Investigating SNOMED CT national extensions for duplicate content

Matt Cordell, Australia Digital Health Agency;
Dr. Olivier Bodenreider, National Library of Medicine.

Background

Many SNOMED International Member Countries (and some affiliates) produce a national SNOMED CT extension that supplements the international release to meet local requirements. A sub-group of the Content Managers Advisory Group undertook a project to investigate the type of content that was being produced in national extensions.
More specifically, we analysed the content of the extensions in terms of size, type of logical definition (primitive vs. defined concepts), and semantics (distribution across SNOMED CT hierarchies). Of particular interest were duplicate concepts across extensions, as they are candidates for inclusion in the international release.

Methods

The Member Forum was instrumental in helping us acquire the extensions from member countries. The extensions were collected into a single relational database for analysis with SQL queries. We focused on the core components (Concepts, Descriptions and (inferred) Relationships). We gathered descriptive statistics. Additionally, each hierarchy was analyzed for duplicate content using several methods:

- Direct comparison of Fully Specified Names (FSNs);
- Identification of non-unique synonyms across extensions (but within a hierarchy); and
- Classification of the content to detect equivalence based on description logic.

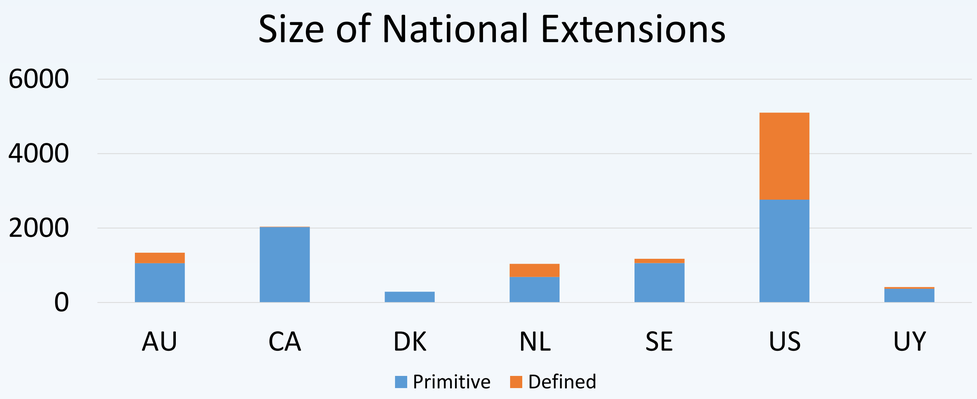
Results

We acquired extensions from 9 member countries: Australia (AU), Canada (CA), Denmark (DK), Lithuania (LI), Netherlands (NL), Sweden (SE), United Kingdom (UK), United States (US), and Uruguay (UY).

Size and Type of logical definitions

There are just over 41,000 concepts created across the 9 extensions that were analysed. The extensions ranged from 290 and over 30,000 concepts. Two other extensions included in the study are not shown below:

- UK – 30 000 concepts, all primitive;
- Lithuania – Translation only, no extension concepts.

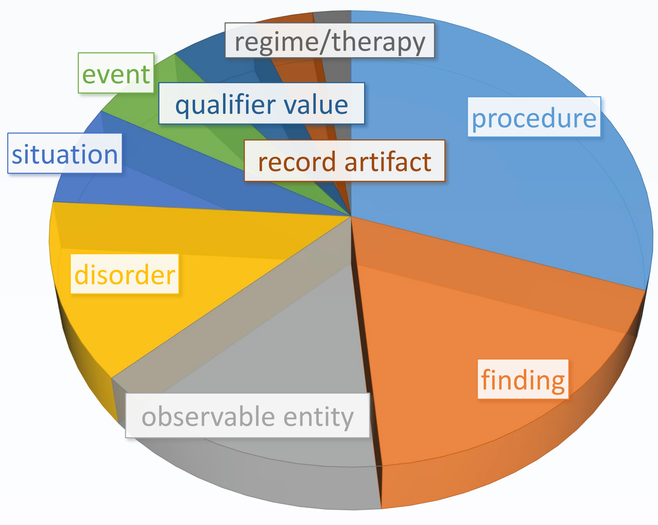


Semantic distribution

Over 90% of relate to 7 of the top level hierarchies, with the Clinical finding (including disorders) and procedures accounting for almost a quarter of the extension content. Unsurprisingly given its size, the UK extension responsible for the majority of extension content across most hierarchies, with the exception of:

- Substances - US (43%), Canada (30%) and Denmark (17%)
- Pharmaceutical / biologic product – US, Canada and Denmark again accounting for 90% of additions.
- Organism – Netherlands dominate with 64%

In general though, most countries had created content across a range of areas. A concentration of content relative to other hierarchies is likely reflective of local priorities; for example Microbiology in the Netherlands.



Duplicated content

Fully Specified Names. The presence of non-unique Fully Specified Names (FSN) was the simplest mechanism for detecting duplicated content. Duplicated content was detected across 10 hierarchies. This overlap was not exclusive to English extensions, with all extensions except Uruguay contributing at least 1 non-unique FSNs.

Hierarchy	Non-Unique FSNs	Hierarchy	Non-Unique FSNs
Clinical Finding	26	Record artefact	8
Organism	2	Situation	8
Procedures	16	Social context	5
Products	12	Special concept	15
Qualifier value	28	Substances	15

Synonyms. The analysis of synonyms within any hierarchy revealed a much greater level of duplication. There were 6400 non-unique synonyms just from the Clinical finding hierarchy. Most of these relate to localising, where countries add synonyms to existing content. For example:

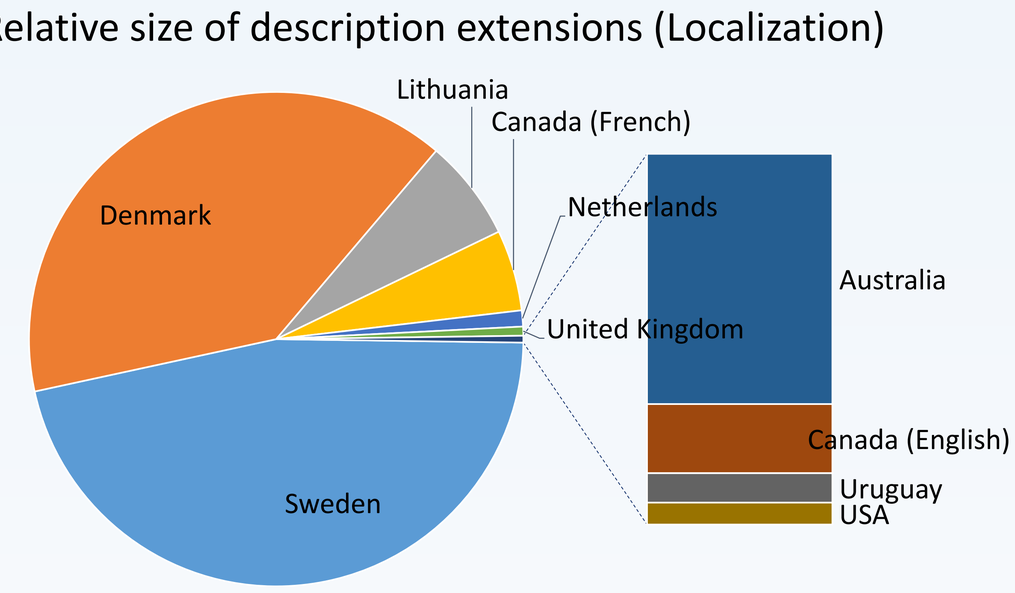
- 371093006 | *Urosepsis (disorder)* | has descriptions in, the extensions from three countries, that are the same as the 'en' description.
- 27830001 | *Brachial radiculitis (disorder)* | has translations in two extensions that are different to the 'en', but differ from each other by the case of the first character.
- 75049004 | *Jeune thoracic dystrophy (disorder)* | has translations in two extensions that appear identical.

When comparison was made case sensitive, the number of duplicate terms halved.

DL classification. We only found a few example of (near) duplicate logical definitions after classifying the various extensions with a description logic (DL) classifier.

Term localization.

All extensions do some form of localization, adding descriptions for international concepts. Predominately these are non-English translations.



Conclusions

The results of this project provide evidence that there is certainly some duplication of content across extensions. Non-unique concepts have value to at least 2 countries, and likely others. Although the process described only discovered 135 concepts, the methods are extremely dependent on extension authors adhering consistently to the same editorial and terming conventions.

Similarly for extension synonyms. The inclusion of these synonyms in extensions improves search effectiveness in the country of origin. All members would benefit if these were promoted to the international release.

Awareness about the existence of content in national extensions is the first step in preventing duplication across extensions (e.g., by promoting content to the international release), promoting collaboration among NRCs, and ultimately lowering the burden of content development by the NRCs. Our investigation demonstrates the feasibility of collecting and integrating SNOMED CT extensions and represents an initial effort towards analyzing their content for duplication.

Complete results available online: https://confluence.ihtsdotools.org/display/cmag/Results+of+Analysis+of+SNOMED+CT+Extensions