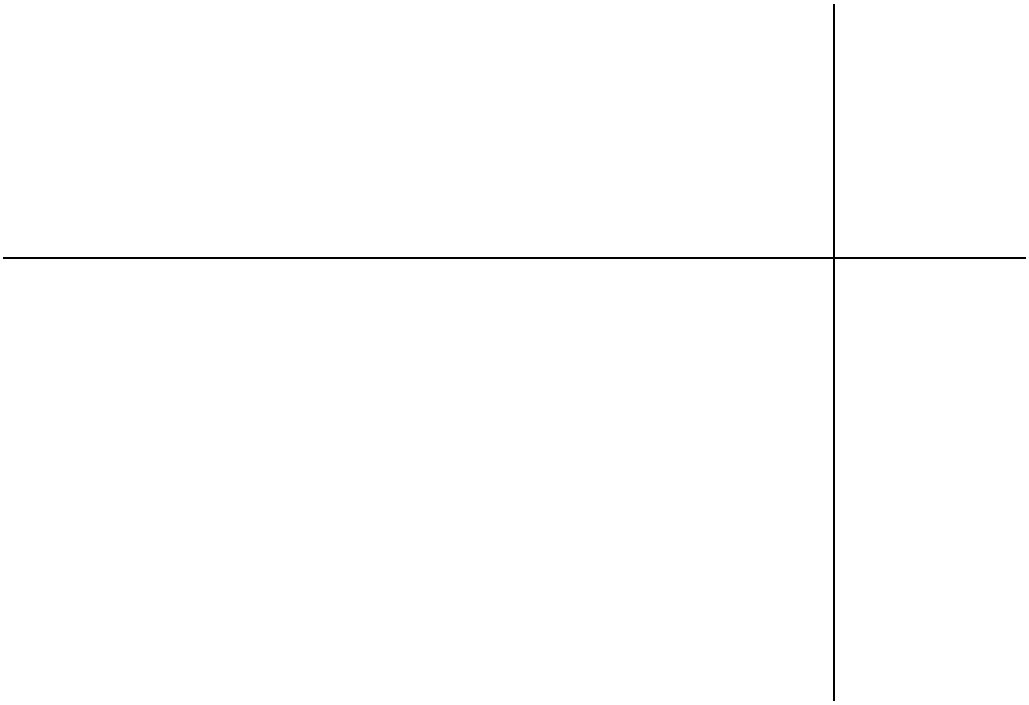
SNOMED Clinical Terms®
Lexical Resource Guide

Draft 2015-05-15



Document History

Version	Notes
2015-05-15	Draft based on extract of elements from the former Toolkit Guide relevant to the ExcludedWord and WordEquivalents tables. Added relevant content from 2013 report on Lexical Resources produced as SNOMED CT Implementation Advisor assignment by Alejandro Lopez Orsornio.

Table of Contents

1 Introduction

1.1 Purpose

This document describes the file structure of two files that originated in the SNOMED CT Developers Toolkit for Release Format 1. Both files are potentially useful example resources that are also relevant to implementers using Release Format 2. Therefore this document has been prepared as short guide to those files.

1.2 Who should read this guide?

The intended audience for this document is any individual or any organization that wishes to develop or use systems that will use SNOMED Clinical Terms and is interested in resources that may support more effective searches.

1.3 Scope and Format

This document uses material from the SNOMED CT technical specifications that were used to create the work.

2 Supporting effective SNOMED CT Searches

2.1 Overview

Effective implementation of SNOMED CT depends on the ease and speed with which users can locate the terms and Concepts that they wish to use. An essential contribution to meeting this requirement is the ability to perform rapid and flexible text searches.

2.2 Search Requirements

Development of user-interfaces that facilitate rapid and appropriate access to SNOMED CT is a legitimate area for competition between application suppliers.

The IHTSDO publishes two guides that are directly relevant to search and data entry requirements:

- SNOMED CT Search and Data Entry Guide
- SNOMED CT Members' Browser Requirements

3 Excluded Words Table

3.1 Introduction

The Excluded Words file is available as part of the Lexical Resource archive file.

This file includes a list of words that should be excluded from searches, also called stop-words in other contexts. Languages are identified in the table by "language code". The words are included in the "keyword" format, a capitalized, 8 characters long string, that is used by the index table generator to exclude words from word indexes.

ExcludedWords File			
Each row in the Excluded Words Table is a word excluded from the list of possible keywords and dualkeys. Words are excluded if they are frequently used and are so limited in semantic specificity that they impair rather than enhance searches. Word exclusions are language or dialect specific.			
Key Fields	**Field Type**	**Permitted Characters**	**Length**
LanguageCode	*String*	*0 to 9, a to z, A to Z and* dash "-"	*1 to 8*
A string identifying a language and or dialect in the word is excluded from keyword generation. Consists of a code and optionally a sub-code. If a sub-code is present it is separated from the code by a dash ("-"). ❖ The code is the ISO639 language code, which is a string of two lower-case letters. ISO639 is the International Standard for "Codes for the representation of names and languages." ❖ The sub code is a string of upper-case letters. This will either be: ✧ A two-letter ISO3166 country code. ISO3166 is the International Standard for "Codes for the representation of names of countries." ✧ A string of more than two letters, which is registered with IANA as a sub code for the language. IANA is the Internet Assigned Numbers Authority. This structure follows Internet conventions. Examples: "en" for "English," "es" for Spanish, "enUS" for United States English, "en-GB" for British English.			
Keyword	*String*	*0 to 9, A to Z and dash* "-"	*1 to 8*
A word used in Descriptions in the Descriptions Table but excluded from keyword generation. Words are represented using only upper case letters and words of more than eight characters are represented as their first eight characters only.			

3.2 Usage Notes

The version of the table current in May 2015 has not been updated since 2007. However, the content of stop-words table is general stable so the lack of udpates is not a significant issue.

The current table only contains English terms. If tools and applications use this table in non-English speaking countries they would need to add stop words that are relevant in the language of use.

This file is designed to support index generation and searches. It could also be used for matching with natural language queries using sub-string matching, this table cannot be used to configure the list of stop-words on most common search engine products.

Various search tools include a default set of stop-words but these are often not appropriate to clinical term searches. Tools that support configurable stop-word lists can make use of the relevant data in this file.

Configurations based on Lucene, like SOLR (http://lucene.apache.org/solr/) require a stop-words list with full words, in some cases the 8 character limit applied to this table will lead to limitation.

The excluded words index could be used to create a simple normalized index for descriptions, simple SQL code could be provided to create a normalized index table, and sample search SQL query based on that index.

4 Word Equivalents

4.1 Introduction

The Word Equivalent file is available as part of the Lexical Resource archive file. It supports enhanced searches that take into account semantically related words such as KIDNEY and RENAL. It includes 4 types of semantic relationships:

- Word form variant (e.g. "abdomen", "abdominal")
- Word equivalents (e.g. "renal", "kidney")
- Abbreviation or acronym (e.g. "MI" → "myocardial infarction")
- Equivalent phrase (e.g. "MI" → "myocardial infarction")

Even when SNOMED CT provides a structure for preferred and acceptable descriptions (synonyms), not all possible equivalent words are used to describe SNOMED CT concepts; editorial guidelines favor the use of some words over other equivalent ones.

Using the Word Equivalents table as a complement of the Descriptions table enables the developers to design advanced search strategies. This approach will match proper SNOMED CT content even when the user enters words that are not included in the descriptions, like special variants or acronyms.

The structure also includes a “Word role” column, with the goal of supporting the selection of related concepts used to create a post-coordinated expression based on a text entry.

4.2 Word Equivalents Tables – Summary

Key Fields	
WordBlockNumber	A 32-bit integer shared by a set of equivalent words or phrases. The WordBlockNumber links together several rows that have an identical or similar meaning.
WordText	A word, phrase, acronym or abbreviation that is equivalent to the WordText of other rows that share the same WordBlockId.
Data Fields	
WordType	An integer indicating the type of equivalence.
WordRole	An integer indicating the usual role of this word. This should be considered if attempting to find a post-coordinated combination of Concepts that matches a phrase.

Word Equivalents Table

Each row in this table represents the potential equivalence between a word, phrase, or abbreviation and other words, phrases or abbreviations.

Key Fields	Field Type	Permitted Characters	Length
WordBlockNumber	*Integer*	*0 to 9*	*1 to 10*
A 32-bit integer shared by a set of equivalent words or phrases. The words, phrases and abbreviations that share a common WordBlockId value are interchangeable for the purposes of searches. *Example:* An equivalent block could contain the following: “TB”, “tuberculosis”, “tuberculous” *Note:* WordBlockId is not maintained as a unique identifier across releases. It should only be regarded as an index to link equivalents in the context of a particular release.			
WordText	*String*	*0 to 9 only, A to Z and dash “-”*	*1 to 50*
A word, phrase or abbreviation that is equivalent to the WordText of other rows that share the same WordBlockId. *Note:* If a word or phrase has two or more possible meanings it may be represented by more than one row in this table. Each row containing the same WordText must be associated with a different WordBlockId value. Note that earlier specifications of this field incorrectly show its maximum length as 8 characters. In practice since 2002 the released file has contained many rows with longer text			
Data Fields	**Field Type**	**Permitted Characters**	**Length**
WordType	*Enumerated*	*See listed values*	*2*
An integer indicating the type of equivalence *Values* 0 unspecified 1 word form variant (e.g. "abdomen", "abdominal") 2 word equivalents (e.g. "renal", "kidney") 3 abbreviation or acronym (e.g. "MI" → "myocardial infarction") 4 equivalent phrase (e.g. "MI" → "myocardial infarction")			
WordRole	*Enumerated*	*See listed values*	*2*
An integer indicating the usual role of this word. This should be considered if attempting to find a post-coordinated combination of Concepts that matches a phrase. *Values* 0 unspecified 1 general qualifier 2 topography 3 topography qualifier 4 object (including organism or substance) 5 action 6 unit of measure *Note:* All rows with the same WordBlockId value must have same WordRole			

4.3 Usage notes

Support for the use of local jargon, dialects or acronyms in searches has a great impact on the terminology usability by end users, and it should be a high priority feature in any implementation project.

This file has a potential value in helping implementation projects to enhance search features.

The other option that can used to resolve local language variations is to add new descriptions to the concepts in a local SNOMED CT extension.

In simple implementations, the option of using the word equivalents table to represent local terms in searches has many advantages. Adding an acronym in the table will automatically match with all occurrences of Word Equivalents in SNOMED CT, using new descriptions this may require to add hundreds of descriptions in the local extension.

As shown in Table 1, adding descriptions for all possible synonyms and acronyms for every concept is liable to create combinatorial explosion of descriptions. The Word Equivalents file allows similar functionality to be delivered more efficiently and reproducibly.

Table 1. Example of Word Equivalents for renal and calculus

WORDBLOCKNUMBER	WORDTEXT	WORDTYPE	WORDROLE
9852	calculus	0	0
9852	lithous	0	0
9852	stone	0	0
10724	kidney	0	2
10724	nephric	0	2
10724	nephritic	0	2
10724	renal	0	2

Table 2 shows that of the twelve possible combinations of words meaning kidney and stone four are found in terms for active concept in the SNOMED CT International Edition (2015-01-31). A total of 37 concept match one or more of these combinations. The number of concepts found for each combinations varies do to the different synonyms applied to each concept.

Table 2. Active concepts in International Edition 2015-01-31 with one or or more descriptions containing terms a word that means stone and a word that means kidney

		Count	% total (37)
calculus	kidney	8	22%
calculus	renal	23	62%
stone	kidney	10	27%
stone	renal	6	16%

None of the combinations returns more than 62% of the potential concept matches. Expanding the search with word equivalents could return all 27 concepts.

Although combinations involving ‘nephric’ and ‘lithous’ do not occur as separate words, they are used in combination ‘nephrolith’ and ‘nephrolithiasis’. These alternatives are not currently represented in the Word Equivalents file.