UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

Departamento de Electrónica y Computación

TESIS DOCTORAL

MÉTODOS SEMÁNTICOS AUTOMATIZADOS DE APOYO A LA GESTIÓN Y A LA INTEROPERABILIDAD DE LA INFORMACIÓN CLÍNICA

Presentada por:

Jose Luis Iglesias Allones

Dirigida por:

María Jesús Taboada Iglesias

Santiago de Compostela, Septiembre de 2014

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Dna. María Jesús Taboada Iglesias,

Profesora Titular de Universidad del Área de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Santiago de Compostela

Como directora de la Tesis Doctoral titulada **Métodos semánticos automatizados de apoyo a la gestión y a la interoperabilidad de la información clínica**

Presentada por **Jose Luis Iglesias Allones**

Alumno del Programa de Doctorado en Investigación en Tecnologías de la Información.

Autoriza la presentación de la tesis indicada, considerando que reúne los requisitos exigidos en el artículo 34 del reglamento de Estudios de Doctorado y que como directora de la misma no incurre en las causas de abstención establecidas en la ley 30/1992

María Jesús Taboada Iglesias
Directora de la tesis

Jose Luis Iglesias Allones
Autor de la tesis

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

A mis padres y a Tamara

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Agradecimentos

Agradecer a mi directora de tesis, Chus, su dedicación y sus buenos consejos durante todo el transcurso de la tesis. También me gustaría agradecer a María Meizoso, Diego Martínez, Raimundo Lozano y María Jesús Sobrido su colaboración durante la tesis y a Helen Parkinson y a los integrantes del grupo de investigación *Functional Genomics Production Team* por la gran acogida que me dieron durante mi estancia de investigación predoctoral en el *European Bioinformatics Institute* y por la oportunidad de conocer su trabajo.

Expresar mi agradecimiento al apoyo económico recibido durante estos años. Me gustaría destacar la financiación recibida por el Ministerio de Economía y Competitivad a través de una beca de Formación de Personal Investigador (FPI); y también debo hacer mención a dos proyectos que han financiado esta investigación: *OntoNeuroPhen* (FIS2012-PI12/00373) del Instituto de Salud Carlos III y *Gestión de Terminologías Médicas para Arquetipos* (TIN2009-14159-C05-05) del Ministerio de Economía y Competitivad.

Por último, quería agradecer especialmente a mis padres y a Tamara su paciencia y apoyo durante estos años.

Santiago de Compostela, Septiembre de 2014

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Índice general

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Capítulo 1

Introducción

1.1. Antecedentes

La historia clínica electrónica (HCE) es el conjunto de documentos en formato electrónico que contiene los datos, valoraciones e informaciones de cualquier índole, sobre la situación y la evolución clínica de un paciente a lo largo del proceso asistencial. La implantación de los sistemas de HCE en las instituciones sanitarias es cada vez mayor, lo que abre nuevas oportunidades relacionadas con la explotación de la información clínica presente en dichos sistemas.

Desgraciadamente, en la actualidad es muy frecuente que en las instituciones sanitarias la información clínica esté fragmentada en diversos sistemas independientes no interoperables entre sí. Esta situación dificulta el acceso completo a la información de un paciente desde un único sistema o institución, ocasionando eventualmente una atención médica inadecuada.

La interoperabilidad de los sistemas de salud se define como la habilidad de los sistemas para intercambiar y entender la información relacionada con la salud de los pacientes [121]. La necesidad de interoperabilidad entre los diversos sistemas de HCE, como soporte a la continuidad asistencial, está plenamente establecida desde hace ya bastante tiempo y la aplicación de estándares se ha perfilado como la principal estrategia para conseguirla.

En las últimas dos décadas varias investigaciones financiadas por la Unión Europea han tratado el tema de la interoperabilidad de la HCE [30, 29, 121]. El estudio EHR IMPACT [29] ha analizado en profundidad el impacto socio-económico de una HCE interoperable. El informe SemanticHEALTH [121] ha definido una hoja de ruta con recomendaciones y acciones encaminadas a lograr la interoperabilidad semántica de la HCE, entre las que se

incluyen: el uso de una arquitectura de información para la HCE basada en un modelo dual; la creación y el uso compartido de repositorios de modelos de datos clínicos consensuados por expertos (concretamente, se recomienda usar los llamados arquetipos clínicos); y finalmente el uso consistente de terminologías clínicas dentro de la HCE, principalmente mediante la creación de enlaces entre ítems de información clínica, presentes en los registros y modelos de datos de la HCE, con conceptos estándar de las terminologías.

Varias organizaciones (entre ellas el Comité Europeo de Normalización [102] y la Fundación openEHR [98]) han estado trabajando en los últimos años en el diseño de una arquitectura de información rigurosa y estable, basada en un modelo dual, para la comunicación e intercambio de la HCE. Esta arquitectura ha quedado reflejada en la norma ISO EN 13606 [38, 39], único estándar completo sobre interoperabilidad de HCE a nivel internacional. La arquitectura de modelo dual propuesta en esta norma organiza la HCE en dos niveles conceptuales: el modelo de referencia y de arquetipos. El modelo de referencia detalla el conjunto de entidades que forman los bloques de construcción genéricos de la HCE. Los arquetipos son modelos clínicos de conocimiento definidos para capturar de forma ordenada y sistemática la información de pacientes en escenarios clínicos determinados (p.e. el informe de alta o la medición de la presión sanguínea de un paciente). Los distintos escenarios clínicos son modelados con los arquetipos mediante las entidades básicas, los atributos, y la estructura lógica del modelo de referencia.

La característica más destacada de la arquitectura de modelo dual es la separación entre la información (implementada en el modelo de referencia) y el conocimiento (definido en los arquetipos por expertos clínicos) que maneja un sistema de HCE. El uso de un modelo de referencia común en los sistemas de información de distintas instituciones posibilita el intercambio de un extracto de HCE sin necesidad de acuerdo previo del contenido clínico. Mientras que el uso de arquetipos, como especificaciones de modelos de datos clínicos, facilita la captura y el intercambio de información de forma sistemática en escenarios clínicos determinados, favoreciendo la interoperabilidad.

En la última década, importantes instituciones (entre ellas NEHTA [103], NHS [104], Centre for eHealth de Suecia [101] y la fundación openEHR [98]) han estado desarrollando arquetipos openEHR para modelar formalmente contenido clínico. Paralelamente, han surgido repositorios para dar soporte a la creación, revisión y gestión de los arquetipos [92, 90]. Un repositorio puede ser visto como una librería de arquetipos con funcionalidades avanzadas

(búsqueda, autoría, etc.) que facilita la colaboración de expertos del dominio clínico en el desarrollo de arquetipos.

Por otra parte, la informática médica ha estado trabajando intensamente en la definición y actualización de terminologías clínicas. Estas pueden definirse como listas de términos empleados en el ámbito médico. Normalmente, tienen algunas características ontológicas para describir formalmente los términos y sus relaciones. Las terminologías clínicas han surgido para ser usadas por los sistemas de información para capturar, procesar y transferir los datos clínicos de una forma consistente y estandarizada. Las terminologías además son claves en varios escenarios: en la integración de diversos sistemas de información, en la conexión de la HCE con los entornos de soporte a la decisión y en la reutilización de la información clínica (generada durante el proceso asistencial de los pacientes) para otros fines, como puede ser la investigación, la gestión hospitalaria o la evaluación de la calidad. En la actualidad, SNOMED-CT es la terminología más completa para codificar todos los aspectos de la HCE [100].

Actualmente existe un importante consenso en que la integración de las terminologías clínicas en la HCE, y más concretamente en los modelos de datos clínicos estructurados (tales como los arquetipos), es un paso clave en el camino hacia la interoperabilidad semántica [29, 121]. La Fundación openEHR [98], surgida para desarrollar especificaciones abiertas en sistemas de HCE, es una de las organizaciones más comprometidas con el proceso de integración de terminologías y modelos clínicos. Las especificaciones openEHR han incorporado una funcionalidad para que los ítems de información incluidos en sus modelos clínicos (i.e. los arquetipos openEHR) sean definidos por los expertos con términos del lenguaje natural y puedan también ser enlazados a conceptos de terminologías clínicas externas.

1.2. Problemas y retos en la gestión y en la interoperabilidad semántica de la HCE

A pesar del creciente interés que se ha experimentado en la definición de terminologías clínicas y en el modelado formal de contenido clínico, todavía existen varias cuestiones que dificultan la gestión eficiente y la interoperabilidad semántica de la HCE, entre ellas destacamos las siguientes: (1) gran parte de la información clínica de la HCE todavía se define y recopila en lenguaje natural sin enlazar con las terminologías clínicas estándar, (2) las instituciones encargadas del modelado de arquetipos clínicos (llamados a ser elementos claves en

la HCE en un futuro próximo) no han seguido un proceso riguroso que asegure la calidad de los mismos, por lo que los actuales repositorios pueden contener arquetipos con estructuras no validadas o con contenidos clínicos solapados; y, finalmente, (3) los repositorios de arquetipos no cuentan con sistemas avanzados de búsqueda para facilitar el acceso y reuso de los arquetipos clínicos.

1.2.1. Ausencia de enlaces entre datos clínicos y terminologías estándar

En la actualidad, gran parte de la información clínica registrada en la HCE está definida en lenguaje natural, y no presenta enlaces con conceptos de SNOMED-CT ni con otras terminologías clínicas estándar. Una mayor presencia de enlaces terminológicos en la HCE permitiría capturar la información clínica de una forma consistente y estandarizada, facilitando el intercambio, la gestión y el acceso eficiente a dicha información.

La creación de los enlaces terminológicos (también llamados mappings o bindings) implica localizar en SNOMED-CT (o en cualquier otra terminología clínica) los conceptos equivalentes a los términos clínicos definidos en la HCE. En teoría, cabe pensar que la tarea de enlazado debería ser realizada por expertos (idealmente, por profesionales médicos con experiencia en SNOMED-CT) asistidos por herramientas (llamadas frecuentemente navegadores) para visualizar y acceder al contenido de SNOMED-CT. Sin embargo, en la práctica esta metodología 'semi-manual' (en la que gran parte del trabajo y la responsabilidad del proceso de mapping lo tienen los expertos y no las herramientas) no ha resultado ser operativa. R. Qamar, en su trabajo [105], expone que se han realizado varios intentos de mapping o enlazado con esta metodología, y se ha comprobado que el proceso es muy costoso, ya que exige mucho tiempo a personal altamente cualificado, originando además enlaces imperfectos e incompletos.

Hay que destacar que existen varias cuestiones que dificultan especialmente la búsqueda de conceptos en SNOMED-CT y por ende los procesos de enlazado o mapping con esta terminología. En primer lugar, el gran tamaño y granularidad de esta terminología. SNOMED-CT contiene más de 300.000 conceptos, siendo en la actualidad la terminología clínica de mayor tamaño. En segundo lugar, una inmensa mayoría de los profesionales médicos no están lo suficientemente familiarizados con la terminología SNOMED-CT para llevar a cabo búsquedas efectivas de conceptos en ella. En tercer lugar, SNOMED-CT se actualiza semestralmente. En

cada actualización surgen y desaparecen conceptos, por lo que los enlaces existentes pueden quedar desactualizados y necesitan ser revisados.

Limitaciones de los navegadores actuales de SNOMED-CT

Además de las anteriores cuestiones, hay otro aspecto clave que merece especial atención que está dificultando la creación de enlaces terminológicos: las funcionalidades de búsqueda de los navegadores actuales (usados por los expertos para acceder a SNOMED-CT) [33, 60, 58, 93] son en la actualidad bastante básicas y no están lo suficientemente maduras para llevar a cabo búsquedas eficientes en una terminología del tamaño de SNOMED-CT [107, 70, 108]. A continuación, se detallan varias limitaciones que presentan los navegadores actuales.

Los sistemas de búsqueda de los navegadores actuales usan esencialmente técnicas clásicas de similitud de cadenas de texto con el objetivo de localizar un conjunto de conceptos similares léxicamente al término de búsqueda. Dichos sistemas han sido desarrollados para hacer una primera criba o selección (frecuentemente de decenas o cientos de conceptos) dado un término de búsqueda, más que para obtener un sólo concepto exacto final. Por tanto, todavía es necesaria una alta implicación de los expertos para seleccionar los conceptos más precisos.

Los navegadores actuales apenas han incluido técnicas de normalización léxica para combatir las pequeñas diferencias léxicas (frecuentes en el lenguaje natural) que evitan, en no pocas ocasiones, la detección del concepto correcto. Los navegadores tampoco han incorporado técnicas de expansión de consultas. La expansión de consultas es una técnica que ha sido ampliamente estudiada y usada en las disciplinas del Procesamiento del Lenguaje Natural y la Recuperación de la Información. Estas técnicas permiten reformular y expandir los términos de búsqueda con términos alternativos con significados similares. El uso de sinónimos, procedentes de recursos lingüísticos, es una de las metodologías más habituales para la generación de los términos alternativos. Hay que considerar que normalmente múltiples términos pueden hacer referencia a un mismo concepto médico y que la sinonimia definida en SNOMED-CT todavía no es, en absoluto, completa. Por tanto, la aplicación de estas técnicas podría incrementar notablemente el rendimiento de las búsquedas en SNOMED-CT.

Los navegadores no han usado en ningún momento las relaciones semánticas de SNOMED-CT durante la búsqueda de conceptos relevantes; estos consideran la terminología SNOMED-CT como una enorme lista de conceptos no conectados. SNOMED-CT cuenta con una rica red de relaciones semánticas creada para definir lógicamente el significado de los conceptos.

En nuestra opinión las funcionalidades de búsqueda de los navegadores deberían usar las relaciones semánticas, junto con sus descripciones textuales de los conceptos, para disponer de más información sobre el significado de los conceptos.

Estado actual de los enlaces terminológicos en los arquetipos clínicos

Un informe elaborado por el instituto EuroRec para evaluar el estado actual de los arquetipos ha reconocido que los enlaces entre los ítems/términos de los arquetipos y los conceptos terminológicos todavía son muy escasos y ha destacado la necesidad de aumentar dichos enlaces en los próximos años [63]. Esta situación, junto con las limitaciones actuales de los navegadores de SNOMED-CT para realizar búsquedas eficientes, ha motivado la reciente aparición de una línea de investigación dedicada al desarrollo de herramientas para automatizar el mapping entre arquetipos clínicos y SNOMED-CT [135, 105].

El sistema MoST [105] es una de las metodologías más avanzadas que ha surgido para enlazar de forma semi-automática los términos clínicos de los arquetipos openEHR a conceptos de SNOMED-CT. MoST incorpora dos etapas: un proceso automático de búsqueda de conceptos candidatos y un proceso manual de selección de los mappings finales realizado por expertos clínicos. La evaluación de MoST, realizada con más de 100 términos clínicos de 4 arquetipos openEHR, obtuvo una media de 5.5 conceptos candidatos por término. La elección final del concepto correcto, realizada por un grupo de expertos, determinó que aproximadamente el 70 % de las veces el código correcto está entre los conceptos candidatos sugeridos por la herramienta. Estos resultados demuestran que todavía queda bastante camino para automatizar los procesos de mapping con SNOMED-CT. Estos todavía demandan una alta participación de expertos para desambiguar y seleccionar los conceptos más precisos.

El sistema MoST y otras herramientas orientadas a enlazar arquetipos con SNOMED-CT han mejorado en algunos aspectos a los navegadores actuales de SNOMED-CT, incluyendo algunas técnicas lingüísticas adicionales y de normalización léxica. Sin embargo, todavía presentan ciertas limitaciones en común con los navegadores, tales como la falta de técnicas de expansión de consultas y de uso de las relaciones semánticas de SNOMED-CT. Además, en nuestra opinión estas herramientas no han explotado lo suficiente la estructura de los arquetipos para extraer contexto de sus términos; prácticamente han considerado cada término del arquetipo de forma aislada. Las herramientas podrían considerar varias cuestiones sobre el contexto de los términos que deben ser mapeados a SNOMED-CT: el tipo de arquetipo donde está ubicado el término buscado (p.e. considerar si es un arquetipo para registrar observacio-

nes o es un arquetipo para indicar procedimientos o acciones futuras), la ubicación del término en el arquetipo y los términos vecinos dentro del mismo. Esta información de contexto sobre los términos podría ser aprovechada por las herramientas para obtener mappings más precisos y para desambiguar cuando haya varios conceptos candidatos.

1.2.2. Falta de procesos formales para modelar arquetipos y de criterios de calidad

En la actualidad, otra problemática que presentan los arquetipos actuales, es que las instituciones encargadas de su desarrollo no han seguido un proceso riguroso que asegure la calidad de los mismos [63]. Sin embargo, para que los arquetipos sean adoptados ampliamente en los sistemas de información sanitarios deben tener una calidad contrastada. Recientemente, se han empezado a estudiar criterios para evaluar la calidad de los arquetipos, así por ejemplo, el informe elaborado por el instituto EuroRec ha destacado la necesidad de métodos formales para la validación de la estructura y el contenido de los arquetipos y ha recomendado reducir los solapamientos existentes entre los arquetipos de los repositorios [64].

1.2.3. Ausencia de sistemas avanzados para gestionar y buscar información en los repositorios de arquetipos

Varios estudios han destacado que las funcionalidades existentes para gestionar y buscar contenido clínico en los repositorios de arquetipos son insuficientes [119, 18]. Los repositorios actuales, dado un término de búsqueda, se limitan a buscar equiparaciones exactas de dicho término en las definiciones de los ítems de los arquetipos. La falta de sistemas avanzados de búsqueda dificulta la tarea de localizar arquetipos relevantes para entornos o aplicaciones concretas, obstaculiza la conversión de modelos propietarios (presentes aún en muchas instituciones sanitarias) a modelos basados en arquetipos y tiene un impacto negativo en la reutilización de los arquetipos; y hay que considerar que el uso compartido de un mismo conjunto de arquetipos entre distintas instituciones ha sido establecido como una de las medidas necesarias para lograr la interoperabilidad [121]. Los mismos estudios han destacado la necesidad de sistemas de búsqueda más avanzados en los repositorios, capaces de localizar contenido clínico en los arquetipos de forma más automatizada y eficiente [119, 18]. En nuestro opinión, estos sistemas no deberían depender sólo de la coincidencia léxica de los

términos, sino que deberían apoyarse también en los conceptos, en la sinonimia y en las relaciones semánticas definidas en las terminologías clínicas, especialmente en SNOMED-CT, para optimizar y enriquecer las búsquedas.

1.3. Objetivos

1.3.1. Objetivos generales

Las terminologías clínicas han surgido para capturar la información clínica de forma consistente y estandarizada, favoreciendo con ello la gestión eficiente de dicha información, la interoperabilidad semántica entre diferentes instituciones sanitarias, y en última instancia la calidad asistencial de los pacientes. Sin embargo, en la actualidad la información clínica de la HCE está mayoritariamente definida en lenguaje natural y apenas está enlazada con conceptos terminológicos, por lo que no se está aprovechando todo el potencial de las terminologías. Los **objetivos generales** de esta tesis son: aportar metodologías automáticas para **facilitar el enlazado de datos clínicos con terminologías estándar del ámbito clínico** y mejorar la búsqueda y el acceso a conjuntos de datos clínicos explotando la semántica de las terminologías.

1.3.2. Objetivos específicos

El objetivo general puede desglosarse en varios objetivos específicos:

- Diseñar y desarrollar un método de búsqueda automático en SNOMED-CT, centrado en localizar conceptos relevantes dada una frase o término clínico de búsqueda, sin ningún tipo de contexto adicional. Este método tiene el propósito de mejorar el rendimiento de las búsquedas (en términos de precisión y automatización) respecto a los sistemas de búsqueda implementados en los navegadores actuales de SNOMED-CT.

- Reducir al máximo la participación de los expertos en los procesos de mapping y codificación con SNOMED-CT. En la actualidad, existen varias propuestas y herramientas que ayudan a los expertos en estos procesos. Sin embargo, muchas de estas no han conseguido automatizar demasiado el proceso ya que sólo incluyen técnicas elementales de similitud de cadenas de texto para detectar equiparaciones en SNOMED-CT.

- Diseñar y desarrollar una metodología automática para enlazar información textual de modelos clínicos estructurados (concretamente arquetipos openEHR) con conceptos de

la terminología SNOMED-CT. El método debe considerar y aprovechar las características y la estructura de los modelos para extraer información de contexto y mejorar así el enlazado con SNOMED-CT.

– Mejorar la gestión y el manejo de los repositorios de arquetipos mediante el uso de los conceptos y las relaciones semánticas de la terminología SNOMED-CT. Especialmente, se pretende optimizar y enriquecer las búsquedas en los repositorios para favorecer la (re)utilización de los arquetipos y la detección de solapamientos en el contenido clínico de estos.

1.4. Estructura de la memoria

Capítulo 2: Estado del arte

Este capítulo presenta el contexto en el que se sitúa el trabajo de investigación. En primer lugar, se introduce el concepto de Historia Clínica Electrónica, junto con los estándares y roles implicados en promover su interoperabilidad. A continuación, se describen detalladamente los arquetipos openEHR y la terminología SNOMED-CT, elegidos para la investigación. Posteriormente, se hace una revisión literaria de los principales trabajos relacionados a la codificación de datos clínicos con SNOMED-CT. Finalmente, se exponen las principales técnicas existentes para encontrar automáticamente correspondencias entre conceptos, así como los principales trabajos de segmentación ontológica centrados en SNOMED-CT.

El núcleo del trabajo de investigación es descrito en los capítulos 3,4,5 y 6. El capítulo 3 recopila e introduce de forma general el conjunto de métodos propuestos en la tesis doctoral para la búsqueda de conceptos en la terminología SNOMED-CT. Los capítulos 4,5 y 6, con una estructura similar al de un artículo de investigación, presentan tres aplicaciones concretas en las que fundamentalmente se usan los métodos propuestos en el capítulo 3 para enlazar automáticamente términos clínicos procedentes de diferentes fuentes a la terminología SNOMED-CT. Estos capítulos muestran el modo en que los métodos fueron combinados y ajustados apropiadamente para optimizar los resultados de las búsquedas en SNOMED-CT en cada una de las aplicaciones consideradas. Además, a lo largo de estos capítulos también se presentan varios servicios que explotan la semántica de las terminologías para mejorar la búsqueda y el acceso a conjuntos extensos de datos clínicos. A continuación, se comenta más en detalle el contenido de cada uno de estos capítulos.

Capítulo 3: Métodos de búsqueda semántica en SNOMED-CT

Este capítulo recopila el conjunto de técnicas propuestas en esta tesis doctoral para la búsqueda de conceptos en la terminología SNOMED-CT. Muchas de las técnicas presentadas en este capítulo son (re)usadas en las distintas herramientas de mapping descritas en los capítulos 4,5 y 6. El capítulo 3 puede verse como un catálogo de técnicas de búsqueda orientadas a SNOMED-CT, el cual puede ser de utilidad, como fuente de consulta, para el desarrollo de futuras herramientas de mapping centradas en SNOMED-CT e incluso en otras terminologías clínicas con similares características.

Capítulo 4: Mapping de términos clínicos a SNOMED-CT

Este capítulo presenta una herramienta de búsqueda centrada en localizar conceptos relevantes de SNOMED-CT dado un término de búsqueda del que no se dispone de ningún contexto. El capítulo incluye una evaluación en la que se compara el rendimiento de la herramienta con las funcionalidades de búsqueda de dos de los navegadores más populares de SNOMED-CT.

Capítulo 5: Mapping entre arquetipos OpenEHR y SNOMED-CT

Este capítulo realiza un análisis de los arquetipos clínicos presentes en repositorios públicos para examinar cómo organizan la información clínica. Describe un método para enlazar automáticamente la información clínica de los arquetipos a conceptos de la terminología SNOMED-CT. El capítulo además realiza una evaluación del método e incluye una extensa discusión sobre las dificultades detectadas durante el mapping.

Capítulo 6: Gestión de arquetipos dirigida por subontologías de SNOMED-CT

Este capítulo presenta una aproximación encaminada a añadir funcionalidades semánticas a los repositorios de arquetipos. El capítulo en primer lugar describe una metodología para anotar semánticamente los arquetipos mediante segmentos de SNOMED-CT. Posteriormente, se presentan dos servicios que demuestran las ventajas de la anotación semántica de los arquetipos mediante segmentos SNOMED-CT. El primer servicio está orientado a identificar relaciones y solapes entre arquetipos, mientras que el segundo gestiona búsquedas semánticas de conceptos clínicos en los arquetipos. Finalmente, el capítulo incluye una evaluación del servicio de búsqueda semántica en la que se hace una comparativa con los buscadores presentes en los actuales repositorios de arquetipos.

Capítulo 7: Conclusiones y trabajo futuro

En el último capítulo se presentan las principales contribuciones y conclusiones de la tesis y se proponen posibles líneas de trabajo futuro.

1.5. Publicaciones

A continuación se muestra la lista de artículos publicados durante el transcurso de esta tesis doctoral:

Publicaciones en revistas JCR

- Jose L. Allones, Diego Martínez, María J. Taboada. "Automated mapping of clinical terms into SNOMED-CT. An application to codify procedures in pathology". *Journal of medical systems*, Vol:38, DOI: 10.1007/s10916-014-0134-x (disponible online), 2014.
- Jose L. Allones, María J. Taboada, Diego Martínez, Raimundo Lozano y María J. Sobrido. "SNOMED CT module-driven clinical archetype management". *Journal of biomedical informatics* 46, no. 3: 388-400, 2013.
- María Meizoso, Jose L. Allones, Diego Martínez y María J. Taboada. "Semantic similarity-based alignment between clinical archetypes and SNOMED CT: an application to observations". *International journal of medical informatics* 81, no. 8: 566-578, 2012.

Publicaciones en congresos

- Jose L. Allones, David Penas, María J. Taboada, Diego Martínez y Serafín Tellado. "A study of semantic proximity between archetype terms based on SNOMED CT relationships". En *Joint Workshop on Process-oriented Information Systems in Healthcare (ProHealth'12) and Knowledge Representation for Health Care (KRH4C'2012)*, Tallinn, Estonia. Disponible en Process Support and Knowledge Representation in Health Care, págs. 98-112. Springer Berlin Heidelberg, 2013.
- Jose L. Allones, María Meizoso, María J. Taboada, Diego Martínez y Serafín Tellado. "Combining lexical and structure-based methods to align clinical archetypes to SNOMED CT". En *16th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*. Disponible en Advances in Smart Systems Research, Workshop Papers from KES Conferences, Vol. 2 No. 1, págs. 27-32. Future Technology Publications, 2012.

– Jose L. Allones, María J. Taboada, María Meizoso, Diego Martínez y Serafín Tellado. "Combining mapping methods to align clinical archetypes to SNOMED CT". En *10th Terminology and Knowledge Engineering Conference (TKE 2012)*. Madrid, España, 2012.

– María Meizoso, Jose L. Allones, María J. Taboada, Diego Martínez y Serafín Tellado. "Automated mapping of observation archetypes to SNOMED CT concepts". En *4th international conference on Interplay between natural and artificial computation*. Disponible en Foundations on Natural and Artificial Computation, págs. 550-561. Springer Berlin Heidelberg, 2011.

Capítulo 2

Estado del arte

2.1. Interoperabilidad de la Historia Clínica Electrónica

La historia clínica electrónica (HCE) es el conjunto de documentos en formato electrónico que contiene los datos, valoraciones e informaciones de cualquier índole, sobre la situación y la evolución clínica de un paciente a lo largo del proceso asistencial. La progresiva adopción de la HCE provocará un impacto beneficioso en muchas áreas médicas como son: la calidad de la atención al paciente, la investigación clínica o la investigación en salud pública [121].

En las últimas dos décadas varios proyectos de investigación [11, 49] han ido perfilando los requisitos necesarios para el desarrollo de la HCE. Estos requisitos han sido aprobados por ISO[1] en la especificación técnica ISO 18308 en el año 2011 [36]. Los requisitos están formulados para garantizar que el sistema de HCE sea clínicamente válido y fiable, sea éticamente adecuado, cumpla con los requisitos legales vigentes, apoye las buenas prácticas clínicas y facilite el análisis de datos para una multitud de propósitos.

La interoperabilidad de los sistemas de información de la salud se define como la habilidad de los sistemas para intercambiar y entender el conocimiento y la información relacionada con la salud de los pacientes sin que los usuarios necesiten conocer las características particulares de dichos sistemas [121]. La necesidad de interoperabilidad en los sistemas de salud como base de la continuidad asistencial está plenamente establecida desde hace ya bastante tiempo y la estandarización se ha situado como la estrategia más destacada para conseguirla [89]. A medida que se ha profundizado en la aplicación de la interoperabilidad en el ámbito

[1] International Organization for Standardization. `www.iso.org/`

de la salud, se ha ido descubriendo que es necesario establecer vínculos o relaciones entre las organizaciones a distintos niveles. En consecuencia, también han surgido diferentes tipos de interoperabilidad para abordar cada uno de estos niveles; estas son la interoperabilidad técnica, sintáctica, semántica y organizativa [89]. A continuación, se describen las características de cada uno de los tipos de interoperabilidad y se citan algunas normas y estándares que han surgido para dar soporte a dichos tipos.

– Interoperabilidad técnica: es la base en la que se sustenta la conexión entre sistemas. La interoperabilidad técnica define los interfaces, tanto físicos como lógicos, como pueden ser los protocolos de comunicación y el formato técnico de los datos, que permiten que los sistemas puedan intercambiar información. La interoperabilidad técnica se encuentra muy avanzada, ya que no es exclusiva del escenario sanitario y su desarrollo ha sido necesario en otros ámbitos.

  Algunas de las normas que dan soporte a la interoperabilidad técnica son: XML para describir el formato de los ficheros, SOA para desplegar servicios y TCP/IP para el protocolo de comunicaciones.

– Interoperabilidad sintáctica: permite la transferencia de mensajes y documentos entre sistemas. Define los tipos de datos y la estructura y el formato de los mensajes y documentos intercambiados y permite que los sistemas de información puedan interpretar correctamente los datos recibidos aunque no puedan entender completamente el contenido. Los sistemas que únicamente implementan este tipo de interoperabilidad actúan como mensajeros sin intervenir en el contenido del mensaje y sin poder responder dependiendo del mismo. Las especificaciones para tipos de datos de la ISO 21090 [37] o los modelos de referencia de HL7 V3 [26], de ISO 13606-1 [38] o de openEHR [128] son ejemplos de normas que podrían usarse para lograr la interoperabilidad sintáctica.

– Interoperabilidad semántica: se consigue cuando la información circula entre distintos sistemas sin que el significado original de dicha información se vea alterado y cada uno de los sistemas entiende por sí mismo lo que el otro le envía y puede actuar en consecuencia de forma automática.

  En el contexto actual, donde se busca la continuidad asistencial, la interoperabilidad semántica tiene el objetivo de que la información dispersa de un paciente, generada en distintas fuentes y momentos, pueda ser agregada y compartida y esté a disposición

de los profesionales allí donde se necesite o pueda ser utilizada fácilmente en usos secundarios como la salud pública o la investigación.

En los últimos años se ha estado trabajando principalmente en dos ámbitos para lograr la interoperabilidad semántica [89]:

- Se han estado desarrollado terminologías clínicas, como SNOMED-CT o LOINC[2], para expresar y codificar la información clínica mediante un vocabulario común.
- Se han creado modelos clínicos para formalizar y compartir la información que se debe recoger para cada concepto clínico (qué debe contener un informe de alta o el índice de Barthel, etc.). Ejemplos de estos modelos son los arquetipos definidos en la ISO 13606-2 [39] y en la especificación de openEHR [127].

– Interoperabilidad organizativa: se consigue cuando las organizaciones comparten un contexto común en sus procedimientos y flujos de trabajo. En este ámbito queda mucho trabajo por realizar. La ISO 13940 [40] ha dado los primeros pasos en este sentido, definiendo un sistema de conceptos que permite definir, entre otras cosas, los flujos de trabajo y los servicios de las organizaciones. Esta norma facilita la creación de contextos comunes entre las organizaciones facilitando en gran medida la interoperabilidad.

2.1.1. Proyectos de investigación sobre interoperabilidad de la HCE

En los últimos años varias investigaciones financiadas por la Unión Europea trataron el tema de la interoperabilidad de la HCE. El proyecto RIDE [30] estableció un conjunto de recomendaciones y acciones para alcanzar la interoperabilidad de la HCE a nivel europeo. El estudio EHR IMPACT [29] analizó el impacto socio-económico de una HCE interoperable y de sistemas de prescripción electrónica. El informe SemanticHEALTH [121] definió una hoja de ruta con recomendaciones y acciones necesarias en distintos ámbitos, como son los registros electrónicos, las terminologías y la salud pública, para lograr la interoperabilidad semántica de la HCE. Además, el informe recomienda abiertamente el uso de una arquitectura basada en el modelo dual (ISO EN 13606), compartir repositorios de modelos de datos clínicos (arquetipos) y el uso consistente de terminologías clínicas (preferiblemente SNOMED-CT) como medidas necesarias para lograr la interoperabilidad semántica.

[2]Logical Observation Identifiers, Names, and Codes. `http://loinc.org/`

2.1.2. Principales actores en la interoperabilidad

ISO EN 13606

La norma ISO EN 13606, desarrollada por un comité técnico del CEN[3], es el único estándar completo sobre interoperabilidad de HCE a nivel internacional. Este estándar define una arquitectura de información rigurosa y estable, basada en un modelo dual, para la comunicación e intercambio de la HCE.

El estándar consta de 5 partes. Las dos primeras partes corresponden a los dos niveles del modelo dual: modelo de referencia y de arquetipos. La parte 3 define un vocabulario de términos que pueden ser utilizado al instanciar el modelo de referencia. También define un conjunto de arquetipos de referencia que se corresponden con las estructuras de datos de los modelos de referencia de openEHR y de HL7 V3. Las partes 4 y 5 especifican principios sobre seguridad en el acceso a la HCE y sobre interfaces de servicio y mensajes para la comunicación de la HCE.

La arquitectura de modelo dual propuesta en la ISO EN 13606 organiza la HCE en dos niveles conceptuales: el modelo de referencia y de arquetipos. El modelo de referencia detalla el conjunto de entidades que forman los bloques de construcción genéricos de la HCE. El modelo de arquetipos establece cómo representar los conceptos clínicos que han de registrarse en la HCE usando los bloques de construcción genéricos del modelo de referencia. Dicho de otro modo, los conceptos del modelo de referencia representan las características estables del dominio y establecen entre sí las relaciones válidas que posibilitan conformar los conceptos del dominio (arquetipos).

La característica más destacada de la arquitectura de modelo dual es la separación entre la información (implementada en el modelo de referencia) y el conocimiento (definido en los arquetipos por expertos clínicos) que maneja un sistema de HCE. La adopción del modelo dual permite a los expertos definir el dominio sin necesidad de conocimientos técnicos, y también facilita que los sistemas HCE pueden adaptarse a los cambios en la práctica clínica.

Fundación OpenEHR

La Fundación openEHR es una organización que promueve el desarrollo de especificaciones abiertas, recursos y herramientas para sistemas de HCE [98]. La arquitectura de información de openEHR, al igual que la ISO EN 13606, consta de dos niveles conceptuales: un

[3]Comité Europeo de Normalización. `https://www.cen.eu/`

modelo de referencia y de arquetipos. Comparte con la norma ISO EN 13606 el modelo de arquetipos. El modelo de referencia de openEHR es más amplio y detallado que el propuesto por ISO EN 13606, incluyendo más tipos de entradas clínicas, como son las observaciones, instrucciones, acciones y evaluaciones. Sin embargo, los tipos de datos elementales de openEHR son distintos a otros estándares, lo que dificulta en gran medida la interoperabilidad. Se espera que esto se solucione con la futura introducción de los tipos básicos propuestos por ISO.

Las especificaciones openEHR son el resultado de varios proyectos de investigación [59, 11] más que de un proceso formal de estandarización. A pesar de ello, sus especificaciones han influido notablemente en el desarrollo de estándares en las principales organizaciones de estandarización: CEN, HL7 e ISO. De hecho, la norma ISO 13606 se puede considerar un subconjunto de la especificación de openEHR [111].

HL7

HL7 (Health Level Seven) es una organización de estandarización para el ámbito de la salud [99]. Opera a nivel internacional y su misión es proveer estándares para facilitar el intercambio electrónico de información clínica. El término HL7 es usado tanto para referirse a la organización como a sus estándares de mensajería.

La organización HL7 cuenta con un proceso para definir un conjunto de herramientas de interoperabilidad (mensajes, documentos electrónicos y modelos de referencia). Esto ha dado origen a varios estándares que facilitan los procesos de intercambio de información de salud. Algunos de los más destacados son los estándares HL7 versión 2, HL7 versión 3 y HL7 CDA. Estas normas definen cómo se empaqueta y transfiere la información entre sistemas, estableciendo el lenguaje, la estructura y los tipos de datos necesarios para el intercambio efectivo de información entre sistemas. Así, por ejemplo, el estándar HL7 CDA (Clinical Document Architecture) define la estructura y semántica de los documentos clínicos intercambiados entre sistemas de información sanitarios [31].

Terminologías

Una terminología puede definirse como una lista de términos utilizados en ámbitos delimitados para favorecer la comunicación entre sistemas. Normalmente, las terminologías tienen algunas características ontológicas para describir formalmente los términos y sus relaciones; además también suelen incluir un sistema de identificadores o códigos.

Las terminologías clínicas contienen listas de términos empleados en el ámbito médico o sanitario. Estas son requeridas por los sistemas de información para capturar, procesar y transferir los datos clínicos de una forma consistente y estandarizada. El uso de terminologías también juega un papel fundamental en varios escenarios: en la integración de sistemas de información, en la conexión de la HCE con los sistemas de soporte a la decisión y en la reutilización de la información clínica (recopilada durante el cuidado de los pacientes) para otros fines, como puede ser la investigación, la gestión hospitalaria o la evaluación de la calidad.

Existe un importante consenso en que el uso de las terminologías clínicas y su alineamiento correcto con los modelos clínicos es un factor clave para aportar significado a la información clínica de los modelos, favoreciendo así la interoperabilidad semántica de la HCE [121].

Algunas de las terminologías clínicas más destacadas en la actualidad son ICD[4], LOINC[5] y SNOMED-CT. Estas tienen distintos fines. ICD es una terminología, impulsada por la Organización Mundial de la Salud, para la clasificación de diagnósticos clínicos. Se usa especialmente en estudios de epidemiología y en la gestión administrativa y sanitaria. LOINC es una terminología formada por un conjunto de identificadores, nombres y códigos para la anotación y documentación de observables clínicas y de laboratorio. Los códigos de LOINC se han creado para representar las preguntas u observables de un test o medición, más que sus valores o respuestas. SNOMED-CT es la terminología de referencia a nivel mundial en el dominio de la HCE. De hecho, varios organismos de estandarización, como ISO o HL7, han seleccionado SNOMED-CT como terminología de referencia.

Existen todavía varios problemas en torno a las terminologías clínicas. En los últimos años ha habido un aumento importante en el tamaño y la complejidad de las terminologías clínicas, debido a que en el ámbito clínico continuamente se van haciendo nuevos descubrimientos. A pesar de esta creciente complejidad, actualmente todavía no existen medios sistemáticos para la elección de los conceptos de terminológicos adecuados a usos específicos en modelos de datos clínicos. También ha habido un incremento en el número de terminologías, y sin embargo no existen guías para recomendar que terminología se debería usar en cada caso. Hay también el problema de que las comunidades de desarrollo de modelos de datos clínicos han elaborado sus propios vocabularios propietarios para protegerse a sí mismos contra el diná-

[4]International Classification of Diseases. `http://www.who.int/classifications/icd/en/`
[5]Logical Observation Identifiers Names and Codes. `http://loinc.org/`

mico mundo de las terminologías clínicas externas. Con el fin de mantener la fiabilidad y la calidad de los datos, existe una urgente necesidad de garantizar que todos los sistemas de información accedan a un conjunto común y acordado de terminologías estándar para codificar y registrar todos los datos clínicos.

2.2. OpenEHR

OpenEHR propone una arquitectura dual formada por un Modelo de Referencia y un Modelo de Arquetipos. El Modelo de Referencia (también llamado Modelo de Información) representa las propiedades genéricas de la información de los registros clínicos. Este es un modelo lógico, abstracto y de alto nivel. Los Arquetipos son modelos basados en el Modelo de Referencia creados para representar las características específicas de un escenario clínico determinado (p.e. el informe de alta o la medición de la presión sanguínea de un paciente).

Los Arquetipos openEHR han sido seleccionados como los modelos de datos clínicos preferidos para la investigación realizada en esta tesis. En las siguientes secciones explicamos las características más relevantes del Modelo de Referencia y de Arquetipos, así como los motivos para la elección de estos modelos de datos clínicos.

2.2.1. Modelo de Referencia de openEHR

El Modelo de Referencia de openEHR define el conjunto de entidades genéricas de la HCE reutilizables por los modelos clínicos o arquetipos. La figura 2.1 muestra una visión global de la estructura de paquetes del Modelo de Referencia de openEHR. A continuación, se comentan brevemente los paquetes más relevantes, comenzando desde el paquete de más bajo nivel.

- El paquete *Support* contiene los conceptos más básicos requeridos por el resto de paquetes. Comprende los subpaquetes *Definitions*, *Identification*, *Terminology* and *Measurement*, así como los tipos de datos primitivos (Integer, Real, etc).
- El paquete *Data_types* ofrece una serie de tipos de datos genéricos y específicos del ámbito clínico requeridos para modelar información clínica de más alto nivel. Entre los tipos incluidos en este paquete está texto, cantidades, fechas, etc.
- El paquete *Data_structures* define una serie de estructuras de datos genéricas (p.e. listas, tablas, árboles) para ser usadas en arquetipos.

– El paquete *Common* consta de paquetes con conceptos abstractos y patrones de diseño usados en modelos de nivel más alto en openEHR.

– El paquete *Security* define el control de acceso y las opciones de privacidad para la información en openEHR.

– El paquete *EHR* contiene la estructura de más alto nivel, un registro (EHR). Está formada por los siguientes componentes:

  - *EHR_access*: un objeto que controla las opciones de acceso al registro.
  - *EHR_status*: un objeto con información de control y estado del registro.
  - *Compositions*: los contenedores con todo el contenido clínico y administrativo del registro.
  - *Directory*: una estructura jerárquica de carpetas para organizar lógicamente los objetos *Compositions*.

– El subpaquete *Entry* contiene los modelos de información claves para esta investigación. Estos definen las estructuras de información necesarias para capturar en la HCE los datos clínicos generados durante un encuentro clínico. El documento de especificaciones *Information Model* de openEHR [129] establece que toda la información generada durante un evento clínico es registrado como una instancia de una subclase del modelo *Entry*.

  El modelo *Entry* tiene 4 subclases: *Observation*, *Evaluation*, *Instruction* y *Action*. A continuación, brevemente describimos cada una de ellas:

  - La subclase *Observation* es usada para modelar la observación de cualquier fenómeno o estado de interés relacionado con el paciente. Sólo registra la información relacionada con el estado o situación del paciente, y no lo que se hace realmente con el paciente. Esta clase se usa entre otras cosas para registrar los resultados patológicos o de cualquier tipo de prueba, la historia familiar y las circunstancias sociales del paciente.
  - La subclase *Evaluation* es usada para modelar evaluaciones de problemas y diagnósticos, establecimiento de objetivos, así como recomendaciones para afrontar el cuidado de un paciente. Esta clase sirve para registrar opiniones de un médico, y es, por tanto, subjetivo, en contraste con la clase *Observation* que contiene hechos objetivos.

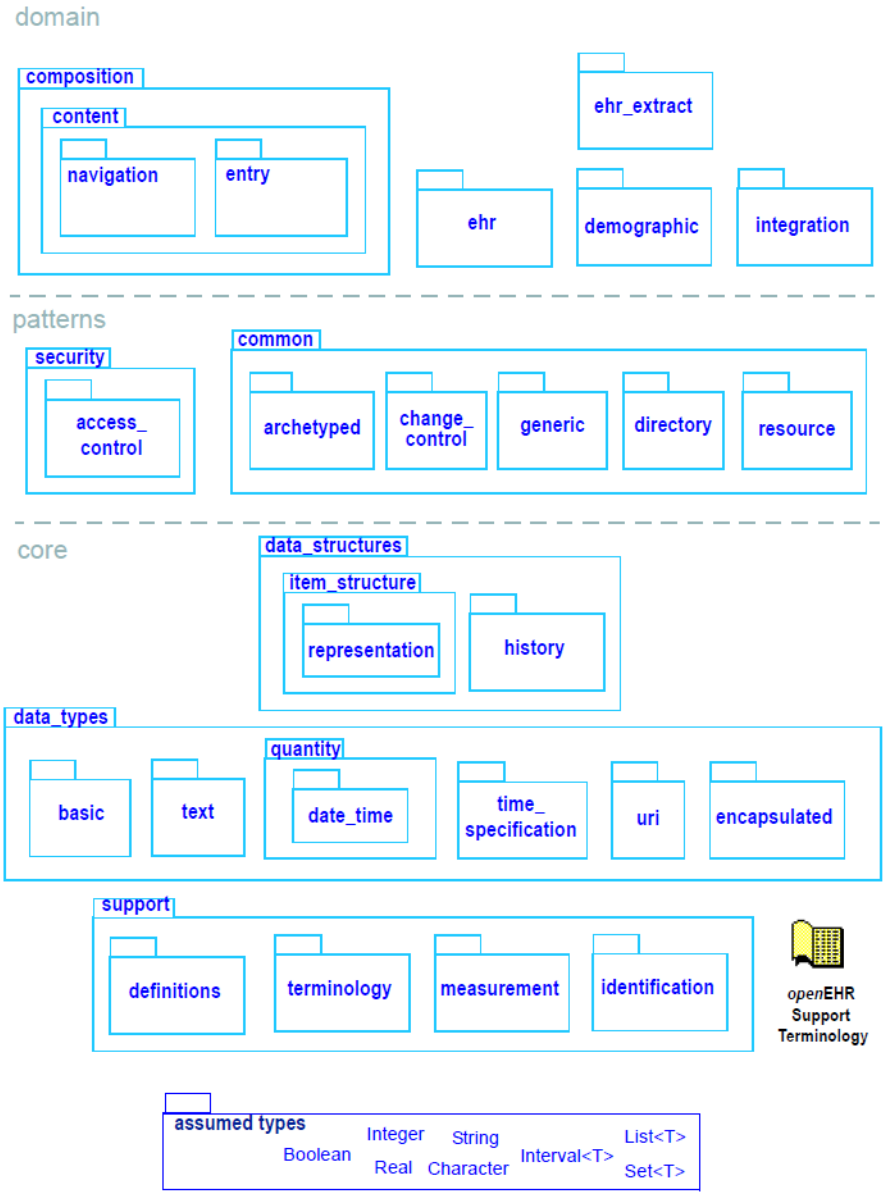

Figura 2.1: Estructura de paquetes del modelo de referencia de openEHR (figura tomada de [128])

- La subclase *Instruction* es usada para especificar tareas relacionadas con el cuidado médico para ser realizadas en el futuro. Estas tareas son especificadas con el suficiente detalle como para ser realizadas directamente por una enfermera o por el propio paciente.
- La subclase *Action* es usada para modelar la información registrada debido a la ejecución de alguna *Instruction* realizada por algún agente.

2.2.2. Arquetipos openEHR

La RAE define arquetipo como una 'representación que se considera modelo de cualquier manifestación de la realidad'. En openEHR, un arquetipo es un modelo de conocimiento que establece la estructura de la información utilizada en un escenario clínico determinado (p.e. el informe de alta o la medición de la presión sanguínea de un paciente) usando las entidades básicas, los atributos, y la estructura lógica del Modelo de Referencia.

Propósito de los arquetipos

Los Arquetipos son la piedra angular de la arquitectura openEHR ya que permiten que médicos y expertos del dominio participen en el diseño de especificaciones de contenido clínico normalizadas para la HCE sin necesidad de tener conocimientos técnicos. Los arquetipos pueden ser incorporados a la HCE para documentar observaciones clínicas, evaluaciones, instrucciones, o acciones.

Los arquetipos también ofrecen un mecanismo para capturar la información de una forma sistemática en escenarios clínicos determinados, facilitando la interoperabilidad.

Características de los arquetipos

Aunque los arquetipos han sido diseñados para ser ampliamente re-usables, estos pueden ser especializados para incluir singularidades locales [127]. Un arquetipo también puede reusar bloques de información de otros arquetipos o incluso arquetipos completos.

Los arquetipos pueden evolucionar y actualizarse de forma segura para tratar con el cambiante y dinámico entorno clínico gracias a la arquitectura dual definida en openEHR [127].

Los arquetipos incluyen micro-vocabularios de términos en lenguaje natural [45]. Estos términos son definidos por los expertos del dominio para especificar la información clínica requerida en el arquetipo. Con ello se pretende que los arquetipos no dependan directamen-

te de ninguna terminología clínica. Los términos en lenguaje natural también ofrecen a los expertos un gran nivel de flexibilidad y expresividad para especificar la información clínica, sobrepasando algunas limitaciones de los enfoques basados en terminologías [45]. A pesar de la apuesta de openEHR por la creación de términos en lenguaje natural para la definición de la información clínica, los arquetipos también incluyen mecanismos para enlazar dichos términos a conceptos de terminologías clínicas externas. Esta característica es conocida como *term binding* (enlazado/mapeo de términos) en la comunidad openEHR. En las siguientes secciones se exponen más detalles sobre esta característica.

Lenguaje de definición de arquetipos

El lenguaje de definición de arquetipos (ADL) [126] es un lenguaje formal para la definición de arquetipos desarrollado por la Fundación openEHR y recientemente incluido en la norma ISO EN 13606-2 [39].

Los arquetipos expresados en ADL se asemejan a archivos de lenguaje de programación y tienen una sintaxis definida. El lenguaje ADL aglutina dos sintaxis: cADL y dADL. La sintaxis cADL es utilizada en la definición del contenido clínico del arquetipo; mientras que la sintaxis dADL se emplea para representar las instancias del modelo de referencia incluidas en el arquetipo, la metainformación del arquetipo y la descripción de los términos usados en la definición.

La figura 2.2 muestra la estructura y las distintas secciones que componen un arquetipo. Las secciones *archetype*, *concept*, *language* y *description* componen la cabecera del arquetipo. Estas definen el identificador, el lenguaje y metainformación como el autor, el propósito, la fecha, etc. La sección *definition* o cuerpo del arquetipo incluye la estructura y las restricciones de la información clínica del arquetipo. La sección *ontology* contiene dos subsecciones: *term definitions* y *term bindings*. La primera incluye las descripciones textuales en uno o más idiomas del vocabulario empleado en la definición del arquetipo. La sección *term bindings* permite enlazar el vocabulario del arquetipo con conceptos de terminologías clínicas externas, como SNOMED-CT o LOINC.

Ejemplo de arquetipo

La figura 2.3 muestra parte del fichero ADL correspondiente al arquetipo ‘respiration’. La cabecera del fichero ADL muestra que el arquetipo es de tipo *Observation*, por tanto su definición se ajustará a las restricciones de la entidad *Observation* del Modelo de Referencia.

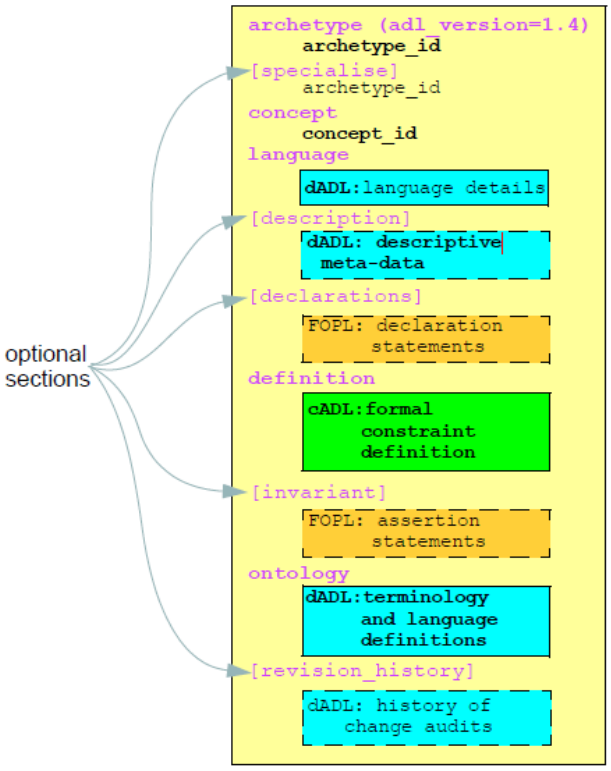

Figura 2.2: Estructura de un fichero ADL (figura tomada de [126])

La cabecera también muestra el propósito del arquetipo: registrar la frecuencia, el ritmo y las características de la respiración de un paciente.

La definición del arquetipo de la figura 2.3 incluye las restricciones y la estructura de los ítems de información clínica necesarios para modelar y registrar las características de la respiración de un paciente. El arquetipo incluye varias observables relacionadas con la respiración (p.e. ritmo y frecuencia respiratoria) que deben ser medidas o evaluadas por el médico. También incorpora un conjunto de hallazgos o valores predefinidos para estas observables (p.e. ritmo irregular, episodios de apnea).

En la definición del arquetipo los ítems de información sólo tienen identificadores locales (tales como at0000 o at0016). La subsección *term definitions* de la sección ontología incluye

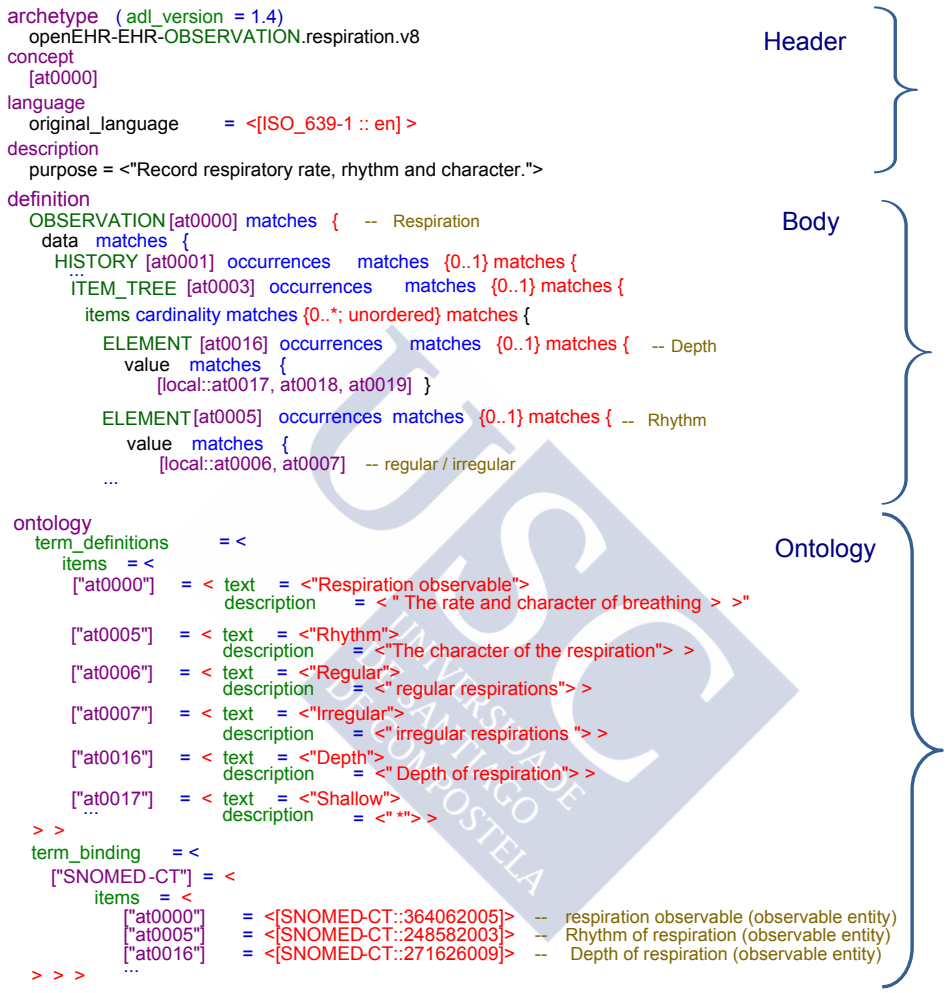

Figura 2.3: Fichero ADL del arquetipo 'respiration'

un micro-vocabulario del arquetipo con nombres/etiquetas locales y descripciones para cada uno de estos ítems. Así, por ejemplo, en la figura 2.3 podemos ver que los identificadores locales at0000 y at0016 tienen asociadas las etiquetas 'Respiration observable' y 'Depth of Respiration', respectivamente. Los ítems de información, incluidos en la definición del arquetipo, son instancias de alguna entidad del Modelo de Referencia. En la figura 2.3 se puede ver que el ítem at0000 es una instancia de la entidad *Observation*, mientras que el ítem at0016 es

una instancia de la entidad *Element*. Es importante destacar que no todos los ítems incluidos en la definición modelan contenido clínico, algunos tienen sólo una función estructural. Por ejemplo, el ítem at0003 es una instancia de la entidad *Item_Tree* y su misión es agrupar varios ítems de tipo *Element*.

La definición del arquetipo agrupa los ítems de forma jerárquica. La figura 5.1 muestra la jerarquía asociada al arquetipo 'respiration'. En la figura sólo se han incluido las etiquetas/términos de los ítems con significado clínico. El ítem raíz at0000 (con etiqueta 'Respiration observable') acoge, entre otros, los ítems at0016 y at0005 (con etiquetas 'Depth of Respiration' y 'Rhythm'). A su vez, el ítem at0016 incluye los ítems at0017, at0018 y at0019 (con etiquetas 'Shallow', 'Normal' y 'Deep', respectivamente).

Enlace de arquetipos a terminologías clínicas externas

La sección *term_binding* de los ficheros ADL permite enlazar los ítems de información del arquetipo con conceptos de terminologías clínicas externas, como SNOMED-CT o LOINC. En la parte inferior de la figura 2.3 se muestran 3 enlaces o mappings entre ítems del arquetipo y conceptos de la terminología SNOMED-CT. Cada mapping es definido con la siguiente información: el identificador local del ítem del arquetipo, el nombre de la terminología usada y el identificador del concepto SNOMED-CT seleccionado. Así, por ejemplo, el primer mapping del ejemplo enlaza el ítem at0000 ('Respiration observable') con el concepto SNOMED-CT 364062005 cuya descripción es 'Respiration observable (observable entity)'.

Repositorios de arquetipos

En la última década, gobiernos de diferentes países e importantes instituciones[6] (NEHTA, NHS, Centre for eHealth de Suecia y la fundación openEHR) han desarrollado arquetipos openEHR para modelar formalmente contenido clínico [98, 103, 104, 101]. Paralelamente, han surgido repositorios para dar soporte a la creación, revisión y gestión de los arquetipos [92, 90]. Un repositorio puede ser visto como una librería de arquetipos con funcionalidades avanzadas (búsqueda, autoría, etc.) que facilita la colaboración de expertos del dominio clínico en el desarrollo de arquetipos.

Actualmente, el repositorio más relevante es el openEHR Clinical Knowledge Manager (CKM) promovido por la fundación openEHR [92]. Cualquier usuario interesado puede vi-

[6] `http://www.openehr.org/who_is_using_openehr/governments`

sualizar o participar en la creación y/o revisión de un conjunto internacional de arquetipos openEHR. El número de arquetipos publicados en este repositorio empieza a ser considerable (en mayo de 2014 tiene 384 arquetipos).

Uso de arquetipos openEHR en los sistemas de información clínica

Las arquitecturas basadas en un modelo dual, como openEHR y la arquitectura definida en la norma ISO EN 13606, están llamados a ser elementos importantes en los sistemas de información clínica. La arquitectura dual promueve la separación entre el modelo de información (implementado en el software y en la base de datos), y los conceptos del dominio (creados por expertos del dominio al margen del software). Esta separación permite crear sistemas más fáciles de mantener y adaptar a cambios en las prácticas clínicas.

Tapuria et. al. [124] han recopilado las principales contribuciones de los arquetipos clínicos y también algunos de las cuestiones que todavía deben resolverse respecto a estos modelos clínicos.

En los últimos años se ha hecho un importante esfuerzo de modelado de arquetipos openEHR, sin embargo la experiencia de usar estos arquetipos en los sistemas de información clínica existentes (p.e. en sistemas desplegados en centros primarios) es todavía bastante limitada [20]. Una de las principales dificultades para desplegar de forma exitosa los arquetipos openEHR es que la mayoría de los sistemas actuales de información clínica cuentan con modelos de datos propietarios, creados a medida, para representar el contenido clínico. La migración de estos modelos hacia arquetipos openEHR no es en absoluto trivial. Varias investigaciones han propuesto soluciones para manejar este problema [20, 32, 15, 88, 119, 18]. El estudio de R. Chen et. al. realizó una conversión entre un modelo propietario utilizado en hospitales de Suecia y un modelo basado en arquetipos openEHR y viceversa [20]. El estudio desarrolló un método automático para realizar la conversión, concluyendo que los arquetipos ya son suficientemente expresivos para representar el modelo propietario. R. Chen et. al. defienden una migración lenta e incremental entre sistemas basados en modelos propietarios y sistemas basados en arquetipos y estándares. Sostienen que es importante que convivan los modelos propietarios (para conservar el significado de toda la información clínica existente creada y almacenada con estos modelos) con modelos basados en arquetipos y estándares para facilitar el intercambio de contenido clínico con sistemas de información externos.

En los últimos años se han desarrollado varias herramientas y prototipos (EHRGen [97], GastrOS [8]) que automatizan la generación de interfaces de usuario a partir de arquetipos

openEHR. Estas interfaces tratan de facilitar la captura estandarizada de información clínica a los médicos.

Calidad en arquetipos

Los arquetipos desarrollados hasta la fecha, principalmente por la Fundación openEHR y por NHS, no han seguido un proceso riguroso que asegure la calidad de los mismos [63]. Sin embargo, para que los arquetipos sean adoptados ampliamente en los sistemas de información sanitarios deben tener una calidad contrastada.

Recientemente, se ha empezado a estudiar criterios para evaluar la calidad de los arquetipos y se han definido guías y recomendaciones para el modelado de arquetipos. Un informe elaborado por el instituto EuroRec ha establecido un conjunto de requisitos administrativos, técnicos y clínicos relacionados con la calidad de los arquetipos [64]. El informe también recomienda algunos requisitos de funcionamiento de los repositorios, enlazar los términos locales de los arquetipos con conceptos de terminologías clínicas, y reducir los solapamientos existentes entre los arquetipos de los repositorios. El informe del proyecto SemanticHEALTH también ha incluido una serie de acciones necesarias para mejorar calidad de los arquetipos [121]. Las más importantes son la creación de una guía de buenas prácticas de modelado de arquetipos, metodologías para enlazar arquetipos con terminologías clínicas y la creación de herramientas de validación de arquetipos. La tesis de Qamar también ha recogido algunas deficiencias en el modelado de los arquetipos, así como algunas recomendaciones encaminadas a mejorar la calidad de los mismos [105].

2.2.3. Razones para seleccionar arquetipos openEHR

Varias razones nos han llevado a seleccionar arquetipos openEHR en esta investigación: las recomendaciones de importantes estudios a nivel Europeo, el fácil y libre acceso a arquetipos openEHR, y la apuesta de openEHR por el uso de terminologías clínicas.

En los últimos años varios estudios financiados por la Unión Europea han tratado el tema de la interoperabilidad de la HCE [30, 29, 121]. Los estudios han definido varias recomendaciones y medidas necesarias para lograr la interoperabilidad semántica en los próximos años, entre ellas destacan: el uso de una arquitectura basada en un modelo dual en los sistemas de información clínica, la creación de repositorios comunes de arquetipos clínicos y el uso consistente de terminologías clínicas (preferiblemente SNOMED-CT).

Los arquetipos openEHR son fácilmente accesibles desde repositorios públicos [92, 90]. En cambio, otros estándares muy difundidos, como puede ser HL7, prácticamente no cuentan con modelos de datos clínicos disponibles libremente.

La fundación openEHR está haciendo una clara apuesta para integrar las terminologías clínicas en los modelos de arquetipos, prueba de ello es el programa[7] de colaboración que ha acordado con IHTSDO para estudiar formas de usar conjuntamente arquetipos openEHR y SNOMED-CT. Actualmente, los arquetipos openEHR ya disponen de una característica que permite crear fácilmente los enlaces semánticos entre los términos locales del arquetipo y conceptos de terminologías clínicas (ver sección 2.2.2).

La investigación se ha centrado esencialmente en arquetipos openEHR de tipo 'Observation'. Estos arquetipos son usados para capturar hallazgos de exploraciones, resultados de pruebas y síntomas de un paciente, y en general cualquier condición o estado de un paciente. La investigación se ha centrado en estos arquetipos ya que son actualmente los más numerosos en los repositorios, y los más maduros y revisados por la comunidad de modeladores de openEHR. Además, se ha comprobado que son los arquetipos que tienen definido una mayor cantidad de contenido clínico, lo que incrementa la necesidad de enlazado con terminologías clínicas.

2.3. SNOMED-CT

SNOMED-CT (Systematized Nomenclature of Medicine – Clinical Terms) es la terminología clínica integral, multilingüe y codificada de mayor amplitud, precisión e importancia desarrollada en el mundo. SNOMED-CT es producto de la fusión entre SNOMED Reference Terminology (SNOMED RT), creada por el College of American Pathologists (CAP) y Clinical Terms Version 3 (CTV3), desarrollada por el National Health Service (NHS) del Reino Unido. Actualmente, esta terminología es mantenida y distribuida por la IHTSDO[8].

SNOMED CT proporciona la terminología general básica para la HCE incluyendo más de 300.000 conceptos activos con significado único y definiciones basadas en lógica formal. Cuando SNOMED-CT se integra en las aplicaciones que gestionan la HCE, permite a los profesionales de la salud representar la información clínica de forma adecuada, precisa e inequívoca, facilitando así la recuperación, intercambio y análisis de dicha información. La terminología se constituye principalmente por conceptos, descripciones y relaciones.

[7] http://www.openehr.org/news_events/industry_news.php?id=38

[8] International Health Terminology Standards Development Organisation. http://www.ihtsdo.org/

2.3.1. Componentes de SNOMED-CT

El contenido de SNOMED-CT es representado, principalmente, a través de 3 componentes:

- Conceptos: representan ideas o significados clínicos organizados en jerarquías.
- Descripciones: términos o nombres asignados a un concepto de SNOMED CT.
- Relaciones: conectan los conceptos a otros conceptos relacionados.

Estos componentes son suplementados por conjuntos de referencia que permiten configurar la terminología SNOMED-CT para manejar requerimientos específicos.

Conceptos

Cada concepto representa un significado o idea clínica y tiene asociado un identificador numérico (ConceptID). El identificador proporciona una referencia única inequívoca a cada concepto sin revelar ninguna información sobre la naturaleza del concepto.

Descripciones

Las descripciones de los conceptos son los términos o nombres asignados a un concepto de SNOMED-CT. En este contexto, término significa una frase utilizada para nombrar un concepto. Varias descripciones pueden asociarse con un concepto. Las descripciones proveen la representación legible por humanos de un concepto. Existen tres tipos de descripciones en SNOMED-CT: descripción completa (*en inglés Fully Specified Name, FSN*), término preferido (*preferred term*) y sinónimo (*synonym*).

La descripción completa constituye una forma no ambigua de nombrar a un concepto. No necesariamente representa la frase más utilizada para describir a ese concepto. Cada descripción completa termina con una etiqueta semántica entre paréntesis que expresa la categoría semántica a la que pertenece el concepto (por ejemplo: hallazgo clínico, organismo, parte del cuerpo, etc.).

El término preferido representa la palabra o frase más habitual utilizada para describir un concepto. A diferencia de la descripción completa, los términos preferidos no necesariamente son únicos. Por lo que, ocasionalmente, el término preferido para un concepto también puede ser el sinónimo o el término preferido de otro concepto.

Los sinónimos representan otros términos utilizados para describir un concepto.

A modo de ejemplo, a continuación se muestran algunas de las descripciones en español asociadas al concepto SNOMED-CT con identificador 22298006:

- Descripción completa: *Infarto de miocardio (trastorno)*
- Término preferido: *Infarto de miocardio*
- Sinónimos: *Infarto cardíaco* *Ataque al corazón* *Infarto de corazón*

Relaciones

Las relaciones representan asociaciones entre dos conceptos. Son usadas para definir lógicamente el significado de un concepto. Además, de los dos conceptos relacionados, un tercer concepto participa en las relaciones para representar el tipo de relación, esto es, el significado de la asociación. Esencialmente, existen dos tipos de relaciones importantes dentro de SNOMED-CT: jerárquicas (también llamadas de subtipo o *IS_A*) y lógicas (también llamadas de atributo).

- **Relaciones jerárquicas**

  Las relaciones jerárquicas son el tipo de relación más frecuente en SNOMED-CT. Son usadas para asociar un concepto con otro más general. En otras palabras, establecen que el concepto origen de la relación tiene un significado clínico más específico que el concepto destino. La figura 2.4 muestra un ejemplo de relación jerárquica entre el concepto *diabetes mellitus type 2* (origen de la relación) y *diabetes mellitus* (destino).

  Este tipo de relaciones componen las jerarquías de SNOMED-CT. El nivel de detalle clínico de los conceptos aumenta con la profundidad de las jerarquías. La figura 2.5 muestra la jerarquía de SNOMED-CT desde el concepto *diabetes mellitus type 2* hasta el concepto raíz de SNOMED-CT.

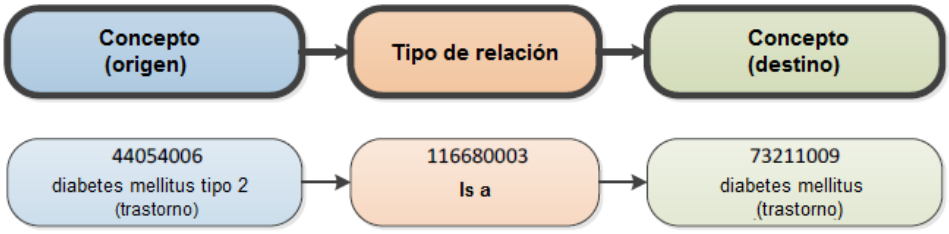

Figura 2.4: Ejemplo de relación jerárquica en SNOMED-CT (imagen tomada de [117]).

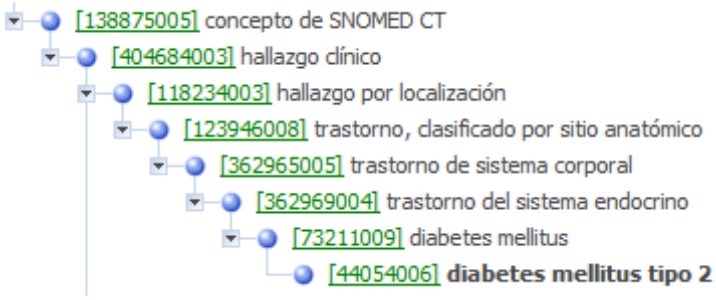

Figura 2.5: Jerarquía de SNOMED-CT desde el concepto *diabetes mellitus type 2* hasta el concepto raíz de SNOMED-CT

– **Relaciones de atributo**

Las relaciones de atributo contribuyen a la definición lógica de los conceptos de SNOMED-CT, especificando una característica de los conceptos origen de las relaciones. La característica es especificada por el tipo de relación y el valor es definido por el concepto destino de la relación.

SNOMED-CT contiene más de 50 tipos de relaciones de atributo. La aplicabilidad de cada tipo de relación está limitada a un determinado dominio y rango. El dominio se refiere a las categorías semánticas válidas para actuar como concepto origen en las relaciones de atributo. Mientras que el rango se refiere a las categorías permitidas en el destino de las relaciones. Por ejemplo, la relación de atributo *interpreta* (*interprets* en inglés) enlaza un hallazgo clínico con la entidad observable evaluada o interpretada por dicho hallazgo, o bien, con el procedimiento seguido para llegar al hallazgo. Por tanto, el dominio en este tipo de relación incluye conceptos de la jerarquía *hallazgo clínico*, mientras que el rango incluye conceptos de las jerarquías *entidad observable* o *procedimiento*. La figura 2.6 muestra dos ejemplos de la relación de atributo *interpreta*.

En esta investigación se han tenido muy en cuenta las relaciones de atributo para mejorar el proceso de mapping entre términos clínicos y SNOMED-CT. Los tipos de relaciones de atributo más usadas han sido: *interprets*, *procedure site - direct* y *method*. El tipo *procedure site* describe la parte de cuerpo en la que se realiza un procedimiento clínico; mientras que *method* representa la acción que se realiza durante un procedimiento clínico.

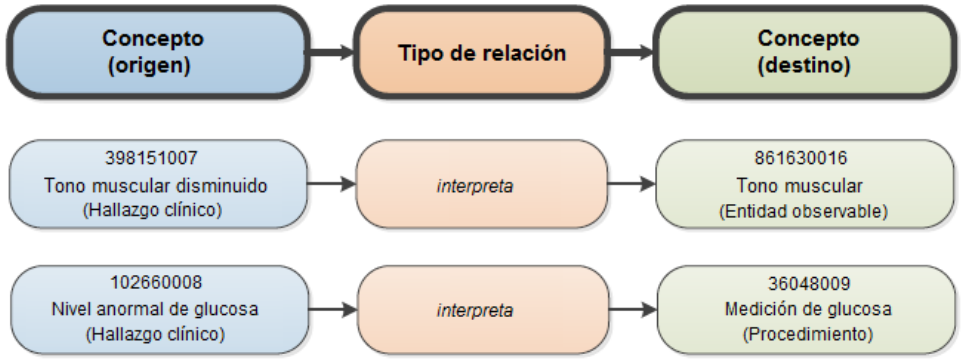

Figura 2.6: Ejemplos de relaciones de atributo en SNOMED-CT

2.3.2. Jerarquías de conceptos de SNOMED-CT

Todos los conceptos de la terminología SNOMED-CT pertenecen a alguna de las 19 jerarquías de alto nivel definidas en SNOMED-CT. El único concepto que no cuelga de estas jerarquías es el concepto raíz de SNOMED-CT. Las categorías de alto nivel son enumeradas en la tabla 2.1.

Las 19 jerarquías pueden ser consideradas categorías semánticas en las cuales los conceptos SNOMED-CT son divididos. Para esta investigación, las categorías semánticas más relevantes han sido: *Entidad Observable*, *Hallazgo Clínico*, *Procedimiento*, *Parte del cuerpo*, *Objetivo físico* y *Calificador*. La tabla 2.2 muestra una breve descripción y ejemplos de cada categoría semántica relevante. Las descripciones y ejemplos fueron tomados de la documentación oficial de SNOMED-CT[9].

2.3.3. Características de SNOMED-CT

SNOMED-CT tiene una amplia cobertura de temas relacionados con la salud. Se puede utilizar para describir la historia médica de un paciente, los detalles de un procedimiento, la propagación de epidemias, y mucho más. Además, la terminología tiene una profundidad sin precedentes, que permite a los médicos codificar la información clínica en el nivel adecuado de granularidad.

[9] http://ihtsdo.org/fileadmin/user_upload/doc/

Tabla 2.1: Listado de jerarquías de conceptos de SNOMED-CT

Hallazgo clínico	Fuerza física
Procedimiento	Evento
Entidad observable	Ambiente o localización geográfica
Estructura corporal	Contexto social
Organismo	Situación con contexto explícito
Sustancia	Estadificaciones y escalas
Producto farmacéutico / biológico	Objeto físico
Espécimen	Calificador
Concepto especial	Elemento de registro
Concepto de enlace	

Tabla 2.2: Descripción y ejemplos de las jerarquías/categorías semánticas de SNOMED-CT más relevantes en esta investigación

Jerarquía	Descripción
Hallazgo clínico	Los conceptos de esta jerarquía representan el resultado de una observación, evaluación o juicio clínico e incluyen estados clínicos normales y patológicos. Ejemplos: *Ruidos respiratorios normales (hallazgo)* *presión arterial diastólica anormal (hallazgo)*
Procedimiento	Los conceptos de esta jerarquía representan las actividades realizadas durante el proceso de la atención de la salud (incluyen, entre otros, procedimientos invasivos o administración de medicamentos) Ejemplos: *Ecografía de mama (procedimiento)* *Apendicectomía (procedimiento)*
Entidad observable	Los conceptos de esta jerarquía representan una pregunta que puede producir una respuesta o un resultado. Por ejemplo, el concepto *Presión diastólica final del ventrículo izquierdo (entidad observable)* podría interpretarse como la pregunta ¿Cuál es la presión diastólica del ventrículo izquierdo?
Estructura corporal	Los conceptos de esta jerarquía incluyen estructuras anatómicas normales y anormales, usados normalmente para especificar la región corporal afectada por una enfermedad o por un procedimiento. Ejemplos: *estructura aórtica (estructura corporal)* *estructura del riñón (estructura corporal)*
Objeto físico	Los conceptos de esta jerarquía incluyen objetos naturales y fabricados por el hombre Ejemplos: *Implante, dispositivo (objeto físico)* *respirador mecánico (objeto físico)*
Calificador	Esta jerarquía contiene algunos de los conceptos utilizados como valores para los atributos de SNOMED-CT que no se encuentran en otras partes de la terminología. Ejemplos: *Izquierdo (calificador)* *Punción - acción (calificador)*

Algunas aplicaciones que manejan información clínica tienden a centrarse en un conjunto restringido de términos o conceptos. SNOMED-CT ofrece mecanismos para crear subconjuntos o conjuntos de referencia incluyendo partes pertinentes de la terminología, dependiendo del contexto clínico y los requisitos locales. Por ejemplo, se podrían crear subconjuntos con los hallazgos más habituales en las distintas especialidades.

Cuando las organizaciones sanitarias tienen necesidades más allá de lo que está reflejado en una terminología clínica global, estas pueden crear extensiones locales o nacionales de SNOMED-CT con términos específicos para los requisitos propios de la organización o país.

SNOMED-CT es una terminología multinacional y multilingüe. Tiene una arquitectura capaz de gestionar diferentes idiomas y dialectos. Hoy en día, SNOMED-CT está disponible en inglés (británico y americano), español, danés y sueco y traducciones a otros idiomas están en proceso.

2.3.4. Razones para seleccionar SNOMED-CT

Varias razones nos han llevado a seleccionar SNOMED-CT frente a otras terminologías en esta investigación: su amplia cobertura, las recomendaciones de importantes estudios a nivel europeo y de organismos de estandarización, la colaboración estrecha entre IHTSDO y openEHR, y el fácil acceso a la terminología.

Actualmente, SNOMED-CT es la terminología más extensa (con más de 300.000 conceptos y 1.300.000 relaciones) y más completa (cubre todos los aspectos necesarios para codificar la información de la HCE). Por tanto, SNOMED-CT es claramente la terminología que puede dar una mayor cobertura para codificar la información clínica de pacientes.

En el dominio de la HCE, hay un importante consenso en que la terminología de referencia a nivel mundial es SNOMED-CT. Varios estudios financiados por la Unión Europea [30, 29, 121], tratando el tema de la interoperabilidad de la HCE, han recomendado el uso de terminologías clínicas y especialmente SNOMED-CT. Además, varios organismos de estandarización, como ISO o HL7, actualmente ya están usando SNOMED-CT.

IHTSDO, la institución que gobierna SNOMED-CT, tiene un creciente interés en promover el uso de SNOMED-CT en los modelos de datos clínicos. De hecho, ha firmado un acuerdo[10] de colaboración con la fundación openEHR para avanzar en la integración de arquetipos openEHR y SNOMED-CT.

Actualmente, más de 25 países han firmado acuerdos con la IHTSDO para acceder libremente a todo el contenido de la terminología. El formato en el que SNOMED-CT es publicado facilita considerablemente su uso. Básicamente, el contenido de SNOMED-CT es difundido a través de 3 ficheros de texto plano, fácilmente exportables a una base de datos. Una vez importado el contenido de SNOMED-CT en una base de datos, esta puede ser usada, por ejemplo,

[10] `http://www.openehr.org/news_events/industry_news.php?id=38`

para extraer rápidamente subjerarquías de SNOMED-CT o todas las descripciones textuales de los conceptos a través de consultas SQL[11].

2.3.5. Uso de SNOMED-CT

Una revisión bibliográfica sobre el uso de SNOMED-CT, incluyendo 488 artículos desde 2001 hasta 2012, revela que el número de publicaciones anuales sobre SNOMED-CT se incrementa año a año. Sin embargo, sólo hay unos pocos estudios (en torno al 10 % de los artículos revisados) que exponen el uso o la implementación de SNOMED-CT dentro de la práctica clínica [74]. La mayoría de las publicaciones se centra en escenarios no operativos de SNOMED-CT, esto es, en analizar aspectos técnicos de SNOMED-CT. Los temas más frecuentes en estas publicaciones son: la auditoría del contenido y estructura de SNOMED-CT [132, 62, 133], la comparación y mapping de SNOMED-CT con otras terminologías clínicas [16, 17, 46] y el análisis de la cobertura de SNOMED-CT para codificar términos clínicos en contextos específicos [25, 34, 61].

El alto número de publicaciones sobre comparación y mapping de SNOMED-CT con otras terminologías se debe a que actualmente existe un gran número de terminologías fragmentadas con dominios o ámbitos de aplicación solapados, lo que claramente representa una barrera a la codificación consistente y reutilización de la información clínica [74]. La gran cantidad de estudios sobre este tema pone de manifiesto la necesidad de armonización de las terminologías clínicas.

Las publicaciones analizando la cobertura de SNOMED-CT tratan de medir el grado o la proporción con la cual SNOMED-CT puede codificar correctamente términos locales en contextos específicos (p.e. para representar problemas en recién nacidos [61] o procedimientos de tomografía computarizada [25]). La gran cantidad de publicaciones sobre este tema se justifica en que la codificación de términos locales a SNOMED-CT y el análisis de cobertura es uno de los primeros requisitos o pasos necesarios para adoptar y utilizar SNOMED-CT en la práctica clínica.

Los estudios que han implementado SNOMED-CT en escenarios operativos de la práctica clínica, se han centrado mayoritariamente en desarrollar estrategias para capturar datos con SNOMED-CT, sin todavía alcanzar la madurez suficiente para usar los datos capturados [74]. La sofisticación de las implementaciones de SNOMED-CT para capturar los datos clínicos varía enormemente. Algunas implementaciones mapearon los términos clínicos de las plantillas

[11] Structured Query Language

e interfaces gráficas a conceptos SNOMED-CT en segundo plano, manteniendo los términos locales a la vista de los usuarios [114, 138]. Otras implementaciones incluyeron funcionalidades de búsqueda en SNOMED-CT en las interfaces o formularios de entrada de datos para sugerir un conjunto de conceptos relevantes [96, 137]. En estos casos, la responsabilidad final para seleccionar el concepto a indexar en el registro será del usuario.

Sólo 8 trabajos de los 488 evaluados en la revisión bibliográfica de SNOMED-CT realizada por D. Lee et. al., han usado o experimentado con los datos capturados en SNOMED-CT. Por ejemplo, tres de estos trabajos han desarrollado funcionalidades de apoyo a las decisiones clínicas para detectar efectos adversos de medicamentos [19] y gestionar casos de obesidad [76] y de úlceras por presión [67]. Claramente, hasta el momento ha habido pocos estudios que aprovechen los datos capturados en SNOMED-CT, y menos aún, que evalúen o cuantifiquen formalmente el valor de SNOMED-CT en escenarios operacionales.

La revisión literaria de D. Lee et. al. también advierte que la mayoría de las implementaciones de SNOMED-CT todavía no han sido publicadas en la literatura científica ya que son bastante recientes (algunas todavía están en fase de desarrollo o como parte de un proyecto piloto), por lo que aclara que es necesario indagar más allá de la literatura científica para tener una imagen real de la implementación de SNOMED-CT en escenarios operativos de la práctica clínica [74].

Precisamente, D. Lee et. al. han realizado un estudio sobre la implementación de SNOMED-CT en 13 centros de atención sanitaria [73]. En el momento del estudio, las implementaciones todavía no habían sido publicadas en la literatura científica. Para el estudio, los autores realizaron entrevistas directamente con los individuos encargados de implementar SNOMED-CT, centrándose en los siguientes aspectos: cómo extrajeron subconjuntos relevantes de SNOMED-CT, cómo la información clínica es capturada con SNOMED-CT, qué tipo de información clínica es capturada con SNOMED-CT y el uso de la post-coordinación.

2.4. Trabajo relacionado: mapping y búsqueda en SNOMED-CT

Un paso importante para promover la adopción de SNOMED-CT en los sistemas de información clínica es la existencia de herramientas avanzadas que automaticen, o al menos faciliten, la codificación de datos clínicos con SNOMED-CT.

En muchas ocasiones, la codificación con SNOMED-CT exige buscar correspondencias o mappings entre términos definidos en lenguaje natural (procedentes de diversas fuentes) y

conceptos SNOMED-CT. Esta búsqueda no es, en absoluto, un proceso trivial debido al gran tamaño, complejidad y granularidad de SNOMED-CT.

Esta sección incluye una revisión de las principales herramientas y trabajos que tratan la búsqueda y el mapping con SNOMED-CT. En la revisión distinguimos 3 categorías de trabajos según el tipo o la procedencia de los datos a mapear: información textual, modelos de datos clínicos (semi)estructurados y terminologías clínicas.

2.4.1. Mapping de información textual a SNOMED-CT

La conversión automática de texto libre en formas más estructuradas (como por ejemplo conceptos de terminologías estándar) puede permitir el acceso computacional a una gran cantidad de información actualmente "bloqueada" dentro de las notas clínicas e informes de pacientes.

En esta sección se exponen varios trabajos y herramientas que han abordado el mapping de información textual a conceptos de SNOMED-CT. Se distinguirán dos categorías o formas de información textual: textos clínicos y términos clínicos. Ambos contienen información clínica en lenguaje natural. La diferencia entre ambos es que los textos clínicos tienen una mayor extensión que los términos clínicos. Así, por ejemplo, un texto clínico podría ser un informe médico (no estructurado) de un paciente, el *abstract* de un artículo o una guía clínica. Mientras que un término clínico puede verse como una frase que hace referencia a un concepto médico, por ejemplo, 'neoplasia del lóbulo inferior derecho del pulmón'. Los diccionarios, glosarios y algunos registros médicos son fuentes habituales de términos clínicos.

Anotación de textos clínicos con conceptos SNOMED-CT

El trabajo de L. Ahmadian ha investigado si SNOMED-CT puede utilizarse para formalizar la información textual de guías clínicas [1]. La conversión de la información textual de las guías a reglas codificadas con conceptos SNOMED-CT fue realizada de forma manual en su totalidad. En primer lugar, los expertos seleccionaron y extrajeron de las guías los extractos referentes a las recomendaciones. Posteriormente, seleccionaron de los extractos los términos clínicos y construyeron reglas de tipo "SI *condición* ENTONCES *acción*". Finalmente, mapearon manualmente los términos a conceptos SNOMED-CT. L. A pesar de que Ahmadian et. al. encontraron algunas dificultades para representar algunos términos ambiguos de las guías con SNOMED-CT, concluyeron que la formalización de guías clínicas usando SNOMED-CT es factible.

Existen varios sistemas y herramientas que han demostrado tener éxito en la anotación automática de textos clínicos con SNOMED-CT [34, 108, 94, 120, 10, 53, 10, 9, 54]. P. Elkin et. al. han desarrollado un sistema de procesamiento de lenguaje natural para anotar textos con conceptos SNOMED-CT [34]. En un trabajo posterior, evaluaron la precisión del sistema para la identificación automática de pacientes con neumonía usando un corpus de informes textuales radiológicos [35]. El sistema, que también se apoyó en un conjunto de reglas basadas en conceptos SNOMED-CT, obtuvo una precisión muy elevada (superior al 95 %) detectando los pacientes con neumonía.

P. Ruch et. al. desarrollaron una herramienta para dar soporte a la anotación automática de contenidos textuales con conceptos de SNOMED-CT [108]. La herramienta combina dos módulos: el primero aplica varias estrategias de normalización léxica y expresiones regulares para buscar equiparaciones, mientras que el segundo usa un motor de recuperación de información genérico basado en espacios vectoriales.

El sistema propuesto por J. Patrick et. al. identifica automáticamente conceptos médicos de SNOMED-CT en textos clínicos [94]. Este ha sido desarrollado para capturar los datos de los pacientes con conceptos SNOMED-CT en tiempo real en un hospital. El sistema incluye varios módulos: extracción de las frases del texto, normalización léxica, un algoritmo de equiparación entre frases y conceptos SNOMED-CT basado en la coincidencia de tokens consecutivos (similar al algoritmo n-grams) y un detector de frases negadas.

Mapping de términos clínicos a conceptos SNOMED-CT

En la actualidad, hay una gran cantidad de trabajos publicados en la literatura [25, 44, 75, 79, 70, 68], que han descrito la codificación de términos clínicos (procedentes de diccionarios, glosarios y registros médicos) con conceptos SNOMED-CT como un proceso tedioso y manual, en el cual uno o varios expertos en terminologías clínicas son los responsables de seleccionar los conceptos relevantes, asistidos únicamente por navegadores de SNOMED-CT o simples algoritmos de equiparación léxica. A continuación, comentamos tres de estos trabajos de codificación con SNOMED-CT. Posteriormente, se expondrá brevemente qué son los navegadores y cuáles son sus funcionalidades u opciones de búsqueda.

El trabajo de T. De Silva et al. [25] recolectó una lista de términos sobre procedimientos en tomografía computarizada usados en los sistemas de información de varios hospitales de Canadá y evaluó la habilidad de SNOMED-CT para representar dichos términos. La búsqueda de conceptos SNOMED-CT equivalentes a los términos fue realizada por un médico

con formación de postgrado en informática médica, asistido únicamente por el navegador de SNOMED-CT CliniClue [23]. El proceso de mapping llevado a cabo en el trabajo fue en gran medida manual, por lo que los resultados del mapping dependen de la formación, capacidad y tiempo del experto responsable.

Recientemente, la Sociedad Española de Anatomía Patológica[12] (SEAP) ha publicado un listado[13] de términos frecuentes en español de procedimientos en patología, junto con los conceptos SNOMED-CT equivalentes, asignados por expertos [44]. Para esta tarea los expertos no dispusieron de herramientas de búsqueda más allá de los navegadores de SNOMED-CT tradicionales.

El trabajo de D. Lee et al. [75] recolectó un conjunto de términos usados en un sistema de información clínica de atención paliativa y evaluó si SNOMED-CT es útil para codificar dichos términos. La codificación tuvo lugar en dos etapas. La primera aplicó un algoritmo básico de equiparación léxica entre los términos y las descripciones de los conceptos SNOMED-CT, incluyendo además una etapa previa de normalización léxica. Los términos que no fueron mapeados algorítmicamente fueron buscados manualmente, en una segunda etapa, por expertos con el navegador CliniClue. A pesar de que este trabajo incluyó una etapa automática para agilizar la codificación, los resultados de la etapa automática requirió revisión manual, por lo que como indican los autores fue un proceso tedioso y costoso temporalmente.

Herramientas de búsqueda genérica: Navegadores y servidores terminológicos

Los navegadores de SNOMED-CT son herramientas o aplicaciones que permiten visualizar de forma organizada los componentes de la terminología. Frecuentemente, contienen vistas para navegar por las jerarquías, vistas de presentación de relaciones no jerárquicas y soporte para la búsqueda de conceptos relevantes dada una frase o un término clínico. Algunos navegadores también incluyen funcionalidades de edición de la terminología, permiten seleccionar elementos para construir subconjuntos de referencia para usos específicos e incluso combinar conceptos para expresar significados complejos que no se encuentran en términos precoordinados.

[12] https://www.seap.es/

[13] http://www.seap.es/enlaces-de-interes/-/asset_publisher/h3M6/content/id/114697

En la actualidad existe un número importante de navegadores para acceder al contenido de SNOMED-CT. La Biblioteca Nacional de Medicina de EEUU[14] en su web[15] mantiene un listado actualizado de los navegadores más importantes de SNOMED-CT. A continuación, vamos a describir las funcionalidades de búsqueda de conceptos incluidas en los navegadores de SNOMED-CT más populares [33, 60, 58, 23, 131, 116, 93]. La tabla 2.3 muestra resumidamente las funcionalidades implementadas en cada uno de los navegadores.

- Búsqueda exacta: es la funcionalidad más básica, recupera conceptos que tienen asociados descripciones exactamente iguales al término de búsqueda.
- Búsqueda semi-exacta basada en coincidencia de palabras: separa las palabras del término de búsqueda, y recupera los conceptos conteniendo todas las palabras en cualquier orden.
- Búsqueda aproximada basada en coincidencia de palabras: separa las palabras del término de búsqueda, y recupera los conceptos conteniendo al menos una de esas palabras. Normalmente, genera un número excesivo de candidatos.
- Búsqueda basada en distancia de edición: cuantifica como de similares son las cadenas de texto (asociadas al término de búsqueda y al concepto) contabilizando el número mínimo de operaciones de edición (p.e. inserción o sustitución) requeridas para transformar una cadena en la otra. Esta opción de búsqueda devuelve aquellos conceptos que requieren un menor número de operaciones de edición.
- Búsqueda basada en expresiones regulares: permiten introducir patrones para realizar búsquedas complejas en SNOMED-CT.
- Búsqueda para localizar conceptos en frases o textos: esta funcionalidad está pensada para localizar conceptos en términos de búsqueda largos, tales como frases o textos cortos.
- Autocompletado dinámico: esta funcionalidad sugiere dinámicamente conceptos a medida que el usuario va escribiendo el término de búsqueda, similar a los motores de búsqueda web actuales.
- Filtrado por categoría semántica: esta funcionalidad permite limitar los resultados de las búsquedas a los conceptos pertenecientes a determinadas jerarquías o categorías semánticas definidas en SNOMED-CT (p.e. entidad observable o hallazgo clínico).

[14] National Library of Medicine (NLM)

[15] `http://www.nlm.nih.gov/research/umls/Snomed/snomed_browsers.html`

Tabla 2.3: Comparativa de funcionalidades de búsqueda en los navegadores SNOMED-CT más populares

Funcionalidad / Navegador	Nav. NLM [33]	Nav. ITServer [60]	Nav. IHTSDO [58]	Nav. CliniClue [23]	Nav. VTSL [131]	Nav. SnoFlake [116]	Nav. Snon OWL [93]
Búsqueda exacta	X	X	X	X	X	X	X
Búsqueda semi-exacta	X		X	X	X	X	X
Búsqueda aproximada		X			X		X
Distancia de edición		X					X
Expresiones regulares			X				
Búsqueda en frases/textos			X			X	
Autocompletado	X	X	X	X		X	X
Filtrado semántico	X	X	X	X	X	X	X

Limitaciones de las búsquedas en los navegadores de SNOMED-CT

En la actualidad existen bastantes navegadores para acceder al contenido de SNOMED-CT. Sin embargo, todavía no incluyen opciones de búsqueda avanzadas capaces de automatizar el proceso de codificación con SNOMED-CT. A continuación se enumeran problemas y limitaciones de las funcionalidades de búsqueda de los navegadores actuales:

– La mayoría de los navegadores SNOMED-CT aplican sólo técnicas de equiparación léxica para buscar conceptos relevantes dado un término de consulta. Es decir, los navegadores sólo consideran la similitud léxica para buscar conceptos relevantes, sin contemplar prácticamente la similitud semántica. La única función semántica que suelen incluir los navegadores es el filtrado de los resultados de las búsqueda por categoría semántica. Además, los navegadores suelen aplicar técnicas básicas de equiparación léxica, tales como la equiparación exacta o aproximada. Por una parte, la aplicación de la equiparación léxica exacta obtiene generalmente una precisión alta en los resultados de las búsqueda, pero un recall bajo, ya que busca descripciones de conceptos exactamente iguales al término buscado. Por otra parte, la aplicación de la equiparación léxica aproximada obtiene un recall alto y una precisión baja ya que devuelve conceptos con palabras o subcadenas comunes con el término buscado, por lo que podría sugerir cientos o miles de conceptos candidatos para un término de búsqueda.

– Pocos navegadores incluyen técnicas de normalización léxica (p.e. eliminación de stopwords y de caracteres o detección de faltas ortográficas) para combatir las pequeñas

diferencias léxicas (frecuentes en el lenguaje natural) que evitan, en no pocas ocasiones, la detección del concepto correcto. Tampoco, suelen incorporar técnicas de expansión de consultas, para enriquecer el término introducido por el usuario con términos alternativos formados con sinónimos del término original. Hay que considerar que normalmente múltiples términos pueden hacer referencia a un mismo concepto médico y que la sinonimia definida en SNOMED-CT todavía no es, en absoluto, completa. Sin duda la ausencia de técnicas que traten este problema está limitando los resultados en las búsquedas de los navegadores.

– Los navegadores durante las búsquedas de conceptos relevantes no usan en ningún momento las relaciones semánticas de SNOMED-CT. Estos consideran la terminología SNOMED-CT como una enorme lista de conceptos no conectados. En nuestra opinión los navegadores deberían usar las relaciones semánticas de los conceptos, ya que estas definen características de los conceptos y pueden aportar información sobre el significado de los conceptos, no presente en sus descripciones textuales.

– En nuestra opinión, los sistemas de búsqueda implementados en la mayoría de los navegadores de SNOMED-CT están pensados, para a partir de un término de búsqueda, hacer una primera criba o selección de conceptos relevantes, facilitando a los expertos la tarea de encontrar el concepto equivalente al término buscado. Claramente estos sistemas no fueron diseñados para ser herramientas automáticas de codificación, sino como una simple ayuda para los expertos. Por tanto, el proceso de mapping con las herramientas actuales sigue demandando una alta implicación de los expertos.

En sintonía con nuestra opinión, el trabajo de J. Rogers et al. [107] evaluó las características de 17 navegadores de SNOMED-CT y concluyó que las funcionalidades de búsqueda varían mucho entre los distintos navegadores, siendo en general insuficientes y demasiado genéricas. Además, muchos trabajos publicados en la literatura sobre codificación con SNOMED-CT han expuesto que el proceso de codificación con las herramientas actuales es altamente costoso e ineficiente y que es necesario el desarrollo de nuevas herramientas de búsqueda avanzada para agilizar y mejorar la calidad de la codificación con SNOMED-CT [68, 21, 70, 108].

2.4.2. Mapping de modelos de datos clínicos a SNOMED-CT

El enlace entre modelos de datos clínicos y terminologías es reconocido como un requisito necesario en el camino hacia la interoperabilidad semántica de la HCE [121, 47, 106].

Algunos retos dificultan el enlazado de los modelos de datos con terminologías (particularmente con SNOMED-CT). En primer lugar, los modelos de datos y SNOMED-CT presentan estructuras diferentes, por lo que no es trivial la aplicación de técnicas estructurales y de contexto, frecuentemente usadas en otro tipo de problemas, tales como el alineamiento de ontologías. En segundo lugar, las guías de modelado y los criterios de calidad para la especificación de los modelos de datos son todavía infrecuentes, por lo que los modelos de una misma especificación (p.e. arquetipos openEHR) presentan todavía estructuras diversas. Esto dificulta la aplicación de técnicas estructurales y de contexto, sobretodo de forma automática. En tercer lugar, SNOMED-CT es en la actualidad la terminología clínica más completa y con mayor granularidad. La navegación a través de SNOMED-CT y la recuperación eficiente de conceptos relevantes resulta extremadamente tediosa y difícil.

En los últimos años, se han realizado algunos intentos para hacer el mapping entre modelos clínicos y terminologías de forma manual y se comprobó que este proceso exige mucho tiempo a personal altamente cualificado, lo que originó mappings imperfectos e incompletos [105]. Además, se han desarrollado algunas herramientas que facilitan el enlazado [135, 105]. Sin embargo, estas herramientas todavía están lejos de conseguir realizar el mapping de forma automática y precisa. A continuación, se exponen algunas propuestas que han surgido para enlazar modelos openEHR y HL7 con SNOMED-CT.

Mapping de arquetipos openEHR a SNOMED-CT

Actualmente la mayoría de los arquetipos openEHR, disponibles públicamente, apenas están enlazados con terminologías. Esta situación ha motivado la investigación en técnicas semi-automáticas que faciliten esta tarea.

La herramienta de mapping desarrollada por Yu et al. [135] ha aplicado técnicas de recuperación de información para codificar los términos locales de los arquetipos con conceptos SNOMED-CT. La herramienta usa Lucene[16], una popular librería de código abierto para la recuperación de información, para obtener un ranking ordenado de 10 conceptos para cada uno de los términos. Esta herramienta tiene algunas limitaciones: no incluye ninguna técnica de

[16] `http://lucene.apache.org/core/`

normalización léxica, por lo que, pequeñas diferencias léxicas entre los términos del arquetipo y conceptos SNOMED-CT impiden que la herramienta obtenga enlaces/mappings correctos. Tampoco usa ninguna información de contexto del arquetipo para mejorar el mapping. La elección final del concepto correcto se deja en manos del usuario o experto. La evaluación de la herramienta ha demostrado que aproximadamente sólo la mitad de las veces el código correcto está entre los 10 conceptos sugeridos por la herramienta, por lo que la otra mitad de las veces el experto tiene que usar navegadores de SNOMED-CT u otras herramientas externas para hacer nuevas búsquedas para encontrar el concepto correcto.

Berges et al. [13] elaboraron un estudio para evaluar el rendimiento de 9 técnicas léxicas (basadas en la similitud de cadenas de texto) orientadas al enlazado de términos de arquetipos con conceptos de SNOMED-CT. Obtuvieron un recall del 25 % si sólo un concepto SNOMED-CT es recuperado para un término del arquetipo. El recall aumentó hasta un 50 % y 75 % cuando recuperaron 10 y 100 conceptos respectivamente.

El sistema MoST [105] es una metodología semi-automática para enlazar los términos clínicos de los arquetipos openEHR a conceptos de SNOMED-CT. El sistema incorpora dos etapas: un proceso automático de búsqueda de conceptos candidatos y un proceso manual de selección de los mappings finales realizado por expertos clínicos.

La búsqueda de conceptos candidatos incluye un amplio abanico de métodos: un procedimiento para normalizar los términos, técnicas léxicas y lingüísticas. Además, usa los servicios terminológicos de UMLS [113], cierta información contextual del arquetipo y algunas reglas de post-filtrado semántico para mejorar la calidad del mapping.

La metodología de MoST obtiene una media de 5.5 conceptos candidatos por término. La elección final del concepto correcto se realiza por consenso entre un grupo de expertos. La evaluación de MoST concluyó que aproximadamente el 70 % de las veces el código correcto está entre los conceptos sugeridos por la herramienta. En nuestra opinión la limitación principal de la metodología MoST es la ausencia de técnicas semánticas que aprovechen en mayor medida el contexto del arquetipo y las relaciones semánticas de SNOMED-CT para reducir el número de conceptos candidatos.

Al inicio de esta tesis, posiblemente MoST fuese la herramienta de mapping más avanzada para arquetipos. Sin embargo, MoST no ha conseguido automatizar demasiado el proceso de mapping, este todavía depende en gran medida de la participación de expertos.

Mapping de modelos HL7 a SNOMED-CT

El grupo de trabajo Terminfo, perteneciente a la organización de estandarización para el ámbito de la salud HL7, ha estado trabajando para posibilitar que la información clínica contenida en varios estándar HL7 v3 pudiese ser representada mediante conceptos SNOMED-CT [50]. De forma similar a los arquetipos openEHR los modelos de datos HL7 usan un vocabulario interno para definir los atributos de información del modelo, pero también soporta el uso de terminologías externas para codificar dicha información. El grupo de trabajo ha publicado un informe detallado incluyendo guías y reglas para usar SNOMED-CT en los modelos HL7, así como, algunas correspondencias semánticas entre los componentes del modelo HL7 y las jerarquías de SNOMED-CT [50]. Desde la publicación de este informe ha habido algunas contribuciones en la integración de HL7 y SNOMED-CT. A. Ryan et. al. [109] han mapeado de forma manual modelos HL7 de observaciones clínicas con conceptos de SNOMED-CT siguiendo las recomendaciones publicadas en [50]. S. Heymans et. al. [52] presentaron una estrategia para validar automáticamente las guías y reglas recogidas en [50] sobre el uso de SNOMED-CT en modelos HL7.

Limitaciones de los sistemas para mapear modelos clínicos a SNOMED-CT

El desarrollo de herramientas específicas de mapping entre modelos clínicos y terminologías es todavía un campo poco explorado. Las herramientas que existen en la actualidad todavía están lejos de conseguir un mapping automático y preciso. A continuación se enumeran varios problemas y limitaciones de las herramientas actuales:

– Las herramientas usan esencialmente técnicas de equiparación léxica junto con alguna técnica lingüística, por lo que diferencias léxicas (p.e. sinónimos, hipónimos, abreviaciones) entre los términos de los modelos y los conceptos SNOMED-CT pueden dificultar que la herramienta obtenga mappings correctos. Mejores técnicas de normalización y de expansión de términos (con sinonimia) podrían contribuir a un mapping de mayor calidad.

– Las herramientas actuales no son capaces de realizar un mapping automático. Esto es, no devuelven un único concepto para cada término, más bien sugieren un conjunto de candidatos. Por tanto, el proceso de mapping todavía demanda una alta participación de los expertos.

– No explotan lo suficiente la estructura de los modelos para extraer contexto de los términos; prácticamente consideran cada término del modelo de forma aislada. La información de contexto podría ser útil para obtener mappings más precisos y para desambiguar cuando haya varios conceptos candidatos.

– No aplican técnicas de similitud estructural entre los modelos y las terminologías. Si bien es cierto que los modelos y terminologías tienen estructuras diferentes, la aplicación de este tipo de técnicas podría ser útil para mejorar la calidad del mapping.

– Las herramientas no explotan las relaciones semánticas de SNOMED-CT, "ven" la terminología SNOMED-CT como un conjunto de conceptos no conectados. Las relaciones semánticas definen las características de los conceptos. Por tanto, es interesante que las herramientas las consideraran, junto con las descripciones textuales, para tener más información del significado de los conceptos.

Por tanto, hoy en día hay una importante necesidad de herramientas que logren un mayor nivel de automatización en el mapping entre modelos de datos y terminologías clínicas. Las limitaciones expuestas anteriormente pueden guiar el desarrollo de futuras herramientas.

2.4.3. Alineamiento de SNOMED-CT con otras terminologías clínicas

El alineamiento o mapping de SNOMED-CT con otra terminología clínica implica determinar las correspondencias entre los conceptos de ambas terminologías. A cada una de estas correspondencias se le suele llamar mapping.

En los últimos años IHTSDO ha estado colaborando activamente con otras organizaciones para armonizar SNOMED-CT con otras terminologías, principalmente mediante la creación de alineamientos entre ellas. La armonización entre las terminologías clínicas favorece la interoperabilidad entre sistemas y permite reutilizar los datos clínicos para distintos propósitos. A continuación, se comentan los principales proyectos de armonización de SNOMED-CT.

Una colaboración[17] entre IHTSDO y WHO[18] ha resultado en más de 20.000 mappings entre conceptos SNOMED-CT y términos de la terminología ICD-10. Ambas terminologías han sido diseñadas con propósitos distintos, SNOMED-CT cubre todos los aspectos necesarios para codificar la información de la HCE con mucho grado de detalle, mientras que ICD-10

[17] http://www.ihtsdo.org/about-ihtsdo/governance-and-advisory/harmonization/who/

[18] World Health Organization

es una terminología que organiza el contenido en categorías significativas para dar soporte a estudios estadísticos y de epidemiología. La creación de un mapping entre ambas permite reutilizar los datos clínicos, así por ejemplo, la información clínica de un paciente puede ser recogida con gran nivel de granularidad con SNOMED-CT y gracias al mapping con ICD-10 esta información es clasificada en una categoría de alto nivel pudiendo ser utilizada en estudios estadísticos.

Recientemente, Regenstrief (organización sin ánimo de lucro que desarrolla y mantiene LOINC) e IHTSDO han firmado un acuerdo de colaboración[19] para construir un alineamiento entre LOINC y SNOMED-CT. El objetivo del acuerdo es reducir la duplicación de esfuerzos y hacer los registros electrónicos de salud más efectivos en la mejora de la atención médica.

A principios del 2014, se ha publicado una tabla de equivalencias[20] entre conceptos de la clasificación *International Classification for Nursing Practice* (ICNP) y SNOMED-CT.

El proceso de alineamiento entre SNOMED-CT y otras terminologías clínicas ha sido realizado en gran parte de forma manual. En el proceso se usaron algunas herramientas de apoyo, principalmente para la edición de los mappings. Sin embargo, las correspondencias entre las terminologías fueron creadas por un un grupo de expertos en mapping y terminologías. El artículo de K. Giannangelo y J. Millar comenta en detalle el proceso seguido por un grupo de expertos para alinear SNOMED-CT y la clasificación ICD-10 [46]. Hay que tener en cuenta que se requiere un alto nivel de conocimiento del dominio para crear las correspondencias entre terminologías clínicas, por lo que no es un proceso fácilmente automatizable. Aun así, en los últimos años han surgido algunos trabajos proponiendo enfoques semi-automáticos para alinear y validar terminologías a gran escala en el ámbito médico [72, 69].

2.5. Técnicas de mapping

Algunas disciplinas (p.e. recuperación de información, alineamiento ontológico o procesamiento de lenguaje natural) han estado trabajando activamente en el desarrollo de técnicas para encontrar mappings entre conceptos de diversas terminologías y ontologías y entre términos en lenguaje natural y conceptos [77, 123, 115, 65, 22].

J. Euzenat y P. Shvaiko han recopilado y clasificado las principales técnicas orientadas a encontrar automáticamente correspondencias entre conceptos [41]. A continuación, exponemos las técnicas de mapping más usadas de acuerdo a la clasificación de J. Euzenat y P.

[19] `https://loinc.org/collaboration/ihtsdo`
[20] `http://goo.gl/OVb7uX`

Shvaiko: técnicas léxicas, lingüísticas, estructurales, y en menor medida de aprendizaje automático.

2.5.1. Técnicas léxicas

Las técnicas léxicas son las más básicas. Usan los nombres o descripciones de los conceptos o entidades para buscar correspondencias. Estas técnicas se basan típicamente en el siguiente principio: cuanto más similares son los nombres o las etiquetas de dos entidades, más posibilidades de que estos denoten el mismo concepto. Existen varios problemas que restan precisión a estas técnicas:

- **La existencia de sinónimos en el lenguaje**. Un sinónimo es una palabra (o expresión) que tiene el mismo o muy parecido significado que otra. En ocasiones dos términos o etiquetas pueden ser muy poco similares y sin embargo denotar el mismo concepto. Las técnicas léxicas por sí solas no son suficientes para detectar correspondencias en estos casos. Cabe destacar que los conceptos de la terminología SNOMED-CT suelen tener asociados varias descripciones sinónimas, sin embargo la cobertura de sinónimos de SNOMED-CT está todavía lejos de estar completa.

- **La existencia de homónimos en el lenguaje**. Los homónimos son palabras usadas para nombrar diferentes conceptos. Por tanto, puede suceder que las etiquetas de dos entidades sean muy similares y sin embargo denoten distintos conceptos. En estos casos, las técnicas léxicas crearán correspondencias incorrectas.

- **La existencia de variaciones léxicas**. Las variaciones léxicas de una palabra a menudo ocurren debido a diferentes ortografías aceptadas, al uso de abreviaciones y al uso de prefijos o sufijos opcionales. Por ejemplo, CD, C.D. , CD-ROM y Disco Compacto son formas consideradas equivalentes para referenciar al mismo concepto.

Hay distintas técnicas de comparar los nombres de las entidades dependiendo en si estos son consideradas como una secuencia de letras o como un conjunto de palabras. A continuación se expone con más detalle las distintas técnicas léxicas:

- **Técnicas léxicas basadas en comparación de cadenas:**

  Estas técnicas consideran las etiquetas o cadenas de las entidades como una secuencia de letras. Existen muchos criterios para comparar dos cadenas como secuencias:

- Equiparación exacta: se produce cuando las cadenas comparadas son idénticas. Previamente a la comparación se podría haber realizado algún tipo de comparación.
- Distancia Hamming: mide la similitud entre dos cadenas contabilizando el número de posiciones en las que las dos cadenas difieren.
- Similaridad N-gram: estima la similitud contando el número de secuencias de N caracteres comunes entre dos cadenas.
- Distancia de edición: cuantifica como de similares son dos cadenas contabilizando el número mínimo de operaciones (p.e. inserción o sustitución) requeridas para transformar una cadena en la otra. Esta métrica suele obtener un valor entre 0 y 1 para estimar el nivel de similitud entre dos cadenas.

La normalización léxica puede ayudar a mejorar los resultados de las comparaciones de las cadenas. La normalización incluye, entre otras acciones: la eliminación de dígitos y signos de puntuación, la normalización de los espacios en blanco y la conversión de todos los caracteres a minúsculas.

– **Técnicas léxicas basadas en el lenguaje:**

En las técnicas basadas en el lenguaje, las cadenas se consideran textos que pueden ser segmentados en palabras. Incluso, las técnicas más avanzadas pueden tener en cuenta la secuencia en la que aparecen las palabras, dicho de otra forma, pueden considerar la estructura gramatical de las cadenas. Estas técnicas se acercarían a las técnicas usadas en el procesado de lenguaje natural.

El primer paso que suelen realizar este tipo de técnicas es aplicar una normalización lingüística a las cadenas. La normalización lingüística suele incluir varias de las siguientes tareas:

- Tokenización: es el proceso de separar una cadena de texto en sus partes constituyentes (tokens), incluyendo palabras, números y caracteres.
- Eliminación de palabras vacías (stopwords): Las palabras vacías son palabras que aparecen con gran frecuencia en los textos, tales como artículos, pronombres, preposiciones, y conjunciones. Este conjunto de palabras no resulta útil para buscar equiparaciones léxicas, ya que apenas aportan información.

- Lematización: es un proceso que consiste en, dada una forma flexionada (es decir, en plural, en femenino, conjugada, etc), hallar el lema correspondiente. El lema es la forma que por convenio se acepta como representante de todas las formas de una misma palabra. Es decir, el lema es la palabra que nos encontraríamos como entrada en un diccionario tradicional: singular para sustantivos, masculino singular para adjetivos, infinitivo para verbos. Por ejemplo, decir es el lema de dije, diré o dijéramos.

Una vez que se aplica el proceso de normalización lingüística a las cadenas, estas pueden ser comparadas con las técnicas basadas en comparación de cadenas presentadas previamente, o con alguna técnica basada en comparación de tokens. Estas últimas miden la similitud entre dos cadenas cuantificando el número de palabras (tokens) comunes entre dichas cadenas.

Existen varios paquetes de software con implementaciones de las métricas léxicas citadas previamente: SimMetrics[21], SecondString[22], Alignment API[23] y SimPack[24].

2.5.2. Técnicas basadas en recursos lingüísticos

Las técnicas basadas en recursos lingüísticos usan recursos, frecuentemente externos, tales como: bases de datos léxicas, diccionarios o tesauros. Estos recursos proveen relaciones lingüísticas (p.e. sinónimos, hipónimos) que pueden ser explotadas para detectar similitudes o correspondencias entre palabras o términos.

Uno de los recursos más ampliamente usados en la bibliografía es Wordnet. Es una enorme base de datos léxica que agrupa palabras en inglés en conjuntos de sinónimos llamados synsets. Proporciona definiciones cortas y generales, junto con las relaciones semánticas entre los conjuntos de sinónimos. WordNet ha sido ampliamente usado en recuperación de información e integración ontológica para expandir los términos de búsqueda con sinónimos, hipónimos e hiperónimos [130, 57]. Otro recurso que ha sido muy utilizado para la expansión de términos en el ámbito clínico es el metatesauro de UMLS [51, 7, 81].

[21] http://sourceforge.net/projects/simmetrics/
[22] http://secondstring.sourceforge.net/
[23] http://alignapi.gforge.inria.fr/
[24] https://files.ifi.uzh.ch/ddis/oldweb/ddis/research/simpack/index.html

Algunos trabajos han desarrollado métodos, que en vez de usar relaciones lingüísticas de recursos externos, infieren nuevos sinónimos mediante el análisis automático de largas colecciones de conceptos. Por ejemplo, [56] y [55] han propuesto métodos para analizar el metatesauro de UMLS con el objetivo de detectar nuevos sinónimos no explícitos en dicho recurso. La principal ventaja de estos trabajos es que son capaces de aumentar la cobertura de sinónimos del recurso que analizan, favoreciendo así los procesos de búsqueda de correspondencias o recuperación de información.

2.5.3. Técnicas estructurales

Las técnicas estructurales han sido usadas tradicionalmente en procesos de alineamiento entre ontologías o terminologías. Estos procesos implican buscar correspondencias semánticas entre las entidades de dos ontologías o terminologías. Las técnicas estructurales han sido usadas en combinación con técnicas léxicas para aumentar el rendimiento global de los sistemas de alineamiento [43]. Las técnicas estructurales explotan las propiedades estructurales de las ontologías, tales como las relaciones semánticas, para tener más información del significado de los conceptos. Así, por ejemplo, las técnicas estructurales basadas en grafos miden la similitud entre dos entidades (procedentes de dos ontologías distintas) en base a la similitud de sus relaciones y vecinos. Esto implica extraer los conceptos vecinos de cada entidad (recorriendo sus relaciones semánticas) y aplicar métricas, generalmente léxicas, entre los vecinos de ambas entidades. Destacar que las relaciones jerárquicas han sido las relaciones más utilizadas en este tipo de técnicas ya que su significado es siempre el mismo en diferentes ontologías y terminologías. El principio en el que se basan las técnicas basadas en grafos es que si dos entidades son similares, sus vecinos deben de alguna forma también ser similares.

En la actualidad, existen modelos de datos que están bastante estructurados (p.e. los arquetipos openEHR). Los términos de estos modelos están internamente organizados en estructuras de datos (p.e. en forma de árbol), por lo que los términos están relacionados informalmente a otros términos de los modelos. Sin embargo, hasta nuestro conocimiento todavía hay muy pocos trabajos de mapping que aprovechen esto para extraer contexto de los términos y mejorar así la búsqueda de correspondencias terminológicas de dichos términos. La investigación de Khare et. al. [66] es uno de los pocos trabajos que ha explotado la estructura semántica de formularios de entrada de datos clínicos para mejorar el mapping a SNOMED-CT. Khare et. al. proponen un enfoque híbrido basado en información lingüística y estructural. Así, dado un término de búsqueda del formulario, el enfoque explota la estructura semántica del formulario

para derivar contexto del término. Posteriormente, aplica métricas léxicas y lingüísticas para mapear el término a un concepto compatible con el contexto derivado. El enfoque de Khare et. al. se ha centrado especialmente en analizar los formularios en su conjunto para derivar las categorías semánticas esperadas de cada uno de sus términos clínicos. Desgraciadamente, todavía hay pocos enfoques que exploten la información estructural y contextual durante los procesos de mapping. Las siguientes herramientas de mapping orientadas a modelos clínicos claramente deberían explotar en mayor medida dicha información. Incluso, aunque existen diferencias entre las estructuras de las terminologías (basadas en relaciones semánticas formales) y las estructuras de los modelos de datos clínicos (más informales), creemos que merece la pena explorar la aplicación de técnicas de similitud estructural en el mapping entre modelos de datos clínicos y terminologías.

2.5.4. Técnicas de aprendizaje automático

Además de las técnicas presentadas previamente, algunos enfoques también han aplicado técnicas de aprendizaje automático en los procesos de mapping o alineamiento [84, 28]. Estas técnicas aplican una etapa de entrenamiento en la que usan las características de una muestra de alineamientos positivos y negativos para aprender de forma automática cómo clasificar nuevos alineamientos. Las características usadas en el entrenamiento suelen ser las puntuaciones resultantes de la aplicación de las técnicas léxicas, estructurales o lingüísticas.

Las técnicas de aprendizaje automático resultan ser especialmente útiles en procesos de mapping donde se aplican múltiples métricas de similitud entre los conceptos. Dichas técnicas, durante el entrenamiento, analizan las características de los alineamientos positivos y negativos, aprendiendo cuales son las más relevantes para determinar si hay o no hay mapping entre dos conceptos. Esto les permite llevar a cabo con más garantías la selección del mapping más adecuado entre un conjunto de candidatos con múltiples puntuaciones de similitud.

Las redes neuronales, los árboles de decisión y el aprendizaje Bayesiano son tipos de técnicas de aprendizaje usados en procesos de equiparación de conceptos [41].

2.5.5. Resumen de las técnicas de mapping

En esta sección se han discutido las técnicas básicas que pueden ser usadas para buscar correspondencias entre conceptos. Nuestro objetivo no fue presentar todas las técnicas existentes, sino más bien exponer de forma resumida las más usadas.

Consideramos que las técnicas de mapping no deberían ser usadas de forma aislada, sino que deberían tomar ventaja de los resultados proporcionados por las otras técnicas. De hecho, gran parte del éxito de un sistema de búsqueda de mappings depende precisamente de la selección y combinación adecuada de las diferentes técnicas [41].

Cabe destacar que muchas de las técnicas presentadas previamente todavía no han sido explotadas en las herramientas actuales de mapping y codificación orientadas a la terminología SNOMED-CT. De hecho, hemos detectado que la mayoría de las herramientas orientadas a SNOMED-CT usan solamente técnicas léxicas. Por tanto, consideramos que hay un importante campo de mejora en estas herramientas mediante la inclusión y combinación de diferentes técnicas de mapping. Concretamente, sugerimos varias mejoras a las herramientas actuales:

- Las herramientas deberían incluir técnicas lingüísticas orientadas a expandir los términos de búsqueda con términos alternativos o sinónimos. Esto incrementaría las posibilidades de éxito en las búsquedas.

- Las herramientas orientadas a mapear modelos de datos clínicos estructurados y SNOMED-CT deberían explorar las técnicas estructurales, usándolas en combinación con técnicas léxicas. Las técnicas estructurales han sido utilizadas con éxito en procesos de alineamiento ontológico.

- Las herramientas de búsqueda de conceptos en SNOMED-CT (p.e. los navegadores), deberían incluir mejores mecanismos para seleccionar o desambiguar. Actualmente, algunas opciones de búsquedas de los navegadores de SNOMED-CT devuelven cientos o miles de conceptos candidatos a las búsquedas de los usuarios. Reglas heurísticas o técnicas de aprendizaje automático podrían ser usadas para eliminar muchos conceptos candidatos, de forma que el usuario sólo reciba los conceptos candidatos con mayores probabilidades de ser correctos.

2.6. Trabajo relacionado: segmentación en SNOMED-CT

El creciente tamaño y complejidad de las terminologías y ontologías ha provocado el surgimiento de una nueva área de investigación conocida como segmentación ontológica. Esta tiene el objetivo de extraer porciones autocontenidas de ontologías adecuadas a necesidades particulares, normalmente denominadas: módulos, segmentos ontológicos o subontologías. Las subontologías han sido usadas para favorecer el procesamiento de las ontologías por las

aplicaciones (p.e. para aplicaciones de razonamiento o inferencia), para la anotación de datos y para generar vistas de ontologías manejables por humanos [12, 48, 122, 112, 134, 91]. La representación y manejo de los módulos ha sido extensamente detallada en la documentación técnica de SNOMED-CT [118].

En la bibliografía hemos encontrado pocos trabajos sobre segmentación (semi)automática en SNOMED-CT. Sari et al. desarrollaron una metodología para extraer segmentos de SNOMED-CT asociados a arquetipos openEHR [110]. Estos segmentos tienen el objetivo de representar el contenido semántico de los arquetipos. En este trabajo, usuarios expertos seleccionaron los conceptos semilla de partida. Es decir, el método dispone de un conjunto dado de conceptos semilla, como suele ser habitual en la segmentación ontológica [112, 80]. A partir de estos conceptos semilla, técnicas de recorrido de grafos y un conjunto de reglas son aplicadas para obtener un segmento válido de SNOMED-CT.

López-García et al. han propuesto extraer segmentos de SNOMED-CT para anotar informes de alta de cardiología [80]. Manualmente, varios expertos anotaron los informes de alta con más de 400 conceptos SNOMED-CT, los cuales fueron usados como semillas para el proceso de segmentación. La segmentación fue realizada con 5 técnicas diferentes: 4 técnicas heurísticas de recorrido de grafo y una técnica basada en lógica. Posteriormente, los segmentos fueron filtrados con información de frecuencia de MEDLINE. Aunque López-García et al. afirmaron que los segmentos extraídos tienen una cobertura adecuada, reconocieron que estos tenían un tamaño excesivo. Por ejemplo, sus técnicas de recorrido de grafo obtuvieron segmentos de un tamaño medio del 17 % al 51 % respecto al tamaño total de SNOMED-CT.

El trabajo de J. Patrick et al. describe la extracción de un segmento de SNOMED-CT derivado de un enorme corpus de historias clínicas textuales procedentes de un sistema de información clínico de cuidados intensivos [95]. La estrategia seguida en este trabajo implicó los siguientes pasos. En primer lugar, se aplicaron varias estrategias de procesamiento de lenguaje natural para identificar conceptos SNOMED-CT en las historias clínicas. En segundo lugar, se contabilizaron la frecuencia con la que aparecen los distintos conceptos en las historias clínicas y se seleccionaron aquellos con al menos 100 apariciones. Posteriormente, un software computó el mínimo segmento formado por los conceptos seleccionados en la etapa previa. Finalmente, revisaron manualmente el segmento, eliminando los conceptos inadecuados. El segmento resultante estuvo formado por 2700 conceptos, lo que proporcionó una cobertura del 96 % en el corpus y un tamaño inferior al 1 % respecto al tamaño total de SNOMED-CT. Los autores destacaron que el segmento, antes de la revisión manual, contenía

un importante número de conceptos inadecuados, debido principalmente a la debilidad de los métodos de procesamiento automático.

Gran parte del éxito de un proceso de segmentación depende de la selección adecuada de los conceptos semilla. Muchos de los enfoques actuales requieren que expertos seleccionen manualmente dichos conceptos. Este proceso está demostrado que es lento y costoso. En nuestra opinión, el proceso de selección de conceptos semilla debería ser realizado, o al menos apoyado, por técnicas de mapping sofisticadas. Esto permitiría automatizar en mayor medida el proceso de segmentación en SNOMED-CT.

Capítulo 3

Métodos de búsqueda en SNOMED-CT

En la actualidad, existe un importante consenso en que la integración de las terminologías clínicas en la HCE, y más concretamente en los arquetipos clínicos, es un paso clave en el camino hacia la interoperabilidad semántica [29, 121]. De la misma forma, hay un consenso amplio en cuanto a que SNOMED-CT es la terminología más adecuada para representar de forma estandarizada gran parte de la información clínica de la HCE [100].

Sin embargo, hasta ahora la inmensa mayoría de la información clínica se ha registrado y almacenado en la HCE en forma de lenguaje natural y no se ha enlazado con conceptos de SNOMED-CT ni con otras terminologías clínicas estándar. La creación de los enlaces terminológicos implica localizar en SNOMED-CT (o en otra terminología clínica) los conceptos equivalentes a la información clínica que los profesionales de la salud desean registrar en la HCE. La búsqueda de conceptos relevantes en SNOMED-CT es compleja y extremadamente laboriosa debido al gran tamaño de esta terminología (más de 300.000 conceptos) y a la falta de herramientas de búsqueda avanzadas para SNOMED-CT [107, 70, 108].

Este capítulo presenta el conjunto de métodos propuestos en esta tesis doctoral para la búsqueda en SNOMED-CT. Los métodos de este capítulo han sido diseñados para localizar conceptos relevantes de SNOMED-CT dado una frase o un término clínico en lenguaje natural (p.e. 'la presión diastólica es alta'). Un término clínico se suele corresponder con un único concepto de SNOMED-CT, por tanto, idealmente los métodos deberían obtener un único concepto relevante que represente completamente el término buscado. Aclarar que los métodos del capítulo no han sido optimizados para detectar conceptos en textos de entrada largos.

Destacar también que este capítulo trata de describir de forma general los métodos. Los siguientes capítulos, en cambio, detallarán cómo se han especializado y cómo se han combinado estos métodos para llevar a cabo tareas más complejas y específicas de mapping en SNOMED-CT. En cierta forma el presente capítulo puede verse como un catálogo de técnicas de búsqueda orientadas a SNOMED-CT, que puede ser de utilidad, como fuente de consulta, para el desarrollo de futuras herramientas de mapping centradas en SNOMED-CT e incluso en otras terminologías clínicas con similares características.

En el comienzo del presente capítulo describiremos las tareas realizadas para almacenar SNOMED-CT en un formato adecuado para la búsqueda. Seguidamente, expondremos una clasificación de las técnicas de búsqueda implementadas. Posteriormente, explicaremos con más detalle cada una de las técnicas. Dado que la aplicación de las técnicas de búsqueda puede generar más de un concepto candidato para un mismo término, al final del capítulo, presentaremos varias estrategias de desambiguación para seleccionar automáticamente el mejor concepto entre los candidatos.

3.1. Preprocesado de SNOMED-CT y consideraciones iniciales

Las versiones de SNOMED-CT son distribuidas por IHTSDO como un conjunto de ficheros de texto. Cada versión incluye ficheros individuales para cada uno de los tres componentes centrales de SNOMED-CT: conceptos, descripciones y relaciones.

Se han procesado estos ficheros de texto y se ha almacenado la información relevante para las búsquedas en tablas de un base de datos postgreSQL. Dos tablas son especialmente importantes para las búsquedas: una almacena todas las descripciones textuales de los conceptos de SNOMED-CT con los correspondientes identificadores de los conceptos, mientras que la otra contiene las relaciones semánticas de SNOMED-CT.

Destacar que todas las técnicas de búsqueda que se describen a continuación han sido implementadas en JAVA. Cuando se inicia un proceso de búsqueda, dichas técnicas usan el Driver PostgreSQL JDBC para comunicarse con la base de datos que almacena SNOMED-CT y cargar rápidamente en memoria las descripciones textuales de los conceptos.

3.2. Clasificación de las técnicas de búsqueda

En la figura 3.1 puede verse una clasificación de las técnicas desarrolladas en esta tesis doctoral para la búsqueda en SNOMED-CT. Estas han sido diseñadas especialmente para localizar conceptos relevantes de SNOMED-CT dado una frase o un término clínico en lenguaje natural. Distinguimos entre técnicas a nivel elemento y a nivel estructura. Las técnicas a nivel elemento ven la terminología SNOMED-CT como un simple listado de conceptos, por lo que recorren el listado tratando de localizar equiparaciones con el término buscado aplicando algún tipo de métrica de similitud. Estas técnicas no explotan las relaciones semánticas entre los conceptos de SNOMED-CT.

En contraste, las técnicas a nivel de estructura ven SNOMED-CT como una extensa red de conceptos interconectados y tratan de explotar las relaciones semánticas entre los conceptos de SNOMED-CT para mejorar el proceso de búsqueda. Además, estas técnicas también se preocupan por analizar el contexto, si lo hubiese, en el que está ubicado el término clínico buscado. Por ejemplo, si el término clínico está ubicado en un registro o en un modelo de datos trata de analizarlo (p.e. ¿el modelo fue creado para registrar observaciones o para indicar procedimientos o acciones futuras?), la ubicación del término en el modelo y los términos vecinos dentro del mismo. Esta información de contexto favorece el entendimiento del término buscado y permite afinar en bastantes casos los resultados de las búsquedas.

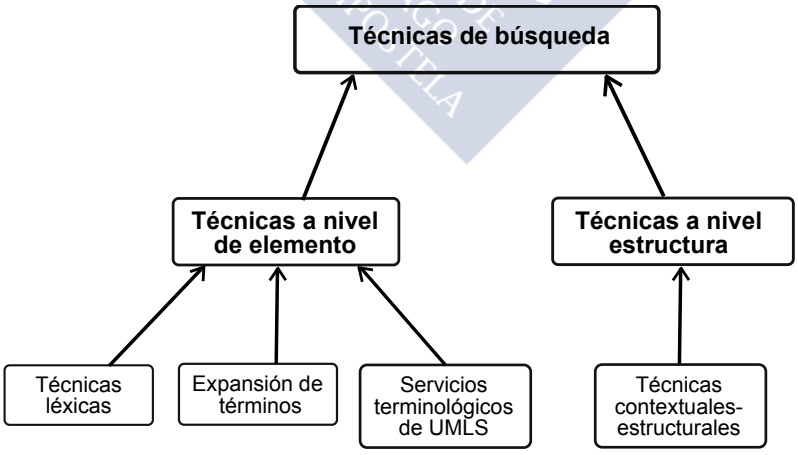

Figura 3.1: Clasificación de las técnicas de búsqueda en SNOMED-CT.

A continuación, se describen las distintas técnicas a nivel de elemento (ver figura 3.1). Posteriormente, se detallarán las técnicas implementadas a nivel de estructura.

3.3. Técnicas léxicas

3.3.1. Normalización léxica

Para aumentar las posibilidades de equiparación léxica se ha aplicado un proceso de normalización que afecta tanto a los términos clínicos como a las descripciones de SNOMED-CT. Esta normalización incluye transformaciones sencillas, tales como convertir las cadenas de texto a minúsculas o eliminación de signos de puntuación; y otras transformaciones más complejas:

– Tokenización

  La tokenización es el proceso de separar una cadena de texto en sus partes constituyentes (tokens), incluyendo palabras, números y caracteres. La tokenización fue realizada usando la siguiente heurística:

  - Los tokens se separan considerando los espacios en blanco de la cadena.
  - Los caracteres y los espacios en blanco no fueron incluidos en la lista de tokens.

– Eliminación de palabras vacías (stopwords)

  Las palabras vacías son palabras que aparecen con gran frecuencia en los textos, tales como artículos, pronombres, preposiciones, y conjunciones. Este conjunto de palabras no resulta útil para buscar equiparaciones léxicas, ya que apenas aportan información a un concepto médico. Por ello, los términos clínicos y las descripciones de SNOMED-CT son procesados para eliminar sus palabras vacías. Así, por ejemplo, el término ‘skin of hand’ es convertido a ‘skin hand’.

3.3.2. Equiparación léxica exacta

La equiparación léxica exacta evalúa si dos cadenas de texto, tras la normalización léxica, son exactamente iguales. Por ejemplo, las dos cadenas: ‘Apgar score at 1 minute’ y ‘APGAR SCORE - 1 MINUTE’ son normalizadas a ‘apgar score 1 minute’ y por tanto equiparan completamente.

3.3.3. Equiparación léxica parcial

La equiparación léxica parcial evalúa si una cadena de texto (tras la normalización léxica) está contenida dentro de otra cadena. Por ejemplo, la cadena 'Body mass index 25-29' equipara parcialmente con la cadena 'Body mass index 25-29 - overweight'.

3.3.4. Equiparación léxica aproximada

Se han usado técnicas para encontrar equiparaciones aproximadas entre cadenas. Estas técnicas son útiles para detectar equiparaciones cuando hay pequeñas diferencias entre las cadenas de texto, tales como problemas ortográficos, por ejemplo 'skin colour' y 'skin color' o que cadenas largas se diferencien por sólo una palabra, por ejemplo 'total resection of large intestine' y 'resection of large intestine'. Las técnicas aproximadas frecuentemente devuelven una puntuación de similitud léxica de 0 a 1 entre dos cadenas. A continuación se detallan las técnicas léxicas aproximadas empleadas:

- Distancia Levenshtein

  La distancia Levenshtein[1] es un tipo de distancia de edición para medir la similitud entre dos cadenas de texto. Esta distancia calcula el mínimo número de operaciones de edición requeridas para convertir una cadena de texto en otra. Las operaciones consideradas incluyen: inserción y supresión de un carácter y reemplazo de un carácter por otro; teniendo todas la misma penalización.

- Similitud de coseno

  La similitud de Coseno es un tipo de distancia basada en tokens, por lo que considera las cadenas de texto como una bolsa o conjunto de palabras. Hemos usado la implementación de similitud de coseno de la librería SimMetrics[2] (una librería con implementaciones de métricas orientadas a detectar equiparaciones entre cadenas de texto).

  SimMetrics incluye una implementación simplificada de lo que frecuentemente se conoce como similitud de coseno. La similitud entre dos cadenas es obtenida con la siguiente fórmula:

[1] http://en.wikipedia.org/wiki/Levenshtein_distance
[2] http://sourceforge.net/projects/simmetrics/

$$SimilitudCoseno = \frac{numPalabrasComunes}{numPalabrasCadena1^{0.5} + numPalabrasCadena2^{0.5}}$$

Siendo:

numPalabrasComunes: el número de palabras que coinciden en ambas cadenas

numPalabrasCadena1: el número de palabras de la primera cadena

numPalabrasCadena2: el número de palabras de la segunda cadena

3.4. Expansión de términos con sinónimos

Se ha desarrollado un método automático para generar términos alternativos mediante el uso de sinónimos procedentes de SNOMED-CT. Dada una cadena o término de búsqueda, el método sigue las siguientes etapas (ver figura 3.2). Primero, se tokeniza la cadena en las palabras constituyentes. Seguidamente, se buscan sinónimos de cada una de esas palabras en SNOMED-CT. Finalmente, el término de búsqueda se expande mediante el reemplazo de sus palabras por los sinónimos generados.

A continuación se exponen con más detalle las etapas clave del método:

– Descubrimiento automático de sinónimos en SNOMED-CT

Este proceso trata de encontrar sinónimos de palabras empleando un corpus formado por todas las descripciones de SNOMED-CT. El proceso se apoya en la siguiente hipótesis: si dos descripciones sinónimas de SNOMED-CT tienen subcadenas comunes, entonces las subcadenas no comunes pueden ser sinónimas. Por ejemplo, considerando que la descripción 'renal biopsy' es sinónima de 'kidney biopsy' en SNOMED-CT, es bastante probable que las subcadenas no comunes, esto es 'renal' y 'kidney', sean sinónimas o estén estrechamente relacionadas. El método de búsqueda de sinónimos para una palabra 'P' en SNOMED-CT sigue las siguientes etapas (ver figura 3.3):

- Extracción de pares de descripciones sinónimas de SNOMED-CT (esto es, asociadas al mismo concepto) cumpliendo que una de ellas contenga la palabra 'P' mientras que la otra no la contenga. Por ejemplo, para la búsqueda de sinónimos de la palabra 'excision' se obtendría, entre otras, el par de descripciones sinónimas: 'excision of muscle' y 'resection of muscle' asociadas al concepto con id 36143002.
- Se procesa cada par de descripciones extraídas en la etapa anterior. El procesamiento implica analizar la secuencia de las dos descripciones, con el objetivo de

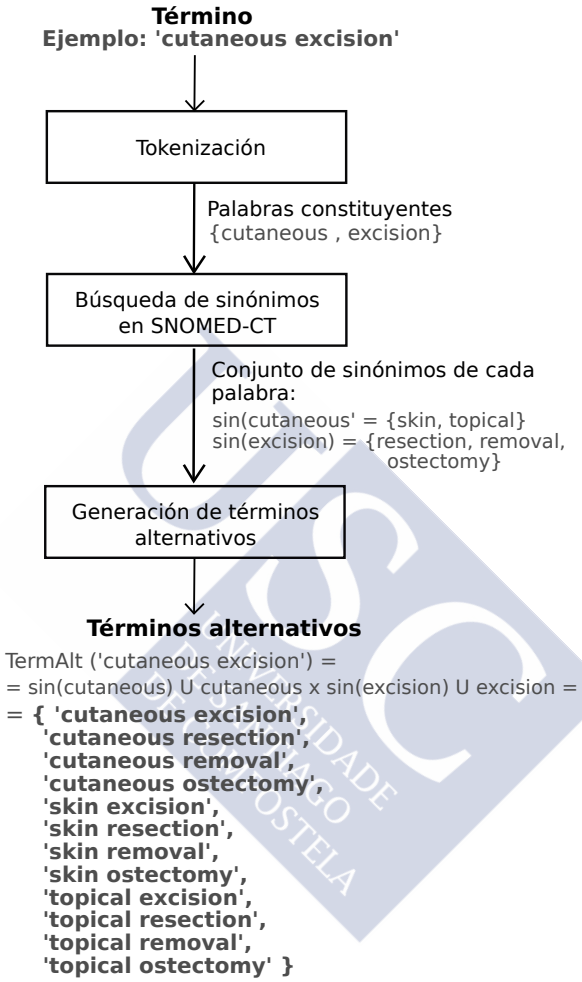

Figura 3.2: Aplicación de la expansión basada en sinónimos. Ejemplo con el término 'cutaneos excision'.

identificar y descartar las subcadenas comunes en ambas descripciones y de extraer las subcadenas no comunes de las descripciones. Si las subcadenas no comunes ocupan la misma posición dentro las descripciones y una de ellas es igual a 'P' entonces la otra (a la que llamaremos 'Psin') es considerada un sinónimo candidato y recibe 1 punto. La asignación de 1 punto registra que 'Psin' fue usado como sinónimo de la palabra 'P' en un par de descripciones sinónimas de SNOMED-

CT. Siguiendo el ejemplo con 'excision' y dado el par de descripciones 'excision of muscle' y 'resection of muscle', el método identifica 'of muscle' como la subcadena común, y 'excision' y 'resection' como las subcadenas no comunes. Ya que 'excision' y 'resection' ocupan la misma posición dentro de las descripciones (son la primera palabra), 'resection' sería considerada un sinónimo candidato de 'excision' y recibiría 1 punto (ver figura 3.3).

- Después de procesar cada par de descripciones, se obtiene un ranking de sinónimos de 'P' ordenado por el número de puntos (esto es, el número veces que fueron identificados como candidatos). Dos condiciones son aplicadas para seleccionar la lista definitiva de sinónimos de 'P' en el ranking: (1) sólo se seleccionan los 3 sinónimos del ranking con un mayor número de puntos y (2) estos deben contener al menos 5 puntos.

– Generación de términos alternativos Los términos alternativos se generan reemplazando y combinando los sinónimos encontrados en la anterior etapa. Teniendo en cuenta que el término tokenizado está formado por P1, P2..Pn donde Pi representa cada una de las palabras constituyentes del término; los términos alternativos se generan a partir de la siguiente expresión:

$$TermAlt(T) = TermAlt(P1, P2..Pn) = Sin(P1) \cup P1 \times Sin(P2) \cup P2... \times Sin(Pn) \cup Pn$$

Siendo:

Sin(Pi) el conjunto de sinónimos seleccionados para la palabra Pi.

En la figura 3.2 pueden verse los términos alternativos generados para el término inicial 'cutaneous excision'. Destacar que no se ha fijado ninguna limitación en cuanto al número de reemplazos por término. Por lo que podrían generarse términos alternativos sin ninguna palabra en común con el término original.

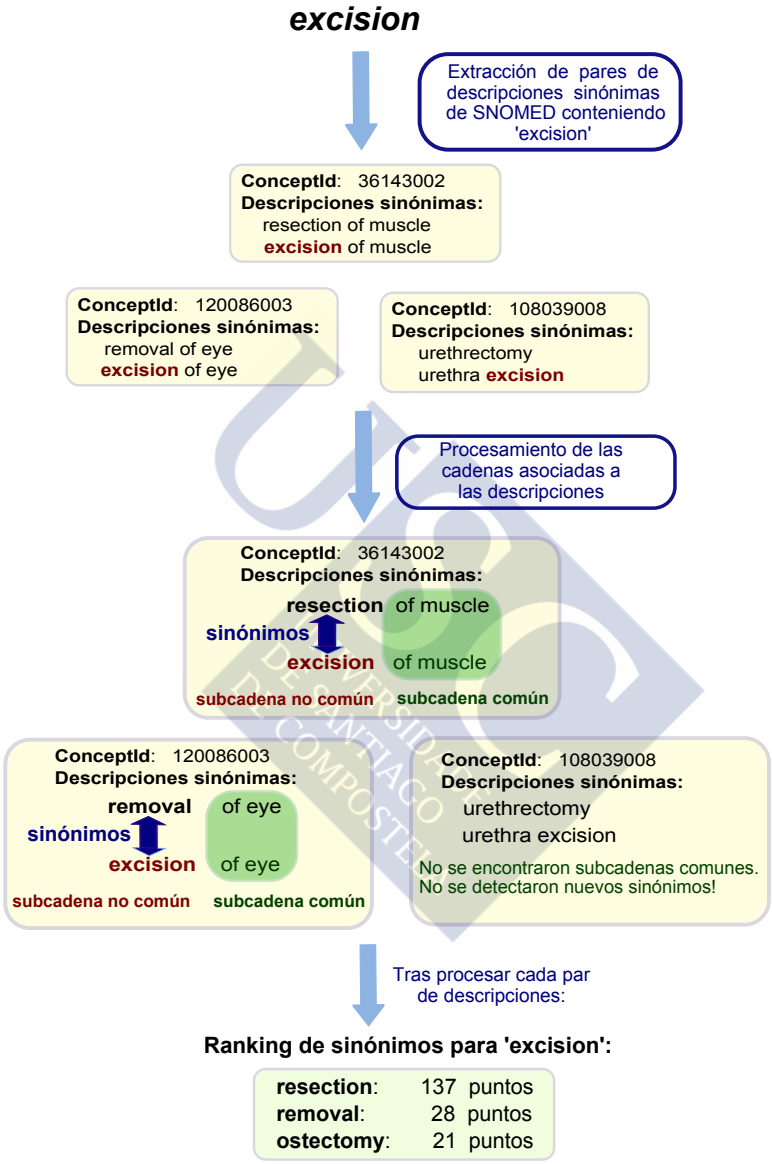

Figura 3.3: Ejemplo de descubrimiento automático de sinónimos para la palabra ‘excision’ en SNOMED-CT

3.5. Servicios terminológicos de UMLS

Se ha usado la API proporcionada por la National Library of Medicine (NLM), llamada UMLS Terminology Services API (UTS API), para llevar a cabo búsquedas de conceptos en el Metatesauro de UMLS [113]. El objetivo principal de este Metatesauro es agrupar conceptos de distintas terminologías con el mismo significado, asociando diferentes nombres o sinónimos para el mismo concepto. La API usada proporciona varios métodos para buscar conceptos UMLS en el Metatesauro dado un término o frase. Nosotros hemos seleccionado los dos métodos que proporcionan una mayor precisión, a cambio de perder algo de cobertura. Estos son:

- Búsqueda exacta (*Exact Match*): obtiene los conceptos que equiparan exactamente al término de búsqueda sin aplicar normalización previa.
- Búsqueda aproximada (*Normalized Word*): el método separa el término en sus palabras constituyentes y elimina un pequeño número de stopwords, así como las variaciones léxicas de cada palabra (tales como mayúsculas y variantes ortográficas). Tras la normalización, el método devuelve los conceptos UMLS que contienen todas las palabras resultantes de la normalización.

Se ha seguido el siguiente procedimiento para buscar conceptos SNOMED-CT: primero UTS API es usado para buscar conceptos UMLS dado el término de búsqueda, a continuación se extraen los conceptos SNOMED-CT asociados a dichos conceptos UMLS (ver figura 3.4).

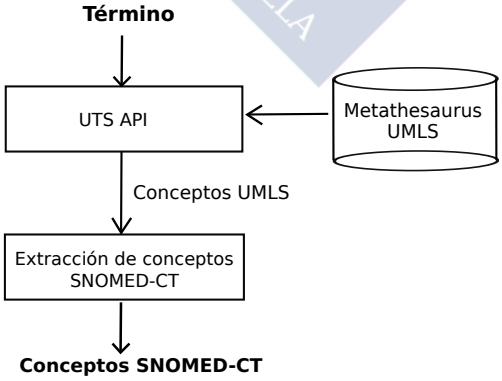

Figura 3.4: Proceso general para buscar alineamientos léxicos en SNOMED-CT con UTS API

3.6. Técnicas estructurales-contextuales

En esta sección se expone un conjunto de técnicas estructurales o de contexto usadas para mejorar la búsqueda en SNOMED-CT. En primer lugar, se describen técnicas que usan las relaciones semánticas de SNOMED-CT para enriquecer las búsquedas léxicas clásicas dentro de la terminología SNOMED-CT. Estas técnicas están destinadas a buscar conceptos en SNOMED-CT para términos clínicos individuales, es decir, aquellos de los que no se tiene contexto ya que no están ubicados dentro de un modelo de datos clínicos (sección 3.6.1). A continuación, se exponen varias técnicas destinadas a mapear términos procedentes de modelos clínicos (semi)estructurados a SNOMED-CT (sección 3.6.2). Estas técnicas aprovechan la estructura interna de los modelos clínicos y de la terminología SNOMED-CT para mejorar la calidad y cobertura del mapping. Cabe destacar que las técnicas estructurales-contextuales no suelen ser aplicadas aisladamente, sino que suelen ser usadas en combinación con técnicas léxicas.

3.6.1. Uso de relaciones semánticas para mejorar las búsquedas léxicas en SNOMED-CT

El enfoque habitual para buscar equiparaciones en SNOMED-CT para un término clínico suele ser el siguiente: primero, se selecciona alguna métrica léxica; a continuación, esta se usa para calcular la similitud del término con cada una de las descripciones de SNOMED-CT; y finalmente se obtiene un ranking de conceptos ordenados por similitud léxica. Con este enfoque, el grado de similitud entre un término y un concepto depende sólo de la semejanza léxica entre el término y las descripciones asociadas a dicho concepto. Un ejemplo de esto puede verse en la figura 3.5, la cual ilustra la información usada para evaluar una posible equiparación entre el término ‘total lung excision’ y el concepto SNOMED-CT con id 49795001 y descripciones ‘total pneumonectomy’ y ‘pneumonectomy’.

En este apartado, se exponen técnicas experimentales que explotan las relaciones semánticas de SNOMED-CT para extraer información adicional de los conceptos. La técnica se apoya en la siguiente hipótesis: los conceptos con los que está relacionado un concepto pueden ser útiles para extraer información adicional para la búsqueda de equiparaciones léxicas. La figura 3.6 muestra un ejemplo de uso de relaciones, en el que el término buscado es ‘total lung excision’ y el concepto evaluado es ‘total pneumonectomy’. Aplicando métricas léxi-

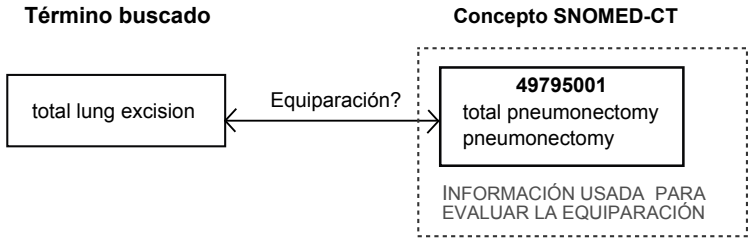

Figura 3.5: Enfoque clásico para buscar equiparaciones léxicas en SNOMED-CT. Ejemplo entre el término de búsqueda 'total lung excision' y el concepto de SNOMED-CT 'total pneumonectomy'.

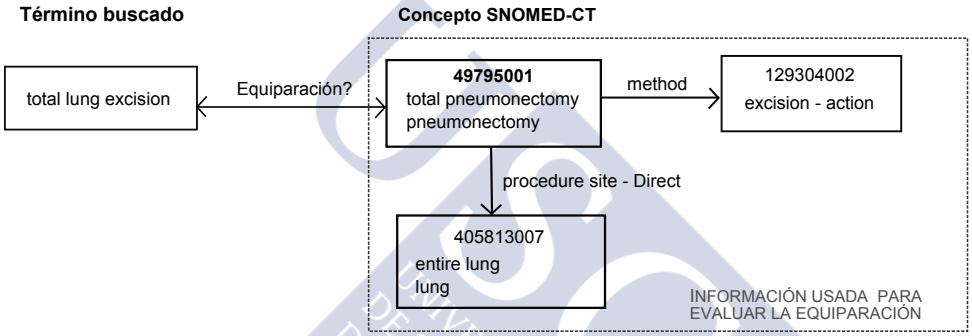

Figura 3.6: Uso de relaciones de SNOMED-CT para extraer contexto adicional de los conceptos SNOMED-CT. Ejemplo con el concepto 'total pneumonectomy'

cas estas entidades tendrían una baja similitud. Sin embargo, si recorremos las relaciones de SNOMED-CT asociadas a 'total pneumonectomy' podemos extraer significado adicional del concepto: 'total pneumonectomy' implica realizar una excisión ('excision - action') en el pulmón ('lung'). Esta información semántica puede ser clave para asignar una correspondencia entre el término buscado 'total lung excision' y el concepto 'total pneumonectomy'.

A continuación, se presentan dos técnicas experimentales que aprovechan las relaciones de SNOMED-CT para extraer información semántica de un concepto. Estas técnicas usan esta información para optimizar las búsquedas de equiparaciones entre el término de búsqueda y los concepto de SNOMED-CT.

Búsqueda de palabras no equiparadas en los conceptos relacionados

Esta técnica comienza realizando una equiparación léxica entre el término buscado y un concepto SNOMED-CT. Si todas las palabras del término están presentes en el concepto se asigna directamente un mapping. En caso de que alguna palabra del término no exista en el concepto, la técnica busca estas palabras en el contexto o entorno del concepto, esto es, en los conceptos con los que está relacionado en SNOMED-CT. La figura 3.7 muestra el esquema general de la técnica. A continuación, se detallan más formalmente los distintos pasos que tienen lugar en la técnica:

– En primer lugar, se realiza una equiparación léxica para extraer el conjunto de palabras del término no presentes en el concepto evaluado. A este conjunto le llamaremos de aquí en adelante C_{term}.

– A continuación, se extraen las descripciones de los conceptos relacionados jerárquicamente y lógicamente al concepto evaluado. Estas descripciones se normalizan, tokenizan y combinan para conseguir un conjunto de palabras que llamaremos C_{rel}.

– Finalmente, se comprueba si cada palabra de C_{term} existe en el conjunto C_{rel}. Si es así, se considera que hay un mapping o equiparación entre el término y el concepto evaluado.

La figura 3.8 muestra un ejemplo de aplicación de esta técnica para el término ‘total lung excision’ y el concepto ‘total pneumonectomy’:

– La equiparación léxica inicial obtiene que C_{term} = lung, excision.

– Los conceptos relacionados al concepto ‘total pneumonectomy’ son extraídos (ver figura 3.8). Las descripciones de estos conceptos son procesadas dando lugar a C_{rel} = excision, action, lung, entire.

– La técnica encuentra correspondencia entre el término de búsqueda ‘total lung excision’ y el concepto ‘total pneumonectomy’ ya que cada palabra de C_{term} está en C_{rel}.

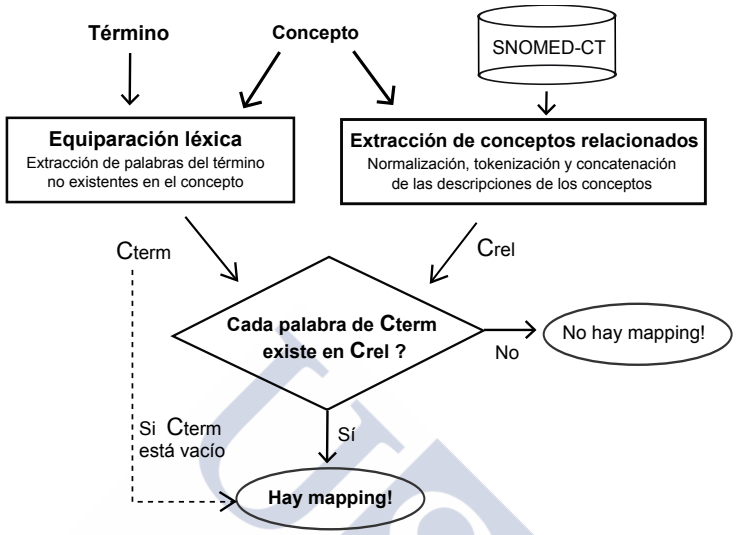

Figura 3.7: Esquema del proceso de búsqueda de palabras no equiparadas en conceptos vecinos.

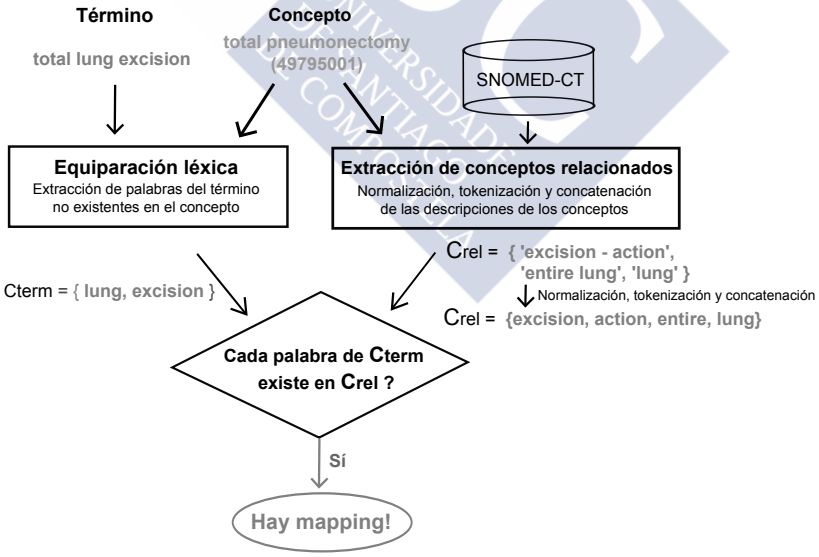

Figura 3.8: Ejemplo de aplicación del proceso de búsqueda de palabras no equiparadas en conceptos vecinos.

Equiparaciones léxicas parciales en los conceptos relacionados.

El enfoque de esta técnica para buscar equiparaciones entre un término 'T' y un concepto 'C' es bastante similar al de la técnica anterior. Pero, en este caso la técnica primero realiza una equiparación léxica aproximada entre 'T' y un concepto 'C', y posteriormente busca equiparaciones léxicas parciales entre el término completo 'T' y los conceptos relacionados a 'C', generando un conjunto de puntuaciones de similitud. La figura 3.9 muestra el esquema general de la técnica. A continuación, se detallan formalmente las etapas más relevantes:

- Primero, se aplica alguna de las métricas léxicas aproximadas descritas en la sección 3.3.3 entre el término 'T' y el concepto evaluado 'C'. Como resultado se obtiene la puntuación principal de similitud que llamaremos P_{TC}.
- A continuación, se extraen todos los conceptos relacionados (C_{rel}) a 'C' en SNOMED-CT.
- Finalmente, se buscan equiparaciones léxicas parciales entre cada uno de los conceptos de C_{rel} y 'T'. Como resultado se obtiene un conjunto de puntuaciones secundarias que llamaremos P_{TCrel}.

A final para cada par término-concepto, se obtendrá una puntuación léxica principal y varias puntuaciones secundarias. Determinar si hay un mapping en función de estas puntuaciones no es trivial. Es necesaria alguna técnica extra de selección o filtrado, que puede ir desde algunas reglas sencillas para establecer umbrales mínimos en las puntuaciones, hasta procesos de aprendizaje automático que aprendan la importancia de cada una de las puntuaciones.

En el capítulo 4 se exponen más detalles de esta técnica en una aplicación concreta de mapping, cuyo objetivo es buscar conceptos SNOMED-CT para un glosario de términos relacionados a procedimientos en patología (ver sección 4.2.3). En esta aplicación, la técnica usa solamente determinadas relaciones de SNOMED-CT que favorecen el mapping, y un conjunto de reglas heurísticas para determinar si hay mapping en base a las puntuaciones léxicas obtenidas.

3.6.2. Mapping entre modelos clínicos (semi)estructurados y SNOMED

Este apartado agrupa un conjunto de técnicas optimizadas para enlazar términos procedentes de modelos clínicos (semi)estructurados con conceptos de SNOMED-CT. Las técnicas

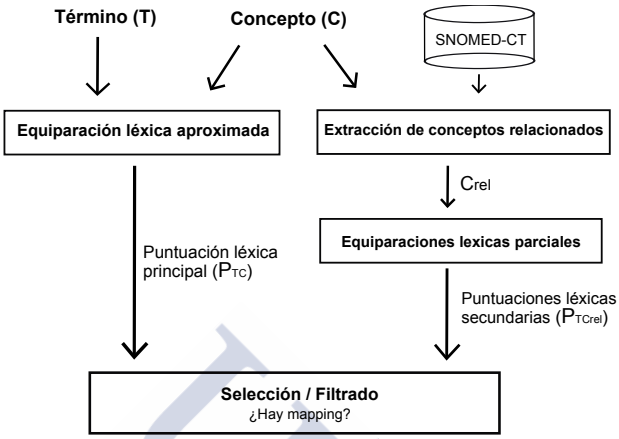

Figura 3.9: Esquema del proceso de búsqueda de equiparaciones parciales en conceptos vecinos.

están orientadas a arquetipos openEHR (descritos ampliamente en la sección 2.2.2), pero son extrapolables a otros modelos clínicos que tengan una cierta estructura.

Los arquetipos openEHR pueden verse como una plantilla o formulario para la recogida sistemática de datos en una situación o instancia clínica determinada, por ejemplo, para la medición de la presión sanguínea de un paciente (ver figura 3.10). Los arquetipos están formados por una agrupación de ítems cuyo propósito es definir aspectos o cuestiones que deben medirse u observarse sobre una situación clínica. Estos ítems tienen asociados un nombre o término en lenguaje natural. Por ejemplo, en el arquetipo de la presión sanguínea se incluyen los términos 'Mean arterial pressure' y 'position'.

Los términos de los arquetipos también pueden agruparse o representarse en forma de jerarquía (ver figura 3.11). Esta jerarquía representa las dependencias entre los términos del arquetipo. Así, por ejemplo, el término 'sitting position' depende de 'position', y este a su vez depende del término 'blood pressure'. No se puede confundir esta agrupación o jerarquía de dependencias con las jerarquías formales de las ontologías y terminologías médicas, donde los conceptos están relacionados a través de relaciones jerárquicas IS A. Se observó que en esta jerarquía de términos, el término raíz representa la situación clínica general (en este caso presión sanguínea). Los términos del segundo nivel normalmente definen aspectos u observaciones que deben recopilarse sobre la situación clínica (p.e. presión sistólica y posición). Mientras que los términos de tercer nivel suelen definir valores o hallazgos predefinidos relacionados (p.e. posición sentada o acostada).

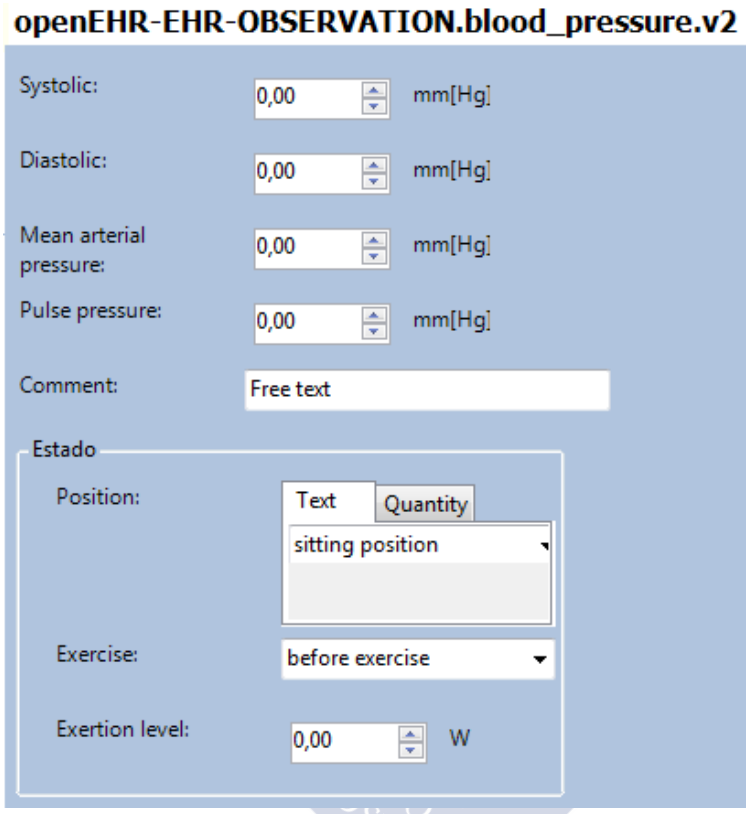

Figura 3.10: Vista en forma de formulario del arquetipo blood pressure.

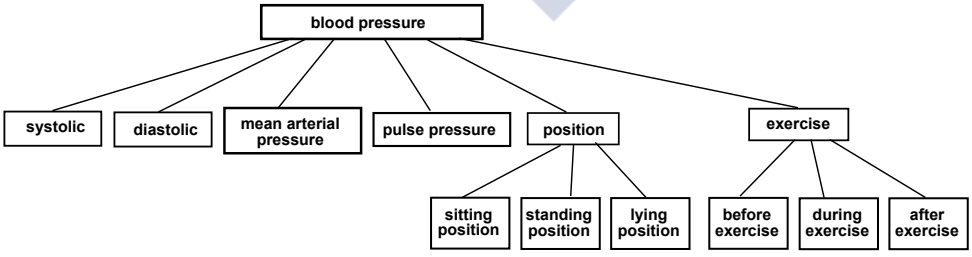

Figura 3.11: Representación jerárquica del contenido clínico del arquetipo blood pressure.

Las técnicas estructurales-contextuales se apoyan en la hipótesis que los términos de un mismo arquetipo están estrechamente relacionados entre sí. En relación a esto, vamos a definir dos principios que usaremos en el proceso de mapping automático:

– Principio de proximidad: si se logra mapear de forma precisa un elemento del arquetipo a un concepto 'c' de SNOMED-CT, entonces es bastante probable que los elementos vecinos del arquetipo mapeen a conceptos semánticamente relacionados a 'c' en SNOMED-CT.

– Principio de similitud estructural: Existen similitudes estructurales entre la estructura de SNOMED-CT y la agrupación de términos de los arquetipos.

A continuación, se expone un conjunto de técnicas estructurales-contextuales orientadas a mapear los términos de los arquetipos a conceptos de SNOMED-CT.

Técnica léxico-contextual

La técnica léxico-contextual trata de aprovechar el principio de proximidad para detectar mappings en un arquetipo. Esta técnica comienza buscando equiparaciones exactas entre los términos de un arquetipo y todos conceptos de SNOMED-CT. Posteriormente, usa los mappings detectados para extraer pequeños subconjuntos de SNOMED-CT formados por conceptos altamente relacionados al arquetipo. Finalmente, aplica técnicas léxicas aproximadas o parciales entre los términos del arquetipo aún no mapeados y los conceptos de los subconjuntos de SNOMED-CT extraídos. A continuación, se describen las principales etapas de la técnica léxico-contextual:

Etapa 1: Búsqueda de equiparaciones léxicas exactas en todo SNOMED-CT

En esta primera etapa se aplican técnicas léxicas para buscar equiparaciones exactas entre los términos del arquetipo y todos conceptos de SNOMED-CT. Como resultado, al final de la etapa se generan mappings candidatos bastante fiables para algunos términos del arquetipo. Por ejemplo, la figura 3.12 muestra a la izquierda la jerarquía de términos del arquetipo 'blood pressure'. Tras la búsqueda de equiparaciones léxicas exactas, se obtiene un mapping candidato para equiparar el término 'blood pressure' con el concepto 'blood pressure (observable entity)'.

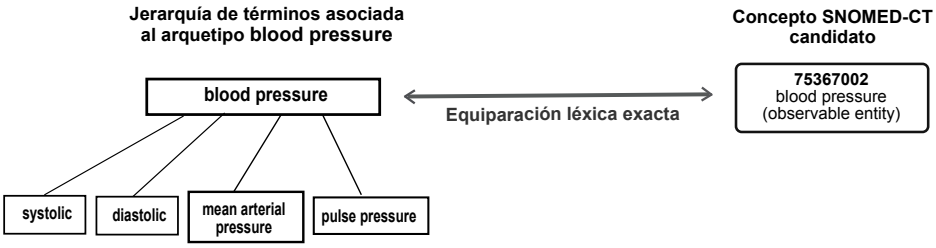

Figura 3.12: Búsqueda de equiparaciones exactas para los términos del arquetipo 'blood pressure'

Etapa 2: Extracción de subconjuntos de SNOMED-CT relevantes

La segunda etapa extrae subconjuntos de SNOMED-CT próximos a los conceptos candidatos generados en la etapa previa. Estos subconjuntos están formados por conceptos relacionados, en SNOMED-CT, a los conceptos candidatos. Por ejemplo, la figura 3.13 muestra parte de un subconjunto de SNOMED-CT asociado al concepto 'blood pressure (observable entity)'. Este subconjunto incluye el ascendiente directo de 'blood pressure', sus descendientes y conceptos relacionados a través de relaciones lógicas.

Los subconjuntos extraídos en esta etapa contienen conceptos muy relacionados con la semántica del arquetipo, por lo que tienen muchas más posibilidades de ser mapeados que un concepto aleatorio de SNOMED-CT.

Etapa 3: Equiparación léxica aproximada en los subconjuntos relevantes de SNOMED

La tercera etapa aplica técnicas léxicas aproximadas/parciales entre los términos del arquetipo aún no mapeados y los conceptos del subconjunto de SNOMED-CT extraído. La figura 3.14 muestra, a la izquierda, la jerarquía de términos del arquetipo 'blood pressure'. El rectángulo marca los términos aún no mapeados. A la derecha, dentro del rectángulo se muestra el subconjunto asociado al concepto 'blood pressure' extraído en la segunda etapa. Tras esta etapa, se encontrarían nuevos mappings correctos entre el término 'systolic' y el concepto 'systolic blood pressure' y entre el término 'diastolic' y el concepto 'diastolic blood pressure'. Si las técnicas léxicas aproximadas/parciales fuesen aplicadas en todo SNOMED-CT, en vez de en subconjuntos relevantes, generarían mappings imprecisos. Por ejemplo, el término 'systolic' podría ser asignado a alguno de los conceptos que contienen la palabra 'systolic', tal como 'systolic phase of uterine contractions' o 'systolic heart failure', los cuales no son candidatos correctos.

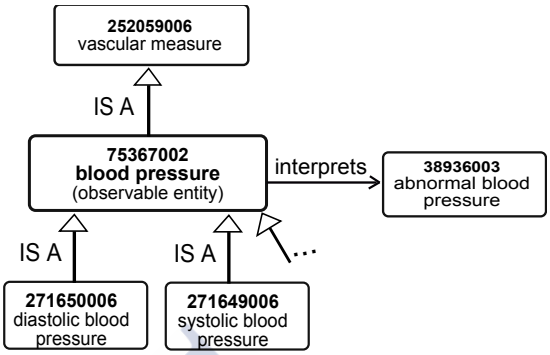

Figura 3.13: Parte de un subconjunto de SNOMED-CT asociado al concepto ‘’blood pressure’

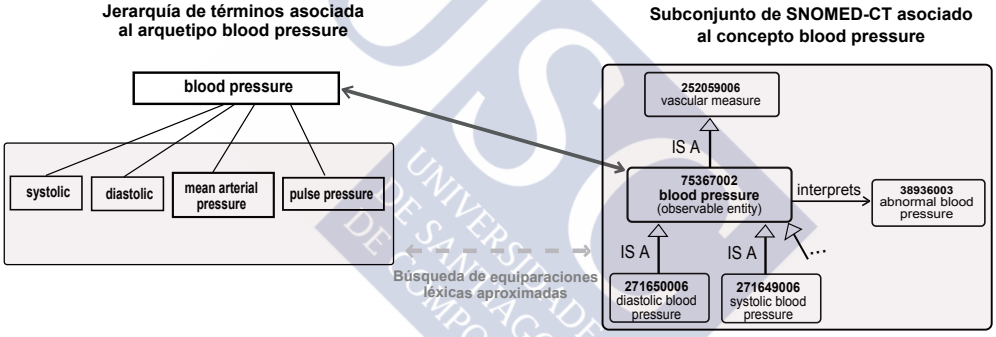

Figura 3.14: Búsqueda de equiparaciones léxicas aproximadas entre los términos del arquetipo ’blood pressure’ y los conceptos del subconjunto de SNOMED-CT relacionado a ’blood pressure’

Combinación de términos basada en la jerarquía interna de los arquetipos

Algunos elementos de los arquetipos contienen un nombre o término poco informativo o ambiguo. Por ejemplo, el término ‘systolic’ del arquetipo ‘blood pressure’ es poco informativo y las técnicas léxicas difícilmente serán capaces de mapearlo al concepto de SNOMED-CT ‘systolic blood pressure’. Cabe destacar que la palabra ‘systolic’ aparece en más de 100 conceptos de SNOMED-CT.

Para aumentar las posibilidades de encontrar un mapping adecuado para este tipo de términos, se ha desarrollado una técnica para enriquecer o expandir los términos usando el contexto del arquetipo. A continuación, se enumeran las etapas para enriquecer un término T:

1. El término T se normaliza y tokeniza.
2. Se extrae el término inmediatamente superior (Tp) en la jerarquía del arquetipo. Tp se normaliza y tokeniza.
3. Se combinan todos los tokens de T con todos los tokens de Tp, evitando tokens repetidos.
4. Se combinan todos los tokens de T con tokens de Tp, excluyendo uno de los tokens de Tp. Este paso se hace para cada token de Tp.
5. Se extrae el término raíz (Tr) en la jerarquía del arquetipo. Tr se normaliza y tokeniza. Si Tr es igual a Tp finaliza la ejecución sin la etapa 6 y 7.
6. Se combinan todos los tokens de T con todos los tokens de Tr, evitando tokens repetidos.
7. Se combinan todos los tokens de T con tokens de Tr, excluyendo uno de los tokens de Tr. Este paso se hace para cada token de Tr.

La figura 3.15 muestra un ejemplo de aplicación de esta técnica para el término 'systolic' del arquetipo 'blood pressure':

1. El término 'systolic' se normaliza y se tokeniza dando lugar a tokens(T) = systolic .
2. El término superior (Tp) de systolic es extraído, normalizado y tokenizado (ver jerarquía en la figura jer), dando lugar a tokens(Tp) = blood, pressure .
3. Combinaciones de T y Tp generan el término 'systolic blood pressure'.
4. Combinaciones de T y Tp (excluyendo un token) generan los términos 'systolic blood' y 'systolic pressure'.
5. Se extrae el término raíz Tr. La ejecución finaliza ya que Tr es igual a Tp.

Los términos obtenidos con esta técnica para el término 'systolic' son: 'systolic blood pressure', 'systolic blood' y 'systolic pressure'. Estos términos son buenas alternativas a 'systolic' para buscar equiparaciones léxicas en SNOMED-CT; y frecuentemente dan lugar a mappings más exactos o precisos.

3.7. Técnicas de desambiguación

El uso de diversas técnicas de búsqueda en SNOMED-CT incrementa las posibilidades de encontrar el mapping correcto para un término clínico. Sin embargo, es probable que en ocasiones se generen varios conceptos candidatos para un mismo término. Además, cada uno

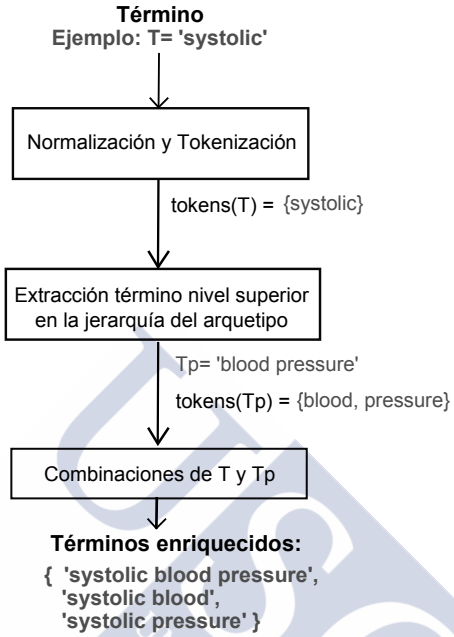

Figura 3.15: Expansión del término 'systolic' con el contexto del arquetipo 'blood pressure'

de estos conceptos puede tener asignado varias puntuaciones de similitud léxica o estructural. Seleccionar automáticamente el concepto correcto sobre el conjunto de candidatos no es trivial. En este apartado se presentan distintas estrategias para llevar a cabo esta selección o desambiguación de conceptos candidatos:

3.7.1. Desambiguación por categoría semántica

Esta estrategia usa las categorías semánticas de SNOMED-CT como criterio para seleccionar un concepto entre varios candidatos. La idea detrás de esta estrategia es que algunas de las categorías semánticas de SNOMED-CT frecuentemente incluyen conceptos más genéricos que otras categorías. Por ejemplo, la categoría 'Qualifier value' agrupa conceptos genéricos que no están asociados a ninguna otra categoría, tal como 'left', 'mild' y 'deep'. En contraste, la categoría 'Clinical finding' incluye conceptos que representan el resultado de una observación o evaluación, por ejemplo: 'deep breathing' y 'mild pain'. Estos conceptos suelen ser bastante más específicos.

Para implementar esta estrategia, se han establecido dos grupos de categorías semánticas: uno al que llamaremos genérico que incluye las categorías 'Qualifier value' y 'Linkage concept' y otro al que llamaremos específico formado por el resto de categorías semánticas de SNOMED-CT.

Para seleccionar los conceptos finales entre los candidatos, esta estrategia sigue una regla sencilla: si entre los candidatos hay conceptos asociados a categorías de los dos grupos, entonces todos aquellos del grupo 'genérico' son descartados, mientras que los conceptos pertenecientes al grupo 'específico' son seleccionados.

3.7.2. Desambiguación por similitud estructural

Esta estrategia es aplicable cuando se dispone de contexto del término de búsqueda, por ejemplo cuando el término está ubicado dentro de un arquetipo. Esta estrategia trata de seleccionar el concepto correcto entre un grupo de candidatos, aprovechando los mappings generados para el resto de términos del arquetipo. La idea central de esta estrategia es favorecer aquellos conceptos candidatos que son semánticamente consistentes con los conceptos asignados a otros términos del arquetipo. Para ello, se contabiliza el número de conexiones semánticas que tiene cada concepto candidato con el resto de conceptos mapeados del arquetipo. Se considera que hay una conexión semántica entre dos conceptos existe cuando estos están conectados a través de relaciones lógicas o jerárquicas en SNOMED-CT. El concepto candidato que contenga más conexiones semánticas es seleccionado como concepto final para mapear, mientras que el resto de conceptos candidatos son descartados.

3.7.3. Desambiguación por reglas heurísticas

Otra estrategia para desambiguar consiste en la creación de un conjunto de reglas heurísticas. Estas se suelen crear específicamente para una aplicación de mapping donde se conoce el tipo de términos clínicos a mapear. Las reglas definen una serie de criterios para automatizar la selección del concepto final entre un conjunto de candidatos. En el capítulo 4 se exponen un conjunto de reglas heurísticas para desambiguar los mappings asociados a términos de procedimientos en patología anatómica (ver sección 4.2.5).

3.7.4. Desambiguación por aprendizaje automático

Esta estrategia implica organizar el problema de mapping como un problema de aprendizaje automático, específicamente como un problema de clasificación binario en el que hay que decidir si hay o no hay mapping entre un término y un concepto en base a un conjunto de métricas o puntuaciones de similitud entre ambas entidades. Esta estrategia es especialmente adecuada cuando se dispone de un grupo grande de características (puntuaciones léxicas, categoría semántica, etc) entre cada concepto candidato y término a mapear. En estas circunstancias resulta complicado crear un conjunto de reglas heurísticas para desambiguar, por lo que el aprendizaje automático puede ser una buena solución. Sin embargo, una limitación importante de estas técnicas es que necesitamos disponer de un 'gold standard' para el proceso de aprendizaje. En este contexto, el 'gold standard' es un conjunto de mappings creados por 'curators' o mapeadores expertos. Lamentablemente, no existen apenas 'gold standard' relacionados al mapping en SNOMED-CT. Además, no hay garantías de que una técnica de desambiguación entrenada con un 'gold standard' formado por un tipo específico de términos clínicos, generalice correctamente y pueda ser usada para cualquier tipo de término clínico. En el capítulo 4 se expone con más detalle el uso del aprendizaje automático para desambiguar los mappings asociados a términos de procedimientos en patología anatómica (ver sección 4.2.5).

Capítulo 4

Enlazado automático de términos clínicos con conceptos SNOMED-CT

Actualmente, el uso de términos locales (definidos en lenguaje natural) está muy extendido en los registros médicos electrónicos. El uso de lenguaje natural para describir la información clínica proporciona un nivel de expresividad completa, pero dificulta el procesamiento computacional, interfiriendo con varios desafíos actuales de la informática médica: la interoperabilidad semántica de los sistemas de salud, el soporte a los Sistemas de Apoyo de Decisiones Clínicas (SADC) y el uso secundario de los datos [121]. En los últimos años están cobrando fuerza las terminologías clínicas como medio para codificar la información clínica de forma precisa y estandarizada, facilitando así la integración e intercambio de información clínica entre diferentes aplicaciones, registros médicos y SADC [105]. SNOMED-CT es, en la actualidad, la terminología más completa para codificar todos los aspectos de los registros electrónicos [100].

Una revisión bibliográfica sobre el uso de SNOMED-CT, incluyendo 488 artículos desde 2001 hasta 2012, ha mostrado que hasta ahora sólo unos pocos estudios han alcanzado un nivel de madurez considerable en el uso de SNOMED-CT dentro de la práctica clínica [74]. Un paso importante para promover el uso de SNOMED-CT en los sistemas de información clínica es facilitar el proceso de búsqueda de correspondencias o mappings entre términos locales de los actuales registros médicos y conceptos SNOMED-CT. Actualmente, existen

navegadores y servicios que permiten buscar mappings de un término clínico en SNOMED-CT [33, 60, 58, 93]. Sin embargo, la mayoría de ellos usan técnicas léxicas sencillas, tales como la búsquedas léxicas exactas o parciales (con comodines) y búsquedas Booleanas. Estás técnicas son insuficientes para tratar con la búsqueda en una terminología del tamaño (más de 300.000 conceptos) y granularidad de SNOMED-CT [82, 108, 68]. Hasta nuestro conocimiento, los sistemas de búsqueda existentes para SNOMED-CT no incluyen técnicas lingüísticas para tratar la sinonimia, ni técnicas estructurales para explotar las relaciones semánticas de las terminologías. En 2008, J. Rogers et al. evaluaron las características de 17 navegadores de SNOMED-CT, concluyendo que las funcionalidades de búsqueda texto-a-concepto de la mayoría de los navegadores son pobres e insuficientes [107]. También, P. Ruch et al. [108] and M. Chiang et al. [21] han expuesto que es necesario el desarrollo de nuevas herramientas de búsqueda para mejorar la calidad de la codificación con SNOMED-CT.

Este capítulo presenta una nueva metodología[1], orientada a la búsqueda de conceptos en SNOMED-CT, que combina técnicas léxicas clásicas con dos técnicas novedosas para descubrir mappings texto-a-concepto en la terminología SNOMED-CT. La primera técnica aplica un sistema de expansión de consultas para reformular y expandir los términos de búsqueda. La segunda explota las relaciones de SNOMED-CT para obtener contexto adicional sobre los conceptos terminológicos. La metodología fue usada para construir una herramienta de mapping con dos configuraciones distintas: una orientada a ser completamente automática - de aquí en adelante será referenciada como HMAS (herramienta de mapping automática para SNOMED-CT) - y la otra con una perfil más semi-automático - de aquí en adelante referenciada como HMSS (herramienta de mapping semi-automática para SNOMED-CT). HMAS fue diseñada para mapear automáticamente términos clínicos a conceptos de SNOMED-CT, sin ninguna intervención de usuarios expertos. Mientras que HMSS tiene el objetivo de ayudar a los expertos durante el proceso de mapping, recomendando varios conceptos SNOMED-CT para cada término clínico buscado.

El capítulo también incluye una evaluación, realizada con un glosario de procedimientos patológicos en español, de la herramienta desarrollada y de otros servicios de búsqueda en SNOMED-CT.

[1] Una descripción detallada de esta metodología ha sido incluida en uno de los artículos de investigación publicados durante la elaboración de la tesis [2]

4.1. Materiales

4.1.1. SNOMED-CT

SNOMED-CT es una terminología muy extensa que proporciona un estándar para codificar la información clínica. Contiene más de 300.000 conceptos, cada uno asociado a un identificador único y a varias descripciones textuales sinónimas (ver figura 4.1). Además, cada concepto tiene una o más relaciones semánticas a otros conceptos. Las relaciones pueden ser clasificadas en jerárquicas y en lógicas (o de atributo):

– Las relaciones jerárquicas *IS A* relacionan un concepto con otro más general.

– Las relaciones lógicas o de atributo asocian dos conceptos especificando una característica de uno de los conceptos. Hay diferentes tipos de relaciones de atributo. Por ejemplo, *procedure site* describe la parte de cuerpo en la que se realiza un determinado procedimiento clínico; mientras que *method* representa la acción que se realiza en un procedimiento clínico (ver ejemplos en la figura 4.1).

En la sección 2.3 se ha incluido una descripción más detallada de la terminología SNOMED-CT.

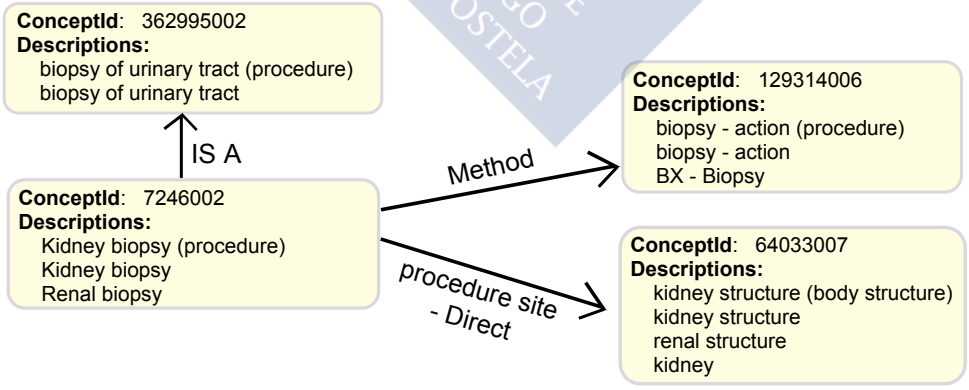

Figura 4.1: Ejemplo de conceptos y relaciones de SNOMED-CT

4.1.2. Navegadores de SNOMED-CT

En la actualidad existe un número importante de navegadores para acceder al contenido de SNOMED-CT (ver sección 2.4.1 para más información). Dos de los más destacados son los navegadores de la NLM [33] y de ITServer [60]. Estos navegadores proporcionan métodos de búsqueda texto-a-concepto basados principalmente en técnicas léxicas. El navegador NLM usa el Metatesauro UMLS para extraer términos sinónimos de los conceptos biomédicos procedentes de múltiples vocabularios y terminologías, tales como SNOMED-CT, MeSH and OMIM. El navegador NLM ofrece varias opciones de búsqueda, dos de las cuales son especialmente apropiadas para búsquedas de términos multi-idioma: *búsqueda exacta* y *búsqueda aproximada*[2] . La opción *búsqueda exacta* recupera conceptos que tienen asociados sinónimos exactamente iguales al término de búsqueda. La opción *búsqueda aproximada* separa las palabras del término de búsqueda, y recupera los conceptos conteniendo al menos una de esas palabras. El navegador ITServer provee un servicio para buscar conceptos SNOMED-CT equivalentes a términos en lenguaje natural (da soporte para términos en inglés y en español). El servicio aplica métricas basadas en distancia de edición y en coincidencia de palabras.

4.2. Métodos

Nuestra herramienta de mapping incluye técnicas léxicas, terminológicas, estructurales y de desambiguación para buscar mappings entre términos clínicos y conceptos SNOMED-CT. La figura 4.2 muestra las distintas fases o etapas de la herramienta. Primero, el término de búsqueda es normalizado. Seguidamente, este es expandido con términos alternativos. A continuación, se aplican técnicas léxicas y estructurales para buscar conceptos equivalentes a los términos alternativos generados. Estas técnicas obtienen un ranking de conceptos candidatos, por lo que finalmente se usan varias estrategias para seleccionar los mappings finales. En las secciones 4.2.1-4.2.4 se detallan las distintas etapas de la herramienta. En la sección 4.2.5 se exponen las diferencias entre los dos perfiles creados dentro de la herramienta: HMAS y HMSS.

[2] En el navegador NLM estas opciones de búsqueda son referenciadas como *Exact Match* y *Word Match*

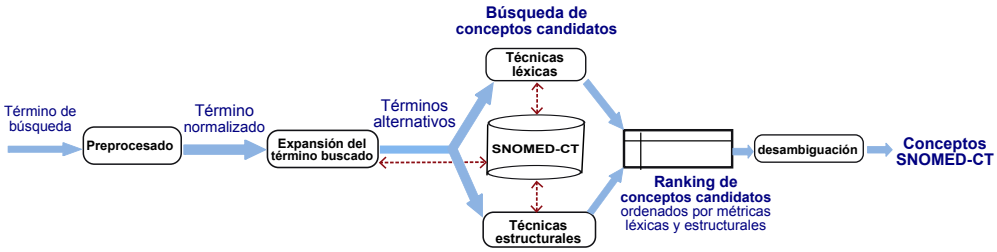

Figura 4.2: Flujo de trabajo de la herramienta de mapping

4.2.1. Preprocesado de términos y descripciones

Antes de la etapa de equiparación, las descripciones de SNOMED-CT y los términos de búsqueda son normalizados con la metodología expuesta en la sección 3.3.1.

4.2.2. Expansión del término de búsqueda

Hemos desarrollado un sistema automático para expandir el término de búsqueda con términos sinónimos alternativos con el propósito de favorecer la etapa de búsqueda de equiparaciones. Hasta nuestro conocimiento, esta técnica no ha sido previamente utilizada en SNOMED-CT.

Dada una cadena o término de búsqueda, nuestro sistema primero tokeniza la cadena en las palabras constituyentes. Seguidamente, busca sinónimos de cada una de esas palabras en SNOMED-CT. Finalmente, el término de búsqueda se expande mediante el reemplazo de sus palabras por los sinónimos generados. La sección 3.4 expone esta técnica en detalle.

4.2.3. Búsqueda de conceptos candidatos

Técnicas léxicas

Se han usado dos técnicas léxicas (*Distancia Levenshtein* y *Similitud de coseno*) de la librería SimMetrics[125] para buscar equiparaciones aproximadas entre las cadenas de textos asociadas a los términos de búsqueda (término original más los alternativos) y todas las descripciones de SNOMED-CT. La *Distancia Levenshtein* es un tipo de distancia de edición que mide la similitud de dos cadenas de texto calculando el mínimo número de operaciones de edición requeridas para convertir una de las cadenas en la otra. La *Similitud de coseno* basa la

similitud en el número de palabras comunes entre las dos cadenas. La sección 3.3.4 incluye más detalles sobre estas técnicas.

Estas técnicas obtienen una puntuación de similitud en el rango [0, 1] entre dos cadenas de texto. Nuestro sistema de búsqueda de equiparaciones, usando estas técnicas, calcula las puntuaciones de similitud entre un término de búsqueda y cada descripción de SNOMED-CT, obteniendo en última instancia un ranking de descripciones ordenado por puntuación. Por tanto, las descripciones de SNOMED-CT más similares léxicamente al término de búsqueda las podemos encontrar en las primeras posiciones del ranking obtenido. Así, por ejemplo, la tabla 4.1 muestra las dos primeras posiciones del ranking obtenido para el término de búsqueda 'excision biopsy of skin'.

Tabla 4.1: Dos primeras posiciones del ranking léxico obtenido para el término 'excision biopsy of skin'

Descripción SNOMED-CT candidata (id del concepto)	Distancia Levenshtein	Similitud de coseno
incision biopsy of skin (282014007)	0.90	0.58
excision biopsy of skin lesion (312968005)	0.74	0.8

Técnicas léxico-estructurales

Las técnicas léxicas obtienen un ranking de conceptos ordenados por similitud léxica. Sin embargo, se ha detectado que en ocasiones las técnicas léxicas recuperan resultados erróneos en las primeras posiciones del ranking. Por ejemplo, dado el término de búsqueda 'excision biopsy of skin', la *Distancia Levenshtein* propone como mejor candidato al concepto 'incision biopsy of skin' (conceptId: 282014007) ya que sólo son necesarias dos ediciones para transformar el término de búsqueda en el concepto propuesto. Sin embargo, este concepto podría no ser el más adecuado semánticamente.

Considerando estas situaciones, hemos diseñado una estrategia automática para prevenir los resultados erróneos de las técnicas léxicas. La estrategia selecciona los mejores conceptos del ranking léxico y analiza su semántica examinando sus relaciones lógicas en la terminología SNOMED-CT. Estas relaciones proveen información adicional de los conceptos.

La figura 4.3 muestra ejemplos de relaciones de atributo entre conceptos de SNOMED-CT. En la figura se puede ver que el concepto 'total pneumonectomy' (conceptId: 49795001) está relacionado a 'excision - action' (conceptId: 129304002) a través de la relación lógica

'method', y al concepto 'lung' (conceptId: 181216001) a través de 'procedure site - Direct'. Gracias a estas relaciones podemos saber que 'total pneumonectomy' es un procedimiento en el que se realiza una excisión y que esta se realiza en el pulmón. En este ejemplo, las relaciones podrían ser usadas para sugerir el concepto 'total pneumonectomy' como candidato para el término de búsqueda 'excision of lung', a pesar de la baja similitud léxica entre ellos. Además, las relaciones también pueden ser usadas para identificar las subcadenas más relevantes de la descripción de un concepto. Por ejemplo, en el concepto 'excision biopsy of skin lesion' las relaciones identificarían dos subcadenas como relevantes: 'excision biopsy' y 'skin' ya que el concepto está relacionado a través de la relación lógica 'method' al concepto 'excision biopsy' (conceptId: 277261002), y a través de la relación 'procedure site - Direct' al concepto 'skin' (conceptId: 39937001)

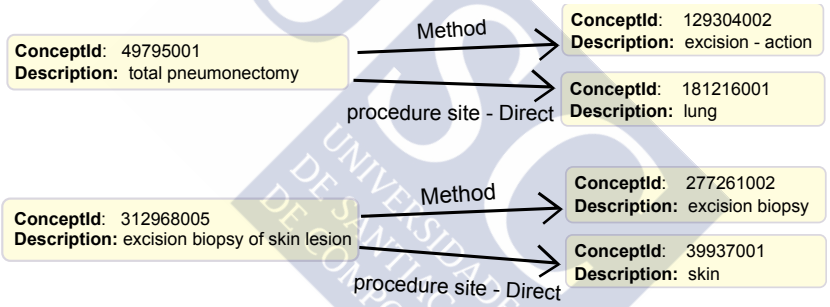

Figura 4.3: Relaciones SNOMED-CT proporcionando contexto adicional para los conceptos 'total pneumonectomy' y 'excision biopsy of skin lesion'

Uso de relaciones lógicas de SNOMED-CT para mejorar las búsquedas léxicas

Nuestra estrategia para tomar ventaja de las relaciones lógicas de SNOMED-CT ha sido expuesta de forma genérica en la sección 3.6.1 *Equiparaciones léxicas parciales en los conceptos relacionados*. Resumidamente, nuestra estrategia, dado un término de búsqueda 'T', primero extrae los mejores conceptos candidatos aplicando las técnicas léxicas expuestas anteriormente (típicamente los 10 mejores). Seguidamente, usa las relaciones lógicas de estos conceptos para extraer sus conceptos asociados. Finalmente, aplica métricas léxicas entre el término 'T' y los conceptos asociados para calcular nuevas puntuaciones léxicas.

Considerando que nuestro conjunto de datos de evaluación incluye términos de procedimientos patológicos (ver sección 4.3.1), hemos realizado algunas adaptaciones a la estrategia

genérica expuesta en la sección 3.6.1 *Equiparaciones léxicas parciales en los conceptos relacionados*:

- La búsqueda inicial de candidatos léxicos se limita a la jerarquía *procedure* de SNOMED-CT.
- Hemos seleccionado dos tipos de relaciones de atributo (*method* y *procedure site - Direct*) especialmente relevantes para conceptos de la jerarquía *procedure* de SNOMED-CT. Así, para cada uno de los procedimientos candidatos, nuestra estrategia usa estas relaciones para extraer dos conceptos de SNOMED-CT: la *acción* y la *parte del cuerpo* asociada al procedimiento.
- Hemos usado técnicas léxicas parciales (ver sección 3.3.3) para comprobar si el término de búsqueda 'T' contiene o no la descripción de los dos conceptos extraídos. Hemos definido dos métricas para reflejar si hay equiparación parcial:
  - *Similitud 'Action'*: métrica que es igual a 1 si todas las palabras del concepto *acción* obtenido con la relación *method* están presentes en el término de búsqueda. Si no es así, la métrica toma el valor 0.
  - *Similitud 'Body'*: métrica que es igual a 1 si todas las palabras del concepto *parte del cuerpo* obtenido con la relación *procedure site - Direct* están presentes en el término de búsqueda. Si no es así, la métrica toma el valor 0.

Vamos a ver un ejemplo de este proceso para el término de búsqueda 'excision biopsy of skin'. Usando las técnicas léxicas, nuestra herramienta primero obtiene un ranking de candidatos ordenados por similitud léxica (la tabla 4.1 muestra las dos primeras posiciones de este ranking). Seguidamente, se extraen los conceptos relacionados a cada candidato con las relaciones *method* y *procedure site - Direct* (ver figura 4.4). A continuación, las técnicas léxicas parciales comprueban si el término de búsqueda 'excision biopsy of skin' contiene la descripción de los conceptos extraídos y asignan nuevas puntuaciones para las métricas *Similitud 'Action'* y *Similitud 'Body'* (ver tabla 4.2). Como podemos ver en la tabla, ahora cada candidato tiene 4 puntuaciones asignadas. El concepto 'excision biopsy of skin lesion' que ocupaba la segunda posición en el ranking, obtuvo una mejor puntuación en la métrica *Similitud 'Action'* y la misma puntuación en *Similitud 'Body'* que el primer candidato 'incision biopsy of skin'.

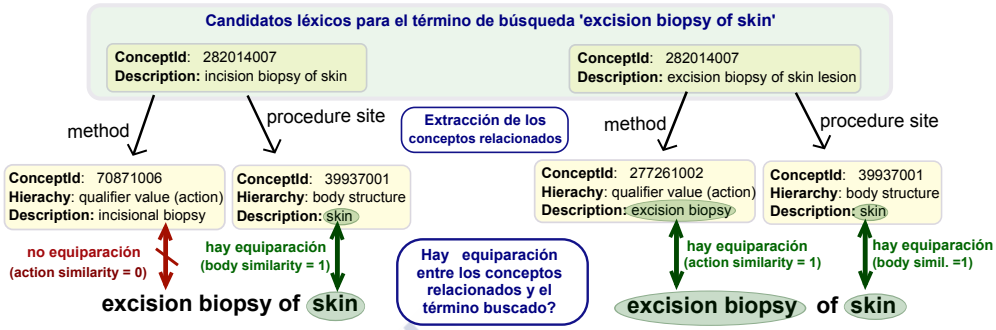

Figura 4.4: Ejemplo de uso de relaciones SNOMED-CT para mejorar la búsquedas de mappings del término 'excision biopsy of skin'

Tabla 4.2: Dos primeras posiciones del ranking para el término 'excision biopsy of skin' incluyendo las puntuaciones léxicas y léxica-estructurales

Descripción SNOMED-CT candidata	Distancia Levenshtein	Similitud de coseno	Similitud 'Action'	Similitud 'Body'
incision biopsy of skin (282014007)	0.90	0.58	0	1
excision biopsy of skin lesion (312968005)	0.74	0.8	1	1

Las dos nuevas métricas asignadas en esta etapa añaden más información o características sobre cada concepto candidato. Sin embargo, en algunos casos pueden dificultar la selección final del mapping, ya que las distintas métricas podrían no "coincidir" en cual es el mejor concepto. Por ello, se han implementado varias técnicas de selección o desambiguación.

4.2.4. Técnicas de desambiguación

Se han seguido dos estrategias para desambiguar el conjunto de candidatos encontrados para un término de búsqueda:

– Reglas heurísticas

  Mediante múltiples pruebas, se han creado manualmente un conjunto de reglas para seleccionar el mapping final dadas las puntuaciones de similitud de un conjunto de conceptos candidatos.

– Aprendizaje automático

Considerando que nuestro conjunto de datos incluye términos de procedimientos patológicos junto con los conceptos SNOMED-CT equivalentes (ver sección 4.3.1), hemos usado un subconjunto de estos datos para entrenar un modelo para predecir cual es el mapping final dadas las puntuaciones de similitud de un conjunto de conceptos candidatos.

Hemos organizado el problema de predicción como un problema de clasificación binaria. El modelo ha sido entrenado con máquinas de vectores de soporte (Support Vector Machines, SVM), específicamente con la implementación de la librería Kernlab[3] del lenguaje estadístico R. Los SVM son un conjunto de algoritmos de aprendizaje supervisado adecuados para problemas de clasificación y regresión. Durante la fase de aprendizaje, los SVM combinan diferentes tipos de características (en nuestro caso métricas de similitud) para tratar de minimizar el error de entrenamiento.

4.2.5. Diferencias entre las configuraciones HMAS y HMSS

Las dos configuraciones creadas dentro de la herramienta usan las técnicas expuestas en las secciones 4.2.1-4.2.4. La principal diferencia entre ellas está en la implementación y configuración de las técnicas de desambiguación. Hay que tener en cuenta que HMAS busca automatizar completamente el proceso de mapping, por lo que la desambiguación intenta reducir al máximo los falsos positivos seleccionando sólo el candidato con mayor probabilidad de ser correcto. En contraste, HMSS trata de lograr una cobertura alta más que una precisión perfecta. Por ello, las técnicas de desambiguación en este caso fueron ajustados para seleccionar los 5 mejores conceptos candidatos para cada términos buscado. A continuación, se exponen más detalles sobre la implementación y configuración de la desambiguación en ambas configuraciones:

Desambiguación con reglas heurísticas para HMAS

Las reglas heurísticas definidas para la desambiguación de la configuración HMAS son:

– Los conceptos candidatos con alguna métrica léxica-estructural (*Similitud 'Action'* y *Similitud 'Body'*) igual a 0 son descartados.

– Los candidatos con una puntuación léxica media (obtenida con la (*Distancia Levenshtein* y *Similitud de coseno*) menor que 0.95 son descartados.

[3] `http://cran.r-project.org/web/packages/kernlab/index.html`

– Entre los conceptos que cumplen las dos anteriores condiciones, sólo se selecciona el concepto con la puntuación léxica media más alta.

– Si dos conceptos tienen la misma puntuación léxica media, se selecciona, si lo hubiese, aquel obtenido con el término original de búsqueda (esto es, sin uso de sinonimia).

Desambiguación con reglas heurísticas para HMSS

Las reglas heurísticas definidas para la desambiguación de la configuración HMSS son:

– Se calcula la puntuación léxica media de cada concepto candidato obtenida con la *Distancia Levenshtein* y *Similitud de coseno*. Los conceptos candidatos con alguna métrica léxica-estructural (*Similitud 'Action'* y *Similitud 'Body'*) igual a 0 decrecen la puntuación léxica media en un 10 %.

– Se seleccionan los 5 conceptos candidatos con mejor puntuación léxica media.

Desambiguación con aprendizaje automático en las configuraciones HMAS y HMSS

Se han usado máquinas de vectores de soporte (SVM) en ambas configuraciones siguiendo las siguientes etapas:

– Se ha creado una tabla inmensa con las 4 métricas de similitud entre cada término del conjunto y cada concepto SNOMED-CT, más una columna extra para establecer si los expertos asignaron que hay mapping (ver ejemplo en la tabla 4.3).

– Aleatoriamente se ha generado un conjunto de entrenamiento y de evaluación.

– Se ha entrenado el modelo usando SVM con el conjunto de datos de entrenamiento.

  - El modelo de **configuración HMSS** fue afinado para devolver los 5 conceptos candidato con mayor valor de confianza.
  - En cambio, el entrenamiento para la **configuración HMAS** fue ajustado para penalizar en gran medida los falsos positivos. El apéndice A incluye más detalles sobre el entrenamiento en ambas configuraciones.

– Se ha usado el modelo creado para clasificar el conjunto de datos de evaluación.

– Se ha evaluado las predicciones del modelo contra los mappings de referencia creados por los expertos.

Tabla 4.3: Extracto de la tabla usada para el entrenamiento del modelo de desambiguación de conceptos candidatos. Contiene las métricas de similitud y la validez del mapping asignada por los expertos.

Término buscado / Concepto	Distancia Levenshtein	Similitud de coseno	Similitud 'Action'	Similitud 'Body'	Clase (Mapping experto?)
Término: excision biopsy of skin Concepto: incision biopsy of skin (282014007)	0.90	0.58	0	1	0 (Negativo)
Término: excision biopsy of skin Concepto: excision biopsy of skin lesion (312968005)	0.74	0.8	1	1	1 (Positivo)
TérminoN - ConceptoN	...	...	...	...	...

4.3. Evaluación

El principal objetivo de la evaluación fue comprobar nuestra hipótesis inicial, esto es, la combinación de técnicas léxicas con técnicas lingüísticas y estructurales mejoran considerablemente los sistemas actuales de búsqueda en SNOMED-CT basados principalmente en técnicas léxicas.

En nuestra evaluación hemos medido el rendimiento de nuestra herramienta de búsqueda y lo hemos comparado con el rendimiento ofrecido por los navegadores de SNOMED-CT de la Librería Nacional de Medicina de EEUU [33] y de la empresa ITServer [60]. Estos dos navegadores han sido seleccionados por varios motivos: (1) son de los navegadores más populares, (2) son de los navegadores que incluyen las opciones de búsqueda más avanzadas, (3) funcionan tanto con términos en inglés como en español y (4) proporcionan servicios web para acceder a las búsquedas facilitando el proceso de evaluación.

4.3.1. Conjunto de datos

Recientemente, la Sociedad Española de Anatomía Patológica[4] (SEAP) ha publicado un glosario[5] de términos frecuentes en español de procedimientos en patología, junto con los conceptos SNOMED-CT equivalentes, asignados por expertos clínicos [44].

[4] `https://www.seap.es/`

[5] `http://www.seap.es/enlaces-de-interes/-/asset_publisher/h3M6/content/id/114697`

Este glosario ha sido usado como un *gold standard* para medir el rendimiento de la herramienta de búsqueda automática. La evaluación incluyó cerca de 300 términos sobre biopsias mapeadas por los expertos a un único concepto SNOMED-CT (ver apéndice A). La tabla 4.4 muestra cuatro de los términos incluidos en la evaluación y los correspondientes conceptos SNOMED-CT asignados por los expertos.

Tabla 4.4: Muestra de los términos incluidos en la evaluación

Término	ID del concepto asignado por experto	Descripción del concepto en español	Descripción del concepto en inglés
Biopsia de miocardio	387828005	biopsia de miocardio	myocardial biopsy
Biopsia de retina	172573007	biopsia de lesión retiniana	biopsy of retinal lesion
Mastectomía total	172043006	mastectomía simple	simple mastectomy
Esofaguectomía parcial	3980006	resección subtotal del esófago	subtotal resection of esophagus

4.3.2. Configuración de los experimentos

Se han diseñado dos experimentos para evaluar el rendimiento de las dos configuraciones de la herramienta de búsqueda, así como de otros servicios de búsqueda similares:

Experimento 1: Evaluación de la tarea de mapping automático

Este experimento está centrado en evaluar la capacidad de las herramientas para mapear automáticamente los términos a SNOMED-CT, sin ninguna intervención de expertos. En este experimento, se han buscado mappings para los 300 términos del conjunto de datos con las siguientes herramientas y configuraciones:

– La opción de *búsqueda exacta* del navegador NLM, limitando las búsquedas a la jerarquía de SNOMED-CT *Procedure*.

– El servicio de búsqueda del navegador ITServer. Este devuelve un ranking de conceptos. Las búsquedas fueron también limitadas a la jerarquía *Procedure*, y sólo los conceptos con una puntuación mayor que 0.85 fueron considerados mappings y tenidos en cuenta en la evaluación[6].

[6]Este umbral fue elegido por prueba y error. Probamos diferentes valores y seleccionamos el umbral que obtuvo mejores resultados

– Cuatro configuraciones de HMAS fueron evaluadas para analizar la contribución de las diferentes técnicas desarrolladas:

  - Configuración 1: HMAS incluyendo preprocesado (sección 4.2.1), técnicas léxicas (sección 4.2.3) y reglas heurísticas para desambiguar (sección 4.2.5).
  - Configuración 2: Configuración 1 + Expansión del término de búsqueda (sección 4.2.2).
  - Configuración 3: Configuración 2 + Técnicas léxico-estructurales (sección 4.2.3).
  - Configuración 4: Igual que Configuración 3 pero usando aprendizaje automático en vez de reglas heurísticas para desambiguar (sección 4.2.5).

Tres medidas clásicas de recuperación de información (precisión, recall y F-measure) fueron calculadas para estimar la calidad de los mappings automáticos, usando como referencia los mappings expertos creados por la SEAP.

En este contexto, la precisión estima la proporción entre el número de mappings automáticos correctos y el número de mappings encontrados por el método. Mientras que el recall estima la proporción entre el número de mappings automáticos correctos y el número de mappings creados por los expertos. F-measure representa una media ponderada entre recall y precisión. Hemos establecido un $\beta = 0.7$ dándole un menor peso al recall que a la precisión ya que esta última es claramente más importante en una tarea de mapping automático.

$$\text{Precisión} = \frac{\text{\# mappings automáticos correctos}}{\text{\# mappings automáticos totales}}$$

$$\text{Recall} = \frac{\text{\# mappings automáticos correctos}}{\text{\# mappings expertos totales}}$$

$$F - Measure = (1+\beta^2) * \frac{\text{Precisión * Recall}}{(\beta^2 * Precisin) + Recall}$$

Experimento 2: Evaluación de la tarea de recomendación o mapping semi-automático

El segundo experimento está centrado en evaluar el rendimiento de las herramientas para recomendar a los expertos varias alternativas de mappings. Este experimento usa los mismos términos de búsqueda que el experimento previo. Pero en este caso, las herramientas han sido configuradas para obtener 5 alternativas de mapping, esto es, los 5 conceptos mejor puntuados por las herramientas.

A continuación, se exponen las herramientas y configuraciones contempladas en este experimento:

- Las opciones de *búsqueda exacta* y *aproximada* del navegador NLM, limitando las búsquedas a la jerarquía de SNOMED-CT *Procedure*. La salida de ambas opciones de búsqueda fue agregada hasta alcanzar 5 conceptos candidato.
- El servicio de búsqueda del navegador ITServer. Los 5 conceptos del ranking con mejor puntuación fueron propuestos como conceptos candidatos.
- Cuatro configuraciones de HMSS fueron evaluadas:
  - Configuración 1: HMSS incluyendo preprocesado (sección 4.2.1), técnicas léxicas (sección 4.2.3) y reglas heurísticas para desambiguar (sección 4.2.5).
  - Configuración 2: Configuración 1 + Expansión del término de búsqueda (sección 4.2.2).
  - Configuración 3: Configuración 2 + Técnicas léxico-estructurales (sección 4.2.3).
  - Configuración 4: Igual que Configuración 3 pero usando aprendizaje automático en vez de reglas heurísticas para desambiguar (sección 4.2.5).

Para este experimento hemos evaluado sólo el recall. Consideramos que en este caso el recall gana mucha importancia ya que proveer el concepto correcto en un conjunto no demasiado amplio de candidatos podría ahorrar mucho tiempo a los codificadores expertos. La precisión no es importante en este experimento ya que los todas las herramientas están configuradas para devolver 5 conceptos candidato.

Para realizar ambos experimentos, hemos aleatoriamente particionado el conjunto de datos en dos partes: aproximadamente 200 términos fueron usados para el proceso de entrenamiento necesario en la configuración 4, mientras que 100 términos fueron usados para la evaluación.

Cada experimento fue repetido 5 veces con particionados aleatorios distintos para obtener una evaluación más fiable.

4.4. Resultados

4.4.1. Resultados del mapping automático (experimento 1)

La tabla 4.5 muestra los resultados del primer experimento obtenidos con HMAS y con los navegadores NLM e ITServer. Los mappings obtenidos durante este experimento pueden ser consultados en el apéndice A.

Todos los enfoques han logrado alta precisión en el mapping, entre el 84.7 % y 91.3 %. En los niveles de recall se ha observado una mayor diferencia entre enfoques. El navegador NLM ha obtenido 32 % de recall, el navegador ITServer un 40 %, mostrando una mejora de un 25 % respecto a NLM. La configuración 3 de HMAS ha alcanzado un 51.4 % de recall, mejorando un 28 % los resultados de ITServer. Los valores de F-measure de NLM, ITServer y HMAS (configuración 3) han sido 54.7 %, 63.6 % y 71.1 %, respectivamente.

Configuraciones de HMAS

A continuación, se enumeran algunos resultados relevantes de las 4 configuraciones de HMAS evaluadas:

- La configuración 2 de HMAS obtuvo una ligera reducción de un 6 % en precisión y un incremento de un 20 % en recall respecto a la configuración 1, gracias a que la primera incluye la técnica de expansión de términos basada en sinonimia.
- La configuración 3 obtuvo un incremento del 2.5 % en precisión y un 4 % en recall respecto a la configuración 2, gracias a que la primera incluye las técnicas léxico-estructurales.
- La desambiguación basada en reglas heurísticas (configuración 3) ha obtenido mejores resultados (un incremento del 8 % en F-Measure) que la basada en aprendizaje automático (configuración 4). Seguramente, el aprendizaje automático no ha obtenido mejores resultados debido al bajo número de ejemplos con los que se realizó el entrenamiento (sólo 200 datos de entrada).

– La configuración que obtiene mejores resultados es la configuración 3, que junto a la configuración 4, es la que incluye el mayor número de técnicas: preprocesado de términos, técnicas léxicas, técnicas léxico-estructurales, expansión del término de búsqueda con sinonimia y técnicas de desambiguación basadas en reglas heurísticas. Por tanto, los experimentos han demostrado que las distintas técnicas añaden valor a la herramienta de mapping.

Tabla 4.5: Resultados del mapping automático (experimento 1)

Herramienta de búsqueda	Recall (%)	Precisión (%)	F-Measure ($\beta = 0.7$)
Navegador NLM (*búsqueda exacta*)	32.0	84.7	54.7
Navegador ITServer	40.0	90.1	63.6
HMAS (Configuración 1)	41.2	91.2	64.9
HMAS (Configuración 2)	49.4	85.8	68.9
HMAS (Configuración 3)	51.4	88.0	71.1
HMAS (Configuración 4)	42.0	91.3	65.6

4.4.2. Resultados del mapping semi-automático (experimento 2)

La tabla 4.6 muestra los resultados del segundo experimento obtenidos con HMSS y con los navegadores NLM e ITServer. Los mappings obtenidos durante este experimento pueden ser consultados en el apéndice A.

Los valores de recall de NLM, ITServer y HMSS (configuración 3) han sido 41.6%, 54.0% y 71.1%, respectivamente. Nuestra herramienta obtuvo una mejora del 31.5% y del 70.6% en el recall respecto a los navegadores ITServer y NLM, respectivamente.

Configuraciones de HMSS

A continuación, se enumeran algunos resultados relevantes de las 4 configuraciones de HMSS evaluadas:

– La configuración 2 de HMSS obtuvo importante incremento de un 21% en recall respecto a la configuración 1, gracias a la técnica de expansión de términos.

– La configuración 3 obtuvo un ligero incremento del 3 % en recall respecto a la configuración 2, gracias a las técnicas léxico-estructurales.

– En este experimento, la desambiguación basada en reglas heurísticas (configuración 3) también ha obtenido mejores resultados que la basada en aprendizaje automático (configuración 4).

Tabla 4.6: Resultados del mapping semi-automático (experimento 2)

Herramienta de búsqueda	Recall (%)
Navegador NLM (*búsqueda exacta y aproximada*)	41.6
Navegador ITServer	54.0
HMSS (Configuración 1)	56.8
HMSS (Configuración 2)	68.8
HMSS (Configuración 3)	71.0
HMSS (Configuración 4)	51.0

4.5. Discusión

4.5.1. Comparativa de HMAS con las otras herramientas de búsqueda evaluadas

En esta sección, comparamos la mejor configuración de HMAS (configuración 3) con las herramientas de búsqueda de los navegadores NLM e ITServer. La tabla 4.7 muestra un resumes de las técnicas usadas en cada uno de los enfoques. HMAS incorpora 3 técnicas adicionales respecto a los otros enfoques: expansión de términos, técnicas léxico-estructurales y técnicas de desambiguación.

Si bien es cierto que todos los enfoques han logrado alta precisión (esencial si buscamos un enfoque sin intervención de usuarios expertos), HMAS ha mejorado notablemente la cobertura o recall del mapping, principalmente debido a la aportación de la técnica de expansión de términos con sinónimos. Cabe destacar, que esta técnica no ha afectado en gran medida a la precisión, tal y como podría esperarse.

Tabla 4.7: Resumen de las técnicas usadas por NLM, ITServer y HMAS para la tarea de mapping automático

Herramienta de búsqueda	Normalización	Técnicas léxicas	Expansión de términos	Técnicas léxico-estructurales	Técnicas de desambiguación
Navegador NLM	✗	✓	✗	✗	✗
Navegador ITServer	✓	✓	✗	✗	✗
HMAS (Configuración 3)	✓	✓	✓	✓	✓

4.5.2. Expansión de términos con sinonimia

Nuestra técnica de expansión de términos[7] ha sido diseñada para inferir sinónimos de palabras usando un *corpus* de más de un millón de descripciones de SNOMED-CT. Durante los experimentos detectamos que la técnica es capaz de identificar: (1) palabras con idéntico o muy similar significado en todos los contextos (p.e. 'total' y 'completo'), (2) palabras con significado similar en el contexto médico (p.e. 'extremidad' y 'miembro') y (3) nombres y adjetivos referenciando al mismo concepto (p.e. 'estómago' y 'gástrico'). La tabla 4.8 muestra una muestra de los sinónimos detectados por nuestra técnica.

Nuestra herramienta de mapping expande los términos de búsqueda reemplazando una o más palabras por los sinónimos inferidos. Los términos alternativos generados han contribuido significativamente al incremento de recall obtenido por nuestra herramienta respecto a las otras herramientas evaluadas[8].

WordNet[9] es una base de datos léxica que agrupa palabras en un conjunto de sinónimos llamados *synsets*. En los últimos años WordNet ha sido utilizado en recuperación de información para expandir términos de búsqueda con sinónimos y otras relaciones lingüísticas [130, 57].

[7] Demo disponible en `http://snomed-synonym-finder.appspot.com/`

[8] El siguiente enlace `http://goo.gl/vJgq3U` incluye algunos ejemplos en que los términos alternativos han facilitado el descubrimiento de mappings correctos en SNOMED-CT

[9] http://wordnet.princeton.edu/

Tabla 4.8: Muestra de sinónimos detectados por nuestra técnica de expansión

Tipo de sinónimos	sinónimos (en español)	sinónimos (en inglés)
Sinónimos en todos los contextos	total ≡ completo parcial ≡ incompleto bucal ≡ oral	total ≡ complete partial ≡ incomplete buccal ≡ oral
Sinónimos en el contexto médico	curetaje ≡ raspado miembro ≡ extremidad extirpación ≡ remoción	curettage ≡ excision member ≡ limb extirpation ≡ removal
Nombres y adjetivos asociados	estómago ≡ gástrico hígado ≡ hepático piel ≡ cutáneo	stomach ≡ gastric liver ≡ hepatic skin ≡ cutaneous

En nuestra herramienta de mapping no hemos usado WordNet para extraer sinónimos debido a que es una base de datos de propósito general, por lo no está especialmente diseñada para el dominio médico. Esto implica, en primer lugar, que no tiene una gran cobertura para el dominio médico. En segundo lugar, incluye muchos sinónimos no relevantes para el dominio médico. Por ejemplo, la palabra 'limb' en WordNet tiene bastantes sinónimos (p.e. 'branch', 'extremity', 'member', 'border', 'portion' y 'arc'), pero muchos de ellos no están relacionados con el ámbito médico, y por tanto podrían generar ruido en nuestra herramienta de búsqueda en SNOMED-CT.

A diferencia de otros enfoques que usan WordNet como una fuente de sinónimos, nuestro enfoque descubre automáticamente sinónimos en SNOMED-CT. De esta forma, nuestro enfoque es capaz de encontrar sinónimos relevantes para el contexto médico, en vez de sinónimos de propósito general.

4.5.3. Análisis de errores en el mapping automático

Hemos analizado los mappings erróneos comparando los resultados de HMAS con los correspondientes mappings expertos. Hemos detectado que en muchas de las situaciones en las que HMAS no acierta en el mapping, obtiene un concepto muy cercano (p.e. un concepto padre o hijo en SNOMED-CT) al concepto correcto.

Hemos clasificado los desaciertos de HMAS en 3 tipos[10]. La tabla 4.9 muestra un ejemplo de cada tipo de desacierto. Las columnas de la tabla muestran: (1) el término de búsqueda, (2) el concepto obtenido por HMAS, (3) el concepto asignado por los expertos (4) y un comentario sobre el error.

El primer tipo de desaciertos incluye mappings de HMAS, que bajo nuestro punto de vista, podrían ser considerados correctos a pesar de que los expertos no los eligieran como el mejor mapping. Por ejemplo, el término de búsqueda 'Extirpation of major salivary gland' fue mapeado por los expertos al concepto 'Excision of salivary gland' y por HMAS al concepto 'Excision of major salivary gland', siendo este último un concepto hijo del concepto elegido por los expertos. En este ejemplo, consideramos que el mapping HMAS podría ser incluso más apropiado que el propio mapping experto. Este tipo de casos deberían ser revisados de nuevo por los codificadores expertos ya que podrían mejorar la calidad de los mappings seleccionados inicialmente por los expertos.

En el segundo tipo de desaciertos, los mappings expertos han resultado ser más específicos que los mappings de HMAS. En estos casos los expertos asumieron que los procedimientos de biopsias se realizan en una parte del cuerpo dañada o anómala. HMAS no fue capaz de asumir esto y por tanto obtuvo mappings más generales. Por ejemplo, el término de búsqueda 'Incisional biopsy of breast' fue mapeado por los expertos al concepto 'Incisional biopsy of breast mass', mientras que HMAS lo mapeó a 'Incisional biopsy of breast'.

El tercer tipo de desaciertos contiene mappings de HMAS que son menos precisos que los mappings expertos. Este tipo de desaciertos si que pueden considerarse errores de HMAS. Muchos de estos errores se deben a que el concepto correcto y el término de búsqueda tienen nombres muy diferentes, dificultando la búsqueda a las herramientas automáticas.

Tabla 4.9: Ejemplos de errores en el mapping automático realizado por HMAS

Término buscado (en inglés y español)	**Mapping HMAS** (Concepto SNOMED en inglés y español)	**Mapping Experto** (Concepto SNOMED en inglés y español)	**Comentario**
Extirpation of major salivary gland Extirpación de glándula salival mayor	Excision of major salivary gland (234937001) Resección de glándula salival mayor	Excision of salivary gland (71735005) Resección de glándula salival	Mapping HMAS no coincide con el experto pero podría ser considerado correcto
Incisional biopsy of breast Biopsia por incisión de mama	Incisional biopsy of breast (237378001) Biopsia por incisión de mama	Incisional biopsy of breast mass (28768007) Biopsia por incisión de tumor mamario	Mapping HMAS no coincide con el experto pero podría ser considerado correcto
Biopsy of globe Biopsia de globo ocular	Biopsy of lesion of globe (231559005) Biopsia de lesión del globo ocular	Biopsy of eye proper (446938009) Biopsia del ojo propiamente dicho	Mapping HMAS no coincide con el experto pero podría ser considerado correcto

[10] El siguiente enlace `http://goo.gl/Yz4i13` incluye varios ejemplos de los 3 tipos de desaciertos detectados

4.5.4. Aplicaciones y alcance de la herramienta de mapping

En este capítulo presentamos una herramienta para descubrir mappings texto-a-concepto en la terminología SNOMED-CT. La herramienta implementa dos perfiles diferentes: uno es completamente automático (HMAS) y el otro es semi-automático (HMSS). Los dos perfiles comparten esencialmente el mismo objetivo, esto es, facilitar el proceso de búsqueda de mapping entre término clínicos y conceptos de SNOMED-CT. HMAS ha sido ajustado para devolver un solo concepto relevante, por lo que debería ser más adecuado para un proceso de mapping en el que no haya expertos disponibles o para una tarea de mapping que requiera un resultado rápido (p.e. para aplicaciones en tiempo real). Por otra parte, HMSS ha sido adaptado para sugerir 5 conceptos candidatos para cada término de búsqueda, por lo que está más orientado a tareas de mapping en las que hay expertos disponibles para revisar y seleccionar el mapping más preciso.

HMAS alcanzó una moderada cobertura (51 %) y una elevada precisión (88 %) en los experimentos; por su parte HMSS logró una importante cobertura (70 %) sugiriendo 5 conceptos candidatos por término. Cabe destacar que algunos estudios han evaluado el rendimiento de médicos en tareas de codificación con SNOMED-CT [68, 21] y han descubierto que este es imperfecto, y que además la tasa de acuerdo en la codificación entre varios médicos es baja. El estudio de S.Young et al. descubrió que médicos con un nivel intermedio de conocimiento en tareas de codificación sólo mapeó correctamente el 62 % de los términos a conceptos SNOMED-CT [68].

Aunque la evaluación se realizó con un glosario de términos en español de procedimientos patológicos, la herramienta fue desarrollada para la búsqueda de cualquier tipo de término en SNOMED-CT. Además, la herramienta puede ser utilizada para mapear términos en español o en inglés ya que durante la fase de preprocesado se han normalizado e indexado todas las descripciones de la edición Internacional y Española de SNOMED-CT. Todas las técnicas implementadas están adaptadas para trabajar sin problemas con términos de los dos idiomas. La herramienta de mapping también puede ser configurada para buscar conceptos en ciertas jerarquías o en todo SNOMED-CT.

¿Cómo de aplicable sería la herramienta de mapping a otras terminologías?

Consideramos que una metodología que combina diferentes técnicas de mapping, tal como hace la herramienta descrita en este capítulo, es un punto clave para obtener buenos resultados en procesos automáticos de mapping. Por tanto, en este sentido, la herramienta sí que podría

ser adecuada para otras terminologías. Cabe destacar que la técnica estructural propuesta está orientada a terminologías clínicas que contenga abundante axiomática, especialmente relaciones lógicas entre conceptos. Respecto a la técnica de expansión de consultas implementada, esta podría ser completamente reusada para generar términos alternativos en otros contextos. Hemos comprobado que las descripciones de SNOMED-CT forman un gran corpus donde buscar e inferir sinónimos. La terminología SNOMED-CT proporciona una buena cobertura de sinónimos en diferentes áreas clínicas, ya que está considerada como la terminología clínica más completa en la actualidad. Consecuentemente, creemos que los términos alternativos generados por la técnica de expansión de consultas (con el corpus de SNOMED-CT) puede claramente contribuir en procesos de búsqueda en otras terminologías.

4.5.5. Aportaciones

Las aportaciones de este capítulo se pueden sintetizar en los siguientes puntos:

- Se propone una herramienta de búsqueda que combina diferentes estrategias de mapping para facilitar la búsqueda de conceptos relevantes en SNOMED-CT. La herramienta propuesta:
  - Incluye varias técnicas innovadoras que hasta nuestro conocimiento no habían sido usadas en los procesos de búsqueda en SNOMED-CT:
    - Una técnica de expansión de consultas avanzada capaz de inferir sinónimos relevantes del ámbito médico analizando el inmenso corpus de descripciones de SNOMED-CT, y con ello, expandir las consultas con términos alternativos con significados similares.
    - Una técnica que explota las relaciones semánticas de SNOMED-CT para mejorar los procesos de búsqueda en SNOMED-CT.
    - Técnicas de desambiguación para reducir el número de conceptos candidatos obtenidos como resultado de procesos de búsqueda en SNOMED-CT. Estas técnicas tratan de seleccionar un único concepto final relevante entre el conjunto de conceptos candidatos.
  - Mejora en más de un 25 % en términos de recall a dos destacados navegadores de SNOMED-CT en tareas de búsqueda automática de conceptos.

4.5.6. Trabajo futuro

Las técnicas estructurales usadas en la herramienta de mapping tienen la desventaja de que para sacar su máximo provecho, estas deben ser configuradas para usar un conjunto de relaciones lógicas relevantes al tipo de término o entidad buscada. Por ejemplo, en la aplicación concreta que presentamos en la evaluación de este capítulo, se seleccionaron relaciones lógicas que pudiesen ser útiles para la búsqueda de conceptos biopsia en SNOMED-CT. Desgraciadamente, la selección de relaciones lógicas requiere un considerable conocimiento de la estructura interna de SNOMED-CT. Recientemente, algunos trabajos de investigación han propuesto enfoques automáticos para analizar el contenido axiomático de SNOMED-CT [87, 27]. Por ejemplo, el *framework RIO* [87], desarrollado por E. Mikroyannidi et al., permite detectar patrones y estructuras repetitivas en los axiomas de entidades similares de SNOMED-CT. Nuestra herramienta de mapping podría tomar ventaja de un *framework* como *RIO* ya que proporciona una rápida abstracción de un conjunto de conceptos similares (p.e. conceptos biopsia), y esto a su vez podría facilitar e incluso mejorar la selección del contenido axiomático (i.e. relaciones lógicas) usado por las técnicas estructurales de nuestra herramienta de mapping.

Nuestra herramienta de mapping no explota las estrategias de post-coordinación para conceptos SNOMED-CT. Nuestras técnicas se han centrado en la búsqueda de un único concepto capaz de representar plenamente el término de búsqueda. En el futuro sería interesante explorar estrategias para detectar términos que no pueden ser representados por un único concepto, a la vez que se desarrollan técnicas para mapear estos términos con expresiones post-coordinadas formadas por varios conceptos.

En el futuro sería interesante analizar el recall de la configuración semi-automática de la herramienta de mapping en función del número de conceptos candidatos sugeridos, tal y como se ha hecho en la evaluación de una herramienta similar [136].

Además, nuestra herramienta de mapping está optimizada para trabajar con términos de entrada cortos, tal como 'biopsia por escisión en la piel'. En el futuro, la herramienta podría ser adaptada para dar soporte a la anotación automática de textos de entrada largos (tal como resúmenes de los artículos o informes de pacientes) con conceptos SNOMED-CT.

Nuestro sistema de expansión de consultas genera un conjunto de términos alternativos, pero no aplica ninguna estrategia para estimar la calidad ni la fiabilidad de cada uno de ellos. En el futuro, se podrían usar dos parámetros para estimar la calidad de los términos: el número

de palabras sustituidas en el término alternativo y una estimación de la calidad de cada una de las palabras sustituidas.

También sería interesante integrar la herramienta de mapping desarrollada en algún navegador SNOMED-CT.

4.6. Resumen

Las terminologías clínicas han surgido para capturar la información clínica de forma consistente y estandarizada, favoreciendo con ello la gestión eficiente de dicha información y la interoperabilidad semántica entre diferentes instituciones sanitarias. Sin embargo, en la actualidad la información clínica de la HCE está mayoritariamente definida en lenguaje natural y apenas está enlazada con conceptos terminológicos, por lo que actualmente no se está aprovechando todo el potencial de las terminologías. Automatizar la búsqueda en las terminologías y los procesos de enlazado entre la información clínica y los conceptos relevantes de las terminologías impulsaría claramente el uso y la integración de las terminologías en la HCE.

En este capítulo se ha presentado una herramienta cuyo objetivo principal es facilitar la búsqueda automatizada de conceptos relevantes en SNOMED-CT. El capítulo ha incluido también una evaluación en la que se compara el rendimiento de la herramienta con las funcionalidades de búsqueda de dos de los navegadores más populares de SNOMED-CT.

La herramienta propuesta combina técnicas léxicas clásicas con técnicas de búsqueda innovadoras en el ámbito de SNOMED-CT. Estas últimas aplican un sistema de expansión de consultas para reformular y expandir los términos de búsqueda, y además explotan las relaciones de SNOMED-CT para obtener información adicional sobre el significado de los conceptos terminológicos. Además, la herramienta ha sido diseñada con dos configuraciones ligeramente diferentes que hemos llamado: HMAS y HMSS. La primera configuración está pensada para mapear automáticamente términos clínicos a conceptos de SNOMED-CT, sin requerir ninguna o muy poca intervención de usuarios expertos. Mientras que HMSS tiene el objetivo de ayudar a los expertos durante el proceso de mapping, recomendando varios conceptos SNOMED-CT para cada término clínico buscado.

En la evaluación de la herramienta se usaron 300 términos de un glosario de procedimientos patológicos. En general, los resultados obtenidos por la herramienta fueron muy positivos. La configuración semi-automática de nuestra herramienta (HMSS) alcanzó el 71.0 % de cobertura sugiriendo 5 conceptos por cada término buscado. La versión automática de nuestra

herramienta (HMAS) logró mapear los términos con un recall del 51.4 % y una precisión del 88.0 %, frente al 32 % y 84.7 % logrado por el buscador de la NLM, y al 40.0 % y 90.1 % del navegador ITServer, dos de las herramientas más populares y avanzadas en la actualidad para la búsqueda en SNOMED-CT. Los experimentos mostraron que la técnica de expansión de consultas implementada mejora la cobertura de nuestra herramienta en más de un 20 %.

En el capítulo se ha demostrado que es posible mejorar el rendimiento de las herramientas de búsqueda existentes en SNOMED-CT (en nuestros experimentos en más de un 25 % en términos de recall) mediante una aproximación que combine diferentes técnicas de mapping. La aportación de la técnica de expansión de consultas incluida en nuestra aproximación ha resultado especialmente importante para mejorar el rendimiento de las búsquedas. Dicha técnica es capaz de inferir sinónimos relevantes del ámbito médico analizando el inmenso corpus de descripciones de SNOMED-CT, y con ello, expandir las consultas con términos con significados similares maximizando las posibilidades de éxito de las técnicas léxicas.

Capítulo 5

Enlazado automático de arquetipos OpenEHR con SNOMED-CT

Actualmente, uno de los objetivos más importantes de la informática médica es el de conseguir la interoperabilidad entre los diferentes sistemas de información clínica. En este sentido, la informática médica ha estado trabajando al menos en dos campos, las terminologías clínicas y los modelos de datos clínicos [121].

Por una parte, las terminologías han sido propuestas como estándar para codificar la información clínica de los pacientes. Estas permiten codificar la información sin ambigüedades, reduciendo así los errores usuales de los registros médicos tradicionales. SNOMED-CT es, en la actualidad, la terminología más completa para codificar todos los aspectos de los registros electrónicos [100].

Por otra parte, varias organizaciones [102, 98] han estado trabajando en el diseño de modelos clínicos de conocimiento orientados a capturar de forma ordenada y sistemática la información de pacientes en escenarios clínicos determinados. Los arquetipos openEHR son en la actualidad los modelos clínicos recomendados a nivel europeo para componer los futuros sistemas de HCE [121].

La integración entre las terminologías clínicas y los modelos de datos clínicos es un paso clave en el camino hacia la interoperabilidad semántica. En este sentido, los arquetipos openEHR han sido diseñados para permitir un cierto grado de integración con terminologías [127]. Los arquetipos incluyen un conjunto de ítems de información relacionados a una cierta situación clínica (por ejemplo, la medición de la presión sanguínea de un paciente). Estos

ítems son definidos mediante términos en lenguaje natural (por ejemplo, 'presión sanguínea' o 'presión diastólica alta'). Opcionalmente, los ítems pueden ser enlazados con conceptos de terminologías clínicas.

Varias organizaciones internacionales, NHS [104], openEHR [98] y NEHTA [103], llevan participando desde hace más de 5 años en el desarrollo de repositorios de arquetipos [92, 90]. Grupos de expertos de diferentes dominios colaboran para modelar y actualizar los arquetipos. En la actualidad, los repositorios cuentan con un número importante de arquetipos, algunos incluso en estado de publicación, que quiere decir que están listos para usar y no requieren más cambios. Sin embargo, los enlaces (bindings) o mappings entre los ítems de los arquetipos y los conceptos terminológicos todavía son muy infrecuentes.

Idealmente, los mappings deberían ser realizados por profesionales médicos con experiencia en SNOMED-CT, asistidos por navegadores y herramientas básicas de búsqueda en SNOMED-CT. Sin embargo, la tarea del mapping es en la actualidad muy lenta y costosa por varias razones: el gran tamaño y granularidad de SNOMED-CT, la continua evolución de SNOMED-CT, la falta de familiaridad de los profesionales médicos con SNOMED-CT y la ausencia de herramientas de búsqueda avanzadas.

Este capítulo presenta un método[1] automático para enlazar términos clínicos de arquetipos con conceptos de SNOMED-CT. Este enfoque aplica una combinación de técnicas léxicas y contextuales para producir y validar las equiparaciones. Las técnicas léxicas buscan equiparaciones comparando los términos con las descripciones de los conceptos de SNOMED-CT, mientras que las técnicas contextuales buscan similitudes estructurales entre la jerarquía del arquetipo y la red de relaciones de SNOMED-CT.

El objetivo del método es reducir y facilitar en la medida de lo posible la participación de los profesionales médicos en la tarea de mapping, de forma que puedan concentrarse en dos tareas: validar los mappings automáticos y crear los mappings en situaciones ambiguas o que requieran experiencia o conocimiento avanzado.

En este capítulo, comenzaremos introduciendo los arquetipos clínicos, especialmente los de tipo 'Observation'. Seguidamente, profundizaremos en las distintas técnicas usadas para buscar conceptos en SNOMED-CT apropiados para los arquetipos. Posteriormente, explicaremos el procedimiento de evaluación seguido. Finalmente, expondremos los resultados y una discusión de los mismos.

[1] Versiones menos avanzadas del método automático han sido publicadas en publicaciones de congreso [85, 4, 3] y en un artículo de investigación [86]

5.1. Materiales

El método propuesto en el presente capítulo ha sido especialmente diseñado para mapear términos clínicos de arquetipos openEHR (ver sección 2.2.2) con conceptos de la terminología SNOMED-CT (ver sección 2.3). A continuación, se profundiza en algunas cuestiones de los arquetipos openEHR que han condicionado el diseño del método de enlazado presentado en el capítulo.

5.1.1. Arquetipos openEHR

Los arquetipos openEHR son modelos de datos formales e interoperables definidos por expertos clínicos. Los arquetipos incluyen fragmentos o ítems de información requeridos para la captura sistemática de datos en una situación o instancia clínica determinada, tal como la observación de la respiración de un paciente.

Como ya se detalló en el capítulo 2, los arquetipos openEHR son definidos con un lenguaje formal llamado ADL (ver sección 2.2.2). La figura 2.3 muestra parte del fichero ADL correspondiente al arquetipo 'respiration'. Los arquetipos constan de 3 secciones destacadas: cabecera (header), cuerpo (body) y ontología (ontology). La cabecera contiene metadata sobre el arquetipo. El cuerpo incluye la estructura y las restricciones de la información clínica del arquetipo. La sección ontología está formada por dos subsecciones: 'term definitions' y 'term bindings'. La primera incluye el nombre y la descripción de cada uno los fragmentos de información presentes en el cuerpo del arquetipo. La sección 'term bindings' permite enlazar los fragmentos de los arquetipos con conceptos de una o más terminologías clínicas. Por ejemplo, en la figura 2.3 el fragmento del arquetipo con identificador local at0016 tiene asociado el nombre 'Depth' y la descripción 'Depth of respiration' en la sección 'term definitions', y a su vez está enlazado al concepto 'Depth of respiration (observable entity) de la terminología SNOMED-CT en la sección 'term bindings'.

Los repositorios de arquetipos incluyen varios tipos de arquetipos (Observation, Action, Evaluation, Instruction). El estudio se ha centrado en arquetipos 'Observation' ya que estos son actualmente los más maduros y los más revisados por la comunidad de modeladores y expertos de openEHR. Estos arquetipos son usados para capturar hallazgos de exploraciones, resultados de pruebas y síntomas de un paciente, y en general cualquier condición del paciente sin interpretar. Por tanto, los arquetipos 'Observation' contienen esencialmente fragmentos de información sobre observables u observaciones clínicas. Los fragmentos sobre observables

definen qué es lo que se está observando del paciente (p.e. la profundidad de la respiración o el ritmo de la respiración), mientras que los fragmentos sobre observaciones modelan posibles valores o hallazgos del paciente (p.e. respiración profunda o respiración irregular).

Análisis de los arquetipos 'Observation'

Los ficheros ADL incluyen detalles sobre el arquetipo que no son relevantes para la tarea del mapping. Por ello, hemos usado un parseador para eliminar la información no relevante y extraer una jerarquía con los fragmentos del arquetipo con significado clínico, mapeables a conceptos SNOMED-CT. La figura 5.1 muestra parte de la jerarquía de fragmentos asociada al arquetipo 'respiration'. Para cada fragmento, se muestra el término asignado en el campo nombre en la sección 'term definitions' y el tipo de fragmento.

De acuerdo a las especificaciones de openEHR, los arquetipos 'Observation' incluyen principalmente 3 tipos de fragmentos o términos con contenido clínico: raíz, elemento y valor. El fragmento raíz representa la situación clínica o el **observable general** de un arquetipo. Por ejemplo, en el arquetipo 'respiration' el fragmento raíz es 'respiration observable' (ver figura 5.1). Los fragmentos elemento definen los aspectos u **observables** que deben recopilarse en el arquetipo (p.e. 'depth' y 'rhythm' son fragmentos elemento en el arquetipo 'respiration'). Los fragmentos valor especifican **hallazgos u observaciones predefinidas** relacionadas a los fragmentos elemento (p.e. 'shallow', 'normal' y 'deep' son fragmentos valor asociados al fragmento 'Depth' en el arquetipo respiration).

Hemos analizado manualmente un grupo amplio de arquetipos 'Observation', examinando como organizan la información clínica. Además, hemos hecho una búsqueda rápida manual en SNOMED-CT para equiparar los fragmentos de los arquetipos con conceptos SNOMED-CT. Por ejemplo, la figura 5.2 muestra los mappings manuales que hemos asignado al arquetipo 'respiration'. Fruto del análisis se han extraído varias reflexiones e ideas útiles para el desarrollo de los métodos automáticos de mapping.

Se han detectado ciertos paralelismos entre los tipos de fragmentos de arquetipos 'Observation' y jerarquías de SNOMED-CT (ver figura 5.3). Se ha verificado que los términos raíz y elemento modelan observables de pacientes, y tienden a equiparar muy frecuentemente a conceptos SNOMED-CT de la jerarquía Observable Entity. También, se ha comprobado que los términos valor modelan observaciones o hallazgos clínicos, siendo muy propensos a equiparar a conceptos SNOMED-CT de la jerarquía Clinical Finding.

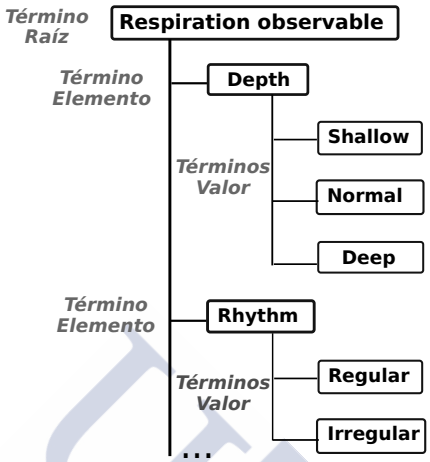

Figura 5.1: Jerarquía de términos asociada al arquetipo respiration

También, se ha descubierto que los conceptos SNOMED-CT asignados a los fragmentos de un mismo arquetipo, tienden a estar semánticamente o lógicamente relacionados, y por tanto están próximos en la terminología SNOMED-CT; incluso en ocasiones están relacionados directamente a través de las relaciones jerárquicas o lógicas de SNOMED-CT. Específicamente, hemos observado dos situaciones en los arquetipos analizados (ver esquema en la figura 5.4). En primer lugar, los conceptos que equiparan a los términos elemento tienden a estar relacionados jerárquicamente al concepto que equipara con la raíz del mismo arquetipo. Por ejemplo, en el arquetipo respiration, los dos conceptos 'depth of respiration' y 'rhythm of respiration' equiparados a términos elemento están relacionados jerárquicamente (a través de la relación IS A) con el concepto 'respiration observable' equiparado al término raíz del arquetipo (ver figura 5.2). En segundo lugar, los conceptos que equiparan a los términos valor están en bastantes ocasiones relacionados lógicamente al concepto que equipara con el término elemento. Por ejemplo, en el arquetipo respiration, tres conceptos 'shallow breathing', 'normal depth of breathing' y 'Deep breathing' equiparados a términos valor están relacionados lógicamente (a través de la relación interprets) con el concepto 'depth of respiration' equiparado al término elemento 'Depth' (ver figura 5.2).

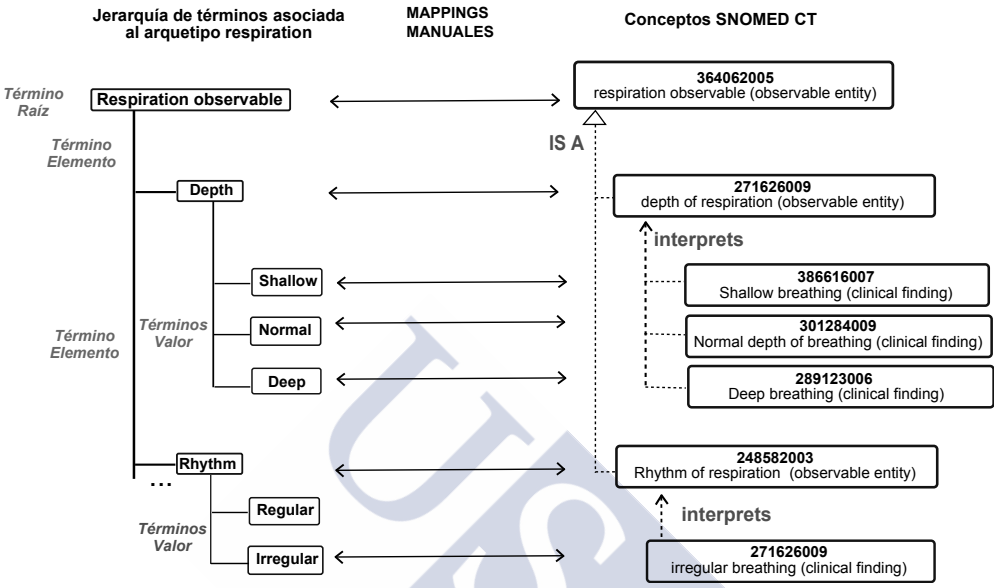

Figura 5.2: Mappings expertos asignados al arquetipo 'respiration'.

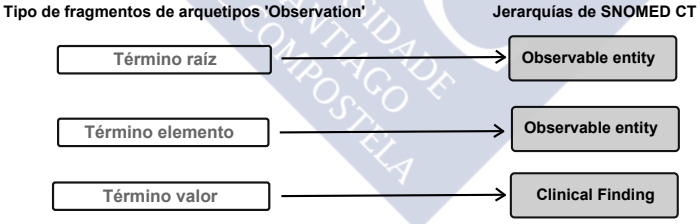

Figura 5.3: Paralelismos encontrados entre tipos de fragmentos de arquetipos 'Observation' y jerarquías de SNOMED-CT

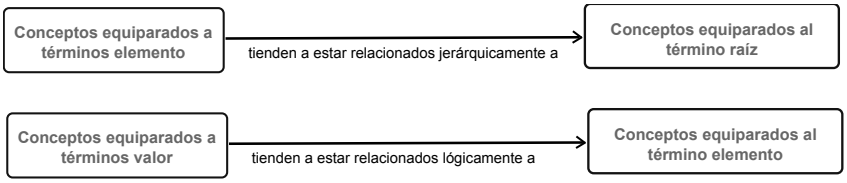

Figura 5.4: Relaciones frecuentes entre los conceptos equiparados a fragmentos de información de arquetipos 'Observation'

5.2. Métodos

En este apartado se describen los métodos utilizados para buscar y validar mappings entre términos de un arquetipo y conceptos de SNOMED-CT. En primer lugar, tiene lugar una etapa de búsqueda de conceptos candidatos para los términos del arquetipo (ver fig. 5.5). Posteriormente, se aplica una etapa de desambiguación para seleccionar los mappings finales entre los candidatos. El resultado de este proceso es un fichero XML con los mappings finales obtenidos por los métodos automáticos. El fichero incluye meta-información del arquetipo (nombre, versión y tipo del arquetipo) e información de los mappings. Para cada mapping se almacena el término del arquetipo, su identificador local, los identificadores de los conceptos SNOMED-CT y la descripción textual de los mismos.

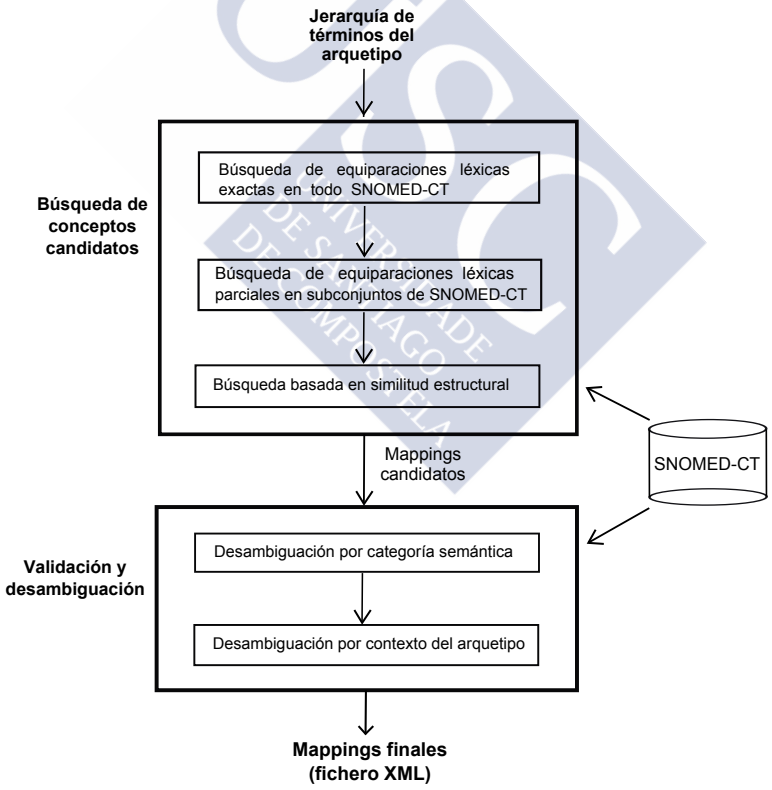

Figura 5.5: Etapas destacadas en el mapping entre arquetipos openEHR y conceptos SNOMED-CT

5.2.1. Búsqueda de conceptos candidatos

En esta etapa se usan varias técnicas introducidas en el capítulo 3 para obtener un conjunto de conceptos candidatos equivalentes a los términos de un arquetipo 'Observation'. Esta etapa está a su vez compuesta de tres etapas que se ejecutan secuencialmente: búsqueda de equiparaciones léxicas exactas en todo SNOMED-CT, búsqueda de equiparaciones léxicas parciales en subconjuntos relevantes de SNOMED-CT y búsqueda basada en similitud estructural. Los mappings de las tres etapas se agregan para obtener un conjunto de conceptos candidatos. A continuación, se exponen más detalles sobre las tres etapas:

Etapa 1: Búsqueda de equiparaciones léxicas exactas en todo SNOMED-CT

En esta etapa inicial se aplican varias técnicas y estrategias expuestas en el capítulo 3 para buscar equiparaciones léxicas exactas entre los términos del arquetipo y todas las descripciones de los conceptos de la terminología SNOMED-CT:

- Normalizaciones léxicas (ver sección 3.3.1)
- Técnicas léxicas exactas (ver sección 3.3.2).
- Métodos de búsqueda 'Exact Match' y 'Normalized Word' de la API de UMLS (ver sección 3.5).
- Combinación de términos basada en la jerarquía interna de los arquetipos (ver sección 3.6.2)

La figura 5.6 muestra los mappings obtenidos para el arquetipo 'respiration' tras la aplicación de estas técnicas. En esta etapa se obtienen conceptos léxicamente muy similares a los términos del arquetipo. La combinación de los términos con otros de nivel superior en la jerarquía del arquetipo, permite en algunos casos obtener mappings muy precisos. Por ejemplo, el término 'Depth' es combinado con el término superior en el arquetipo 'Respiration observable', gracias a lo cual es mapeado al concepto 'Depth of respiration'. Sin embargo, en otros casos, como por ejemplo en los términos 'normal' y 'regular', los mappings son demasiado genéricos y no hacen referencia al contexto del arquetipo, esto es, a la observación de la respiración.

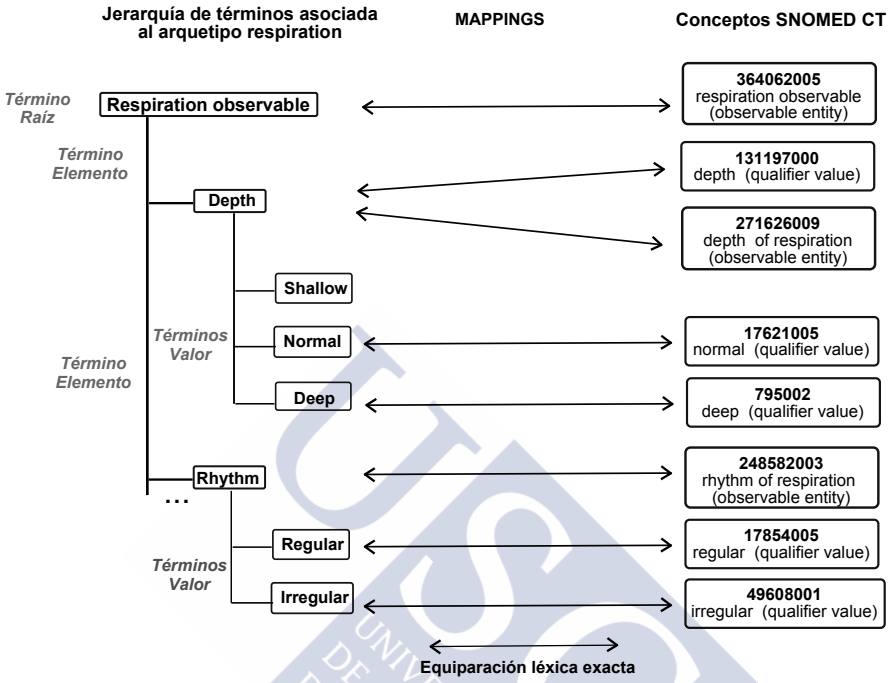

Figura 5.6: Búsqueda de equiparaciones exactas en todo SNOMED-CT para el arquetipo 'respiration'

Etapa 2: Búsqueda de equiparaciones léxicas parciales en subconjuntos relevantes de SNOMED-CT

Esta etapa está inspirada en el **Principio de proximidad** definido en el capítulo 3, el cual decía: *si se logra mapear de forma precisa un término del arquetipo a un concepto 'c' de SNOMED-CT, entonces es bastante probable que los términos vecinos del arquetipo mapeen a conceptos semánticamente relacionados a 'c' en SNOMED-CT*. Además, la presente etapa está también motivada por los hallazgos detectados en el estudio de los arquetipos 'Observation' (sección 5.1.1); concretamente en las relaciones frecuentes que se han detectado en el estudio entre mappings de arquetipos 'Observation' (ver figura 5.4).

En esta etapa se extraen pequeños subconjuntos de SNOMED-CT formados por conceptos semánticamente relacionados a los conceptos SNOMED-CT obtenidos en la etapa 1. Los conceptos de estos subconjuntos son muy relevantes para el arquetipo y tienen muchas más posibilidades de ser mapeados (a algún término del arquetipo) que un concepto aleatorio de

SNOMED-CT. Esto permite flexibilizar la búsqueda léxica sobre estos subconjuntos para detectar nuevos mappings. A continuación, se exponen dos casos concretos en los que se ha aplicado esta idea:

Búsqueda de equiparaciones parciales en el contexto jerárquico del término raíz del arquetipo

Si el término raíz tiene asignado conceptos candidatos de la etapa 1, entonces se extrae el contexto jerárquico de cada uno de ellos. El subconjunto o contexto jerárquico de un concepto está formado por todos sus conceptos descendientes en la jerarquía de SNOMED-CT. A continuación, se aplican técnicas léxicas parciales (expuestas en la sección 3.3.3) para buscar equiparaciones entre los términos elemento del arquetipo y los conceptos pertenecientes a los contextos jerárquicos extraídos. La figura 5.7 muestra un ejemplo de aplicación para el arquetipo respiration. En la figura se puede ver el contexto jerárquico de 'respiration observable (observable entity)', único concepto candidato asociado al término raíz. La figura también muestra las equiparaciones detectadas en esta etapa entre los términos 'Depth' y 'Rhythm' y los conceptos del contexto jerárquico 'depth of respiration' y 'rhythm of respiration'.

Búsqueda de equiparaciones léxicas parciales en el contexto lógico de los términos elemento del arquetipo

En este caso, recorremos los términos elemento del arquetipo. Si estos términos contienen algún concepto candidato, entonces su contexto lógico es extraído. El contexto lógico de un concepto está formado por los conceptos con los que está relacionado lógicamente (es decir, a través de relaciones SNOMED-CT no jerárquicas). A continuación, se aplican técnicas léxicas parciales para buscar equiparaciones entre los términos valor y los conceptos pertenecientes a los contextos lógicos extraídos. Al final de esta etapa, los términos valor podrán incorporar nuevos conceptos candidatos.

La figura 5.8 muestra un ejemplo de aplicación de esta etapa para el arquetipo 'respiration'. En la figura se puede ver el contexto lógico del concepto 'depth of respiration' equiparado al término elemento 'Depth'. La figura muestra también las nuevas equiparaciones detectadas en esta etapa entre los términos valor que cuelgan del término 'Depth', esto es 'Shallow','Normal' y 'Deep', y los conceptos del contexto lógico 'shallow breathing', 'normal depth of breathing' y 'deep breathing', respectivamente. A pesar de que estos conceptos

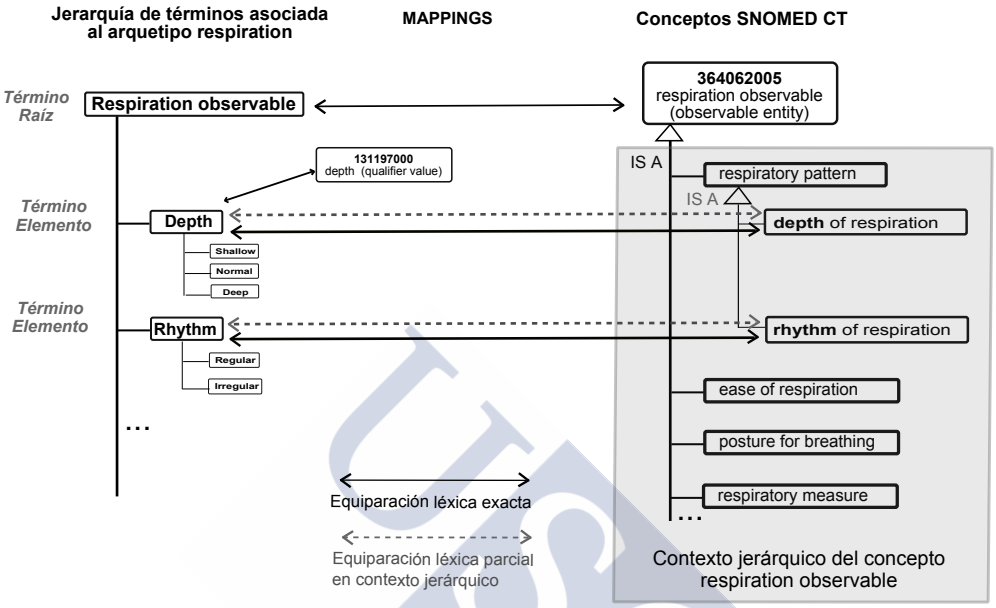

Figura 5.7: Búsqueda de equiparaciones parciales en el contexto jerárquico de la raíz del arquetipo

no son léxicamente iguales a los correspondientes términos, estos son semánticamente muy apropiados para un arquetipo sobre observaciones de la respiración.

Etapa 3: Generación de mappings basada en similitud estructural

Esta etapa tiene el objetivo de buscar nuevos mappings para términos elemento teniendo en cuenta los mappings creados en etapas previas en los términos valor.

En primer lugar, se seleccionan los términos valor asociados a un término elemento. Si estos términos contienen un concepto asignado en una etapa previa, entonces se usan las relaciones lógicas de SNOMED-CT asociadas a dicho concepto para extraer nuevos conceptos. A continuación, se aplica una regla sencilla: si varios términos valor extraen el mismo concepto, entonces este es un nuevo mapping candidato para el término elemento correspondiente. La figura 5.9 muestra un ejemplo de aplicación de esta etapa para el arquetipo 'conscious state'. En este ejemplo, inicialmente disponemos de varios conceptos SNOMED-CT 'unconscious', 'drowsy' y 'disorientated' que equiparan exactamente a los términos valor 'unconscious', 'drowsy' y 'disorientated'. Como se puede ver en la figura, dos de estos conceptos,

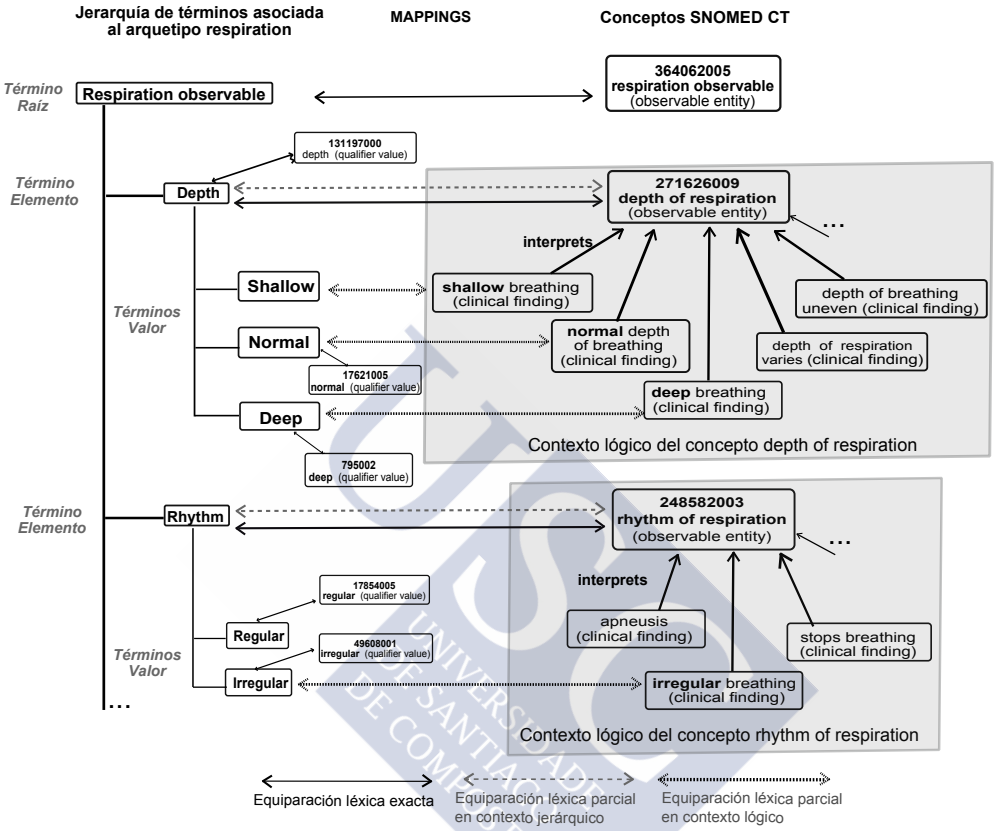

Figura 5.8: Búsqueda de equiparaciones parciales en el contexto lógico de los términos elemento del arquetipo

'unconscious' y 'drowsy', están relacionados lógicamente al concepto 'level of consciousness', mientras que el otro concepto no participa en ninguna relación lógica. En base a la regla que se ha definido, el concepto 'level of consciousness' se añade como nuevo mapping para el término elemento 'Statements regarding conscious state'. En esta etapa no se aplica ninguna técnica de equiparación léxica, el descubrimiento de mappings depende de técnicas de similitud estructural. Por tanto, esta etapa es especialmente adecuada para detectar equiparaciones para términos ambiguos o léxicamente muy diferentes a los conceptos equivalentes en SNOMED-CT.

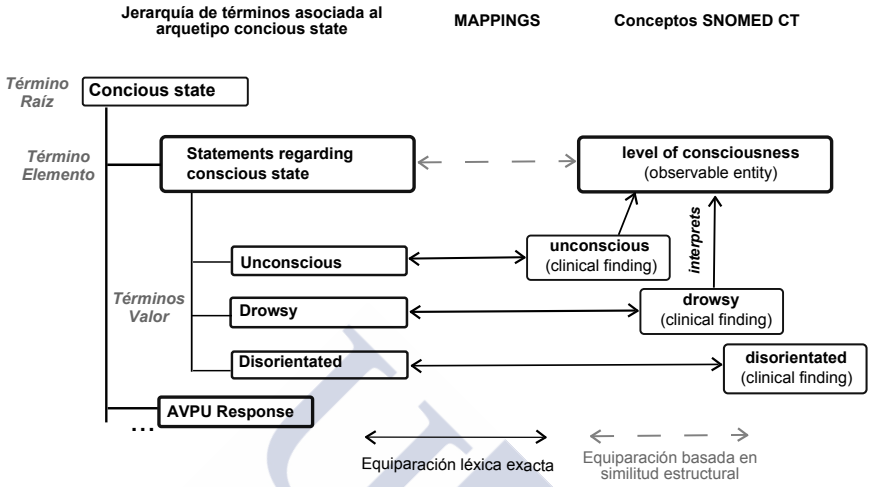

Figura 5.9: Generación de mappings estructurales para términos elemento del arquetipo

5.2.2. Desambiguación de conceptos candidatos en arquetipos Observation

La etapa anterior de búsqueda puede generar más de un concepto candidato para un mismo término. En este apartado se presentan técnicas de desambiguación orientadas a seleccionar automáticamente el mejor mapping entre un grupo de candidatos.

5.2.3. Desambiguación por categoría semántica

Se implementaron dos estrategias para desambiguar por categoría semántica. La primera estrategia asume que hay jerarquías o categorías de SNOMED-CT que son más específicas que otras y que estas son más adecuadas para mapear términos de arquetipos (ver más detalles de esta estrategia en la sección 3.7.1). En la figura 5.8 se puede ver que el término 'deep' en el arquetipo 'respirations' tiene asociado dos conceptos obtenidos con técnicas distintas: 'deep (qualifier value)' y 'deep breathing (clinical finding)'. El concepto 'deep (qualifier value)' pertenece al grupo de categorías genérico, mientras que 'deep breathing (clinical finding)' al grupo 'específico'. Usando esta estrategia el concepto 'deep (qualifier value)' sería descartado, quedando como único concepto 'deep breathing (clinical finding)' para mapear el término 'deep'.

La segunda estrategia es específica para arquetipos Observation y se basa en el hallazgo presentado en la sección 5.1.1 *Análisis de los arquetipos Observation* que expone que "los términos raíz y elemento frecuentemente equiparan a conceptos de SNOMED-CT pertenecientes a la jerarquía Observable Entity. Mientras que los términos valor frecuentemente equiparan a conceptos de la jerarquía Clinical Finding". Esta estrategia establece que si hay varios conceptos candidatos para los términos raíz o elemento, se seleccionan aquellos que pertenecen a la jerarquía Observable entity, mientras que si hay varios candidatos asociados a los términos valor, se escogen los de la jerarquía Clinical Finding. Supongamos que la etapa de búsqueda obtenga dos conceptos candidatos para el término elemento 'Depth' del arquetipo 'respirations': 'depth of respiration (observable entity)' y 'finding of depth of respiration (finding)'. Aplicando esta estrategia se analiza las jerarquías de los dos candidatos y se escoge el concepto 'depth of respiration (observable entity)' como mapping final ya que es un término de tipo elemento que frecuentemente modela preguntas u observaciones del paciente.

5.2.4. Desambiguación por contexto en un arquetipo Observation

Esta estrategia selecciona los conceptos candidatos con un mayor número de relaciones semánticas con el resto de conceptos candidatos del arquetipo. Más detalles de esta estrategia pueden ser consultados en la sección 3.7.2 *Desambiguación por contexto del arquetipo*. Esta estrategia de desambiguación no requirió ninguna adaptación específica para arquetipos de tipo Observation.

Vamos a ver un ejemplo de uso de esta estrategia para desambiguar los conceptos candidatos del término 'Depth' del arquetipo 'respirations'. Tal como podemos ver en la figura 5.8 el término 'Depth' tiene asociado dos conceptos candidatos 'depth (qualifier value)' y 'depth of respiration (observable entity)'. Mientras que el concepto 'depth (qualifier value)' no tiene conexiones semánticas, el concepto 'depth of respiration (observable entity)' tiene 4 conexiones semánticas. Concretamente, está relacionado jerárquicamente con el concepto 'respiration observable (observable entity)' (ver figura 5.7), y lógicamente (a través de la relación 'interprets') con los conceptos 'shallow breathing (clinical finding)', 'normal depth of breathing (clinical finding)' y 'deep breathing (clinical finding)' (ver figura 5.8). Debido a que el concepto 'depth of respiration (observable entity)' tiene un mayor número de conexiones semánticas que el concepto 'depth (qualifier value)', el primero es seleccionado como concepto final para mapear el término 'depth'.

5.3. Evaluación del mapping

5.3.1. Conjunto de datos

La evaluación incluyó 25 arquetipos Observation tomados del repositorio del proyecto NHS Connecting for Health. Muchos de los arquetipos de este repositorio están todavía en una fase de revisión y mejora. Para esta evaluación se procuró seleccionar los 25 arquetipos Observation más maduros y con menos posibilidades de cambio. Los 25 arquetipos constan de 921 términos, de los cuales sólo el 4 % de ellos (37 términos) tienen mappings a conceptos de SNOMED-CT creados por los diseñadores de los arquetipos. La lista de arquetipos seleccionados puede ser consultada en el apéndice B.

5.3.2. Procedimiento de evaluación

Debido a que sólo el 4 % de los términos contienen mappings a conceptos SNOMED-CT, hemos colaborado con dos profesionales médicos (Raimundo Lozano y María Jesús Sobrido) para crear manualmente los mappings para todos los términos de los 25 arquetipos. La sección 5.3.3 muestra más detalles sobre el procedimiento seguido para la creación de los mappings. Fruto de esta colaboración, el 52 % (477 de 921) de los términos fueron mapeados a conceptos de SNOMED-CT. Los 444 términos restantes no fueron mapeados a SNOMED-CT ya que no expresan conceptos clínicos (por ejemplo: 'notes on measurement' y 'Description'), o bien, expresan contenido clínico demasiado específico (por ejemplo 'Feed Times and Volumes within Last 24 Hours' y 'Total Feed Volume per kg within Last 24 Hours'). Estos 477 términos con sus respectivos mappings expertos nos sirvieron de referencia para evaluar los mappings obtenidos de forma automática por los métodos implementados. De los 477 términos, 25 son términos raíz, 188 términos elemento y 264 términos valor. La evaluación consistió en ejecutar los métodos automáticos para estos 477 términos, obtener los mappings automáticos y comparar estos con los mappings creados por los expertos (ver figura 5.10). Dos medidas clásicas de recuperación de información (precisión y recall) fueron calculadas para estimar la calidad de los mappings automáticos. En este contexto, la precisión estima la proporción entre el número de mappings automáticos correctos y el número de mappings encontrados por el método. Mientras que el recall estima la proporción entre el número de mappings automáticos correctos y el número de mappings creados por los expertos (477).

$$\text{Precisión} = \frac{\text{\# mappings automáticos correctos}}{\text{\# mappings automáticos totales}}$$

$$\text{Recall} = \frac{\text{\# mappings automáticos correctos}}{\text{\# mappings expertos totales}}$$

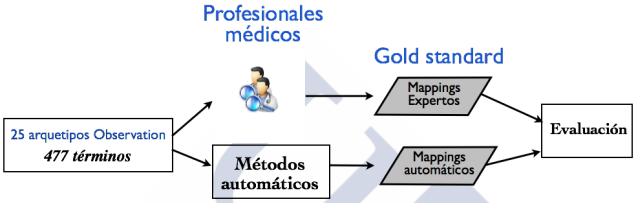

Figura 5.10: Procedimiento de evaluación de los métodos de mapping

5.3.3. Creación de los mappings expertos

Como ya se ha comentado dos profesionales médicos (Raimundo Lozano y María Jesús Sobrido) han creado mappings expertos para los 25 arquetipos seleccionados. Raimundo Lozano es licenciado en medicina e informática. Ejerció durante años como profesional clínico. En la actualidad trabaja como informático médico en proyectos relacionados con arquetipos openEHR y SNOMED-CT. María Jesús Sobrido es neuróloga e investigadora. En la actualidad, ejerce la práctica clínica. Además, tiene experiencia con ontologías y terminologías clínicas. La creación de los mappings expertos tuvo lugar en tres rondas o iteraciones. En la primera, dos personas con perfil informático con experiencia con SNOMED-CT seleccionaron una media de unos 5 conceptos candidatos para mapear los 477 términos. En la segunda ronda, Raimundo seleccionó el concepto más preciso (entre los 5 candidatos) para cada término. Si consideraba que ningún concepto era apropiado, también tenía la posibilidad de buscar directamente en la terminología SNOMED-CT para elegir un nuevo concepto. En la tercera ronda, María Jesús revisó los mappings creados por Raimundo añadiendo nuevos mappings y eliminando los mappings que consideraba oportuno. La salida de esta revisión dio lugar a los mappings expertos finales usados en la evaluación[2].

[2]Hemos creado una web `http://www.usc.es/keam/TermArchetypes/input.html` en la cual están disponibles los 25 arquetipos junto con los mappings a SNOMED-CT creados por los expertos

Durante este proceso se usaron dos herramientas: el navegador de SNOMED-CT CliniClue y una interfaz gráfica creada como parte de esta tesis para facilitar la creación y revisión de mappings entre arquetipos y SNOMED-CT. La sección 5.5.2 incluye más detalles sobre esta interfaz gráfica.

5.4. Resultados

La tabla 5.1 muestra los resultados individuales al aplicar las distintas etapas del método de búsqueda para cada tipo de término. Destacar, que la etapa 1 (esencialmente léxica) se puede aplicar a los 3 tipos de términos, mientras que las etapas 2 y 3 mezclan técnicas contextuales y léxicas y sólo son aplicables a ciertos tipos de términos. En general, se puede ver que la etapa 1 logra un recall bastante alto (100 %, 55.3 % y 70.5 %) para los 3 tipos de términos, con una precisión buena, cercana al 85 %. Las etapas 2 y 3 obtienen un recall más moderado (siempre por debajo del 25 %) y una precisión realmente buena, superando el 95 % en la etapa 2 para términos valor.

Tabla 5.1: Resultados de cada una de las etapas del método de búsqueda aplicadas a los 3 tipos de términos de arquetipos Observation.

Tipo de término	# de términos	Técnica utilizada	# mappings automáticos correctos	# mappings automáticos totales	Precisión %	Recall %
Raíz	25	Equiparación exacta en todo SNOMED (Et. 1)	25	25	86.2	100
Elemento	188	Equiparación exacta en todo SNOMED (Et. 1)	104	124	83.9	55.3
		Equiparación parcial en contexto jerárquico (Et. 2)	51	57	89.5	21.1
		Equiparación por similitud estructural (Etapa 3)	15	17	88.2	8.0
Valor	264	Equiparación exacta en todo SNOMED (Et. 1)	188	223	83.4	70.5
		Equiparación parcial en en contexto lógico (Et. 2)	66	68	97.1	25.0

La tabla 5.2 muestra los resultados considerando todos los términos. La primera fila muestra los resultados de la etapa léxica (etapa 1). La segunda fila muestra los resultados agregados de todas las etapas léxico-contextuales (etapas 2 y 3). La tercera fila muestra los resultados agregando la etapa léxica más las etapas léxicos-contextuales. La última fila muestra los resultados agregados con una etapa extra de desambiguación de los mappings candidatos.

La etapa léxica obtiene 376 mappings, de los cuales 315 (el 83.8 %) son correctos. Los 315 mappings correctos representan el 66 % de los mappings totales creados por los expertos. La etapas léxicas-contextuales recuperan 142 mappings, de los cuales 132 (el 93.0 %) son correctos. Estos 132 mappings correctos cubren el 27.7 % de los mappings expertos. La agregación de las etapas léxicas y léxicas-contextuales genera 430 mappings, de los cuales 366 (85.1 %) son correctos. Destacar que cuando se agregan estas dos etapas hay un importante grado de solape en los mappings, es decir, ambas etapas obtienen bastantes mappings comunes. Sin embargo, los resultados muestran claramente que la agregación es muy beneficiosa obteniendo mejores resultados que las etapas por separado. Por ejemplo, si comparamos los resultados de la agregación con los resultados de la etapa léxica, vemos que la agregación obtiene 51 mappings correctos adicionales. El recall de la agregación alcanza el 76.7 %, lo que supone un incremento de un 16.2 % respecto al recall de la técnica léxica (66.0 %). Además, la precisión también se incrementa en un 1.5 % con la agregación. La aplicación de las técnicas de desambiguación (sobre los mappings candidatos de la agregación) es también muy positiva: logra eliminar 52 mappings candidatos, de los cuales 48 son mappings incorrectos y sólo 4 son mappings correctos. La precisión final sube hasta el 95.8 %, lo que supone un incremento de un 12.6 % respecto a la precisión obtenida por la agregación (85.1 %). El recall desciende muy ligeramente hasta 75.9 %, suponiendo un descenso del 1 % respecto al recall obtenido por la agregación (76.7 %). El apéndice B incluye más información sobre la salida generada por el método automático de mapping.

Tabla 5.2: Resultados del método automático considerando todos los términos

# de términos	Técnica utilizada	# mappings automáticos correctos	# mappings automáticos totales	Precisión %	Recall %
477	**Léxica**: Equiparación exacta en todo SNOMED (Etapa 1)	315	376	83.8	66.0
	Léxica-contextual: Equip. usando contextos (Etapas 2 y 3)	132	142	93.0	27.7
	Agregación: Léxica (Et. 1) + Léxica-contextual(Et. 2 y 3)	366	430	85.1	76.7
	Todas las técnicas: Agregación + Desambiguación	362	378	95.8	75.9

Las técnicas léxico-contextuales y las técnicas de desambiguación basadas en contexto, además de ser útiles para incrementar el recall y la precisión, permiten validar o ratificar algo más del 30 % de los mapping producidos por técnicas léxicas. Esto significa que del total de mappings obtenidos por técnicas léxicas, cerca del 30 % son también obtenidos por las técnicas léxico-contextuales o bien pasan el filtro de la desambiguación basada en contexto, y por tanto claramente estos mappings tienen un mayor grado de confianza que los mappings obtenidos exclusivamente por técnicas léxicas, y por ello se consideran validados.

5.5. Discusión

5.5.1. Análisis de los enlaces expertos

Además de la evaluación del rendimiento de los métodos de mapping, se ha analizado varias características del conjunto de mappings creados por los expertos. Las conclusiones de este análisis pueden contribuir a desarrollar futuras técnicas de mapping para arquetipos. En concreto, se ha analizado:

- Las categorías semánticas de SNOMED-CT más frecuentes en los mappings de los distintos tipos de términos.
- La existencia o no de relaciones semánticas entre los conceptos mapeados de un arquetipo. Es decir, si estos están conectados a través de relaciones jerárquicas o lógicas definidas en SNOMED-CT.

Categoría semántica de los mappings expertos

En primer lugar, se ha analizado qué categorías semánticas de SNOMED-CT son más frecuentes en los mappings expertos de los términos raíz y elemento. La figura 5.11 muestra el número total de conceptos de cada categoría semántica de SNOMED-CT asociados a este tipo de términos. Se puede apreciar que los expertos han mapeado mayoritariamente conceptos de la categoría *Observable Entity*. Otras categorías con cierta presencia son: *Attribute*, *Clinical finding* y *Body structure*.

También, se han analizado los mappings expertos asociados a términos valor. La figura 5.12 muestra el número total de conceptos de cada categoría semántica de SNOMED-CT asociados a los términos valor. Se puede apreciar que los expertos han mapeado cerca de 200 conceptos de la categoría Clinical Finding y 100 de la categoría Qualifier Value. El resto de las categoría son bastante infrecuentes.

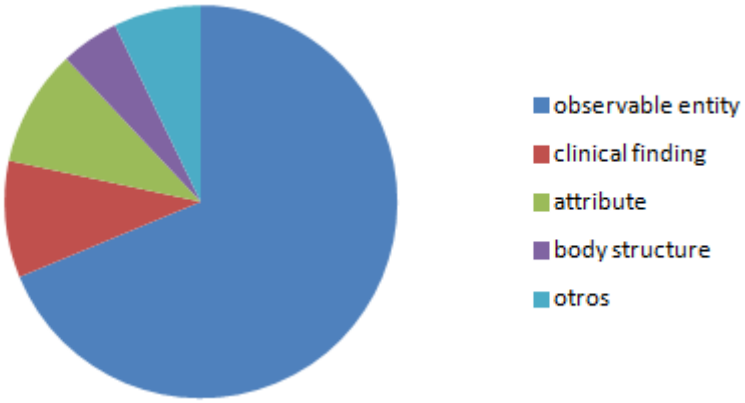

Figura 5.11: Frecuencia de las categorías semánticas de los mappings expertos asociados a términos elemento y raíz

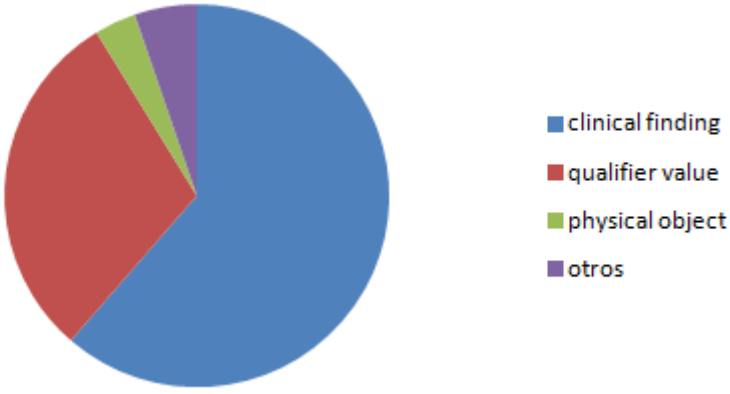

Figura 5.12: Frecuencia de las categorías semánticas de los mappings expertos asociados a términos valor

Los expertos manifestaron que su primera opción al mapear términos valor fue buscar un concepto perteneciente a la categoría Clinical Finding. Cuando no encontraron un concepto adecuado dentro de esta categoría, tendieron a buscar conceptos en la categoría Qualifier Value. Los expertos afirmaron que en general los conceptos Qualifier Value no eran suficientemente específicos. Sin embargo, se vieron obligados a seleccionar conceptos Qualifier value cuando SNOMED-CT no les proporcionaba un concepto más específico.

La figura 5.13 muestra parte de la jerarquía de términos del arquetipo 'fetal movements'. Este arquetipo modela las observaciones sobre los movimientos del feto durante el embarazo. La jerarquía del arquetipo incluye el término elemento 'intensity' y sus tres términos valor: 'weak', 'normal', y 'strong'. Los expertos asignaron dos conceptos de la categoría Clinical Finding: 'weak fetal movements' y 'strong fetal movements' para mapear los términos 'weak' y 'strong', respectivamente. Estos conceptos reflejan de forma precisa y no ambigua el significado de los términos 'weak' y 'strong', teniendo en cuenta que estos están dentro de un arquetipo que modela los movimientos del feto durante el embarazo. Sin embargo, los expertos no encontraron un concepto de la categoría Clinical Finding para el término 'normal', por lo que tuvieron que mapear el término al concepto de la jerarquía Qualifier Value: 'normal'. Este concepto no refleja de forma demasiado precisa el significado del término 'normal' en el contexto de este arquetipo. Pero en este caso, SNOMED-CT no incluye ningún concepto más específico.

Relaciones semánticas entre términos de arquetipos Observation

El diseño de los métodos automáticos de mapping se apoyó en buena medida en la hipótesis de que la información clínica de un arquetipo en bastantes ocasiones está semánticamente relacionada entre sí (ver sección 5.1.1). Dicho de otro modo, los conceptos médicos referenciados dentro de un mismo arquetipo frecuentemente están conectados a través de relaciones

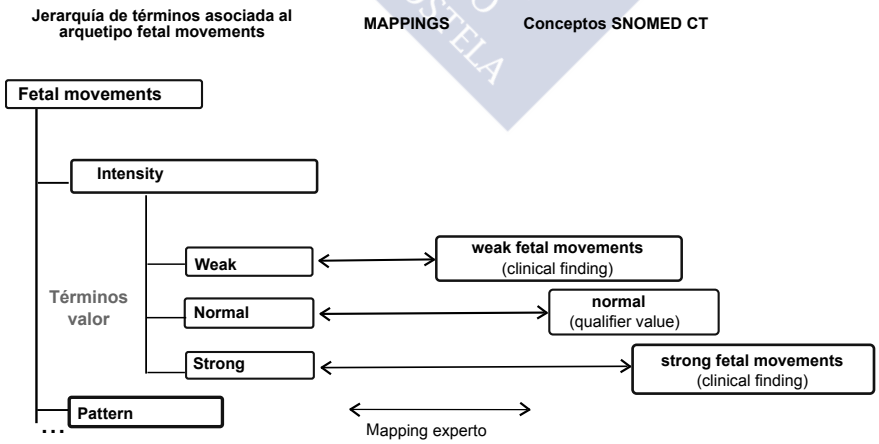

Figura 5.13: Parte de la jerarquía de términos del arquetipo fetal movements, junto con varios mappings creados por expertos.

jerárquicas o lógicas de SNOMED-CT. Con el objetivo de contrastar esta hipótesis, se realizó un estudio para evaluar cuales son las relaciones de SNOMED-CT que conectan los conceptos médicos referenciados dentro de un mismo arquetipo, y con qué frecuencia aparecen.

Diseño del estudio

Para el estudio[3], se usaron los conceptos seleccionados por los expertos para los 25 arquetipos 'Observation'. Los conceptos se dividieron en 3 grupos:

- conceptos raíz: son los conceptos asociados a los términos raíz de un arquetipo.
- conceptos elemento: son los conceptos asociados a los términos elemento.
- conceptos valor: son los conceptos asociados a los términos valor.

Se implementó un procedimiento automático para analizar varios aspectos en cada uno de los 25 arquetipos:

- La frecuencia con la que los conceptos elemento y valor de un arquetipo están conectados a través de relaciones jerárquicas *IS A* (de SNOMED-CT) con el concepto raíz del mismo arquetipo.
- La frecuencia con la que los conceptos elemento y valor de un arquetipo están conectados a través de relaciones lógicas con el concepto raíz del mismo arquetipo.
- La frecuencia con la que los conceptos valor están conectados a través de relaciones jerárquicas *IS A* con el concepto elemento correspondiente.
- La frecuencia con la que los conceptos valor están conectados a través de relaciones lógicas con el concepto elemento correspondiente.

Resultados del estudio

La tabla 5.3 muestra los resultados del estudio. De los 188 conceptos elemento evaluados, 58 (30.8 %) están relacionados jerárquicamente al correspondiente concepto raíz, mientras que tan sólo 2 (1.1 %) están lógicamente conectados al concepto raíz. De los 264 conceptos valor analizados, 2 (0.8 %) están conectados vía relaciones jerárquicas y 29 (11 %) vía relaciones lógicas con el concepto raíz. De los 193 conceptos valor analizados, 12 (6.2 %) están relacionados jerárquicamente al correspondiente concepto elemento, mientras que 78 (40.4 %) están conectados a través de relaciones lógicas con el concepto elemento. De los 78 conceptos

[3] Más detalles sobre el estudio pueden ser encontrados en el artículo [6] publicado en las actas del congreso ProHealth12

valor relacionados con conceptos elemento, 74 lo están a través de la relación interprets. Esta relación de SNOMED-CT se usa para enlazar un concepto de la jerarquía *Clinical Finding* con el concepto de la jerarquía *Observable Entity* o *Procedure* que está siendo evaluado. Por ejemplo, el concepto 'shallow breathing (Clinical Finding)' está relacionado por interprets al concepto 'depth of respiration (Observable Entity)'

Tabla 5.3: Número de relaciones semánticas existentes entre los conceptos mapeados a los arquetipos

Conceptos analizados	Número de análisis	Número de relaciones jerárquicas detectadas	Número de relaciones lógicas detectadas
Concepto Raíz-Concepto Elemento	188	58 (30.8%)	2 (1.1%)
Concepto Raíz-Concepto Valor	264	2 (0.8%)	29 (11.0%)
Concepto Elemento-Concepto Valor	193	12 (6.2%)	78 (40.4%)

Conclusiones del estudio

Aunque tanto los arquetipos clínicos como SNOMED-CT tienen el objetivo de estructurar la información clínica, ambos tienen diferentes puntos de vista respecto a cómo estructurar dicha información. Los arquetipos agrupan la información clínica que debe ser registrada al mismo tiempo durante una instancia o situación clínica. Por ejemplo, el arquetipo 'Presión sanguínea' modela observaciones sobre la presión sanguínea diastólica y sistólica, así como la posición del paciente en el momento de la medición. En contraste, la estructura interna de SNOMED-CT enlaza conceptos clínicos por relaciones semánticas (lógicas o jerárquicas). Por ejemplo, SNOMED-CT establece que el concepto 'presión sanguínea' es padre de los conceptos 'presión diastólica' y 'presión diastólica'.

Se ha detectado que, en muchas situaciones, la información clínica de los arquetipos también está relacionada semánticamente, tal como lo está SNOMED-CT. Especialmente hemos detectado: (1) un alto número de relaciones de hiperonimia-hiponimia existentes entre los términos raíz y elemento, y (2) un importante número de relaciones lógicas o de atributo entre términos elemento y valor de arquetipos 'Observation'.

5.5.2. Interfaz gráfica de apoyo al mapping

Los métodos automatizados son un aspecto muy importante para mejorar el proceso de mapping. Sin embargo, todavía es necesario que los expertos puedan revisar y actualizar fácilmente los mappings de un arquetipo. Por ello, como parte de esta tesis, hemos creado una interfaz gráfica pensada para facilitar a los expertos la creación y la revisión de los mappings en arquetipos. La interfaz gráfica incluye las siguientes funcionalidades:

- Carga de mappings existentes de un arquetipo. La interfaz es capaz de procesar los ficheros XML generados por el método automático y representar de forma amigable los mappings de un arquetipo. La interfaz, en una misma ventana, muestra:
  - A la izquierda la jerarquía completa de los términos del arquetipo.
  - En el centro el listado de mappings seleccionados para cada término.
  - A la derecha más detalles sobre los conceptos candidatos del término del arquetipo seleccionado. Concretamente la interfaz muestra los conceptos ascendientes y descendientes de los conceptos candidatos y las distintas descripciones textuales asociados al concepto.
- Permite añadir/eliminar fácilmente mappings en un arquetipo
- Permite seleccionar a los expertos el mejor mapping cuando el método propone más de un candidato.

La figura 5.14 muestra una captura de la interfaz gráfica tras procesar el XML de mappings del arquetipo 'blood pressure'. Esta muestra los términos del arquetipo blood pressure (a la izquierda), los mappings automáticos asociados a cada término (en el centro), e información detallada sobre uno de los mappings (a la derecha).

5.5.3. Dificultades en el mapping de arquetipos a SNOMED-CT

En este apartado se presentan algunos aspectos que dificultaron el mapping (tanto el manual como el automático) de arquetipos a SNOMED-CT. Adicionalmente, se expone, cuando procede, estrategias para minimizar el impacto de las dificultades y recomendaciones o propuestas de mejora para evitar que se produzcan en el futuro.

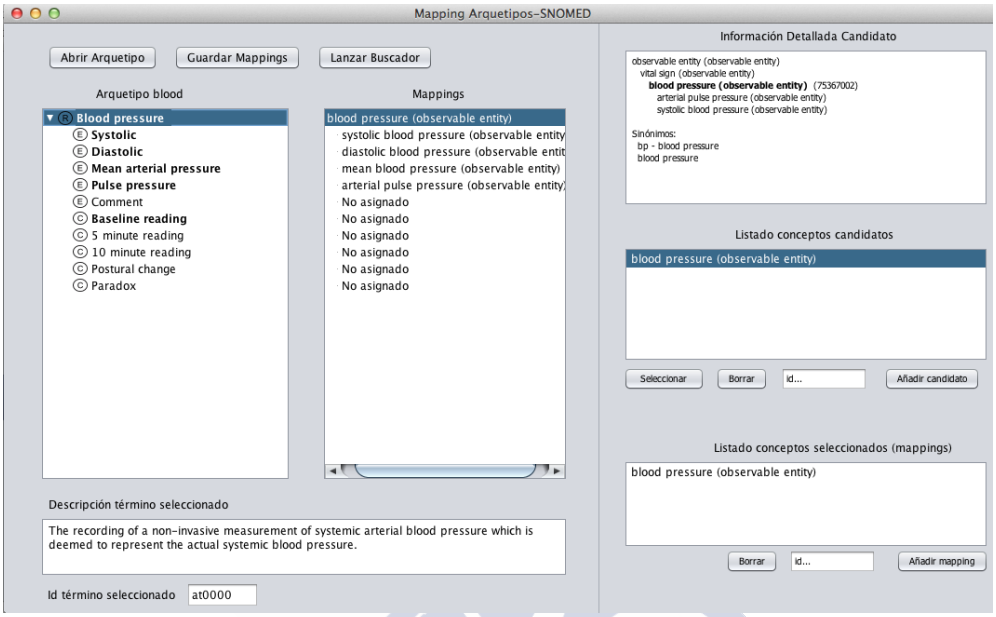

Figura 5.14: Ventana de la interfaz gráfica donde se representan los mappings automáticos asociados al arquetipo blood pressure

Existencia de conceptos muy similares en SNOMED-CT

Los mapeadores expertos han detectado conceptos redundantes o muy similares en SNOMED-CT. Los conceptos redundantes tienen descripciones textuales muy similares entre sí, y no suelen tener definidas relaciones de definición con otros conceptos. Por tanto, es complicado distinguir los matices en el significado entre los conceptos. Por ejemplo, los conceptos 'feeding (observable entity)' y 'feeding observable (observable entity)' son realmente similares. Ambos términos parecen expresar el mismo significado. La palabra 'observable' en el segundo concepto no añade información ya que ambos conceptos pertenecen a la jerarquía 'observable entity', por lo que ya se da por hecho que están definiendo una característica observable de un paciente. Además, los conceptos no tienen asociados ni sinónimos ni otras relaciones de definición en SNOMED-CT. Otro ejemplo detectado de conceptos muy similares es: 'strength uterine contractions (observable entity)' y 'uterine contraction intensity (observable entity)'.

– Propuesta de mejora: La terminología SNOMED-CT no debería incluir conceptos demasiado similares que puedan causar confusión. Por tanto, sería deseable, o bien modificar los conceptos para introducir distintos matices, o bien, eliminar uno de los dos conceptos.

Información de contexto en los arquetipos

Los arquetipos han sido modelados con una estrategia de composición en la que los términos se definen en una jerarquía o árbol de dependencias. Como resultado, los términos de los arquetipos no definen completamente su significado sin referenciar a los términos vecinos del arquetipo. Por tanto, para entender el significado completo de un término debemos considerar el contexto en el que se encuentra dentro del arquetipo. Por ejemplo, en el arquetipo 'respiration' (ver fig. 5.8), el término 'shallow' de forma aislada es demasiado general y no está completamente definido. Por ello, se debe considerar el contexto de 'shallow' en el arquetipo, concretamente los términos de nivel superior 'depth' y 'respiration observable'.

– Propuesta de mejora: De cara a facilitar el mapping, es deseable que cada término sea definido de forma más completa o detallada. Por ejemplo, las descripciones 'shallow breathing' o 'depth of respiration: shallow' son dos alternativas más detalladas para definir el término 'shallow'.

– Estrategia para abordar el mapping: Claramente, es necesario que tanto los expertos como los métodos automáticos de mapping tengan en cuenta la información contextual de los arquetipos para mejorar la precisión de los resultados. En este sentido, esta tesis doctoral propone estrategias para aprovechar la información contextual de los arquetipos (ver secciones 3.6.2 y 5.2.4)

Definición ambigua de términos

Algunos términos de los arquetipos han sido definidos de forma ambigua o confusa, dando lugar a varias interpretaciones. Esto derivó en que los mapeadores expertos encontrasen varios conceptos SNOMED-CT candidatos de diferentes jerarquías y tuvieran muchas dudas para seleccionar el concepto más adecuado. El problema principal con estos términos es que los mapeadores no entendieron con que intención o propósito fueron definidos. La tabla 5.4 muestra algunos ejemplos de los términos ambiguos detectados por los expertos. En la tabla se muestra de izquierda a derecha: el término considerado ambiguo y el arquetipo al

que pertenece, la descripción asignada por el modelador durante la creación del arquetipo, el término "padre"del que depende el término ambiguo, y finalmente los posibles mappings en SNOMED-CT.

Uno de los términos ambiguos detectado por los mapeadores fue biberón (*bootle*) del arquetipo *feeding*. Los mapeadores encontraron tres posibles mappings en SNOMED-CT para este término: 'infant bottle fed (clinical finding)', 'bottle feeding of patient (procedure)', 'bottle, device (physical object)'. Los expertos no saben si el modelador del arquetipo al definir este término estaba haciendo referencia al procedimiento seguido para alimentar a un recién nacido ('bottle feeding of patient (procedure)'), al hallazgo clínico asociado ('infant bottle fed (clinical finding)', o al dispositivo u objeto utilizado para ello ('bottle, device (physical object)').

- Propuesta de mejora: Los modeladores de arquetipos deberían tratar de definir los términos de la forma más clara posible, proporcionando descripciones detalladas.
- Estrategia para abordar el mapping: Esta tesis doctoral propone dos tipos de estrategias para desambiguar automáticamente: por categoría semántica y por contexto del arquetipo (ver sección 5.2.2). En cualquier caso, sería deseable que los mapeadores expertos revisasen los mappings asociados a términos ambiguos.

Tabla 5.4: Términos ambiguos detectados en los arquetipos

Término ambiguo (arquetipo)	Descripción	Término padre en la jerarquía del arquetipo	Posibles mappings en SNOMED-CT
bootle (feeding)	Feeding from the bottle	Type of feeding	infant bottle fed (clinical finding) bottle feeding of patient (procedure) bottle, device (physical object)
faeces (faeces)	For recording faecal output	No tiene	feces (substance) stool observable (observable entity)
saliva (hydration)	Describing state of saliva	Findings	saliva (substance) salivary apparatus observable (observable entity)
tripod (mobility)	tripod	use of mobility aid	tripod (physical object) tripod/quadrupod: walking (clinical finding)

Incompleta sinonimia de SNOMED-CT

La sinonimia en SNOMED-CT es todavía imperfecta e incompleta. Se ha detectado conceptos que sólo incluyen la descripción obligatoria sin definir ninguna descripción sinónima. Además, hemos detectado algunos grupos de conceptos en los que los sinónimos se usan de manera incoherente. La figura 5.15 muestra un ejemplo de uso incoherente de sinónimos. En la figura aparece un conjunto de conceptos SNOMED-CT, algunos de ellos relacionados jerárquicamente. El concepto con identificador 39477002 incluye tres descripciones sinónimas formadas por una palabra: 'feces', 'faeces' y 'stool' para hacer referencia a la substancia heces. Sin embargo, otros conceptos que modelan observaciones sobre esta substancia, no incluyen las tres palabras sinónimas, sino que usan arbitrariamente sólo una o dos de las tres palabras sinónimas. Por ejemplo, el concepto 'stool observable (observable entity)' con identificador 364689004 usa sólo la palabra 'stool'. Sin embargo, uno de sus conceptos descendientes con identificador 167621006 y descripciones sinónimas 'feces quantity' y 'faeces quantity' usa las palabras 'feces' y 'faeces'. Si la sinonimia de SNOMED-CT fuese perfecta, estos conceptos deberían incluir las 3 palabras sinónimas. Así, el concepto con identificador 364689004 definiría las descripciones 'stool observable', 'feces observable' y 'faeces observable', mientras que el concepto con identificador 364689004 incorporaría las descripciones 'feces quantity','faeces quantity' y 'stool quantity'. La falta de sinonimia en SNOMED-CT claramente repercute en el proceso de mapping, especialmente en el mapping automático asistido por técnicas léxicas.

- Propuesta de mejora: La terminología SNOMED-CT debería incluir el máximo nivel de sinonimia posible. Somos conscientes que el extenso tamaño de SNOMED-CT hace casi inviable definir de forma manual una sinonimia completa para todos los conceptos. Creemos que este proceso podría ser asistido por técnicas automáticas. La estrategia que planteamos en la sección 3.4 (para expandir términos de búsqueda con sinónimos) podría ser usada también para proponer nuevas descripciones sinónimas partiendo de las descripciones de conceptos ya existentes.
- Estrategia para minimizar la falta de sinonimia: En la sección 3.4 proponemos una estrategia para expandir los términos buscados con sinónimos. Esta estrategia minimiza el problema de la falta de sinonimia, aumentando las posibilidades de éxito de las técnicas léxicas para encontrar mappings en SNOMED-CT.

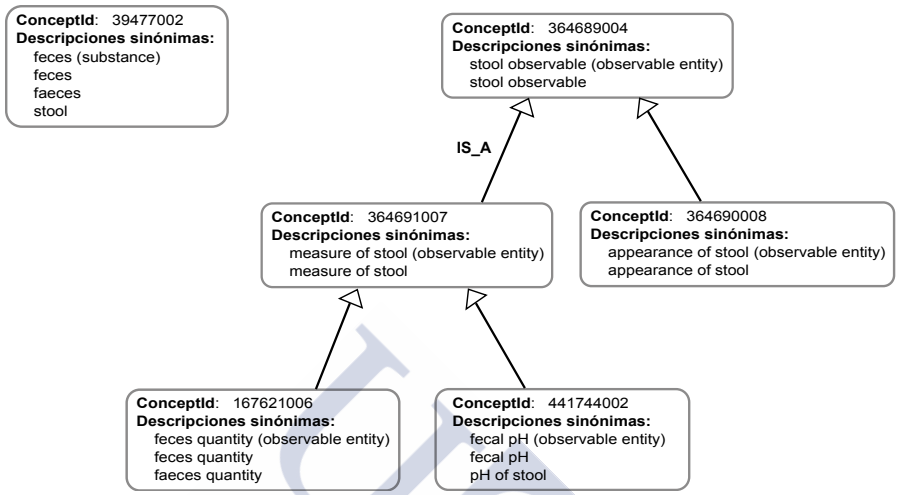

Figura 5.15: Conceptos de SNOMED-CT con poca cobertura de sinónimos

Diferentes grados en hallazgos clínicos

Se han detectado diferencias, entre los arquetipos y SNOMED-CT, en cuanto al nombre y al número de grados/categorías definidos para medir algunas observaciones clínicas. Así por ejemplo, el arquetipo feeding incluye 4 términos para medir el nivel de habilidad de una persona para succionar: 'unable to suck', 'some ability', 'reduced sucking' y 'normal'. En cambio, SNOMED-CT definió 5 conceptos de la jerarquía 'Clinical finding' para medir la misma observación: 'does suck', 'difficulty sucking', 'unable to suck', 'does not suck' y 'able to suck' (ver figura 5.16). Los términos 'unable to suck' y 'normal' se equiparan claramente a los conceptos 'unable to suck' y 'able to suck'. Sin embargo, no está claro cómo equiparar los términos 'some ability' y 'reduced sucking', ya que SNOMED-CT incluye sólo un concepto ('difficulty sucking') para equiparlos. Quizá, en este ejemplo el arquetipo está desgranando más los hallazgos relacionados al nivel de habilidad de una persona para succionar que SNOMED-CT. Esta diferente granularidad dificulta el proceso de mapping entre los términos de los arquetipos y SNOMED-CT.

– Propuesta de mejora: Creemos que hasta ahora la información clínica de los arquetipos fue modelada sin aprovechar y/o reutilizar prácticamente SNOMED-CT (terminología clínica más completa en número de conceptos y relaciones semánticas). Una mayor

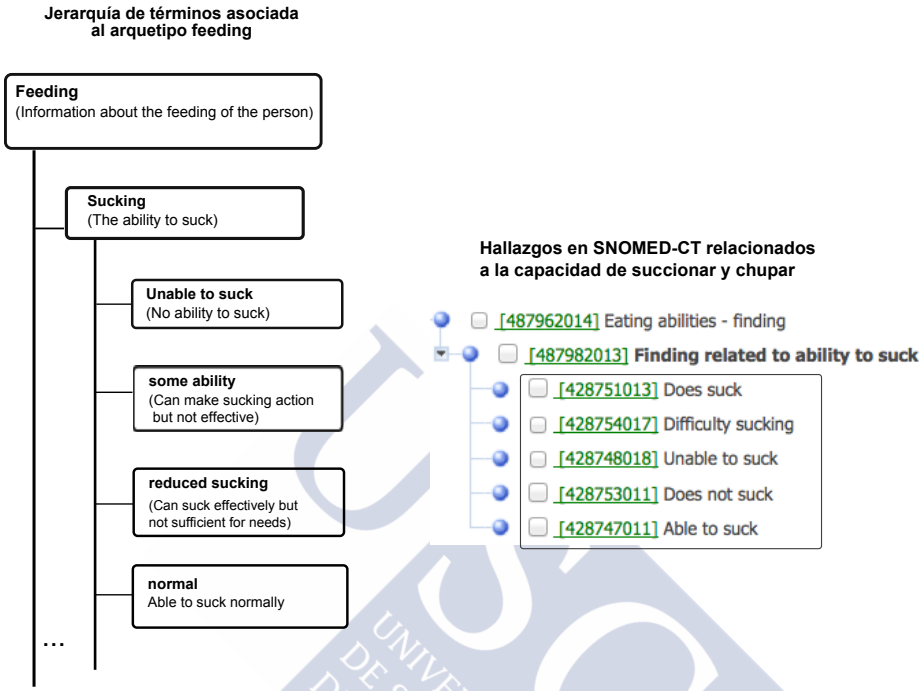

Figura 5.16: Ejemplo de diferencias en la definición de hallazgos clínicos entre el arquetipo 'feeding' y SNOMED-CT

colaboración entre openEHR y SNOMED-CT durante la creación de los arquetipos sería deseable para evitar las situaciones planteadas en este apartado. Consideramos que sería positivo que los modeladores de openEHR tuviesen en mente y accediesen a SNOMED-CT en el momento de crear un arquetipo. Esto les permitiría conocer qué conceptos (relacionados al arquetipo) ya están definidos en SNOMED-CT y reusarlos dentro del arquetipo. Por ejemplo, los modeladores del arquetipo feeding, en el momento de crear los términos para medir el nivel de habilidad de una persona para succionar, podrían acceder a SNOMED-CT para ver qué conceptos hay definidos y reusarlos en el arquetipo, en vez de crear nuevos términos. En el caso de que modeladores no encontrasen conceptos SNOMED-CT adecuados, estos podrían crear nuevos términos en el arquetipo, y al mismo tiempo solicitar a IHTSDO la creación de nuevos conceptos SNOMED-CT para representar estos términos.

En este sentido, sería interesante que los modeladores de openEHR tuviesen una herramienta para extraer de forma automática un subconjunto de SNOMED-CT relevante al arquetipo que están modelando, para facilitar la reutilización de conceptos SNOMED-CT en los arquetipos.

Existencia de términos no representados por un único concepto SNOMED-CT

Se han detectado algunos términos de arquetipos que no pueden ser explícitamente representados por un único concepto de SNOMED-CT. Algunos de estos términos encapsulan más de una idea o concepto clínico y pueden ser representados con expresiones post-coordinadas de varios conceptos SNOMED-CT. La tabla 5.5 muestra dos ejemplos de términos de arquetipos que han sido mapeados por los expertos a expresiones post-coordinadas de SNOMED-CT. El término 'waist and hip circumference' claramente encapsula dos conceptos médicos. Ya que no existe ningún concepto en SNOMED-CT apropiado para mapear este término, los expertos lo han mapeado a dos conceptos: 'waist circumference (observable entity)' y 'hip circumference (observable entity)'.

– Propuesta de mejora: Sugerimos que los modeladores openEHR cuando definan un término no mapeable a un único concepto de SNOMED-CT, manden una solicitud a la IHTSDO para que considere la creación de un nuevo concepto que equipare a dicho término. Estrategia para abordar el mapping: El mapping de este tipo de términos a expresiones post-coordinadas resulta especialmente complejo. Esta tesis doctoral no ha implementado estrategias para mapear automáticamente estos términos.

Tabla 5.5: Expresiones post-coordinadas sugeridas para mapear términos de arquetipos

Término (arquetipo)	Descripción	Expresión post-coordinada formada dos conceptos SNOMED-CT
waist and hip circumference (waist and hip circumference)	The waist (or girth) and hip circumference	waist circumference (observable entity) + hip circumference (observable entity)
total regular 24 hour feed volume (feeding)	- no incluye descripción -	quantity of eating (observable entity) + per 24 hours (qualifier value)

Incompleta cobertura de relaciones lógicas en SNOMED-CT

Se ha detectado que la definición de relaciones lógicas en SNOMED-CT está todavía incompleta. En concreto, se han encontrado bastantes conceptos de la jerarquía Clinical Finding que no tienen definida la relación 'Interprets' (esta enlaza un concepto *Clinical Finding* con el concepto de la jerarquía *Observable Entity* o *Procedure* que está siendo evaluado) y sin embargo si que existe un concepto *Observable Entity* adecuado para crear la relación. Por ejemplo, los 3 conceptos *Clinical Finding*: 'normal strength uterine contractions (clinical finding)', 'strong uterine contractions (clinical finding)' y 'mild uterine contractions (clinical finding)' no tienen definida la relación interprets en SNOMED-CT y claramente deberían estar relacionados al concepto 'Uterine contraction intensity (observable entity)' a través de 'Interprets'. Lo mismo sucede entre los conceptos 'regular contraction uterine contractions (clinical finding)' y 'Regularity of uterine contraction (observable entity)'. La falta de cobertura de relaciones lógicas puede repercutir ligeramente en nuestra estrategia de mapping léxico-contextual de la sección 5.2.1, ya que esta usa relaciones lógicas (y jerárquicas) para extraer subconjuntos relevantes de SNOMED-CT.

5.5.4. Trabajo relacionado

Esta sección incluye un estudio comparativo de nuestro método automático de mapping (entre arquetipos y SNOMED-CT) con los enfoques más relevantes surgidos recientemente con el mismo propósito [135, 78, 105]. Primero, describimos las técnicas usadas en cada uno de los enfoques, y posteriormente presentaremos los resultados más relevantes.

Comparativa de técnicas y estrategias de mapping

La tabla 5.6 muestra un resumen de las técnicas usadas por cada uno de los enfoques analizados. A continuación se comentan brevemente cada uno de los enfoques:

– Trabajo propuesto por Yu et. al.

  La herramienta de mapping desarrollada por Yu et. al. [135] ha aplicado técnicas de recuperación de información. Concretamente, ha usado Lucene para (1) indexar todas las descripciones de los conceptos SNOMED-CT como documentos y para (2) implementar un algoritmo de búsqueda basado en sistema de pesado TF-IDF [83], el cual evalúa

cuán relevante es una palabra para un documento (descripción) considerando el número de veces que aparece en la colección (SNOMED-CT) y dentro de la propia descripción.

Esta herramientas ofrece buenos tiempos de respuesta y un ranking de conceptos ordenados por relevancia. Sin embargo, creemos que la propuesta de Yu et.al tiene varias limitaciones. En primer lugar, consideramos que el sistema de pesado TF-IDF está más orientado a textos largos (tal como puede ser los textos de páginas web) que a textos cortos (conceptos SNOMED-CT). En segundo lugar, la herramienta no usa ninguna técnica de normalización léxica. Por tanto, pequeñas diferencias léxicas entre los términos del arquetipo y conceptos SNOMED-CT impiden que la herramienta obtenga mappings correctos. Tampoco usa ninguna información de contexto del arquetipo para mejorar el mapping.

Tabla 5.6: Técnicas y estrategias usadas en diferentes herramientas de mapping entre arquetipos y SNOMED-CT

Enfoque	Técnicas léxicas	Uso de información de contexto del arquetipo	Uso de contexto jerárquico de SNOMED-CT	Uso de contexto lógico de SNOMED-CT	Técnicas de similitud estructural	Servicios terminológicos de UMLS	Técnicas de desambiguación o filtrado	Técnicas de recuperación de inform. (Lucene)
Yu et. al.								X
Lezcano et. al.		X				X		
Qamar y Rector	X	X				X	X	X
Nuestro enfoque.	X	X	X	X	X	X	X	

– Trabajo propuesto por Lezcano et. al.
La herramienta de mapping desarrollada por Lezcano et. al. [78] usa los servicios terminológicos de UMLS para mapear los términos de los arquetipos a conceptos UMLS. Además, incorpora una estrategia para combinar cada término del arquetipo con el término de nivel superior en la jerarquía del arquetipo para generar así términos más específicos. Sin embargo, esta herramienta no incluye técnicas léxicas, ni de normalización, ni de recuperación de información.

– Trabajo propuesto por Qamar y Rector
La herramienta de Qamar y Rector [105], llamada MoST, incorpora dos etapas: un proceso automático de búsqueda de conceptos candidatos y un proceso manual de selección de los mappings finales realizado por expertos clínicos.

La búsqueda de conceptos candidatos incluye un amplio abanico de métodos: un complejo procedimiento para normalizar los términos, técnicas léxicas, lingüísticas y de recuperación de información (Lucene). Además, usa los servicios terminológicos de UMLS, la información contextual del arquetipo y algunas reglas de post-filtrado semántico para mejorar la calidad del mapping.

– Nuestro enfoque

Nuestro enfoque incluye muchos de los métodos incorporados en la herramienta MoST, y adicionalmente usa técnicas léxico-contextuales. Estas técnicas tratan de aprovechar las relaciones semánticas (jerárquicas y lógicas) de SNOMED-CT durante el mapping. En concreto, estas son usadas para (1) extraer subconjuntos de SNOMED-CT relacionados a la semántica de los arquetipos, con el objetivo de poder hacer búsquedas léxicas más aproximadas en ellos; y para (2) buscar similitudes estructurales entre la agrupación de términos del arquetipo y la estructura de SNOMED-CT.

Comparativa de resultados

En esta sección se expone un resumen de los experimentos y resultados de cada uno de los enfoques analizados. La tabla 5.7 muestra algunas características de los experimentos (número y tipo de arquetipos usados, repositorio de donde fueron seleccionados los arquetipos y procedimiento seguido para realizar la evaluación), así como, resultados de los mismos (recall y número medio de conceptos candidatos por término). Como se puede ver en la tabla, cada enfoque realizó evaluaciones con características distintas (en cuanto al número y tipo de arquetipos e incluso al procedimiento seguido para la evaluación), por tanto, no resulta fácil la comparación de resultados. Hay que tener en cuenta que en este campo no existe todavía gold standards para realizar evaluaciones estandarizadas. A continuación se comentan brevemente los resultados de cada uno de los enfoques:

– Resultados del trabajo de Yu et. al.

En los experimentos realizados por Yu et. al. seleccionaron arquetipos en los que los modeladores ya definieron mappings a conceptos de SNOMED-CT. Se aprovecharon estos mappings para realizar la evaluación. Este trabajo obtuvo el resultado más bajo de recall (55 %), a pesar de obtener 10 conceptos candidatos de media por cada término. La falta de técnicas de normalización léxica y de estrategias para aprovechar el contexto del arquetipo claramente repercutió en los resultados.

– Resultados del trabajo de Lezcano et. al.

  Los experimentos realizados por Lezcano et. al. incluyen un gran número de arquetipos. Sin embargo, no especifican su origen ni tipo, ni tampoco el procedimiento seguido en la evaluación. Este trabajo obtuvo un 60 % de recall y una media de 2.4 conceptos candidatos por término.

– Resultados del trabajo de Qamar y Rector

  Los experimentos de Qamar y Rector incluyeron 4 arquetipos 'Observation' del repositorio openEHR. Varios expertos clínicos evaluaron los mappings automáticos de estos arquetipos, asignando un valor de relevancia entre 0 y 10. Qamar y Rector consideraron correctos los mappings automáticos puntuados por todos los expertos con un valor de relevancia por encima de 5. La herramienta logró obtener conceptos correctos para el 68.9 % de los términos de arquetipos analizados, con una media de 5.5 conceptos candidatos por término.

Tabla 5.7: Características y resultados de los experimentos realizados por los diferentes enfoques analizados

Enfoque	Número de arquetipos	Repositorio de arquetipos	Tipo de arquetipos	Procedimiento evaluación	Recall	Número medio de conceptos candidato por término
Yu et. al.	7	NHS	Observation, Evaluations, Intructions y Actions	Evaluación con los mappings existentes en los arquetipos	55 %	10
Lezcano et. al.	40	OpenEHR	No se especifica	No se especifica	60 %	2.4
Qamar y Rector	4	OpenEHR	OBSERVATIONS	Evaluación por expertos clínicos	68.9 %	5.5
Nuestro enfoque	25	NHS	OBSERVATIONS	Evaluación por expertos clínicos	75.9 %	1.2

– Resultados de nuestro enfoque

  Como ya se mencionó en la sección 5.3, nuestros experimentos tomaron 25 arquetipos 'Observation' del repositorio openEHR. En los experimentos se obtuvieron dos conjuntos de mappings. Por una parte, dos profesionales clínicos crearon los mappings expertos para estos arquetipos. Por otra parte, se ejecutó la herramienta de mapping para obtener los mappings automáticos. La comparación entre estos dos conjuntos de mappings mostró que nuestra herramienta logró mapear correctamente el 75.9 % de los términos de arquetipos analizados, con una media de 1.2 conceptos candidatos por término.

Tal como se menciona en la introducción de este capítulo, el objetivo de las herramientas automáticas de mappings es reducir y facilitar la tarea de los profesionales médicos durante el proceso de mapping. Creemos que nuestra herramienta es la que más se acerca a este objetivo ya que alcanzó el recall más alto de todos los enfoques (un 10 % más elevado que el enfoque de Qamar [105]), logró una precisión muy alta (95.8 %) y obtuvo el menor número medio de conceptos candidatos por término.

5.5.5. Aportaciones

Las aportaciones y hallazgos de este capítulo se pueden sintetizar en los siguientes puntos:

– Un estudio sobre varias características de los arquetipos clínicos disponibles en los repositorios públicos. En concreto, se analiza: (a) la frecuencia de las categorías semánticas de los conceptos enlazados con los distintos tipos de términos de los arquetipos y (b) las relaciones semánticas existentes entre los conceptos enlazados en un mismo arquetipo. Gracias al estudio:

  - Se ha comprobado que en muchas situaciones la información clínica de los arquetipos está relacionada semánticamente, tal como lo está SNOMED-CT.
  - Se ha detectado que determinadas categorías de SNOMED-CT son más propensas para mapear ciertos tipos de términos de arquetipos.

– Un método para enlazar automáticamente la información clínica de los arquetipos a conceptos de SNOMED-CT. El método propuesto:

  - Se diferencia de otras herramientas relacionadas en que pone especial énfasis en el uso de la información contextual y estructural implícita en los arquetipos y en SNOMED-CT para mejorar el enlazado entre ambos. Los experimentos mostraron que la inclusión de las técnicas contextuales y estructurales en el método logró aumentar el recall un 16.2 %.
  - Ha mejorado en más de un 10 % en términos de recall a otras herramientas relacionadas, reduciendo además el número medio de conceptos candidatos necesarios para enlazar correctamente la información clínica de los arquetipos.
  - Ha demostrado que cuando las técnicas contextuales y semánticas cooperan con las técnicas léxicas clásicas es posible enlazar de forma automatizada la informa-

ción clínica de los arquetipos con conceptos SNOMED-CT con una precisión y un recall elevado.

– Una recopilación de deficiencias de los modelos de arquetipos y de SNOMED-CT que han dificultado el enlazado y la integración de estos modelos.

5.5.6. Trabajo futuro

Como ya se mencionó en la sección 5.5.3, se han detectado términos de arquetipos que no pueden ser explícitamente representados por un único concepto de SNOMED-CT y deben ser representados con expresiones post-coordinadas de varios conceptos SNOMED-CT. En el futuro, podría ser interesante implementar estrategias específicas para mapear automáticamente este tipo de términos a expresiones post-coordinadas de SNOMED-CT.

Los experimentos han sido realizados sobre arquetipos openEHR de tipo 'Observation' ya que estos actualmente son los más maduros y revisados. Muchas de las técnicas planteadas en la tesis son independientes al tipo de arquetipo. Sin embargo, algunas de las técnicas contextuales usadas podrían requerir ciertas adaptaciones al tipo de arquetipo. Por tanto, en el futuro, sería interesante realizar experimentos adicionales con otros tipos de arquetipos para afinar y ajustar la herramienta de mapping.

En el futuro, el método propuesto podría ser integrado en los actuales repositorios de arquetipos con el fin de facilitar, en un entorno real y de una forma más efectiva, la creación de enlaces entre los términos clínicos de los arquetipos y los conceptos SNOMED-CT.

El método propuesto podría ser aplicado (con algunas adaptaciones) a otros modelos de datos clínicos, especialmente en aquellos en los que la información esté total o parcialmente estructurada.

5.6. Resumen

Los modelos de datos clínicos y las terminologías tienen un papel importante en el camino hacia la interoperabilidad semántica en los sistemas de información clínica. Una de las tareas a las que se enfrenta actualmente la informática médica es conseguir una mayor integración entre los modelos de datos clínicos y las terminologías. En este sentido, los arquetipos openEHR (modelos de datos clínicos estructurados) han sido diseñados, en teoría, para permitir un cierto grado de integración con terminologías. Así, la información clínica de los arquetipos

(definida con términos de lenguaje natural) puede ser enlazada con conceptos de terminologías clínicas. Sin embargo, en la práctica estos enlaces o mappings actualmente todavía son muy infrecuentes. Esto se debe en gran parte a que la creación de mappings realizada por profesionales médicos es todavía muy lenta y costosa.

En este capítulo hemos planteado un método automático para enlazar la información clínica de los arquetipos a conceptos de la terminología SNOMED-CT, con el objetivo de reducir lo máximo posible la participación de los profesionales médicos en la tarea de mapping.

El método propuesto incorpora un amplio abanico de técnicas y estrategias: normalización léxica de términos, técnicas de equiparación léxica, servicios terminológicos de UMLS, técnicas de desambiguación, técnicas léxico-contextuales y estructurales. Estas dos últimas son las técnicas más innovadoras y las que nos diferencia de otros trabajos relacionados de mapping. Las técnicas léxico-contextuales han sido usadas para extraer subconjuntos de SNOMED-CT relacionados a la semántica de los arquetipos. Esto permitió realizar búsquedas léxicas más aproximadas en los subconjuntos de SNOMED-CT relevantes, obteniéndose nuevos mappings con alta precisión. Las técnicas estructurales han aprovechado las similitudes estructurales entre la agrupación de términos del arquetipo y la estructura de SNOMED-CT para detectar nuevos mappings.

Los experimentos realizados con 25 arquetipos de tipo 'Observation' mostraron resultados prometedores: un 75.9 % de los términos analizados obtuvieron un mapping automático correcto, el 95.8 % de los mappings obtenidos por el método son correctos y el número medio de conceptos candidatos por cada término es bajo (sólo 1.2 conceptos por término). Las técnicas léxico-contextuales y estructurales han contribuido de forma notable en el recall. La inclusión de estas dos técnicas en la herramienta logró aumentar el recall un 16.2 %. Los resultados también muestran que las técnicas de desambiguación fueron muy útiles para aumentar la precisión de la herramienta (más de un 10 %).

Este capítulo mostró que es posible mapear automáticamente términos de arquetipos a SNOMED-CT con alta precisión y recall, mediante la combinación de diferentes técnicas y estrategias de mapping. El trabajo demostró que es necesario que las técnicas contextuales y semánticas cooperen con las técnicas léxicas para alcanzar un mapping automático de mayor calidad.

Capítulo 6

Segmentación automatizada en SNOMED-CT para facilitar la gestión y búsqueda en arquetipos

OpenEHR es un estándar abierto de Historia clínica electrónica (HCE) que describe la estructura, el almacenamiento, recuperación e intercambio de datos clínicos [98]. OpenEHR propone un paradigma de modelado de dos niveles para representar los contenidos de la HCE: el modelo de referencia y el modelo de arquetipos. El primero permite un consenso en los elementos estructurales básicos de los registros electrónicos de salud. Los arquetipos clínicos, haciendo uso de estos elementos estructurales, definen descripciones formales de datos clínicos consensuadas por expertos. Por ejemplo, los arquetipos de tipo 'Observation' definen una agrupación de ítems cuyo propósito es definir aspectos o cuestiones que deben medirse u observarse sobre una situación clínica (p.e. la medición de glucosa de un paciente). El uso de arquetipos en la práctica clínica permite capturar datos de instancias clínicas de forma estándar y sistemática. Una ventaja clave de OpenEHR es que facilita la interoperabilidad, ya que distintos sistemas de información basados en el estándar openEHR pueden intercambiar datos clínicos fácilmente a través de arquetipos.

En los últimos años, expertos de importantes instituciones (NEHTA, NHS, Centre for eHealth de Suecia) y de la comunidad openEHR han estado modelando arquetipos openEHR [98, 103, 104, 101]. Paralelamente, han surgido repositorios online donde los arquetipos son publicados y revisados por diferentes grupos de expertos [92, 90]. En la actualidad, el reposi-

torio más importante es el openEHR Clinical Knowledge Manager (CKM) promovido por la comunidad internacional de openEHR [92]. Este ofrece, a los usuarios interesados en modelar datos clínicos, la oportunidad y los medios para participar en la creación y/o mejora de un conjunto internacional de arquetipos. El número de arquetipos publicados en este repositorio empieza a ser importante. En mayo de 2014 este repositorio está formado por 384 arquetipos.

La experiencia de usar arquetipos en los sistemas de información clínica existentes (p.e. en sistemas desplegados en centros de atención primario) es todavía muy limitada [20]. Uno de los principales obstáculos para desplegar de forma exitosa un paradigma de modelado de dos niveles, como openEHR o ISO EN 13606, es que los sistemas actuales usan modelos propietarios a medida para representar el contenido clínico. Varios estudios han propuesto soluciones para manejar este problema [20, 32, 15, 88, 119, 18]. El estudio de R. Chen et. al. analizó la conversión de un modelo propietario utilizado en hospitales de Suecia a un modelo basado en arquetipos openEHR y viceversa [20]. El estudio desarrolló un método automático para representar el contenido clínico del modelo propietario en arquetipos openEHR, concluyendo que los arquetipos ya son suficientemente expresivos para representar el modelo propietario. R. Chen et. al. defienden una migración lenta e incremental entre sistemas basados en modelos propietarios y sistemas basados en arquetipos y estándares. Sostienen que es importante que convivan los modelos propietarios (principalmente para preservar el significado de toda la información clínica existente creada y almacenada con estos modelos) con modelos basados en arquetipos y estándares para compartir e intercambiar el contenido clínico entre diferentes sistemas de información basados en diferentes modelos propietarios. Otros estudios [32, 15, 88, 119, 18] apuestan por extraer el contenido clínico de los sistemas basados en modelos propietarios mapeándolo a las estructuras de datos definidas en los arquetipos clínicos, prescindiendo totalmente de los modelos propietarios.

Los estudios que han analizado la viabilidad de representar el contenido clínico de modelos propietarios en forma de arquetipos openEHR han reconocido que esta tarea es muy laboriosa y costosa, principalmente por la falta de herramientas avanzadas de búsqueda y de mapping. Hay que tener en cuenta que los estudios, como primera opción, han intentado reusar los arquetipos de los principales repositorios para representar el contenido clínico de los modelos propietarios. Esto requiere buscar equiparaciones entre los términos locales del modelo propietario y los términos locales de todos los arquetipos de los repositorios. Sin embargo, en el momento de los estudios los únicos procedimientos para buscar estas equiparaciones fueron: (1) revisión manual del contenido de todos los arquetipos, o bien, (2) uso de

los buscadores de los repositorios de arquetipos basados en palabras clave. Estos buscadores no son efectivos cuando un mismo concepto clínico es expresado con diferentes palabras en el modelo propietario y en los arquetipos. Por ejemplo, si buscásemos el término de un modelo propietario 'respiratory frequency' en el buscador del repositorio de openEHR, este no será capaz de detectar el término del arquetipo 'rate of respiration', cuando claramente representa el mismo concepto. Estos procedimientos para buscar equiparaciones claramente son insuficientes, teniendo un impacto negativo en el reuso de los arquetipos estándar, causando además la creación de nuevos arquetipos inconsistentes y repetitivos. Una solución apuntada por estos estudios es desarrollar sistemas de búsqueda y mapping más avanzadas en los repositorios de los arquetipos para agilizar la conversión de modelos propietarios a arquetipos openEHR [119, 18].

Este capítulo presenta una aproximación encaminada a añadir funcionalidades semánticas a los repositorios de arquetipos. Nuestra aproximación apuesta por anotar semánticamente los arquetipos mediante segmentos de SNOMED-CT. Los segmentos (o subontologías) son partes autocontenidas de ontologías, que se caracterizan porque en sí mismos son ontologías válidas sin necesidad de referenciar a las ontologías fuente.

Nuestra aproximación incluye un método automático que busca conceptos SNOMED-CT equivalentes a los términos locales de los arquetipos, y a continuación, partiendo de estos conceptos, extrae automáticamente segmentos de SNOMED-CT semánticamente relacionados a los arquetipos clínicos. Consideramos que la representación del contenido clínico de los arquetipos con SNOMED-CT (específicamente, en forma de segmentos SNOMED-CT) mejora la navegación, gestión y búsqueda en extensos repositorios de arquetipos, ya que permite aprovechar el conocimiento léxico (sinonimia) y estructural (relaciones) representado en SNOMED-CT. A su vez, estas mejoras facilitan el mapping entre términos locales de modelos propietarios y términos locales de los arquetipos.

El capítulo también presenta dos aplicaciones [1] que demuestran las ventajas de la anotación semántica de los arquetipos mediante segmentos SNOMED-CT. El primer servicio está orientado a identificar relaciones y solapes entre arquetipos, mientras que el segundo gestiona búsquedas semánticas de conceptos clínicos en los arquetipos.

[1] Estas aplicaciones y el método de segmentación han sido descritos en uno de los artículos de investigación publicados durante la elaboración de la tesis [5]

6.1. Materiales

El método de segmentación propuesto en el presente capítulo ha sido diseñado para anotar semánticamente arquetipos openEHR (ver sección 2.2.2) con segmentos de la terminología SNOMED-CT (ver sección 2.3).

6.2. Métodos

En este apartado, primero se describe el método automático para extraer segmentos de SNOMED-CT asociados a arquetipos. A continuación, se expone el funcionamiento de los dos servicios que usan los segmentos de SNOMED-CT para facilitar la gestión y búsqueda en repositorios de arquetipos.

6.2.1. Extracción de segmentos SNOMED-CT

El método de extracción de segmentos de SNOMED-CT está inspirado en el algoritmo de segmentación propuesto por J. Seidenberg et. al. [112]. Este algoritmo considera SNOMED-CT como un grafo (los conceptos representan los nodos y las relaciones los arcos), y aplica técnicas de recorrido de grafo para extraer segmentos de SNOMED-CT dado un conjunto de conceptos iniciales o semilla.

Para cada arquetipo, nuestro método obtiene dos tipos de segmentos SNOMED-CT: mínimo y enriquecido. El segmento mínimo trata de capturar la semántica del arquetipo, por lo que incluye los conceptos mapeados a los términos locales del arquetipo más algún concepto adicional necesario para formar segmentos válidos. Al igual que otros trabajos que extraen segmentos SNOMED-CT [110], consideramos que un segmento es válido cuando todos sus conceptos están al menos conectados a otro concepto del segmento.

El segmento enriquecido trata de capturar todos los conceptos relacionados a un determinado arquetipo (aunque estos no estén definidos explícitamente en el arquetipo). Por ello, el segmento enriquecido extiende al mínimo con conceptos vecinos, es decir, con conceptos relacionados jerárquicamente o lógicamente en SNOMED-CT.

El proceso de extracción de segmentos SNOMED-CT asociados a arquetipos incluye dos etapas: un proceso de mapping para obtener los conceptos semilla asociados al arquetipo y una etapa de segmentación en SNOMED-CT (ver figura 6.1). Destacar que la etapa de mapping es común para los dos tipos de segmentos, mientras que la etapa de segmentación es diferente. A continuación, se expone la etapa de mapping, de segmentación mínima y de segmentación enriquecida.

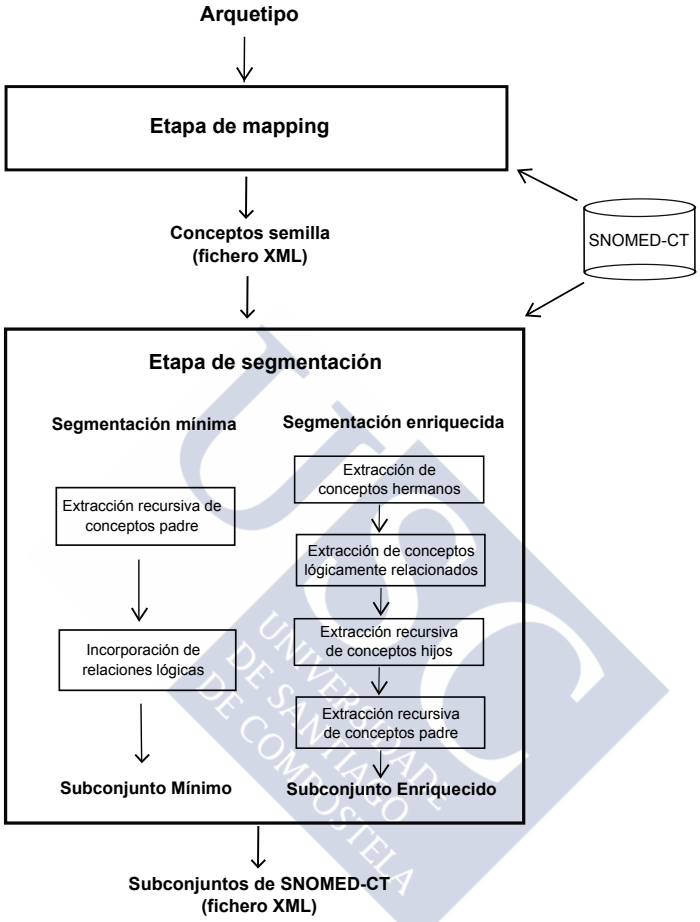

Figura 6.1: Etapas destacadas en la extracción de segmentos mínimos y enriquecidos asociados a arquetipos clínicos

Etapa de mapping para la generación de conceptos semilla

En esta etapa se ha reutilizado el método completo de mapping propuesto en el capítulo 5 (ver sección 5.2), para buscar y validar mappings entre términos de un arquetipo y conceptos de SNOMED-CT. Los conceptos obtenidos en esta etapa son utilizados como conceptos semilla en la etapa de segmentación. En la figura 6.2 se muestran los conceptos semilla obtenidos para el arquetipo 'faeces' tras la aplicación de esta etapa.

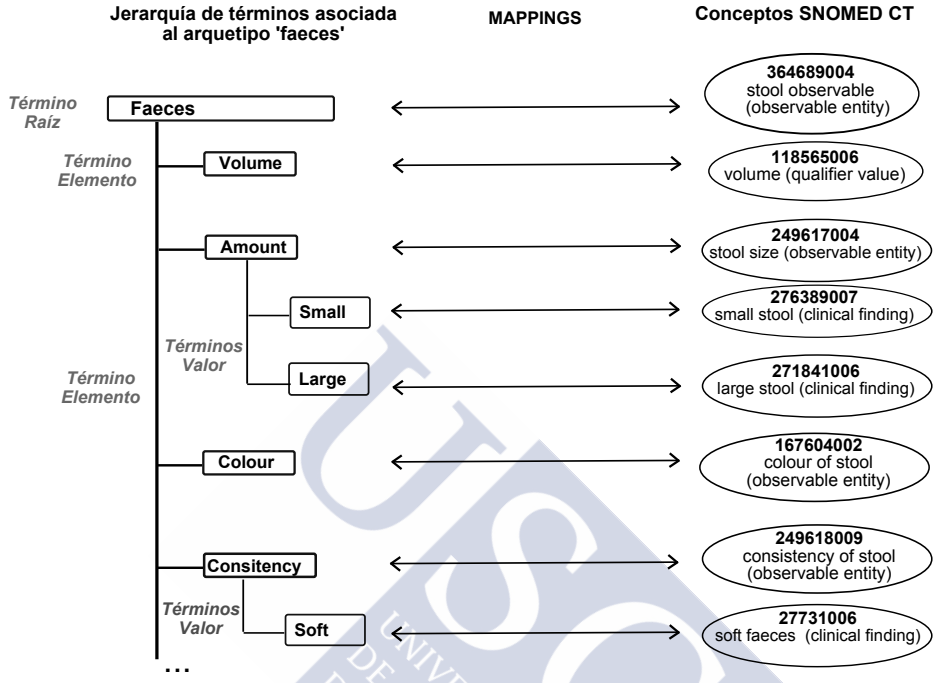

Figura 6.2: Conceptos semilla obtenidos para el arquetipo 'faeces' tras la etapa de mapping.

Segmentación mínima en SNOMED-CT

El objetivo de esta etapa es extraer un subconjunto de SNOMED-CT mínimo (es decir, que contenga sólo los conceptos más próximos semánticamente a los términos locales de un arquetipo) y válido (es decir, que esté bien formado y no contenga conceptos aislados) dado un conjunto de conceptos semilla. Incluye dos subetapas:

Subetapa 1: Extracción recursiva de conceptos padre

La jerarquía de SNOMED-CT es recorrida, por todas las rutas posibles, desde los conceptos semilla al concepto raíz de SNOMED-CT ('SNOMED-CT concept'). Los conceptos semilla, junto con los conceptos y relaciones jerárquicas recorridas durante esta etapa se almacenan en el segmento mínimo de SNOMED-CT. La figura 6.3 muestra el resultado de la segmentación mínima dado el conjunto de conceptos semilla del arquetipo 'faeces'. Todos los conceptos con fondo blanco y todas las relaciones jerárquicas del segmento mínimo han sido

añadidos en esta etapa. Por ejemplo, esta etapa, para el concepto semilla 'volume' extrae los conceptos 'measurement property', 'qualifier value' y 'SNOMED-CT concept'.

Subetapa 2: Incorporación de relaciones lógicas

El método comprueba si en SNOMED-CT existen relaciones lógicas (de atributo) entre los conceptos pertenecientes al segmento. Si es así, estas son añadidas al segmento. En la figura 6.3 se pueden ver varias relaciones de atributo añadidas en esta etapa al segmento mínimo, por ejemplo, entre 'large stool' y 'stool size' y entre 'soft faeces' y 'consistency of stool'.

Al final del proceso, el segmento mínimo contiene un conjunto de conceptos mapeados a términos locales del arquetipo (representados con fondo gris en la figura 6.3) y un conjunto de conceptos enlazados a estos a través de relaciones de SNOMED-CT (representados con fondo blanco).

Los segmentos mínimos son almacenados en forma de ficheros XML, en los que se incluyen metainformación del arquetipo de partida (nombre, versión y tipo del arquetipo) y el conjunto de conceptos y relaciones que forman el segmento extraído.

Segmentación enriquecida en SNOMED-CT

El objetivo de esta etapa es extraer un segmento de SNOMED-CT (más amplio que el segmento mínimo) formado por porciones de SNOMED-CT relacionadas al arquetipo. Esta etapa está formada por cuatro subetapas ejecutadas secuencialmente:

Subetapa 1: Extracción de conceptos hermanos

En esta etapa se extraen los conceptos hermano de los conceptos semilla, es decir, los conceptos que comparten padre en SNOMED-CT con los conceptos semilla.

Subetapa 2: Extracción de conceptos lógicamente relacionados

En esta etapa se extraen los conceptos enlazados a través de relaciones de atributo a los conceptos semilla.

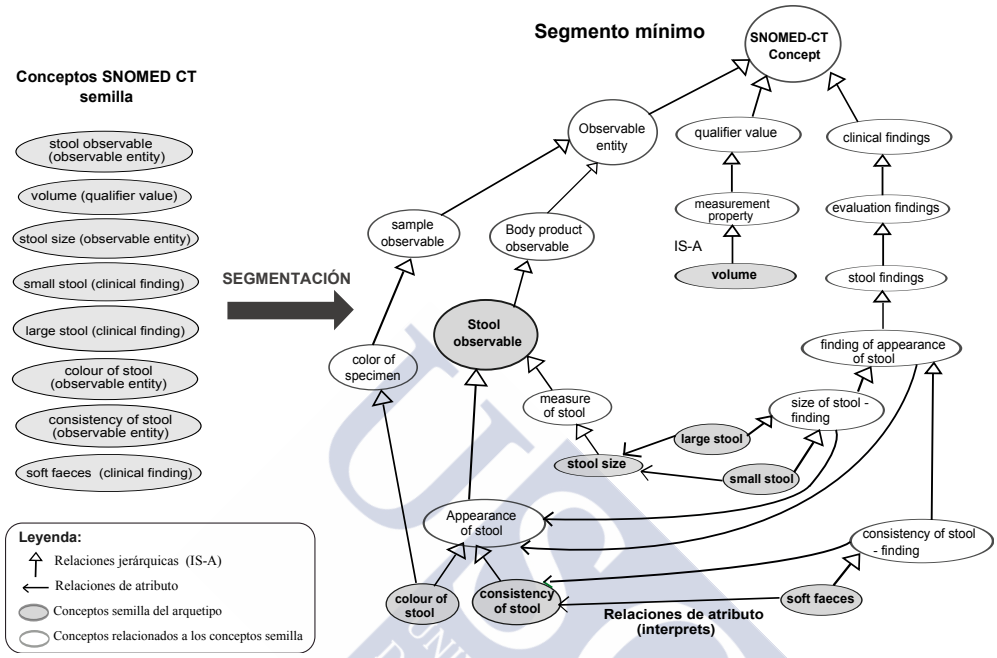

Figura 6.3: Segmento mínimo de SNOMED-CT asociado al arquetipo 'faeces'.

Subetapa 3: Extracción recursiva de conceptos hijos

La jerarquía de SNOMED-CT es recorrida, por todas las rutas posibles, desde los conceptos semilla a conceptos hojas de SNOMED-CT. Se aplica tanto a los conceptos semilla como a los conceptos extraídos en las etapas anteriores.

Subetapa 4: Extracción recursiva de conceptos padre

Esta etapa es exactamente igual que en la segmentación mínima. Se aplica tanto a los conceptos semilla como a los conceptos extraídos en las etapas anteriores. Los conceptos semilla, junto con los conceptos y relaciones jerárquicas y de atributo recorridas en estas 4 etapas son incluidos en el segmento enriquecido de los arquetipos.

Además, se han aplicado filtros en la subetapa 2 y 3 para evitar que el segmento crezca excesivamente. En la subetapa 2, si un concepto semilla está asociado a un número significativamente alto de conceptos (a través de relaciones de atributo), entonces esta subetapa no

se aplica sobre dicho concepto, por lo que no se añaden nuevos conceptos al segmento enriquecido. Hemos establecido un umbral de 50 conceptos para aplicar este filtro, determinado heurísticamente. Nuestros experimentos mostraron que usando este umbral, más del 90 % de conceptos descartados son poco relevantes para el arquetipo. En la subetapa 3, si el concepto de partida está en los primeros 4 niveles superiores de la jerarquía SNOMED-CT, o bien, si tiene más de 100 conceptos descendentes en SNOMED-CT, entonces esta subetapa no se aplica sobre dicho concepto, por lo que no se añaden nuevos conceptos al segmento enriquecido. Estos dos filtros tratan de evitar la expansión en SNOMED-CT desde conceptos genéricos ya que es probable que introduzca un elevado número de conceptos, muchos de los cuales genéricos o poco relevantes para el arquetipo.

6.2.2. Servicios basados en segmentos

En esta sección presentamos dos aplicaciones o servicios orientados a dotar a los repositorios de arquetipos de funcionalidades semánticas.

Para el desarrollo y evaluación de estos servicios hemos trabajado con un grupo de arquetipos del repositorio internacional de openEHR [92]. Para cada arquetipo seleccionado, hemos aplicado el método de extracción de segmentos expuesto en la sección 6.2.1; con ello hemos obtenido un segmento mínimo y enriquecido asociado a cada arquetipo. Estos segmentos pueden verse como la representación o la anotación semántica de los arquetipos.

Los dos servicios usan estos segmentos para implementar funcionalidades semánticas no incluidas actualmente en los repositorios de arquetipos.

Servicio de comparación semántica de arquetipos

Este servicio tiene el objetivo de ayudar a usuarios y modeladores de arquetipos a buscar solapes en el contenido clínico entre un grupo de arquetipos, y también, a detectar relaciones semánticas entre arquetipos, es decir, arquetipos que modelan contenido clínico relacionado.

Consideramos que dos arquetipos están solapados, cuando estos modelan los mismos conceptos clínicos. En cambio, consideramos que dos arquetipos están relacionados cuando estos modelan conceptos clínicos enlazados a través de relaciones definidas en SNOMED-CT.

Funcionamiento del servicio

El servicio trabaja con un grupo de arquetipos del repositorio internacional de openEHR y con sus segmentos enriquecidos asociados. A continuación, se expone el caso de uso típico de este servicio:

- En primer lugar, el usuario del servicio elige uno de los arquetipos con la intención de conocer que otros arquetipos tienen solapes o están relacionados con el arquetipo seleccionado.
- Seguidamente, el segmento enriquecido del arquetipo seleccionado es comparado con el resto de segmentos enriquecidos.
    - Por una parte, el servicio, con el objetivo de detectar arquetipos solapados, busca coincidencias entre los conceptos 'semilla' del segmento asociado al arquetipo seleccionado y el resto de conceptos 'semilla' de los otros segmentos.
    - Por otra parte, el servicio, con el objetivo de detectar arquetipos relacionados, comprueba si algún concepto del segmento del arquetipo seleccionado está enlazado a través de alguna relación jerárquica o lógica de SNOMED-CT con algún concepto de los otros segmentos.
- Finalmente, el servicio recopila toda la información sobre los arquetipos solapados y relacionados mostrándola de forma amigable al usuario.

Servicio de búsqueda semántica en arquetipos

Este servicio tiene el propósito de buscar arquetipos relacionados a un término clínico (p.e. una enfermedad o un hallazgo clínico). El servicio trabaja internamente con conceptos y relaciones de SNOMED-CT, y como veremos a continuación, implementa funcionalidades de búsqueda semántica, superando así a los buscadores basados en palabras clave de los repositorios actuales.

Funcionamiento del servicio

Al igual que el anterior servicio, el servicio de búsqueda trabaja con un grupo de arquetipos del repositorio de openEHR y con los segmentos enriquecidos asociados. A continuación, se expone el funcionamiento general del servicio (ver figura 6.4):

– En primer lugar, el usuario introduce un término clínico con la intención de buscar arquetipos relevantes a dicho término.

– Posteriormente, técnicas de mapping son aplicadas para mapear el término de consulta a uno o varios conceptos equivalentes de SNOMED-CT. Hemos usado técnicas de normalización léxica (ver sección 3.3.1) y una librería de indexación y búsqueda llamada Lucene[2] que permite obtener con gran rapidez una lista de conceptos SNOMED-CT candidatos a mapear el término de consulta. El usuario selecciona de esta lista el concepto más preciso (podría seleccionar más de un concepto).

– Seguidamente, el servicio busca el concepto seleccionado por el usuario en los segmentos enriquecidos de todos los arquetipos. Aquellos segmentos que contengan el concepto son considerados relevantes. Destacar, que el servicio considera más relevantes aquellos segmentos que tienen el concepto buscado entre sus conceptos 'semilla'.

– Finalmente, el servicio muestra al usuario el listado de arquetipos relevantes al término de búsqueda, justificando en cada caso el motivo de la relevancia.

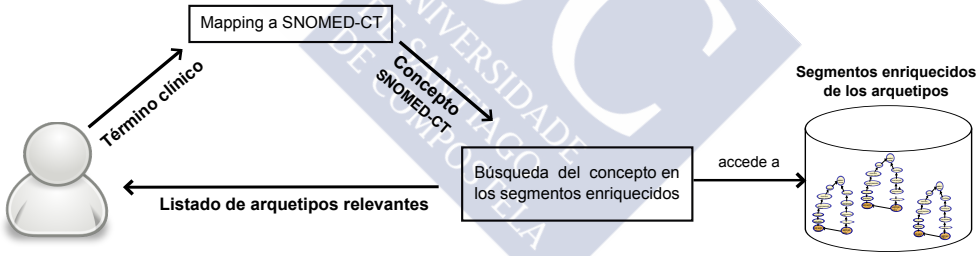

Figura 6.4: Funcionamiento del servicio de búsqueda semántica

6.3. Procedimiento de evaluación

Se han evaluado cuantitativamente dos aspectos: la calidad de los segmentos mínimos de SNOMED-CT y el rendimiento del servicio de búsqueda semántica en arquetipos.

[2] `http://lucene.apache.org/core/`

6.3.1. Procedimiento de evaluación de los segmentos mínimos

Evaluar la relevancia o calidad de los segmentos ontológicos es extremadamente difícil [48, 24]. D'Aquin et. al. concluye que no hay una forma universal de modularizar o segmentar una ontología y que la elección de la técnica de segmentación debería estar guiada por los requerimientos de la aplicación que demande el uso de segmentos ontológicos. Además, no suele haber *gold standards* para evaluar la calidad de los segmentos ontológicos.

Con el objetivo de poder evaluar los segmentos mínimos automáticos (extraídos por el método) asociados a los arquetipos, hemos creado un conjunto de segmentos mínimos de referencia. Aunque este tipo de evaluación no es la ideal, esta intenta medir aproximadamente la capacidad de los segmentos mínimos automáticos para representar el contenido semántico de los arquetipos.

Arquetipos seleccionados

En esta evaluación hemos trabajado con 25 arquetipos Observation del repositorio NHS Connecting for Health. Son los mismos arquetipos que los usados en la evaluación del mapping del capítulo 5 (ver sección 5.3.1). La lista de arquetipos puede ser consultada en el apéndice B.

Procedimiento seguido para la creación de los segmentos mínimos de referencia

Para crear los segmentos mínimos de referencia, hemos reutilizado los mappings expertos creados en la evaluación del mapping del capítulo 5 (ver sección 5.3). Los conceptos SNOMED-CT mapeados por los expertos fueron usados como conceptos 'semilla' para esta evaluación. Partiendo de estos conceptos semilla, hemos aplicado el método de segmentación mínima (expuesto en la sección 6.2.1) para obtener los segmentos mínimos de referencia de los 25 arquetipos seleccionados.

Medidas cuantitativas de la evaluación

Dos medidas clásicas de recuperación de información (precisión y recall) fueron calculadas para estimar la similitud entre los segmentos mínimos y de referencia. En este contexto, la precisión estima la proporción entre el número de conceptos (relaciones) correctos del segmento automático y el número total de conceptos (relaciones) que forman el segmento auto-

mático. Mientras que el recall estima la proporción entre el número de conceptos (relaciones) correctos del segmento automático y el número total de conceptos (relaciones) que forman el segmento de referencia. Consideramos que un concepto del segmento automático es correcto, cuando este también está en el segmento de referencia del mismo arquetipo.

6.3.2. Procedimiento de evaluación de los segmentos enriquecidos

El contenido de los segmentos enriquecidos no ha sido evaluado de forma directa tal como se ha hecho con los segmentos mínimos, debido a la gran dificultad que entraña a los expertos crear segmentos enriquecidos de referencia. Sin embargo, ya que los segmentos enriquecidos juegan un papel clave en nuestro servicio de búsqueda semánticas, la evaluación de dicho servicio nos proporciona también una buena medida de la utilidad de los segmentos enriquecidos.

6.3.3. Procedimiento de evaluación del servicio de búsqueda semántica

El principal objetivo de esta evaluación fue comprobar si nuestro servicio de búsqueda semántica mejora el rendimiento de los buscadores presentes en los actuales repositorios de arquetipos. Concretamente, hemos seleccionado el buscador del repositorio internacional de openEHR como buscador de referencia con los que comparar resultados de búsquedas, ya que este es el más usado actualmente.

Buscador del repositorio internacional de openEHR

El repositorio internacional de openEHR [92] incluye un buscador en el que los usuarios introducen un término clínico con la intención de obtener los arquetipos del repositorio relevantes. El buscador trata de encontrar equivalencias léxicas exactas entre el término buscado y los términos locales definidos en todos los arquetipos del repositorio. Además, ofrece opciones de búsqueda booleana de tipo ‘AND’ y ‘OR’. La opción ‘OR’ recupera arquetipos que incluyan al menos una de las palabras del término buscado, mientras que la opción ‘AND’ obtiene arquetipos que contienen todas las palabras del término de búsqueda.

Arquetipos seleccionados para los experimentos de búsqueda

De los 25 arquetipos que hemos usado en la anterior evaluación (procedentes del repositorio NHS), seleccionamos sólo los 16 que tienen su equivalente en el repositorio openEHR. Por tanto, los experimentos de búsqueda, en nuestro servicio y en el buscador de openEHR, fueron realizados y evaluados sólo con en estos 16 arquetipos. El apéndice C incluye un listado con los 16 arquetipos seleccionados.

Construcción del conjunto de datos de evaluación

Para los experimentos, hemos construido un conjunto de datos de evaluación, formado por términos de búsqueda junto con sus correspondientes respuestas correctas, esto es, un listado de arquetipos relevantes (a dichos términos), seleccionados entre los 16 arquetipos que toman parte en los experimentos.

Los términos de búsqueda han sido seleccionados de dos diccionarios médicos online: MedlinePlus[3] y MediLexicon[4]. MedlinePlus contiene términos para más de 900 temas de salud, mientras que MediLexicon provee más de 100.000 términos médicos. Hemos revisado manualmente estos diccionarios seleccionando 55 términos relacionados al contenido clínico modelado en los 16 arquetipos usados en los experimentos.

Posteriormente, un profesional médico (María Jesús Sobrido), asignó arquetipos que consideró relevantes para cada uno de los 55 términos de búsqueda. El apéndice C incluye más información sobre el conjunto de datos de evaluación obtenido.

Experimentos y medidas cuantitativas de la evaluación

Hemos realizado el siguiente experimento: buscamos los 55 términos clínicos seleccionados con nuestro servicio de búsqueda semántica y con el buscador del repositorio openEHR (con los dos operadores 'OR' y 'AND'). Comparamos los resultados obtenidos con las respuestas correctas asignadas por los expertos, y finalmente, calculamos la precisión y el recall para cada sistema.

El recall refleja la proporción entre las consultas correctas y el número total de consultas realizadas (55). Consideramos una consulta correcta cuando esta devuelve al menos un arquetipo considerado relevante por el experto.

[3] http://www.medilexicon.com/medicaldictionary.php
[4] http://www.nlm.nih.gov/medlineplus/healthtopics.html

La precisión mide la proporción entre el número de consultas correctas (sin obtener ningún arquetipo irrelevante) y el número total de consultas que obtienen resultados.

6.4. Resultados

6.4.1. Resultados de la segmentación mínima en SNOMED-CT

La tabla 6.1 muestra los resultados de la evaluación de los segmentos mínimos automáticos para cada uno de los 25 arquetipos seleccionados para los experimentos. De la segunda a la quinta columna de la tabla podemos ver el número de conceptos y relaciones de los segmentos mínimos automáticos y de referencia. En muchos arquetipos, el número de conceptos y relaciones en los dos tipos de segmentos es muy similar: de media, los segmentos de referencia contienen 83 conceptos y 121 relaciones, mientras que los segmentos automáticos están formados por 81 conceptos y 116 relaciones. Sin embargo, hemos detectado que en algunos casos los segmentos automáticos contienen un número de conceptos bastante superior en relación a los segmentos de referencia. Por ejemplo, los segmentos automáticos de los arquetipos 'blood pressure' y 'respirations' tienen un 169 % y un 78 % más de conceptos que el segmento de referencia correspondiente, respectivamente. En cambio, otros segmentos automáticos contienen un número de conceptos bastante inferior respecto a los segmentos de referencia. Por ejemplo, los segmentos automáticos de los arquetipos 'speech' y 'perineum' tienen un 59 % y un 19 % menos de conceptos que el segmento de referencia correspondiente, respectivamente.

Las últimas cuatro columnas de la tabla muestran los valores de recall y precisión para conceptos y relaciones en cada arquetipo. Vemos que de media el método de segmentación alcanzó: 70.6 % y 67.3 % de precisión para conceptos y relaciones, y un 69.1 % y 64.6 % de recall para conceptos y relaciones, respectivamente. Dicho de otra forma, aproximadamente dos tercios de los conceptos y relaciones obtenidos por la segmentación automática son relevantes para los arquetipos, y también aproximadamente dos tercios de los conceptos y relaciones de referencia han sido extraídos por la segmentación automática.

El apéndice C muestra más información sobre los segmentos mínimos y enriquecidos obtenidos durante los experimentos e incluye enlaces para descargarlos.

Tabla 6.1: Resultados de la segmentación mínima

Arquetipo	# de conceptos		# de relaciones		Precisión %		Recall %	
	Segmento mínimo de referencia	Segmento mínimo automático	Segmento mínimo de referencia	Segmento mínimo automático	Conceptos	Relaciones	Conceptos	Relaciones
Apgar Score	76	70	97	83	68.6	69.9	63.2	59.8
Baby Observations	199	164	276	196	77.4	79.1	63.8	56.2
Blood Pressure	16	43	18	53	37.2	34.0	100.0	100.0
Body Mass	25	25	39	39	100.0	100.0	100.0	100.0
Body Weight	11	17	10	18	47.1	38.9	72.7	70.0
Concious State	62	81	85	126	71.6	62.7	93.5	92.9
Faeces	33	33	46	58	78.8	65.5	78.8	82.6
Feeding	192	215	268	287	64.2	60.6	71.9	64.9
Fetal Movement	33	38	41	47	68.4	70.2	78.8	80.5
Head Circumference	8	8	7	7	100.0	100.0	100.0	100.0
Heart Rate	51	47	65	62	78.7	77.4	72.5	73.8
Height	7	8	6	7	75.0	71.4	85.7	83.3
Hydration	78	81	94	97	77.8	80.4	80.8	83.0
Mobility	77	88	104	147	58.0	44.2	66.2	62.5
Palpation Breast	57	45	86	72	80.0	69.4	63.2	58.1
Perineum	266	215	537	375	84.2	85.9	68.0	60.0
Postnatal Mother	439	360	644	584	70.0	61.3	57.4	55.6
Respiration	69	123	92	174	52.0	48.9	92.8	92.4
Speech	17	10	17	9	60.0	55.6	35.3	29.4
Substance Tobacco	112	107	139	142	59.8	59.2	57.1	60.4
Urine Output	16	27	17	29	55.6	55.2	93.8	94.1
Uterine Contractions	84	83	107	106	97.6	98.1	96.4	97.2
Visual Acuity	83	69	144	96	65.2	64.6	54.2	43.1
Waist Hip	9	9	8	8	100.0	100.0	100.0	100.0
Wellbeing	61	71	73	78	69.0	69.2	80.3	74.0
Average	**83.2**	**81.5**	**120.8**	**116.0**	**70.6**	**67.3**	**69.1**	**64.6**

6.4.2. Resultados de la segmentación enriquecida en SNOMED-CT

El tamaño de los segmentos enriquecidos extraídos automáticamente es considerablemente mayor que el de los segmentos mínimos. De media, los segmentos enriquecidos contienen 511 conceptos y 923 relaciones, frente a los 81 conceptos y 116 relaciones de los segmentos mínimos.

6.4.3. Resultados del servicio de búsqueda semántica

La tabla 6.2 muestra los resultados de los experimentos realizados en nuestro servicio de búsqueda semántica y en el buscador del repositorio openEHR [92]. Nuestro servicio de búsqueda semántica logró recuperar arquetipos relevantes para 38 de las 55 búsquedas, alcanzando un 97.4 % de precisión y un 69.1 % de recall, mientras que el buscador del repositorio de openEHR (con los operadores OR y AND) logró 95.5 % y 80.0 % de precisión y 40.0 % y 9.1 % recall, respectivamente. Por tanto, nuestro enfoque mejoró considerablemente el recall (más de un 70 % de mejora) respecto al buscador openEHR, manteniendo la precisión muy

similar. El apéndice C muestra más detalles sobre los resultados obtenidos por el servicio de búsqueda semántica y por el buscador del repositorio openEHR.

Hemos detectado que la extracción de los segmentos enriquecidos ha sido esencial para mejorar el recall en el servicio de búsqueda. La extracción de conceptos hermanos, de conceptos lógicamente relacionados, ascendentes y descendentes durante la etapa de segmentación enriquecida, ha contribuido a realizar 8, 6, 9 y 7 búsquedas de forma correcta, respectivamente.

En vista de que nuestro sistema alcanzó un recall alto sin sacrificar la precisión, se puede concluir que los segmentos enriquecidos representan adecuadamente el contenido semántico de los arquetipos, y por tanto, el grado de expansión implementado durante la segmentación enriquecida ha sido apropiado y suficiente.

Tabla 6.2: Resultados de los experimentos con el servicio de búsqueda semántica y con el buscador textual del repositorio openEHR

Método de búsqueda	Recall	Precisión
Buscador textual del repositorio openEHR (operador AND)	5/55 (9.1%)	4/5 (80.0%)
Buscador textual del repositorio openEHR (operador OR)	22/55 (40.0%)	21/22 (95.5%)
Nuestro servicio de búsqueda semántica	38/55 (69.1%)	37/38 (97.4%)

Se ha creado una aplicación web que permite probar las búsquedas semánticas en el grupo de arquetipos seleccionados para los experimentos[5]. A continuación, exponemos dos ejemplos de consultas, incluidas en los experimentos, en los que nuestro servicio de búsqueda semántica ha demostrado ser más eficaz que el buscador de openEHR.

Ejemplo 1: Mejorando la búsqueda del término clínico 'respiratory frequency' con sinonimia de SNOMED-CT

El experto, durante la creación del conjunto de datos de evaluación, asignó el arquetipo 'respirations' como relevante ya que identificó un término local en el arquetipo equivalente al término de búsqueda 'respiratory frequency'. El término local encontrado tiene id 'at0004', nombre 'Rate' y descripción 'The rate of respirations'.

Ahora, explicamos brevemente qué sucede cuando buscamos el término clínico 'respiratory frequency' en ambos sistemas:

[5] `http://archetypes-finder.appspot.com/`

El buscador de openEHR se limita a buscar el término ‘respiratory frequency’ en las definiciones de términos locales de los arquetipos del repositorio. El buscador no es capaz de detectar ninguna equiparación léxica, por lo que no propone ningún arquetipo relevante al usuario. En este caso, este buscador no es eficaz ya que el término de búsqueda es muy distinto léxicamente al término ‘Rate’, marcado por el experto como equivalente.

Nuestro servicio de búsqueda semántica, siguiendo el flujo de la figura 6.5, es capaz de determinar que el arquetipo ‘respirations’ es relevante para el término ‘respiratory frequency’.

– Primero, el servicio mapea el término consulta al concepto SNOMED-CT con id ‘86290005’. Dicho concepto tiene varias descripciones sinónimas asociadas: ‘respiratory rate’, ‘breathing rate’, ‘respiratory frequency’, la última es exactamente igual al término buscado.

– Seguidamente, el servicio busca el concepto con id ‘86290005’ en los segmentos enriquecidos de los arquetipos, encontrándolo en el segmento del arquetipo ‘respirations’. Este concepto está presente en dicho segmento ya que ha sido equiparado al término local con id ‘at0004’ y nombre ‘Rate’ del arquetipo ‘respirations’ durante la etapa de mapping de la segmentación (ver sección 6.2.1 para más detalles sobre la etapa de mapping).

– Finalmente, el servicio notifica al usuario los arquetipos relevantes. La figura 6.6 muestra la salida que ofrece la aplicación web de búsqueda semántica en este ejemplo. Vemos que la aplicación informa al usuario de que el concepto mapeado al término buscado, a su vez, ha sido equiparado al término ‘rate’ del arquetipo ‘respirations’.

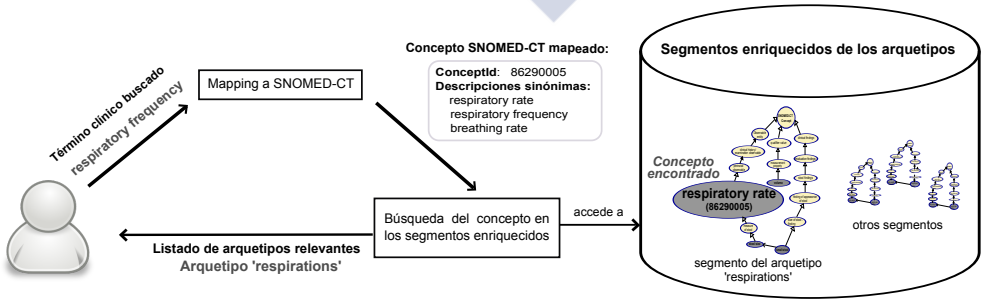

Figura 6.5: Funcionamiento interno del servicio de búsqueda semántica con el término de búsqueda ‘respiratory frequency’

Figura 6.6: Salida que ofrece la aplicación web de búsqueda semántica con el término 'respiratory frequency'

Ejemplo 2: Mejorando la búsqueda del término clínico 'hypotonia' con relaciones semánticas

El experto seleccionó el arquetipo 'apgar score' como el más relevante para el término clínico 'hypotonia'. El arquetipo 'apgar score' es usado para registrar 5 aspectos de la salud de un recién nacido: esfuerzo respiratorio, frecuencia cardiaca, reflejos, tono muscular y color de la piel. El experto considera relevante el arquetipo 'apgar score' ya que 'hypotonia' es un posible hallazgo del tono muscular.

El buscador openEHR no propone ningún arquetipo relevante ya que el término 'hypotonia' no aparece explícitamente en ningún arquetipo.

Nuestro servicio de búsqueda semántica, siguiendo el flujo de la figura 6.7, es capaz de determinar que el arquetipo 'apgar score' es relevante para el término 'hypotonia':

– Primero, el servicio mapea el término buscado al concepto SNOMED-CT con id '398152000'. Dicho concepto tiene una descripción preferida 'poor muscle tone (finding)' y varias descripciones sinónimas asociadas: 'low muscle tone' y 'hypotonia'.

– Posteriormente, el servicio busca el concepto mapeado en todos los segmentos enriquecidos de los arquetipos, encontrándolo en el segmento del arquetipo 'apgar score'.

  En este ejemplo, el concepto 'poor muscle tone (finding)' con id '398152000' no está mapeado directamente a ningún término local del arquetipo 'apgar score', pero este ha sido añadido al segmento de 'apgar score' debido a que está conectado a través de la relación lógica 'interprets' al concepto 'muscle tone', el cual sí que está mapeado directamente al término local del mismo nombre (ver etapa de extracción de conceptos relacionados en la sección 6.2.1 para más detalles).

– Finalmente, la aplicación web de búsqueda semántica muestra al usuario el resultado de la búsqueda con el término 'hypotonia' (ver figura 6.8). Vemos que la aplicación propone como relevante el arquetipo 'apgar score', incluyendo además la siguiente justificación: el concepto mapeado al término buscado ('poor muscle tone (finding)') está conectado a través de la relación de SNOMED-CT 'interprets' a un mapping del arquetipo ('muscle tone (observable entity)').

En este ejemplo, nuestro enfoque logró realizar correctamente la búsqueda gracias al uso de las relaciones semánticas de SNOMED-CT. Estas relaciones le permitieron descubrir que el término buscado, 'hypotonia', es un hallazgo clínico asociado al concepto observable 'muscle tone', el cual está explícitamente modelado en el arquetipo 'apgar score'.

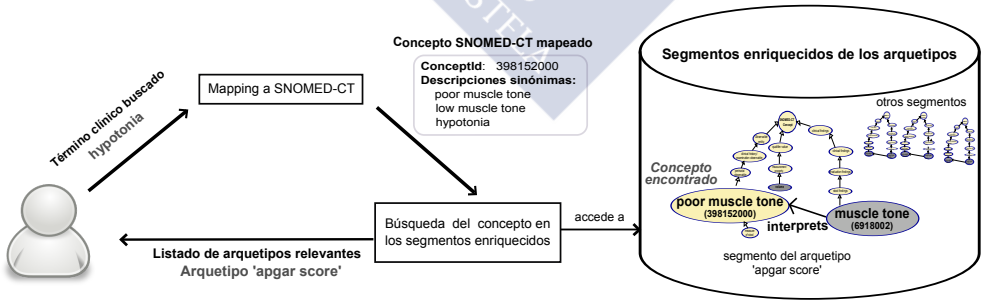

Figura 6.7: Funcionamiento interno del servicio de búsqueda semántica con el término de búsqueda 'hypotonia'

Figura 6.8: Salida que ofrece la aplicación web de búsqueda semántica con el término 'hypotonia'

6.4.4. Servicio de comparación semántica

Se ha creado una sencilla interfaz web[6] para ofrecer al usuario el resultado de las comparaciones semánticas entre arquetipos. La aplicación muestra un informe sobre los arquetipos solapados y relacionados a un arquetipo seleccionado por el usuario.

Vamos a ver un ejemplo de nuestro servicio de comparación semántica con el arquetipo 'baby observations'. Nuestro servicio, compara el segmento enriquecido de este arquetipo con el resto de segmento enriquecidos, buscando: conceptos semilla comunes, relaciones entre conceptos semilla, y conceptos comunes en los segmentos. La figura 6.9 muestra las relaciones y solapes detectados por nuestro servicio entre el segmento de 'baby observations' y el resto de segmentos.

Por una parte, el servicio ha identificado dos arquetipos solapados ('faeces' y 'wellbeing') con el arquetipo 'baby observations' debido a que comparten conceptos semilla en sus respectivos segmentos. El arquetipo 'faeces' y 'baby observations' comparten el concepto semilla:

[6] `http://medicalconcepts-finder.appspot.com/`

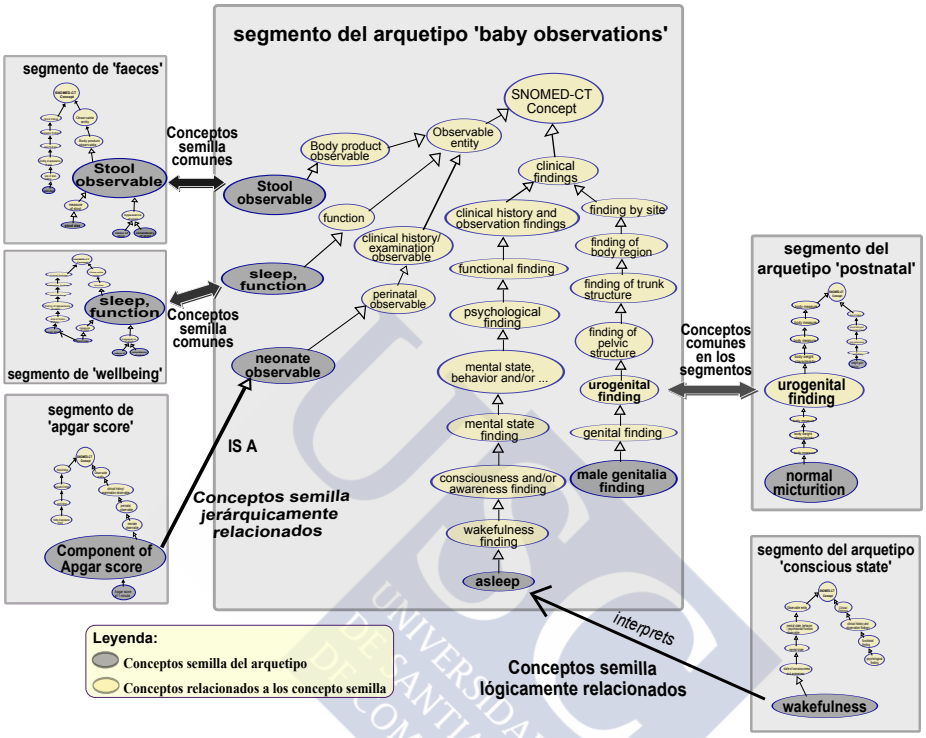

Figura 6.9: Relaciones y solapes detectados por el servicio de comparación semántica entre el segmento de 'baby observations' y el resto de segmentos.

'stool observable', 'soft stool' y 'liquid stool'. Mientras que 'wellbeing' y 'baby observations' comparten 'sleep function'.

Por otra parte, el servicio ha identificado varios arquetipos relacionados: entre ellos 'apgar score' y 'urine output'. El arquetipo 'apgar score' se considera relacionado a 'baby observations' ya que los segmentos enriquecidos de ambos arquetipos contienen conceptos semilla que están relacionados jerárquicamente. Concretamente, el concepto 'component of apgar score' (perteneciente al segmento de 'apgar score') es hijo de 'neonate observable' (perteneciente al segmento de 'baby observations').

La figura 6.10 muestra parte del informe textual generado por la aplicación web sobre los arquetipos solapados y relacionados al arquetipo 'baby observations'.

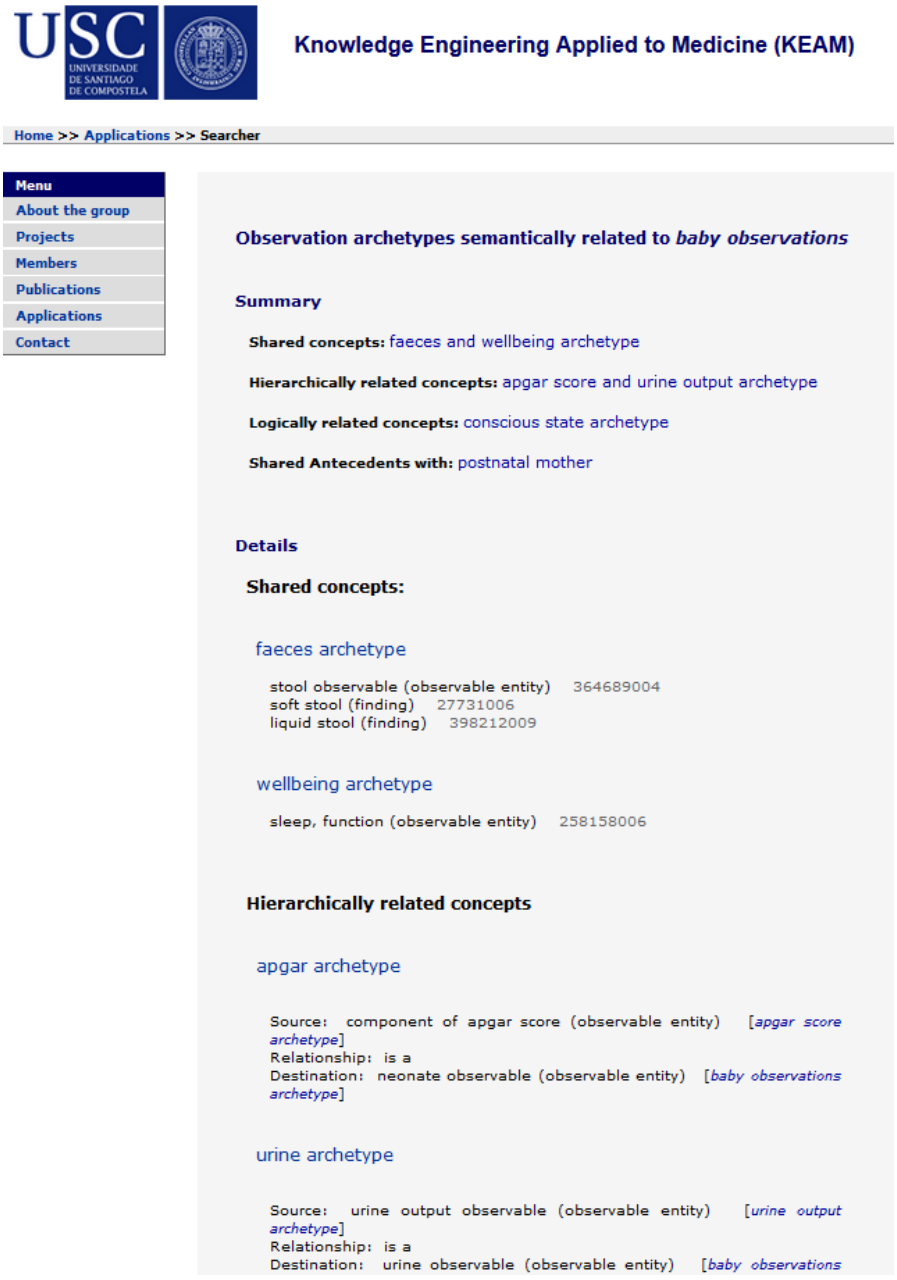

Figura 6.10: Salida generada por la interfaz web sobre los arquetipos solapados y relacionados al arquetipo 'baby observations'

6.5. Discusión

6.5.1. Trabajo relacionado

Trabajo relacionado en segmentación ontológica

El crecimiento, en tamaño y complejidad, de las terminologías y ontologías ha provocado el surgimiento de un área independiente de investigación conocida como segmentación o modularización ontológica. Esta tiene el objetivo de extraer porciones autocontenidas de ontologías adecuadas a necesidades particulares. Estas porciones de ontologías suelen denominarse sub-ontologías, módulos o segmentos ontológicos y tienen la característica de ser ontologías válidas en sí mismas, sin referenciar a la ontología fuente. Las sub-ontologías han sido usadas para favorecer el procesamiento de las ontologías por las aplicaciones (p.e. para aplicaciones de razonamiento o inferencia), para la anotación de datos y para generar vistas de una ontología comprensibles por humanos [48, 122, 112, 134, 91].

Varias metodologías, entre ellas PROMPT [91], MOVE [14] y la investigación de Seidenberg y Rector [112], han sido propuestas para la segmentación ontológica, produciendo porciones autocontenidas de ontologías desde un ontología fuente. Todas ellas comparten la noción de clausura transitiva de un concepto. Sin embargo, sólo la investigación de Seidenberg y Rector [112] usa la meta-información de la semántica de la ontología para automatizar en mayor grado el proceso de extracción y para producir un mejor segmento. Una comparación más detallada sobre estas metodologías puede ser consultada en [112].

Sari et al. usaron MOVE como framework para extraer sub-ontologías de SNOMED-CT con el objetivo de representar el contenido semántico de los arquetipos [110]. En este trabajo, un usuario experto ha seleccionado los conceptos semilla de partida. Es decir, el método dispone de un conjunto dado de conceptos semilla, como suele ser habitual en la segmentación ontológica [112, 80]. Sin embargo, en nuestro trabajo las semillas no estaban disponibles, por lo que adicionalmente se han aplicado técnicas de mappings para obtenerlas, mediante la equiparación de términos locales de arquetipos a conceptos SNOMED-CT. Nuestras técnicas de mapping han demostrado ser efectivas en esta tarea, logrando un 95.8 % de precisión y 75.9 % de recall (ver sección 5.4 para más detalles sobre el rendimiento de las técnicas de mapping). Por tanto, nuestro enfoque tiene la ventaja de no requerir la intervención de un experto en ninguna etapa de la segmentación.

López-García et al. han propuesto extraer segmentos de SNOMED-CT para anotar informes de alta de cardiología [80]. Los segmentos fueron extraídos con 5 técnicas diferentes:

4 técnicas heurísticas de recorrido de grafo y una técnica basada en lógica. Posteriormente, los segmentos fueron filtrados con información de frecuencia de MEDLINE. Aunque López-García et al. afirmaron que los segmentos extraídos son valiosos, reconocieron que estos tenían un tamaño excesivo. Por ejemplo, sus técnicas de recorrido de grafo obtuvieron segmentos de un tamaño medio del 17 % al 51 % respecto al tamaño total de SNOMED-CT. Nuestro trabajo también aplicó técnicas de recorrido de grafo y sin embargo extrajo subconjuntos mucho más manejables (inferiores de media al 1 % de SNOMED-CT). Dos razones podrían explicar la diferencia en el tamaño de los segmentos: (1) nuestras técnicas incluyen filtros heurísticos para descartar conceptos poco relevantes o muy genéricos (ver sección 6.2.1) y (2) hemos comprobado que los concepto semilla de un arquetipo están fuertemente relacionados, por lo que tienden a estar concentrados en las mismas jerarquías de SNOMED-CT. Esto favorece a la extracción de segmentos más pequeños.

Trabajo relacionado en búsqueda semántica en el ámbito clínico

Recientemente, algunos estudios han usado ontologías biomédicas para explorar la búsqueda semántica en el ámbito clínico [71, 42]. Koopman et al. [71] propusieron un enfoque novedoso para buscar información dentro de los EHRs basado en la equiparación de conceptos en lugar de la equiparación de palabras clave (enfoque clásico). En este enfoque, Koopman et al. transformaron las consultas y el texto plano de los documentos de los EHRs a conceptos SNOMED-CT, usando herramientas de UMLS, e implementaron un motor de recuperación basado en la búsqueda de equiparaciones de conceptos entre las consultas y los documentos. Una evaluación realizada con una colección real de EHRs mostró que su sistema basado en conceptos mejoró la precisión en un 30 % con respecto a un sistema basado en palabras clave. La principal diferencia de nuestro enfoque con respecto al trabajo de Koopman et al. es que nosotros además de mapear los términos clínicos de los EHRs a conceptos SNOMED-CT, usamos las relaciones de SNOMED-CT para expandir dichos conceptos con otros del entorno cercano de SNOMED-CT. Esto nos ha permitido aumentar la cobertura de nuestro servicio de búsqueda sin pérdida de precisión.

Fernandez-Breis et al. también desarrollaron un sistema para realizar búsquedas semánticas en arquetipos [42]. Sin embargo, el sistema requiere una costosa etapa previa de anotación manual de los arquetipos con conceptos SNOMED-CT. Nosotros hemos propuesto un método automático para anotar los arquetipos con conceptos SNOMED-CT y crear segmentos en torno a estos conceptos.

Trabajo relacionado en comparación semántica de arquetipos

Lezcano et al. han propuesto un método para categorizar los arquetipos por medio de conceptos UMLS [78]. Lezcano et al. han usado las relaciones del metathesaurus de UMLS para identificar las intersecciones y similitudes entre arquetipos. La salida de su método es un grafo para reflejar la conectividad o similitud entre cada par de arquetipos de un repositorio. Nuestro enfoque anota los arquetipos mediante segmentos SNOMED-CT en vez de con conceptos UMLS y está más centrado en encontrar arquetipos relacionados a una dado, mostrando información detallada sobre los tipos de similitudes semánticas detectadas.

6.5.2. Aportaciones

Las aportaciones y hallazgos de este capítulo se pueden sintetizar en los siguientes puntos:

- Un método automático para representar y anotar semánticamente los arquetipos clínicos mediante la extracción de subconjuntos relevantes de SNOMED-CT.
  - El método propuesto tiene una ventaja importante respecto a la mayoría de los enfoques de segmentación existentes: incluye técnicas de mapping avanzadas capaces de generar automáticamente y con gran precisión los conceptos semilla, necesarios para iniciar la segmentación. La gran mayoría de enfoques existentes parten de un conjunto de conceptos semilla dados por un experto [80, 112].
- Aplicaciones que usan los subconjuntos de SNOMED-CT asociados a los arquetipos para añadir funcionalidades avanzadas a los repositorios de arquetipos. Una de las aplicaciones propuestas plantea un sistema de búsqueda semántica que es capaz de superar algunas de las limitaciones de los buscadores existentes en los repositorios de arquetipos. Dichos buscadores tratan de localizar información mediante una búsqueda textual basada en palabras clave. En contraste, nuestro enfoque es capaz de representar tanto la consulta como los arquetipos mediante conceptos de SNOMED-CT. El sistema de búsqueda semántica propuesto:
  - Mejora en un 70 % en términos de recall a los sistemas de búsqueda basados en palabras clave en tareas de recuperación de información dentro de un repositorio de arquetipos.
  - Aporta dos ventajas importantes:
    - Se aprovecha la **sinonimia** incluida en SNOMED-CT para optimizar las búsquedas y no depender totalmente de la coincidencia de palabras clave. Así,

por ejemplo, si un usuario busca el término 'respiratory frequency' nuestro sistema podría identificar que el término local 'rate of respiration' del arquetipo 'respirations' es equivalente, ya que ambos términos son descripciones sinónimas definidas en el concepto de SNOMED-CT con id 86290005 (ver ejemplo 1 en la sección 6.4.3).

- Se aprovechan las **relaciones semánticas** definidas en SNOMED-CT para enriquecer las búsquedas. Nuestro sistema asigna relevancia a un arquetipo si este tiene términos locales relacionados semánticamente al término de búsqueda. Así, por ejemplo, si un usuario busca el término 'hypotonia' nuestro sistema identifica que el arquetipo 'apgar score' es relevante, ya que incluye términos locales (p.e. 'muscle tone' y 'normal tone') relacionados lógicamente o jerárquicamente en la red de SNOMED-CT a 'hypotonia' (ver ejemplo 2 en la sección 6.4.3).

6.5.3. Trabajo futuro

Los experimentos han demostrado que los segmentos de SNOMED-CT son útiles para añadir funcionalidades de búsqueda y comparación semántica en un grupo los arquetipos. En el futuro sería interesante integrar nuestra metodología de segmentación en los repositorios de arquetipos actuales.

Además, hemos identificado tres potenciales nuevas aplicaciones de la segmentación en SNOMED-CT para arquetipos clínicos: (1) recomendación de nuevo contenido para arquetipos existentes, (2) ayuda para la creación de nuevos arquetipos y (3) agrupación automática de los arquetipos de un repositorio.

Hemos detectado que los segmentos enriquecidos extraídos pueden ser utilizados para recomendar extensiones en los arquetipos existentes. Específicamente, comprobamos que los conceptos pertenecientes a los segmentos enriquecidos que no han sido modelados directamente en los arquetipos son muy buenos candidatos para ampliar o enriquecer los arquetipos. Por ejemplo, en el arquetipo 'respirations' el término 'Depth' tiene sólo asociado 3 valores: 'shallow', 'normal' y 'deep' (ver figura 5.8). Sin embargo, el segmento enriquecido asociado a este arquetipo, gracias al uso de las relaciones SNOMED-CT, contiene una gran cobertura de hallazgos clínicos relacionados a la profundidad respiratoria, entre ellos: 'cannot breathe deeply enough', 'depth of respiration varies' o 'excessively deep breathing', los cuales podrían ser añadidos al arquetipo. En el futuro podría ser interesante implementar herramientas

automáticas para recomendar mejoras concretas en los arquetipos, aprovechando el conocimiento representado en SNOMED-CT.

También, creemos que sería interesante explorar formas de usar el conocimiento médico representado en SNOMED-CT para facilitar la creación de nuevos arquetipos. En este sentido, consideramos que sería útil para los modeladores de arquetipos disponer de un pequeño segmento de SNOMED-CT sobre el concepto clínico a modelar en el arquetipo. Por ejemplo, si un modelador tiene que crear un nuevo arquetipo sobre el colesterol, se podría extraer un segmento SNOMED-CT partiendo del concepto semilla ' cholesterol (substance)', con la metodología propuesta en la sección 6.2.1. El modelador podría consultar dicho segmento para descubrir y reusar los conceptos y relaciones semánticas de SNOMED-CT asociadas al concepto colesterol.

Hemos identificado que los segmentos extraídos pueden ser útiles para agrupar automáticamente el conjunto de arquetipos. La coincidencia de ciertos conceptos genéricos entre diferentes segmentos puede ser una primera aproximación para crear clústers. Por ejemplo, los segmentos correspondientes a varios arquetipos ('height', 'body weight', 'body mass', 'head circumference' y 'waist and hip') comparten el concepto SNOMED-CT 'body measure' por lo que podría crear un clúster. Mientras que otro clúster podría estar formado por los arquetipos 'Urine Output', 'Uterine contractions' y 'Perineum' ya que comparten el concepto 'Urogenital observable'.

6.6. Resumen

En los últimos años han surgido repositorios online donde los arquetipos clínicos son publicados y revisados colaborativamente por diferentes usuarios [92, 90]. El número de arquetipos en los repositorios ha estado constantemente creciendo y todo apunta a que este crecimiento se mantendrá en los próximos años. A pesar de este gran impulso en el modelado de arquetipos, los repositorios todavía no cuentan con herramientas avanzadas para gestionar y buscar información clínica en los arquetipos. Prácticamente, se limitan a ofrecer búsquedas léxicas dentro de las definiciones textuales de los arquetipos.

La falta de herramientas de búsqueda avanzada en los repositorios dificulta la reutilización de arquetipos, causando la creación de nuevos arquetipos solapados y repetitivos. También, complica el proceso de migración y mapping entre modelos de datos clínico propietarios (existentes en muchos sistemas de información clínica actuales) y arquetipos clínicos.

En este capítulo hemos planteado una aproximación para añadir algunas funcionalidades semánticas a los repositorios de arquetipos. En nuestra aproximación, los arquetipos clínicos son anotados semánticamente mediante sub-ontologías de SNOMED-CT. Además, se han implementado dos aplicaciones concretas que usan estas sub-ontologías para identificar relaciones y solapes entre arquetipos, y para gestionar búsquedas semánticas de conceptos clínicos en los arquetipos.

La aplicación de búsqueda semántica fue diseñada con el propósito de buscar arquetipos relacionados a un término clínico (p.e. una enfermedad o un hallazgo clínico). Para ello, la aplicación, en primer lugar, convierte automáticamente el término clínico buscado a un concepto SNOMED-CT, y a continuación lo busca en los segmentos asociados a todos los arquetipos. Es decir, la búsqueda de información se realiza mediante la equiparación de conceptos clínicos, más que mediante la equiparación de palabras clave (método usado en los repositorios actuales).

Los experimentos con la aplicación de búsqueda semántica, realizados con 16 arquetipos clínicos y 55 términos consulta, mostraron resultados prometedores: 38 de las 55 consultas realizadas obtuvieron un resultado relevante, con un recall del 69.1 % y una precisión del 97.4 %, frente al buscador de openEHR que alcanzó 40.0 % de recall y 95.5 % de precisión.

Este capítulo ha demostrado que la anotación semántica de los arquetipos con sub-ontologías SNOMED-CT puede mejorar considerablemente la gestión y búsqueda en los actuales repositorios de arquetipos. Además, se ha planteado que la segmentación en SNOMED-CT con arquetipos clínicos también podría ser útil para otras aplicaciones: en la recomendación de nuevo contenido para arquetipos existentes, en la creación de nuevos arquetipos y en la agrupación automática de los arquetipos de un repositorio.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Capítulo 7

Conclusiones y Trabajo Futuro

Uno de los retos actuales de la informática médica es lograr la interoperabilidad semántica entre los sistemas de información de distintas instituciones sanitarias [121]. La interoperabilidad permitirá que los sistemas de información intercambien y comprendan automáticamente los datos clínicos de los pacientes, facilitando el acceso completo a dichos datos desde cualquier institución sanitaria.

Uno de los pasos necesarios hacia la interoperabilidad semántica es que las terminologías clínicas se incorporen progresivamente en la HCE para representar de una forma estandarizada y consistente la información clínica de los pacientes [29, 121].

Aunque se han hecho algunos avances en el uso de la terminologías, la realidad es que actualmente una gran parte de la información clínica de la HCE es registrada en lenguaje natural y no es enlazada con terminologías del ámbito clínico. El gran tamaño y complejidad de estas terminologías y la falta de herramientas avanzadas que automaticen los procesos de enlazado y de búsqueda de conceptos relevantes son dos de las principales causas que han impedido un mayor uso e integración de las terminologías en la HCE.

7.1. Contribuciones y hallazgos empíricos

En esta tesis, se ha intentado proporcionar una solución para resolver la falta de métodos avanzados para enlazar de forma automática la información clínica de la HCE con terminologías estándar. También, se han planteado aplicaciones que demuestran las ventajas de integrar una terminología de referencia en repositorios de datos clínicos.

Más específicamente, las aportaciones de la tesis se pueden sintetizar en los siguientes puntos:

1. Una extensa recopilación de técnicas orientadas a la búsqueda de conceptos en SNOMED-CT. Esta recopilación puede ser de interés como fuente de consulta para el diseño de futuras herramientas de búsqueda en extensas terminologías clínicas.

2. Una herramienta de búsqueda que combina diferentes estrategias de mapping para facilitar la búsqueda de conceptos relevantes en SNOMED-CT dado un término clínico. La herramienta propuesta:
   - Incluye técnicas innovadoras que hasta nuestro conocimiento no habían sido usadas en los procesos de búsqueda de SNOMED-CT:
     - Una técnica de expansión de consultas avanzada capaz de inferir sinónimos relevantes del ámbito médico analizando el inmenso corpus de descripciones de SNOMED-CT, y con ello, expandir el término buscado con otros términos alternativos con significados similares.
     - Una técnica que explota las relaciones semánticas de SNOMED-CT para adquirir más información sobre el significado de los conceptos.
   - Mejora en más de un 25 % en términos de recall a dos destacados navegadores de SNOMED-CT en tareas de búsqueda automática de conceptos. Gran parte de esta ganancia se obtuvo gracias a la técnica de expansión de consultas.

3. Un método para enlazar automáticamente la información clínica de arquetipos openEHR con conceptos de SNOMED-CT. El método propuesto:
   - Se diferencia de otras herramientas relacionadas en que pone especial énfasis en el uso de la información contextual y estructural implícita en los arquetipos y en SNOMED-CT para mejorar el enlazado entre ambos. Los experimentos mostraron que la inclusión de las técnicas contextuales y estructurales en el método logró aumentar el recall en algo más de un 15 %.
   - Mejora entre un 10 % y un 35 % el recall de otras herramientas relacionadas. Además, logra una mayor precisión ya que reduce considerable el número medio de conceptos candidatos necesarios para enlazar correctamente los arquetipos.
   - Incluye una interfaz gráfica pensada para facilitar a los expertos la revisión y edición de los enlaces creados por el método automático.

4. Un método automático para representar y anotar semánticamente los arquetipos clínicos mediante la extracción de subconjuntos relevantes de SNOMED-CT. El método propuesto tiene una ventaja importante respecto a muchos de los enfoques de segmentación existentes: incluye técnicas de mapping avanzadas capaces de generar automáticamente y con gran precisión los conceptos semilla, necesarios para iniciar la segmentación. La mayoría de los enfoques existentes requieren que un experto seleccione los conceptos semilla [80, 112].

5. Varias aplicaciones que usan los subconjuntos de SNOMED-CT asociados a los arquetipos para añadir funcionalidades avanzadas a los repositorios de arquetipos. Una de las aplicaciones propuestas plantea un sistema de búsqueda semántica orientado a mejorar la recuperación de información en un repositorio de arquetipos. El sistema de búsqueda semántica propuesto:
   - Mapea el término clínico buscado a un concepto SNOMED-CT y usa los segmentos asociados a los arquetipos para realizar las búsquedas.
   - Aporta dos ventajas importantes:
     - Se aprovecha la sinonimia incluida en SNOMED-CT para optimizar las búsquedas y no depender totalmente de la coincidencia de palabras clave.
     - Se aprovechan las relaciones semánticas definidas en SNOMED-CT para enriquecer las búsquedas. Nuestro sistema es capaz de asignar relevancia a un arquetipo si este tiene términos locales relacionados semánticamente al término de búsqueda.
   - Mejora en un 70 % en términos de recall a los sistemas de búsqueda basados en palabras clave en tareas de recuperación de información dentro de un repositorio de arquetipos.

6. Un estudio sobre varias características de los arquetipos clínicos disponibles en los repositorios públicos. En concreto, el estudio incluye un análisis sobre: (1) la frecuencia de las categorías semánticas de los conceptos enlazados con los distintos tipos de fragmentos de los arquetipos, y sobre (2) las relaciones semánticas existentes entre los conceptos enlazados en un mismo arquetipo.

7. Una discusión que recoge varias deficiencias de los modelos de arquetipos y de SNOMED-CT, detectadas durante el transcurso de esta tesis, que han dificultado el enlazado y la integración de estos modelos.

7.2. Conclusiones

La presente tesis proporciona metodologías prometedoras para el enlazado automático de datos clínicos con la terminología SNOMED-CT. Dichas metodologías han incorporado técnicas innovadoras para la búsqueda de conceptos en SNOMED-CT, han integrado eficientemente diferentes tipos técnicas y lo más importante, han conseguido resultados muy positivos, mejorando en diferentes experimentos los resultados de otras herramientas relacionadas, tanto las comerciales, como las desarrolladas por importantes instituciones públicas.

En la actualidad hay un consenso claro que apunta a que las terminologías clínicas deben integrarse en mayor medida en la HCE [29, 121]. Creemos que las metodologías presentadas en esta tesis pueden aportar un significativo valor en futuros procesos de enlazado de información clínica, reduciendo en gran medida la carga de trabajo de los expertos en terminologías.

La tesis también muestra, mediante la implementación de varias aplicaciones piloto, que la terminología SNOMED-CT puede ser usada como un recurso léxico y semántico para facilitar la gestión y la búsqueda en extensos repositorios de datos clínicos.

A continuación, se enumeran algunas lecciones y conclusiones extraídas durante el transcurso de la tesis:

- La combinación de diferentes tipos de técnicas y estrategias de mapping es fundamental para mejorar el rendimiento de las herramientas de búsqueda de conceptos en extensas terminologías clínicas.

- Las técnicas de expansión de consultas mejoran notablemente los resultados de las búsquedas en SNOMED-CT, especialmente cuando no se dispone de información de contexto del término buscado. Se ha comprobado que SNOMED-CT es una gran fuente de conocimiento en la que inferir sinónimos específicos del ámbito médico.

- Las técnicas contextuales y estructurales son especialmente útiles cuando se trata de enlazar términos clínicos procedentes de modelos estructurados, tales como arquetipos openEHR. Se ha demostrado que es factible automatizar el enlazado entre términos de arquetipos y terminologías clínicas como SNOMED CT con una precisión y recall elevada, si se hace uso de la información contextual y estructural implícita en los arquetipos y en SNOMED CT.

– El empleo de técnicas estructurales y de desambiguación permite validar automáticamente los enlaces creados con técnicas léxicas clásicas y reducir de esta manera la participación de expertos en la revisión de los enlaces.

– Se ha detectado que en muchas situaciones la información clínica de los arquetipos está relacionada semánticamente, tal como lo está en SNOMED-CT. Este hallazgo debería ser considerado en el diseño de futuras herramientas de enlazado de arquetipos con terminologías clínicas.

– El contenido semántico de los arquetipos puede ser representado exitosamente con segmentos de SNOMED-CT. Además, es factible que estos segmentos sean generados de forma completamente automática mediante una metodología basada en técnicas de mapping y de segmentación ontológica.

– El uso de segmentos de SNOMED-CT, como forma de representación de arquetipos, facilita el manejo y la gestión de extensos repositorios de arquetipos.

– La búsqueda semántica, apoyada en el uso de conceptos y relaciones SNOMED-CT, mejora notablemente la recuperación de información dentro de repositorios de arquetipos frente a sistemas de búsqueda basados en palabras clave.

7.3. Limitaciones y trabajo futuro

– Las herramientas de enlazado desarrolladas en la tesis (capítulos 4 y 5) no han explotado las estrategias de post-coordinación para conceptos SNOMED-CT. Se han centrado en la búsqueda de un único concepto capaz de representar completamente el término clínico de búsqueda. En el futuro, sería interesante implementar estrategias específicas para mapear automáticamente este tipo de términos a expresiones post-coordinadas de SNOMED-CT.

– Las herramientas desarrolladas están optimizadas para mapear términos de entrada cortos con conceptos de SNOMED-CT. En el futuro, las herramientas podrían ser adaptadas para dar soporte a la anotación automática de textos de entrada largos (p.e. resúmenes de artículos científicos o informes no estructurados de pacientes) con conceptos SNOMED-CT.

– También sería interesante incorporar e integrar las herramientas creadas en entornos relevantes donde se pueda sacar el máximo partido a sus funcionalidades. Por ejemplo, la inclusión de las herramientas a un navegador de SNOMED-CT podría mejorar las funcionalidades de búsqueda de conceptos de dicho navegador. Mientras que la incorporación de las herramientas a un repositorio de arquetipos facilitaría la creación de enlaces entre los términos clínicos de los arquetipos y los conceptos SNOMED-CT.

– Las herramientas de enlazado han sido evaluadas con conjuntos de datos concretos (arquetipos de tipo 'Observación' y términos procedentes de un glosario de procedimientos patológicos). Sin duda sería positivo realizar experimentos adicionales con diversos tipos de datos clínicos para medir el rendimiento de las herramientas desarrolladas en nuevos escenarios.

– Durante la tesis se han identificado nuevas aplicaciones de la segmentación en SNOMED-CT en el ámbito de los arquetipos clínicos (capítulo 6). En concreto, hemos descrito que la extracción y el uso de subconjuntos relevantes de SNOMED-CT podría: (1) ayudar en la creación de nuevos arquetipos, (2) recomendar nuevo contenido clínico en arquetipos existentes y (3) facilitar la agrupación automática de los arquetipos de un repositorio. En el futuro podría ser interesante explorar estas aplicaciones con el objetivo de simplificar la creación y gestión de los arquetipos.

Apéndice A

Evaluación del enlazado de términos clínicos y SNOMED-CT

Este apéndice incluye información extra sobre la evaluación y los resultados obtenidos por la herramienta descrita en el capítulo 4, centrada en localizar conceptos relevantes de SNOMED-CT dado un término clínico.

Detalles sobre la etapa de entrenamiento de la técnica de desambiguación

Se ha usado la implementación de SVM de la librería Kernlab del lenguaje R para entrenar la técnica de desambiguación de las configuraciones HMAS y HMSS. A continuación, se muestra la función y los parámetros concretos utilizados para llevar a cabo los procesos de entrenamiento en ambas configuraciones:

– Parámetros para el entrenamiento de HMAS:

  modeloHMAS <- ksvm(x, y, type='C-svc', kernel='rbf',kpar=list(sigma= 1.0),C= 50.0)

– Parámetros para el entrenamiento de HMSS:

  modeloHMSS <- ksvm(x, y,type='eps-svr', kernel='rbf',kpar=list(sigma= 1.0),C= 20.0)

Términos seleccionados para la evaluación

Para la evaluación de la herramienta de búsqueda en SNOMED-CT se seleccionaron 300 términos de un glosario de procedimientos en patología publicado por la Sociedad Española

de Anatomía Patológica. Dicho glosario incluye además los conceptos SNOMED-CT equivalentes asignados por expertos clínicos [44]. Los términos pueden ser consultados en el siguiente fichero PDF[1]. El fichero contiene 4 columnas con la siguiente información: el término clínico en formato textual, el identificador del concepto SNOMED-CT asociado al término, la descripción textual de dicho concepto en español y la descripción en inglés.

Resultados del mapping automático (experimento 1)

Los mappings creados durante el experimento 1 (centrado en evaluar la capacidad de las herramientas para mapear automáticamente los términos a SNOMED-CT) pueden ser consultados en el siguiente fichero PDF[2]. El fichero incluye los resultados de 5 ejecuciones independientes con 100 términos clínicos seleccionados aleatoriamente. El fichero contiene las siguientes 5 columnas de izquierda a derecha: el término clínico buscado, el concepto de SNOMED-CT asignado por los expertos (el gold standard), el concepto encontrado por el navegador UMLS de la NLM, el concepto seleccionado por el navegador ITServer y finalmente el concepto sugerido por HMAS, nuestra herramienta de mapping automático.

Resultados del mapping semi-automático (experimento 2)

Los mappings creados durante el experimento 2 (centrado en evaluar la capacidad de las herramientas para recomendar a los expertos varias alternativas de mapping) pueden ser consultados en el siguiente fichero PDF[3]. El fichero incluye los resultados de 5 ejecuciones independientes con 100 términos clínicos seleccionados aleatoriamente, y contiene las siguientes 5 columnas de izquierda a derecha: el término clínico buscado, el concepto de SNOMED-CT asignado por los expertos (el gold standard) y los 5 conceptos sugeridos por los navegadores NLM e ITServer y por nuestra herramienta de mapping semi-automático (HMSS).

[1] Términos seleccionados para la evaluación: `http://goo.gl/T3IWzf`
[2] Resultados del experimento de mapping automático: `http://goo.gl/e32GrV`
[3] Resultados del experimento de mapping semi-automático: `http://goo.gl/W32l1y`

Apéndice B

Evaluación del mapping entre arquetipos openEHR y SNOMED-CT

Este apéndice incluye información extra sobre la evaluación y los resultados obtenidos por el método automático descrito en el capítulo 5, orientado a enlazar términos clínicos de arquetipos con conceptos de SNOMED-CT.

Arquetipos seleccionados para la evaluación

Las evaluaciones realizadas en los capítulos 5 y 6 incluyeron arquetipos Observation procedentes de un repositorio surgido a raíz de un proyecto piloto del Servicio de Salud Nacional de Reino Unido (NHS). En la actualidad, este repositorio ya no está disponible. Aunque cabe mencionar que muchos de los arquetipos que poblaban el repositorio NHS pueden encontrarse hoy en día en el repositorio CKM de openEHR [92] (en algunos casos los arquetipos podrían haber sufrido pequeñas modificaciones y actualizaciones). A continuación se enumeran los 25 arquetipos seleccionados para los experimentos:

openEHR-EHR-OBSERVATION.apgar.v4.adl
openEHR-EHR-OBSERVATION.baby_general_observations.v1.adl
openEHR-EHR-OBSERVATION.blood_pressure.v2.adl
openEHR-EHR-OBSERVATION.body_mass_index.v3.adl
openEHR-EHR-OBSERVATION.body_weight.v3.adl
openEHR-EHR-OBSERVATION.conscious_state.v4.adl
openEHR-EHR-OBSERVATION.faeces.v2.adl

openEHR-EHR-OBSERVATION.feeding.v3.adl
openEHR-EHR-OBSERVATION.fetal_movement.v2.adl
openEHR-EHR-OBSERVATION.head_circumference.v2.adl
openEHR-EHR-OBSERVATION.heart_rate.v3.adl
openEHR-EHR-OBSERVATION.height.v4.adl
openEHR-EHR-OBSERVATION.hydration.v3.adl
openEHR-EHR-OBSERVATION.mobility.v4.adl
openEHR-EHR-OBSERVATION.palpation_breast.v1.adl
openEHR-EHR-OBSERVATION.perineum.v1.adl
openEHR-EHR-OBSERVATION.postnatal_mother.v2.adl
openEHR-EHR-OBSERVATION.respiration.v8.adl
openEHR-EHR-OBSERVATION.speech.v2.adl
openEHR-EHR-OBSERVATION.substance_use-tobacco.v7.adl
openEHR-EHR-OBSERVATION.urine_output.v2.adl
openEHR-EHR-OBSERVATION.uterine_contractions.v1.adl
openEHR-EHR-OBSERVATION.visual_acuity.v3.adl
openEHR-EHR-OBSERVATION.waist_hip.v1.adl
openEHR-EHR-OBSERVATION.wellbeing.v3.adl

Ya que el repositorio de NHS ya no está disponible, hemos habilitado una web[1] donde se pueden consultar y descargar los arquetipos enumerados previamente. En la web incluimos dos versiones de los arquetipos: una contiene los arquetipos originales del repositorio NHS con los escasos mappings creados por los propios modeladores de arquetipos. La otra versión incluye estos arquetipos con abundantes mappings a SNOMED-CT elaborados por expertos clínicos. Estos mappings expertos han sido creados como parte de esta tesis con el objetivo de servir de referencia para la evaluación de los métodos automáticos de mapping.

Salida generada por el método automático de mapping

El método automático de mapping genera automáticamente un fichero XML con los mappings encontrados para cada arquetipo. El fichero incluye meta-información del arquetipo (nombre, versión y tipo del arquetipo) e información de los mappings. Para cada mapping se almacena el término del arquetipo mapeado, su identificador local, los identificadores de los

[1] `http://www.usc.es/keam/TermArchetypes/input.html`

conceptos SNOMED-CT, la descripción textual de los mismos y la técnica usada para detectar el mapping. En el fichero XML, también se han incluido los mappings expertos de referencia para facilitar la evaluación tanto manual como automática. Así, en un mismo fichero están los mappings de referencia creados por expertos y los mappings asignados por el método automático. Hemos creado una web[2] para la consulta y descarga de estos ficheros XML.

[2] http://www.usc.es/keam/TermArchetypes/output.html

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Apéndice C

Evaluación de la segmentación y de la búsqueda semántica en arquetipos

Este apéndice incluye información extra sobre la evaluación y los resultados obtenidos por el método de segmentación de SNOMED-CT y por el servicio de búsqueda semántica descrito en el capítulo 6.

Arquetipos seleccionados para los experimentos de búsqueda

openEHR-EHR-OBSERVATION.apgar.v4.adl
openEHR-EHR-OBSERVATION.blood_pressure.v2.adl
openEHR-EHR-OBSERVATION.body_mass_index.v3.adl
openEHR-EHR-OBSERVATION.body_weight.v3.adl
openEHR-EHR-OBSERVATION.faeces.v2.adl
openEHR-EHR-OBSERVATION.feeding.v3.adl
openEHR-EHR-OBSERVATION.fetal_movement.v2.adl
openEHR-EHR-OBSERVATION.heart_rate.v3.adl
openEHR-EHR-OBSERVATION.height.v4.adl
openEHR-EHR-OBSERVATION.respiration.v8.adl
openEHR-EHR-OBSERVATION.speech.v2.adl
openEHR-EHR-OBSERVATION.substance_use-tobacco.v7.adl
openEHR-EHR-OBSERVATION.urine_output.v2.adl
openEHR-EHR-OBSERVATION.uterine_contractions.v1.adl

openEHR-EHR-OBSERVATION.visual_acuity.v3.adl
openEHR-EHR-OBSERVATION.waist_hip.v1.adl

Conjunto de datos de evaluación

La tabla C.1 muestra los 55 términos de búsqueda seleccionados para los experimentos, junto con sus correspondientes respuestas correctas, esto es, un listado de arquetipos relevantes seleccionados por un experto entre los 16 arquetipos que han tomado parte en los experimentos. El experto ha distinguido dos niveles de relevancia: arquetipos muy relevantes y arquetipos con alguna relevancia para los términos buscados.

Tabla C.1: Conjunto de datos de evaluación formado por los términos de búsqueda y los arquetipos relevantes seleccionados por un experto

Término	Arquetipos con mucha relevancia	Arquetipos con alguna relevancia
Cardiac Diseases	body mass, body weight	feeding, height, waist and hip
Infant nutrition	feeding	height, body mass, body weight
Obesity	body mass, body weight	feeding, height, waist and hip
Swallowing Disorders	feeding	body weight
Overweight	body mass, body weight	feeding, height, waist and hip
Underweight	body mass, body weight	feeding, height, waist and hip
Parenteral Nutrition	feeding	body weight
Diet	feeding	height, body mass, waist and hip, body weight
Weight control	body weight	feeding, height, body mass, waist and hip
High blood pressure	blood pressure	heart rate
Cardiac Diseases	heart rate	blood pressure
Irregular Heart beat	heart rate	blood pressure, apgar score
Tachycardia	heart rate	blood pressure, apgar score
Bradycardia	heart rate	blood pressure, apgar score
Breathing Problems	respiration	apgar score, tobacco
Dyspnea	respiration	apgar score
Tachypnea	respiration	apgar score
Respiratory Failure	respiration	
Diarrhea	faeces	feeding
Uterine diseases	uterine contractions	fetal movements
Passive Smoking	tobacco	respiration
paradoxical respiration	respiration	
apneustic breathing	respiration	
sleep-disordered breathing (SDB)	respiration	
stertorous breathing	respiration	
shallow breathing	respiration	apgar score
respiratory sounds	respiration	
respiratory rate	respiration	
respiratory insufficiency	respiration	tobacco
respiratory frequency	respiration	
eating disorders	feeding, body weight	body mass
feeding aid	feeding	body weight
ideal body weight (IBW)	body mass, body weight	feeding, height, waist and hip
soft diet	feeding	faeces
smooth diet	feeding	faeces
cyanosis	apgar score, respiration	
acrocyanosis	apgar score, respiration	
neonatal screening	apgar score	
transient tachypnea of the newborn	apgar score, respiration	
respiratory distress syndrome in the newborn	apgar score, respiration	uterine contractions
blue baby	apgar score, respiration	uterine contractions, heart rate
hypotonia	apgar score	fetal movements, uterine contractions
arterial hypotension	blood pressure	
fatty stool	faeces	feeding
currant jelly stool	faeces	
fetal kick counts	fetal movements	
familial short stature	height	
dwarfism	height	body mass, waist and hip
staccato speech	speech	
slurring speech	speech	
black urine	urine	
incontinence urine	urine	
premature uterine contraction	uterine contractions	
visual agnosia	visual acuity	
accommodation of eye	visual acuity	

Segmentos de SNOMED-CT obtenidos

Los segmentos obtenidos son almacenados en forma de ficheros XML, en los que se incluyen varias secciones con:

- Metainformación del arquetipo de partida (nombre, versión y tipo del arquetipo).
- El conjunto de conceptos que forman el segmento extraído. Para cada concepto se almacena su identificador, sus descripciones textuales y la razón para ser incluido en el segmento (puede ser un concepto semilla, o puede estar conectado a algún concepto semilla a través de relaciones lógicas o jerárquicas de SNOMED-CT).
- El conjunto de relaciones semánticas de SNOMED-CT existentes entre los conceptos del segmento.

Hemos creado varios enlaces para la descarga directa de los segmentos mínimos automáticos[1] y de referencia[2] y para los segmentos enriquecidos[3] obtenidos durante los experimentos.

Información sobre los experimentos de búsqueda en arquetipos

Hemos generado un fichero excel[4] con los resultados de las búsquedas en cada sistema para cada uno de los 55 términos consulta seleccionados para los experimentos. El excel contiene columnas incluyendo la siguiente información:

- El término buscado.
- Los resultados obtenidos por nuestro sistema de búsqueda semántica, incluyendo el concepto SNOMED-CT mapeado y los arquetipos considerados relevantes.
- Los arquetipos considerados relevantes por el sistema de búsqueda de openEHR con los operadores AND y OR.
- Los arquetipos considerados relevantes por un experto.

[1] Segmentos mínimos automáticos: `http://goo.gl/KAlNDq`
[2] Segmentos mínimos de referencia: `http://goo.gl/RrcFJq`
[3] Segmentos enriquecidos: `http://goo.gl/v7KJit`
[4] Disponible en esta web: `http://goo.gl/OeMkSb`

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Bibliografía

[1] Ahmadian, Leila, Ronald Cornet y Nicolette F de Keizer: *Facilitating pre-operative assessment guidelines representation using SNOMED CT*. Journal of biomedical informatics, 43(6):883–890, 2010.

[2] Allones, JL, D Martinez y M Taboada: *Automated Mapping of Clinical Terms into SNOMED-CT. An Application to Codify Procedures in Pathology*. Journal of Medical Systems, 38(10):1–14, 2014.

[3] Allones, JL, María Meizoso, María Taboada, D Martínez y S Tellado: *Combining lexical and structure-based methods to align clinical archetypes to SNOMED CT*. En *Advances in Smart Systems Research, Workshop Papers from KES Conferences*, volumen 2, páginas 27–32. Future Technology Publications, 2012.

[4] Allones, JL, María Taboada, María Meizoso, D Martínez y S Tellado: *Combining mapping methods to align clinical archetypes to SNOMED CT*. En *10th Terminology and Knowledge Engineering Conference (TKE 2012)*, 2012.

[5] Allones, JL, Maria Taboada, D Martinez, R Lozano y María Jesús Sobrido: *SNOMED CT module-driven clinical archetype management*. Journal of biomedical informatics, 46(3):388–400, 2013.

[6] Allones, Jose Luis, David Penas, María Taboada, Diego Martinez y Serafín Tellado: *A study of semantic proximity between archetype terms based on SNOMED CT relationships*. En *Process Support and Knowledge Representation in Health Care*, páginas 98–112. Springer, 2013.

[7] Aronson, Alan R y Thomas C Rindflesch: *Query expansion using the UMLS Metathesaurus.* En *Proceedings of the AMIA Annual Fall Symposium*, página 485. American Medical Informatics Association, 1997.

[8] Atalag, Koray, Hong Yul Yang, Ewan Tempero y Jim Warren: *Model driven development of clinical information sytems using openEHR.* Studies in health technology and informatics, 169:849–853, 2010.

[9] Barrett, Neil, Jens H Weber-Jahnke y Vincent Thai: *Engineering natural language processing solutions for structured information from clinical text: extracting sentinel events from palliative care consult letters.* Studies in health technology and informatics, 192:594–598, 2012.

[10] Batool, Rabia, Asad Masood Khattak, Tae Seong Kim y Sungyoung Lee: *Automatic extraction and mapping of discharge summary's concepts into SNOMED CT.* En *Conference proceedings:... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, volumen 2013, páginas 4195–4198, 2013.

[11] Beale, Thomas: *The GEHR software architecture for a reliable EHR.* En *Toward an Electronic Health Record Europe*, volumen 99, páginas 328–339, 1999.

[12] Benson, Tim: *Principles of health interoperability HL7 and SNOMED.* Springer, 2010.

[13] Berges, I, J Bermudez y A Illarramendi: *Binding SNOMED CT Terms to Archetype Elements. Establishing a Baseline of Results.* Methods of information in medicine, 53(4), 2014.

[14] Bhatt, Mehul, Carlo Wouters, Andrew Flahive, Wenny Rahayu y David Taniar: *Semantic completeness in sub-ontology extraction using distributed methods.* En *Computational Science and Its Applications–ICCSA 2004*, páginas 508–517. Springer, 2004.

[15] Bird, Linda, Andrew Goodchild y Zar Zar Tun: *Experiences with a two-level modelling approach to electronic health records.* Journal of Research and Practice in Information Technology, 35(2):121–138, 2003.

[16] Bodenreider, Olivier: *Issues in mapping LOINC laboratory tests to SNOMED CT*. En *AMIA Annual Symposium Proceedings*, volumen 2008, página 51. American Medical Informatics Association, 2008.

[17] Bodenreider, Olivier y Songmao Zhang: *Comparing the representation of anatomy in the FMA and SNOMED CT*. En *AMIA Annual Symposium Proceedings*, volumen 2006, página 46. American Medical Informatics Association, 2006.

[18] Buck, Jasmin, Sebastian Garde, Christian D Kohl y Petra Knaup-Gregori: *Towards a comprehensive electronic patient record to support an innovative individual care concept for premature infants using theopenEHR approach*. International journal of medical informatics, 78(8):521–531, 2009.

[19] Cao, Feng, Xingzhi Sun, Xiaoyuan Wang, Bo Li, Jing Li y Yue Pan: *Ontology-based knowledge management for personalized adverse drug events detection.* Studies in health technology and informatics, 169:699–703, 2010.

[20] Chen, Rong, Gunnar O Klein, Erik Sundvall, Daniel Karlsson y Hans Åhlfeldt: *Archetype-based conversion of EHR content models: pilot experience with a regional EHR system*. BMC medical informatics and decision making, 9(1):33, 2009.

[21] Chiang, Michael F, John C Hwang, C Yu Alexander, Daniel S Casper, James J Cimino y Justin Starren: *Reliability of SNOMED-CT coding by three physicians using two terminology browsers*. En *AMIA Annual Symposium Proceedings*, volumen 2006, página 131. American Medical Informatics Association, 2006.

[22] Choi, Namyoun, Il Yeol Song y Hyoil Han: *A survey on ontology mapping*. ACM Sigmod Record, 35(3):34–41, 2006.

[23] CliniClue, Navegador. `http://www.cliniclue.com/` (último acceso, septiembre 2014).

[24] d'Aquin, Mathieu, Anne Schlicht, Heiner Stuckenschmidt y Marta Sabou: *Criteria and evaluation for ontology modularization techniques*. En *Modular ontologies*, páginas 67–89. Springer, 2009.

[25] De Silva, Thuppahi Sisira, Don MacDonald, Grace Paterson, Khokan C Sikdar y Bonnie Cochrane: *Systematized nomenclature of medicine clinical terms (SNOMED*

CT) to represent computed tomography procedures. Computer methods and programs in biomedicine, 101(3):324–329, 2011.

[26] Definición del modelo de información de referencia HL7 V3. `http://www.hl7.org/implement/standards/product_brief.cfm?product_id=77` (último acceso, septiembre 2014).

[27] Dentler, Kathrin y Ronald Cornet: *Redundant Elements in SNOMED CT Concept Definitions*. En *Artificial Intelligence in Medicine*, páginas 186–195. Springer, 2013.

[28] Doan, AnHai, Jayant Madhavan, Pedro Domingos y Alon Halevy: *Ontology matching: A machine learning approach*. En *Handbook on ontologies*, páginas 385–403. Springer, 2004.

[29] Dobrev, Alexander, Tom Jones, Veli Stroetmann, Karl Stroetmann, Yvonne Vatter y Kai Peng: *Interoperable eHealth is worth it: securing benefits from electronic health records and ePrescribing*. The European Commission, 2010.

[30] Dogac, Asuman, Tuncay Namli, Alper Okcan, Gokce B Laleci, Yildiray Kabak y Marco Eichelberg: *Key issues of technical interoperability solutions in ehealth and the ride project*. Software R&D Center, Dept. of Computer Eng., Middle East Technical University, Ankara, 6531, 2007.

[31] Dolin, Robert H, Liora Alschuler, Sandy Boyer, Calvin Beebe, Fred M Behlen, Paul V Biron y Amnon Shabo Shvo: *HL7 clinical document architecture, release 2*. Journal of the American Medical Informatics Association, 13(1):30–39, 2006.

[32] Duftschmid, Georg, Thomas Wrba y Christoph Rinner: *Extraction of standardized archetyped data from Electronic Health Record systems based on the Entity-Attribute-Value Model*. International journal of medical informatics, 79(8):585–597, 2010.

[33] EEUU, Navegador de SNOMED-CT de la Librería Nacional de Medicina de. `https://uts.nlm.nih.gov/` (último acceso, septiembre 2014).

[34] Elkin, Peter L, Steven H Brown, Casey S Husser, Brent A Bauer, Dietlind Wahner-Roedler, S Trent Rosenbloom y Ted Speroff: *Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical*

problem lists. En *Mayo Clinic Proceedings*, volumen 81, páginas 741–748. Elsevier, 2006.

[35] Elkin, Peter L, David Froehling, Dietlind Wahner-Roedler, Brett Trusko, Gail Welsh, Haobo Ma, Armen X Asatryan, Jerome I Tokars, S Trent Rosenbloom y Steven H Brown: *NLP-based identification of pneumonia cases from free-text radiological reports*. En *AMIA Annual Symposium Proceedings*, volumen 2008, página 172. American Medical Informatics Association, 2008.

[36] Especificación ISO 18308: Requisitos de la arquitectura de la HCE . `http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=52823` (último acceso, septiembre 2014).

[37] Especificación ISO 21090: Tipos de datos normalizados para el intercambio de información. `http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=35646` (último acceso, septiembre 2014).

[38] Especificación ISO13606-1: Definición del modelo de referencia en los registros electrónicos . `http://www.iso.org/iso/catalogue_detail.htm?csnumber=40784` (último acceso, septiembre 2014).

[39] Especificación ISO13606-2: Definición de intercambio de arquetipos . `http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=50119` (último acceso, septiembre 2014).

[40] Especificación ISO/DIS 13940: Definición de un sistema de conceptos para dar soporte a la continuidad asistencial. `http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=58102` (último acceso, septiembre 2014).

[41] Euzenat, Jérôme, Pavel Shvaiko *y cols.*: *Ontology matching*, volumen 18. Springer, 2007.

[42] Fernandez-Breis, Jesualdo Tomas, Marcos Menarguez-Tortosa, Catalina Martinez-Costa, Eneko Fernandez-Breis, Jose Herrero-Sempere, David Moner, Jesus

Sanchez, Rafael Valencia-Garcia y Montserrat Robles: *A Semantic Web-based System for Managing Clinical Archetypes.* En *Conference proceedings:... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, volumen 2008, páginas 1482–1485, 2007.

[43] Fung, Kin Wah, Olivier Bodenreider, Alan R Aronson, William T Hole y Suresh Srinivasan: *Combining lexical and semantic methods of inter-terminology mapping using the UMLS.* Studies in health technology and informatics, 129(Pt 1):605, 2007.

[44] García-Rojo, Marcial, Christel Daniel y Arvydas Laurinavicius: *SNOMED CT in pathology.* Studies in health technology and informatics, 179:123–140, 2011.

[45] Garde, Sebastian, Petra Knaup, Evelyn JS Hovenga y Sam Heard: *Towards Semantic Interoperability for Electronic Health Records–Domain Knowledge Governance for open EHR Archetypes*. Methods of information in medicine, 46(3):332–343, 2007.

[46] Giannangelo, Kathy y Jane Millar: *Mapping SNOMED CT to ICD-10.* En *MIE*, páginas 83–87, 2012.

[47] Giuse, Dario A y Klaus A Kuhn: *Health information systems challenges: the Heidelberg conference and the future*. International journal of medical informatics, 69(2):105–114, 2003.

[48] Grau, Bernardo Cuenca, Ian Horrocks, Yevgeny Kazakov y Ulrike Sattler: *Modular Reuse of Ontologies: Theory and Practice.* J. Artif. Intell. Res.(JAIR), 31:273–318, 2008.

[49] Grimson, William, Damon Berry, Jane Grimson, Gaye Stephens, Eoghan Felton, Peter Given y Rory O'Moore: *Federated healthcare record server—the Synapses paradigm*. International Journal of Medical Informatics, 52(1):3–27, 1998.

[50] Grupo de interés Terminfo de la organización HL7 : *Guía de uso de SNOMED-CT en HL7 V3*. `http://wiki.hl7.org/index.php?title=Using_SNOMED_CT_in_HL7_Version_3;_Implementation_Guide,_Release_1.5` (último acceso, septiembre 2014).

[51] Hersh, William, Susan Price y Larry Donohoe: *Assessing thesaurus-based query expansion using the UMLS Metathesaurus.* En *Proceedings of the AMIA Symposium*, página 344. American Medical Informatics Association, 2000.

[52] Heymans, Stijn, Matthew McKennirey y Joshua Phillips: *Semantic validation of the use of SNOMED CT in HL7 clinical documents.* J. Biomedical Semantics, 2:2, 2011.

[53] Hina, Saman, Eric Atwell y Owen Johnson: *Secure information extraction from clinical documents using snomed ct gazetteer and natural language processing*. En *Internet Technology and Secured Transactions (ICITST), 2010 International Conference for*, páginas 1–5. IEEE, 2010.

[54] Hina, Saman, Eric Atwell y Owen Johnson: *SnoMedTagger: A semantic tagger for medical narratives*. International Journal of Computational Linguistics and Applications, 4(2):81, 2013.

[55] Hole, William T y Suresh Srinivasan: *Discovering missed synonymy in a large concept-oriented Metathesaurus.* En *Proceedings of the AMIA Symposium*, página 354. American Medical Informatics Association, 2000.

[56] Huang, Kuo chuan, James Geller, Michael Halper y James J Cimino: *Piecewise synonyms for enhanced UMLS source terminology integration*. En *AMIA Annual Symposium Proceedings*, volumen 2007, página 339. American Medical Informatics Association, 2007.

[57] Huang, Kuo Chuan, James Geller, Michael Halper, Yehoshua Perl y Junchuan Xu: *Using WordNet synonym substitution to enhance UMLS source integration*. Artificial intelligence in medicine, 46(2):97–109, 2009.

[58] IHTSDO, Navegador de SNOMED-CT desarrollado por. `http://browser.ihtsdotools.org/` (último acceso, septiembre 2014).

[59] Ingram, D: *The good european health record*. Health in the new communication age, MF Laires, MF Ladeira and JP Christensen (Eds), IOS, páginas 66–74, 1995.

[60] ITServer, Navegador de SNOMED-CT de. `http://www.itserver.es/ITServer/Browser/snomedctbrowser.faces`.

[61] James, Andrew G y Kent A Spackman: *Representation of disorders of the newborn infant by SNOMED CT.* Studies in health technology and informatics, 136:833–838, 2007.

[62] Jiang, Guoqian y Christopher G Chute: *Auditing the semantic completeness of SNOMED CT using formal concept analysis*. Journal of the American Medical Informatics Association, 16(1):89–102, 2009.

[63] Kalra, Dipak, Archana Tapuria, Tony Austin y Georges De Moor: *Quality requirements for EHR archetypes.* En *MIE*, páginas 48–52, 2012.

[64] Kalra, Dipak, Archana Tapuria, Gerard Freriks, F Mennerat, J Devlies *y cols.*: *Management and maintenance policies for EHR interoperability resources*. Q-REC Project IST, 27370(3.3), 2008.

[65] Kashyap, Vipul y Amit Sheth: *Semantic and schematic similarities between database objects: a context-based approach.* The VLDB Journal—The International Journal on Very Large Data Bases, 5(4):276–304, 1996.

[66] Khare, Ritu, Yuan An, Jiexun Li, Il Yeol Song y Xiaohua Hu: *Exploiting semantic structure for mapping user-specified form terms to snomed ct concepts*. En *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, páginas 285–294. ACM, 2012.

[67] Kim, Hyun Young y Hyeoun Park: *Development and evaluation of data entry templates based on the entity-attribute-value model for clinical decision support of pressure ulcer wound management*. International journal of medical informatics, 81(7):485–492, 2012.

[68] Kim, Shine Young, Hyung Hoi Kim, Kyung Hwa Shin, Hwa Sun Kim, Jae Il Lee y Byung Kwan Choi: *Comparison of Knowledge Levels Required for SNOMED CT Coding of Diagnosis and Operation Names in Clinical Records*. Healthcare informatics research, 18(3):186–190, 2012.

[69] Kim, Tae Youn: *Automating lexical cross-mapping of ICNP to SNOMED CT*. Informatics for Health and Social Care, (0):1–14, 2014.

[70] Kooij, Judith van der, WT Goossen, AT Goossen-Baremans, Marinka de Jong-Fintelman y Lisanne van Beek: *Using SNOMED CT codes for coding information in electronic health records for stroke patients.* Studies in health technology and informatics, 124:815–823, 2005.

[71] Koopman, Bevan, Peter Bruza, Laurianne Sitbon y Michael Lawley: *Towards semantic search and inference in electronic medical records: an approach using concept-based information retrieval*. The Australasian medical journal, 5(9):482, 2012.

[72] Lalín Rodríguez, María del Rosario: *Alineamiento y validación de terminologías a gran escala en el ámbito médico*. Tesis de Doctorado, 2012.

[73] Lee, Dennis, Ronald Cornet, Francis Lau y Nicolette De Keizer: *A survey of SNOMED CT implementations*. Journal of biomedical informatics, 46(1):87–96, 2013.

[74] Lee, Dennis, Nicolette de Keizer, Francis Lau y Ronald Cornet: *Literature review of SNOMED-CT use*. Journal of the American Medical Informatics Association, 21(e1):e11–e19, 2014.

[75] Lee, Dennis H, Francis Y Lau y Hue Quan: *A method for encoding clinical datasets with SNOMED CT*. BMC medical informatics and decision making, 10(1):53, 2010.

[76] Lee, Nam Ju y Suzanne Bakken: *Development of a prototype personal digital assistant-decision support system for the management of adult obesity*. International journal of medical informatics, 76:S281–S292, 2007.

[77] Lei Zeng, Marcia y Lois Mai Chan: *Trends and issues in establishing interoperability among knowledge organization systems*. Journal of the American Society for information science and technology, 55(5):377–395, 2004.

[78] Lezcano, Leonardo, Salvador Sánchez-Alonso y Miguel Angel Sicilia: *Associating clinical archetypes through UMLS metathesaurus term clusters*. Journal of medical systems, 36(3):1249–1258, 2012.

[79] Liu, Hongfang, Kavishwar Wagholikar y Stephen Tze Inn Wu: *Using SNOMED-CT to encode summary level data–a corpus analysis*. AMIA Summits on Translational Science Proceedings, 2012:30, 2012.

[80] López-García, Pablo, Martin Boeker, Arantza Illarramendi y Stefan Schulz: *Usability-driven pruning of large ontologies: the case of SNOMED CT*. Journal of the American Medical Informatics Association, 19(e1):e102–e109, 2012.

[81] Lu, Kun y Xiangming Mu: *Query expansion using UMLS Tools for health information retrieval*. Proceedings of the American Society for Information Science and Technology, 46(1):1–16, 2009.

[82] Lusignan, Simon de, Tom Chan y Simon Jones: *Large complex terminologies: more coding choice, but harder to find data–reflections on introduction of SNOMED CT (Systematized Nomenclature of Medicine–Clinical Terms) as an NHS standard*. Informatics in primary care, 19(1):3–5, 2011.

[83] Manning, Christopher D, Prabhakar Raghavan y Hinrich Schütze: *Introduction to Information Retrieval. (Capítulo 6)*, volumen 1. Cambridge university press Cambridge, 2008.

[84] Mao, Ming, Yefei Peng y Michael Spring: *Ontology mapping: as a binary classification problem*. Concurrency and Computation: Practice and Experience, 23(9):1010–1025, 2011.

[85] Meizoso, María, JL Allones, María Taboada, D Martínez y S Tellado: *Automated mapping of observation archetypes to SNOMED CT concepts*. En *Foundations on Natural and Artificial Computation*, páginas 550–561. Springer, 2011.

[86] Meizoso García, María, José Luis Iglesias Allones, Diego Martínez Hernández y María Jesús Taboada Iglesias: *Semantic similarity-based alignment between clinical archetypes and SNOMED CT: an application to observations*. International journal of medical informatics, 81(8):566–578, 2012.

[87] Mikroyannidi, Eleni, Robert Stevens, Luigi Iannone, Alan L Rector *y cols.*: *Analysing Syntactic Regularities and Irregularities in SNOMED-CT.* J. Biomedical Semantics, 3:8, 2012.

[88] Muñoz, Adolfo, Roberto Somolinos, Mario Pascual, Juan A Fragua, Miguel A González, Jose Luis Monteagudo y Carlos H Salvador: *Proof-of-concept design and development of an EN13606-based electronic health care record service*. Journal of the American Medical Informatics Association, 14(1):118–129, 2007.

[89] Muñoz Carrero, Adolfo, Arturo Romero Gutiérrez, Gonzalo Marco Cuenca, Icíar Abad Acebedo, Jesús Cáceres Tello, Ricardo Sánchez de Madariaga, Pablo Serrano Balazote, David Moner Cano y José Alberto Maldonado Segura: *Manual práctico de interoperabilidad semántica para entornos sanitarios basada en arquetipos*. Unidad de investigación en Telemedicina y e-Salud. Instituto de Salud Carlos III - Ministerio de Economía y Competitividad, 2013.

[90] NEHTA Clinical Knowledge Manager. `http://dcm.nehta.org.au/ckm/` (último acceso, septiembre 2014).

[91] Noy, Natalya F y Mark A Musen: *Specifying ontology views by traversal*. En *The Semantic Web–ISWC 2004*, páginas 713–725. Springer, 2004.

[92] OpenEHR Clinical Knowledge Manager. `http://www.openehr.org/knowledge/` (último acceso, septiembre 2014).

[93] OWL, Navegador Snow. `http://www.b2international.com/portal/snow-owl` (último acceso, septiembre 2014).

[94] Patrick, Jon, Yefeng Wang y Peter Budd: *An automated system for conversion of clinical notes into SNOMED clinical terminology*. En *Proceedings of the fifth Australasian symposium on ACSW frontiers-Volume 68*, páginas 219–226. Australian Computer Society, Inc., 2007.

[95] Patrick, Jon, Yefeng Wang, Peter Budd, A Rector, S Brandt, J Rogers, R Herkes, A Ryan y B Vazirnezhad: *Developing SNOMED CT subsets from clinical notes for intensive care service*. Health Care & Informatics Review Online. Open Access, 2008.

[96] Patrick, Jon David, Angela Ryan y Robert Herkes: *Introduction of enhancement technologies into the intensive care service, Royal Prince Alfred Hospital, Sydney*. Health Information Management Journal, 37(1):40, 2008.

[97] Pazos, P: *Ehrgen: Generador de sistemas normalizados de historia clínica electrónica basados en openehr*. En *3er Congreso Argentino de Informática y Salud*, 2012.

[98] Página oficial de la fundación openEHR. `http://www.openehr.org/` (último acceso, septiembre 2014).

[99] Página oficial de la organización HL7. `http://www.hl7.org/` (último acceso, septiembre 2014).

[100] Página oficial de SNOMED-CT administrada por IHTSDO. `http://www.ihtsdo.org/snomed-ct/` (último acceso, septiembre 2014).

[101] Página oficial del Center for eHealth in Sweden. `http://www.cehis.se/en` (último acceso, septiembre 2014).

[102] Página oficial del Comité Europeo de Normalización. `https://www.cen.eu/` (último acceso, septiembre 2014).

[103] Página oficial del National E-Health Transition Authority (NEHTA). `http://www.nehta.gov.au/` (último acceso, septiembre 2014).

[104] Página oficial del National Health Service Connecting for Health. `http://www.connectingforhealth.nhs.uk/` (último acceso, septiembre 2014).

[105] Qamar, Rahil: *Semantic mapping of clinical model data to biomedical terminologies to facilitate interoperability*. Tesis de Doctorado, Citeseer, 2008.

[106] Rector, Alan L *y cols.*: *Clinical terminology: why is it so hard?* Methods of information in medicine, 38(4/5):239–252, 1999.

[107] Rogers, Jeremy y Olivier Bodenreider: *SNOMED-CT: Browsing the Browsers.* En *KR-MED*, 2008.

[108] Ruch, Patrick, Julien Gobeill, Christian Lovis y Antoine Geissbühler: *Automatic medical encoding with SNOMED categories*. BMC medical informatics and decision making, 8(Suppl 1):S6, 2008.

[109] Ryan, Amanda, Peter Eklund y Brett Esler: *Toward the interoperability of HL7 v3 and SNOMED CT: a case study modeling mobile clinical treatment*. 2007.

[110] Sari, Anny Kartika, Wenny Rahayu y Mehul Bhatt: *Archetype sub-ontology: Improving constraint-based clinical knowledge model in electronic health records*. Knowledge-Based Systems, 26:75–85, 2012.

[111] Schloeffel, Peter, Thomas Beale, George Hayworth, Sam Heard y Heather Leslie: *The relationship between CEN 13606, HL7, and openEHR*. HIC 2006 and HINZ 2006: Proceedings, página 24, 2006.

[112] Seidenberg, Julian y Alan Rector: *Web ontology segmentation: analysis, classification and use*. En *Proceedings of the 15th international conference on World Wide Web*, páginas 13–22. ACM, 2006.

[113] Servicios terminológicos de UMLS. `https://uts.nlm.nih.gov/home.html` (último acceso, septiembre 2014).

[114] Sherman, Simon, Oleg Shats, Elizabeth Fleissner, George Bascom, Kevin Yiee, Mehmet Copur, Kate Crow, James Rooney, Zubeena Mateen, Marsha A Ketcham *y cols.*: *Multicenter breast cancer collaborative registry*. Cancer informatics, 10:217, 2011.

[115] Shvaiko, Pavel y Jérôme Euzenat: *Ontology matching: state of the art and future challenges*. Knowledge and Data Engineering, IEEE Transactions on, 25(1):158–176, 2013.

[116] Snoflake, Navegador. `http://www.snoflake.co.uk/` (último acceso, septiembre 2014).

[117] SNOMED-CT Starter Guide. `http://ihtsdo.org/fileadmin/user_upload/doc/` (último acceso, septiembre 2014).

[118] SNOMED-CT Technical Implementation Guide. `http://ihtsdo.org/fileadmin/user_upload/doc/download/doc_TechnicalImplementationGuide_Current-en-GB_INT_20140131.pdf?ok` (último acceso, septiembre 2014).

[119] Späth, Melanie Bettina y Jane Grimson: *Applying the archetype approach to the database of a biobank information management system*. International journal of medical informatics, 80(3):205–226, 2011.

[120] Stenzhorn, H, EJ Pacheco, P Nohama y S Schulz: *Automatic mapping of clinical documentation to SNOMED CT*. Studies in health technology and informatics, 150:228, 2009.

[121] Stroetman, V, Dipak Kalra, Pierre Lewalle, Alan Rector, J Rodrigues, K Stroetman, Gyorgy Surjan, Bedirhan Ustun, Martti Virtanen y P Zanstra: *Semantic interoperability for better health and safer healthcare [34 pages]*. The European Commission, 2009.

[122] Stuckenschmidt, Heiner, Christine Parent y Stefano Spaccapietra: *Modular ontologies: concepts, theories and techniques for knowledge modularization*, volumen 5445. Springer, 2009.

[123] Sun, Jennifer Y y Yao Sun: *A system for automated lexical mapping*. Journal of the American Medical Informatics Association, 13(3):334–343, 2006.

[124] Tapuria, Archana, Dipak Kalra y Shinji Kobayashi: *Contribution of Clinical Archetypes, and the Challenges, towards Achieving Semantic Interoperability for EHRs*. Healthcare informatics research, 19(4):286–292, 2013.

[125] texto, SimMetrics: Librería de técnicas de equiparación de cadenas de. `http://sourceforge.net/projects/simmetrics/` (último acceso, septiembre 2014).

[126] Thomas Beale and Sam Heard: *Archetype Definition Language (ADL) 1.4*. `http://www.openehr.org/releases/1.0.2/architecture/am/adl.pdf` (último acceso, septiembre 2014).

[127] Thomas Beale and Sam Heard: *Archetype Definitions and Principles 1.0.1*. `http://www.openehr.org/releases/1.0.2/architecture/am/archetype_principles.pdf` (último acceso, septiembre 2014).

[128] Thomas Beale and Sam Heard: *OpenEHR Architecture Overview 1.0.2*. `http://www.openehr.org/releases/trunk/architecture/overview.pdf` (último acceso, septiembre 2014).

[129] Thomas Beale and Sam Heard: *OpenEHR Information Model 1.0.2*. `http://www.openehr.org/releases/trunk/architecture/rm/ehr_im.pdf` (último acceso, septiembre 2014).

[130] Varelas, Giannis, Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides GM Petrakis y Evangelos E Milios: *Semantic similarity methods in wordNet and their*

application to information retrieval on the web. En *Proceedings of the 7th annual ACM international workshop on Web information and data management*, páginas 10–16. ACM, 2005.

[131] VTSL (Veterinary Terminology Services Laboratory), Navegador de SNOMED-CT del. `http://vtsl.vetmed.vt.edu/` (último acceso, septiembre 2014).

[132] Wang, Yue, Michael Halper, Hua Min, Yehoshua Perl, Yan Chen y Kent A Spackman: *Structural methodologies for auditing SNOMED*. Journal of biomedical informatics, 40(5):561–581, 2007.

[133] Wang, Yue, Michael Halper, Duo Wei, Huanying Gu, Yehoshua Perl, Junchuan Xu, Gai Elhanan, Yan Chen, Kent A Spackman, James T Case *y cols.*: *Auditing complex concepts of SNOMED using a refined hierarchical abstraction network*. Journal of biomedical informatics, 45(1):1–14, 2012.

[134] Wouters, Carlo: *A formalization and application of ontology extraction*. Tesis de Doctorado, La Trobe University, 2005.

[135] Yu, Sheng, Damon Berry y Jesús Bisbal: *An Investigation of Semantic Links to Archetypes in an External Clinical Terminology through the Construction of Terminological Shadows* . 2010.

[136] Yu, Sheng, Damon Berry y Jesus Bisbal: *Performance analysis and assessment of a tf-idf based archetype-SNOMED-CT binding algorithm*. En *Computer-Based Medical Systems (CBMS), 2011 24th International Symposium on*, páginas 1–6. IEEE, 2011.

[137] Zaninelli, M, A Campagnoli, M Reyes y V Rojas: *The O3-Vet project: Integration of a standard nomenclature of clinical terms in a veterinary electronic medical record for veterinary hospitals*. Computer methods and programs in biomedicine, 108(2):760–772, 2012.

[138] Zetterberg, Carina, Karin Ahlzén, Erika Ericsson y Bengt Kron: *An example of a multi-professional process-oriented structured documentation bound to SNOMED CT.* Studies in health technology and informatics, 180:1215–1217, 2011.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Índice de figuras

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Índice de tablas