GOBIERNO DE ESPAÑA
MINISTERIO DE SANIDAD, CONSUMO Y BIENESTAR SOCIAL

Centro Nacional de Referencia de SNOMED CT

Lexicon

Recursos Léxicos para SNOMED CT

Versión: 1.1
Fecha: 01/12/2021

Introducción

El Centro Nacional de Referencia Spain de SNOMED CT incorpora a su cartera de productos **Lexicon**, que contiene el léxico de SNOMED CT con todas las palabras obtenidas en un análisis de las descripciones incluidas en las diferentes ediciones en español de SNOMED CT.

- Spanish Edition (traducción al español de la International Edition): 1.018.773 descripciones.
- Extensión para España: 28.977 descripciones.
- Extensión para España de Medicamentos: 32.181 descripciones.

Los destinatarios de este producto son todos aquellos particulares, empresas y organizaciones de cualquier tipo que son licenciatarios de SNOMED CT y están interesados en recursos que puedan dar soporte a métodos de búsqueda eficientes y rápidos de la terminología.

Las herramientas y recursos de apoyo al uso de idiomas, dialectos y vocabularios de dominio en búsquedas tienen un alto impacto en la usabilidad de la terminología por parte de los usuarios finales, por lo que esta funcionalidad debe ser contemplada de forma prioritaria en los proyectos de implantación de SNOMED CT (para más información SNOMED International publica una Guía con aspectos relevantes para la búsqueda y registro de información: SNOMED CT Search and Data Entry Guide).

Contenido

Este documento describe la estructura de los ficheros que se publican en el paquete de distribución y el contenido que se incluye o puede incluir en los mismos.

Este nuevo producto se encuentra disponible en el Área de Descargas de SNOMED CT (https://snomed-ct.msssi.es/snomed-ct/solicitudLicencia.do). El paquete se libera con el nombre ***Lexicon SNOMED CT*** e incluye un archivo comprimido denominado xder2_lexiconSpain_es-ES_AAAAMMDD, que contiene los siguientes ficheros:

xder2_lexicon_es-ES_AAAAMMDD.txt
xder2_lexicon_equivalents_es-ES_AAAAMMDD.txt
xder2_lexicon_excluded_es-ES_AAAAMMDD.txt

El compromiso del CNR de SNOMED CT para España es publicar nuevas versiones del producto anualmente, actualizando Lexicon con los contenidos de las nuevas versiones de los productos SNOMED CT en español que se vayan publicando.

Tabla Lexicon

Contiene las diferentes palabras que conforman el léxico de SNOMED CT en español. La versión actual incluye 116.705 palabras diferentes y el número de apariciones de cada una de ellas en el conjunto de las descripciones analizadas. Las palabras incluidas distinguen entre formas acentuadas y no acentuadas, o las

que contienen otras marcas diacríticas o caracteres de alfabeto nórdico, y se separan de acuerdo al uso de mayúsculas o minúsculas. Por ejemplo, **tacrolimus** es diferente de **tacrólimus**. Las palabras con grafía incorrecta están incluidas en el léxico. Se estima que un 3% de las palabras tienen grafía incorrecta. De hecho el léxico ya se está utilizando para un proceso de mejora continua de la calidad ortográfica, de acuerdo a las reglas del Español de España.

tipo	campo	detalle
key	token	String. Palabra que forman parte del léxico.
data	n	Integer. Número de apariciones de la palabra en los términos de SNOMED CT.

Tabla Palabras Excluidas

El fichero de palabras excluidas (stop-words) incluye una lista de palabras que deberían ser excluidas de la lista de posibles palabras clave y claves duales en los sistemas de búsqueda. Este contenido está diseñado para soportar búsqueda rápida, aunque puede tener otros usos como por ejemplo ser utilizado en cruces con lenguaje natural para anotación y marcado de entidades nombradas.

tipo	campo	detalle
key	languageCode	String. Identifica el lenguaje o dialecto para el que la palabra es excluida de la generación de palabras claves. Para la codificación de idioma se utiliza la codificación ISO-639-1, con códigos de 2 caracteres en minúsculas.
key	keyWord	String. Palabra excluida, de una longitud máxima de 8 caracteres

La selección de palabras se realiza **manualmente** sobre la base de un conjunto de candidatas: alta frecuencia, longitud corta, POS de alta frecuencia (preposiciones, artículos).

Tabla Palabras Equivalentes

La tabla se estructura como un conjunto de sinónimos que comparten un mismo wordBlockNumber. Este contenido soporta búsquedas mejoradas basadas en palabras relacionadas semánticamente, con posibilidad de incluir diferentes tipos de relaciones semánticas.

tipo	campo	detalle
key	wordBlockNumber	Integer. Compartido por un conjunto de palabras o frases equivalentes, que tienen un significado idéntico o similar.
key	wordText	String. Una palabra, n-grama, acrónimo o abreviación que es equivalente al wordText de otras filas que comparten el mismo wordBlockNumber.

data	wordType	Integer. Indica el tipo de equivalencia. Posibles valores: ● 0 no especificado ● 1 palabras con variaciones ● 2 palabras equivalentes ● 3 abreviaciones o acrónimos ● 4 n-gramas equivalentes
data	wordRole	Integer. Indica el rol usual de esta palabra, información a considerar, por ejemplo, si se intenta buscar una expresión post-coordinada de conceptos que coincida con una frase. Posibles valores: ● 0 no especificado ● 1 calificador general ● 2 topografía ● 3 calificador de topografía ● 4 objeto (incluye organismos o sustancias) ● 5 acción ● 6 unidad de medida Nota: todas las palabras con el mismo wordBlockNumber deben tener el mismo wordRole.

El contenido de este fichero de palabras equivalentes tiene valor en la ayuda a la implantación de búsquedas mejoradas en la terminología en español. Otro uso del mismo es la resolución de variaciones locales del idioma, ya que permite reducir la necesidad de añadir nuevas descripciones y sinónimos en las extensiones locales.

En la versión actual, este fichero se publica sin contenido o conteniendo solamente entradas a modo de ejemplo. En próximas versiones se irá poblando con dos tipos de contenidos:

- Acrónimos y sus descripciones.
- Casi sinónimos: grupos que contengan palabras diferentes pero con lemas relacionados (ej.: hígado, hepático, hepática)

Fin del documento.