted subontologies was compared to STAR modularisation, which is available as part of the OWL API. The two approaches were compared using real clinical refsets as input, where the refset is used to specify the set of focus concepts for extraction. The results in Table 1 indicate that the extracted subontologies are significantly smaller than STAR modules across all cases. The runtime for subontology extraction ranged from 8–266 seconds for the smallest to largest refsets respectively.

Figure 1 provides an in-practice comparison between the navigation of a subontology, the ERA-EDTA subontology, and the full release of SNOMED CT. As seen from the subhierarchies displayed, the extracted subontology includes only those hierarchies that contain concepts that are relevant to the domain of interest specified by the ERA-EDTA refset. Hierarchies such as "Pharmaceutical/biologic product (product)" are excluded from the subontology, as no concept in this hierarchy was found to be necessary to preserve the semantics of the focus concepts in the refset. Additionally, for each of the included subhierarchies, the descendent count is smaller in the subontology compared to the original SNOMED CT ontology.

Table 1. A comparison between the sizes of the extracted subontologies and locality-based (STAR) modules for a collection of refsets.

Refset	**Refset Size**	**Subontology Size**		**STAR Module Size**	
Name	(Concepts)	Axioms	Concepts	Axioms	Concepts
ERA-EDTA	184	485	475	3076	3086
Dentistry	226	455	642	1449	1478
Nursing	1337	2616	2616	5579	5708
Orphanet	5681	9209	9189	27595	27625
IPS	8182	12793	12745	53736	53708
GPS	26159	33970	33907	86167	86374

Refsets Key:

Dentistry	General Dentistry
ERA-EDTA	European Renal Association / Dialysis and Transplant Association
GPS	International Global Patient Set
IPS	International Patient Set
Nursing	Nursing Activities and Nursing Health Issues (combined)
Orphanet	Rare diseases, orphan drugs

In addition to the experiments, a range of subontologies have been computed for standard lists of clinical concepts, including several of the refsets in Table 1. These subontologies, viewable in the browser at https://iaa.snomed.tools, have received qualitative feedback from users (domain experts). The users each answered questions about a subontology that was relevant to their domain of interest, covering their experience of navigating the subontology, the scope of the content contained within them and the potential usefulness in their own work. The feedback indicated that presenting domain specific content via a subontology in the browser was useful, as it made it easier to navigate the relevant content without having to navigate the entirety of SNOMED CT. Additionally, the feedback generated discussion relating to the refsets provided as input. For example, the nursing refset did not contain several concepts that were expected by domain experts, such as those relating to different types of specimens (samples). This was based on navigation of the subontologies, which relies on semantic information retained by subontology extraction such as the definitions of and hierarchy between included concepts. This points to a promising use of subontologies in maintaining domain specific content and assisting with refset curation.

The subontology extraction prototype has already been used in a range of applications within SNOMED International, including the development of a new concept model for anatomy, which is represented as a subontology, and identifying improvements to the modelling of substances by enabling clinical modellers to examine and navigate content via more concise extracts that are compatible with the existing SNOMED CT browser. Subontology extraction is also a core component in the new release of the International Patient Set subontology, which aims to enable more effective use of clinical data analytics and decision support over essential healthcare information. Community content regarding traditional medicine, which is not part of the International Edition of SNOMED CT, will

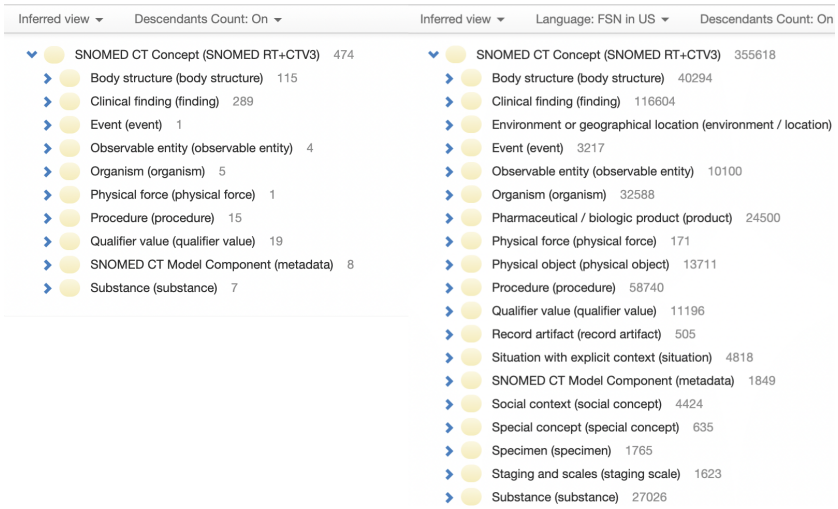

Fig. 1. A screenshot of the top-level subhierarchies of SNOMED CT for the ERA-EDTA subontology at https://iaa.snomed.tools (left) and the full SNOMED CT International Edition (right), viewed using the SNOMED CT browser. The counts beside each concept show the number of inferred subconcepts in the ontology.

also be presented as a subontology to provide a means for users to utilise concepts related to traditional medicines where this is needed in different countries.

References

1. Franz Baader, Sebastian Brandt, and Carsten Lutz. Pushing the EL envelope. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, volume 5, pages 364–369. AAAI Press, 2005.
2. Bernardo Cuenca Grau, Ian Horrocks, Yevgeny Kazakov, and Ulrike Sattler. Modular reuse of ontologies: Theory and practice. *Journal of Artificial Intelligence Research*, 31:273–318, 2008.
3. SNOMED International. Practical guide to reference sets. SNOMED CT Document Library, https://confluence.ihtsdotools.org/display/DOCRFSPG, 2017. [Online; accessed 08-March-2022].
4. Yevgeny Kazakov, Markus Krötzsch, and Frantisek Simancik. The incredible ELK: From polynomial procedures to efficient reasoning with EL ontologies. *Journal of Automated Reasoning*, 53:1–61, 2014.
5. Boris Konev, Carsten Lutz, Dirk Walther, and Frank Wolter. Model-theoretic inseparability and modularity of description logic ontologies. *Artificial Intelligence*, 203:66–103, 2013.
6. Alan L. Rector. Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. In *Proceedings of the 2nd International Conference on Knowledge Capture*, pages 121–128. ACM, 2003.
7. Kent A. Spackman. Normal forms for description logic expressions of clinical concepts in SNOMED RT. In *Proceedings of the AMIA Symposium*, page 627. American Medical Informatics Association, 2001.