GOBIERNO DE ESPAÑA
MINISTERIO DE SANIDAD, CONSUMO Y BIENESTAR SOCIAL

## Centro Nacional de Referencia de SNOMED CT

# Lexicon e Índice de conceptos y descripciones

## Recursos Léxicos para SNOMED CT

Versión: 1.0
Fecha: 23/07/2025

# Tabla de contenidos

# Lexicon

## Introducción

El Centro Nacional de Referencia (CNR) de SNOMED CT para España con el apoyo del grupo de trabajo de Procesamiento de Lenguaje Natural de la Unidad Técnica ha actualizado el **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 (SE), que es traducción al español de la International Edition
- Extensión para España (EE)
- Extensión para España de Medicamentos (EM)

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 en los mismos.

Este recurso se encuentra disponible en el Área de Descarga SNOMED CT (https://snomed-ct.sanidad.gob.es/snomed-ct/solicitudLicencia.do?metodo=volverDeLogin). Este paquete se libera con el recurso **Lexicon e Índice de conceptos y descripciones de SNOMED CT en español** en el que se incluye un archivo denominado Lexicon_SNOMED_CT, que contiene los siguientes ficheros:

- En la carpeta "Content" se encuentran cinco documentos .txt:
    - xder2_lexicon_ES20250723.txt: Contiene las diferentes palabras que conforman el léxico de SNOMED CT en español y el número de apariciones de cada una de ellas.
    - xder2_lexicon_SE-EE_ES20250723.txt: Contiene las diferentes palabras que conforman el léxico de SNOMED CT en español de la SE y la EE y el número de apariciones de cada una de ellas.
    - xder2_lexicon_EM_ES20250723.txt: Contiene las diferentes palabras que conforman el léxico de SNOMED CT en español de la EM y el número de apariciones de cada una de ellas.
    - xder2_lexicon_ excluded_ES20250723.txt: Palabras excluidas o stopwords.
    - xder2_lexicon_equivalents_ES20250723.txt: Conjunto de sinónimos de una sola palabra identificados por el campo wordBlockNumber. En esta tabla también se indica el tag semántico asociado al concepto de cada descripción.

- En la carpeta “Documentation” se encuentra la guía en PDF con el nombre xdoc_LexiconGuide_ES20250723.pdf.

El compromiso del CNR de SNOMED CT para España es publicar nuevas versiones del producto semestralmente, actualizando el recurso 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, en función de la combinación de ediciones y extensiones seleccionadas. Este documento incluye las 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 con tilde y sin tilde, o las que contienen otras marcas diacríticas, y se separan de acuerdo con el uso de mayúsculas o minúsculas. Por ejemplo, tacrolimus es diferente de tacrólimus. Las palabras con grafía incorrecta están incluidas en el léxico. De hecho, el léxico ya se está utilizando para un proceso de mejora continua de la calidad ortográfica, de acuerdo con las reglas del español de España. Las *stopwords* están excluidas del Lexicon.

| Campo | Tipo | Detalle |
|---|---|---|
| word | String | Palabras que forman parte del léxico. |
| n | Integer | Número de apariciones de la palabra en los términos de SNOMED CT. |

## Tabla Palabras Excluidas

El fichero de palabras excluidas (stopwords) incluye una lista de palabras que están excluidas de la lista de 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.

| Campo | Tipo | Detalle |
|---|---|---|
| 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. |
| keyWord | String | Palabra excluida, de una longitud máxima de 8 caracteres |

## Tabla Palabras Equivalentes

La tabla está estructurada como un conjunto de sinónimos identificados por el mismo wordBlockNumber. Se centra en la búsqueda de descripciones de una sola palabra y sus sinónimos, también de una sola palabra.

Este contenido permite realizar búsquedas mejoradas, basadas en palabras relacionadas semánticamente.

| Campo | Tipo | Detalle |
|---|---|---|
| wordBlockNumber | Integer | Compartido por un conjunto de palabras equivalentes, que tienen un significado idéntico o similar. |
| wordText | String | Una palabra, acrónimo o abreviación que es equivalente al wordText de otras filas que comparten el mismo wordBlockNumber. |
| semanticTag | String | Tag semántico perteneciente al Fully Specified Name (FSN) y que relaciona los términos que comparten el mismo wordBlockNumber. |

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 de este es la resolución de variaciones locales del idioma, ya que permite reducir la necesidad de añadir nueva descripciones y sinónimos en las extensiones locales. Además, puede ser revisado para incorporar contenido que actualmente no esté disponible, pero que sí se utilice en la práctica.

# Índice de conceptos y descripciones

## Introducción

El Centro Nacional de Referencia (CNR) de SNOMED CT para España con el apoyo del grupo de trabajo de Procesamiento de Lenguaje Natural de la Unidad Técnica incorpora a su cartera de productos el **Índice de conceptos y descripciones de SNOMED CT en español**, que contiene los conceptos, descripciones y tags semánticos de las diferentes ediciones en español de SNOMED CT.

- Spanish Edition (SE), que es traducción al español de la International Edition
- Extensión para España (EE)
- Extensión para España de Medicamentos (EM)

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 Descarga SNOMED CT (https://snomed-ct.sanidad.gob.es/snomed-ct/solicitudLicencia.do?metodo=volverDeLogin). Este paquete se libera con el recurso ***Lexicon e* Índice de conceptos y descripciones de SNOMED CT en español** en el que se incluye un archivo denominado Índice_SNOMED_CT, que contiene los siguientes ficheros:

- En la carpeta "Content" se encuentra un documento .txt, que contiene información sobre los diferentes conceptos que se encuentran en las ediciones en español de SNOMED CT.

- En la carpeta “Documentation” se encuentra esta guía en PDF con el nombre xdoc_IndexGuide_ES20250723.pdf.

El compromiso del CNR de SNOMED CT para España es publicar nuevas versiones del producto semestralmente, actualizando el recurso con los contenidos de las nuevas versiones de los productos SNOMED CT en español que se publican de manera periódica.

## Tabla Índice

Contiene la información referida a todos los componentes de SNOMED CT en español. Específicamente, este documento incluye los conceptos, descripciones, tipo de descripciones y tags semánticos de cada componente.

| Campo | Tipo | Detalle |
|---|---|---|
| sctid | Integer | Identificador único del concepto de SNOMED CT. |
| descriptor | String | Descripción del componente de SNOMED CT al que el SCTID hace referencia. |
| tipo_descriptor | String | Tipo de descripción, puede ser *sinónimo* o *descripción completa*. |
| tag_semantico | String | Tag semántico perteneciente a la descripción completa y que relaciona los términos que comparten el mismo SCTID. |

Este documento se extrae desde la Implementación de Referencia de SNOMED CT en Base de Datos MySQL (IRBD MySQL), también publicada en el Área de Descarga de SNOMED CT bajo el nombre **IRBD_Multibase_MySQL_snapshot**. A partir de la siguiente *query* se podría extraer la misma información:

```
SELECT
a.id AS sct_concept,
b.term AS sct_descriptor,
c.term AS sct_descriptor_type,
SUBSTRING_INDEX(SUBSTRING_INDEX(d.term, '(', -1), ')', 1) AS semantic_tag
FROM glsc_concept a
JOIN glsd_description_es b ON a.id = b.conceptId
JOIN glsd_description_es c ON b.typeId = c.conceptId
JOIN glsd_description_es d ON a.id = d.conceptId
WHERE
a.active = 1 AND
b.active = 1 AND
b.languageCode = 'es' AND
c.typeId = 900000000000013009 AND
d.typeId = 900000000000003001 AND
d.active = 1 AND
d.languageCode = 'es'
```

El contenido de este fichero resulta valioso para aplicaciones en lingüística computacional y en el entrenamiento de modelos de inteligencia artificial. Este recurso puede ser utilizado para realizar *fine-tuning* de *Large Language Models* (LLM), con el objetivo de mejorar su rendimiento en tareas específicas del ámbito clínico, como la codificación automática de textos médicos utilizando SNOMED CT. Además, puede emplearse como guardarraíl para limitar o guiar las respuestas de los modelos dentro del dominio clínico, asegurando el uso preciso y coherente de terminología clínica estandarizada.