


Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

MÉTODO DE NORMALIZACIÓN DE DATOS Y ABSTRACCIÓN DE CONSULTAS BASADO EN ESTÁNDARES MÉDICOS

Tesis Doctoral

Doctorado en Inteligencia Artificial

Autor: Sergio Paraíso Medina

Tutores:

Víctor Maojo García

David Pérez del Rey

MADRID, 2018


Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

MÉTODO DE NORMALIZACIÓN DE DATOS Y ABSTRACCIÓN DE CONSULTAS BASADO EN ESTÁNDARES MÉDICOS

Tesis Doctoral

Doctorado en Inteligencia Artificial

Autor: Sergio Paraíso Medina

Tutores:

Víctor Maojo García

David Pérez del Rey

MADRID, 2018

Tribunal nombrado por el Magfco. y Excmo. Sr. Rector de la Universidad Politécnica de Madrid el día . . . de . . . . de 2018

Presidente D. . . . . .

Vocal 1º D. . . . . .

Vocal 2º D. . . . . .

Vocal 3º D. . . . . .

Secretario D. . . . . .

Suplente 1º D. . . . . .

Suplente 2º D. . . . . .

Realizado el acto de lectura y defensa de la Tesis el día . . . de . . . . . de 2018 en Madrid.

Calificación: . . . . . .

EL PRESIDENTE EL SECRETARIO

LOS VOCALES

RESUMEN

Los avances producidos durante las últimas décadas en la investigación clínica han provocado un aumento en la cantidad de información y en los recursos informáticos disponibles. La introducción de nuevos marcadores y pruebas moleculares han aumentado las posibilidades diagnósticas y terapéuticas, aunque a costa de incrementar los requisitos necesarios para la realización de ensayos clínicos. Los cambios han sido tan relevantes que actualmente la mayor parte de los ensayos y estudios clínicos tienen que ser realizados en distintas ubicaciones o regiones mediante la colaboración de diversos centros que puedan intercambiar sus datos.

Ante la necesidad de intercambio de datos entre distintos centros surge la oportunidad de investigar nuevos métodos de integración que faciliten la recuperación de la información e interoperabilidad entre distintos sistemas y aplicaciones. En este contexto de interoperabilidad clínica, actualmente destacan propuestas de distintas tecnologías y estándares biomédicos que sirvan como punto común en distintos ámbitos. Hasta el momento presente, esta cuestión no ha sido resuelta. Por ello, se plantea la hipótesis de si es posible enriquecer y corregir la representación de datos clínicos, explotando las ventajas y el conocimiento de terminologías y estándares biomédicos, que mejoren su homogeneización. El desarrollo de métodos basados en estándares clínicos puede facilitar el desarrollo de soluciones adaptables a otros sistemas y contextos de la práctica clínica, lo que permitiría un avance en el campo de la integración e interoperabilidad de datos clínicos.

Como respuesta a esta hipótesis, el presente trabajo plantea la tesis de que es posible crear un nuevo método de interoperabilidad entre sistemas clínicos, que combina el diseño de un método de normalización semántica y un método de abstracción de consultas.

Para evaluar experimentalmente esta propuesta, se ha realizado una implementación sobre un conjunto relevante de estándares clínicos. Dicha implementación se ha integrado dentro de una capa de interoperabilidad semántica utilizada para almacenar conjuntos de datos clínicos reales de diversas instituciones en

entornos de investigación. La evaluación de los métodos propuestos ha constado de un análisis de la representación y acceso a los datos mediante los métodos diseñados, así como de su comparación con otros sistemas similares.

El trabajo propuesto en la presente tesis doctoral se enmarca en el área de la informática médica y se ha evaluado finalmente en el marco de varios proyectos de investigación europeos, que han servido como entorno de pruebas y evaluación de los métodos propuestos. Además, la presente tesis doctoral ha generado diversas publicaciones en revistas científicas de impacto y en congresos internacionales durante los años en los que se ha realizado el trabajo.

ABSTRACT

Clinical research advances during the last decades have led to an increase in the amount of information and the available resources and repositories. The introduction of new biomarkers and molecular tests have greatly increased diagnostic and therapeutic capabilities, although at the expense of increasing the requirement complexity for carrying out clinical trials. Most of the clinical trials and studies have to be currently executed in collaboration among different institutions exchanging data.

Given the need for data exchange between different institutions, there is an opportunity to study new integration methods focused on enabling information retrieval and interoperability between different systems and applications. In the clinical interoperability context, there are different biomedical standards and technologies used as a common point in different areas. However, this question has not been yet solved at this moment.

The present work raises the hypothesis of whether it is possible to enrich and correct clinical data representation, exploiting the advantages and knowledge of biomedical terminologies and standards for improving their homogenization. Novel methods based on well-established standards can facilitate the development of adaptable solutions for other systems and clinical contexts. It will facilitate advances in the field of clinical data integration and interoperability. In response to this hypothesis, this work proposes a new method of interoperability between clinical systems, which combines the design of a semantic normalization method and a query abstraction method.

To experimentally evaluate this proposal, an implementation based on a relevant set of clinical standards has been developed. This implementation has been integrated into a semantic interoperability layer used to store real clinical datasets from various institutions in research projects. The evaluation of the proposed methods has been carried out through an analysis of representation and data access using such methods, as well as a functional comparison with similar systems.

The work proposed in this doctoral dissertation is placed in the area of biomedical informatics. It was finally evaluated within the framework of two European research

projects, which served as an ideal environment for testing and evaluating the proposed methods. In addition, the present doctoral dissertation has already led to several publications in high impact scientific journals and International conferences.

AGRADECIMIENTOS

Realizar esta tesis doctoral ha supuesto un gran esfuerzo que no habría sido posible concluirlo sin la ayuda y los consejos de multitud de gente con la que he coincidido durante estos últimos 4 años de trabajo. Aunque me gustaría hacer una dedicatoria personal para la gran mayoría de estas personas en estas páginas, creo que muchas de ellas preferirán que esto sea en persona. Primero me gustaría agradecer todo el esfuerzo y dedicación a mis directores de tesis; Víctor y David, por todos los consejos que he recibido, por vuestra confianza en mí y por vuestro empuje hasta el último de los momentos. En este aspecto, también me gustaría destacar la ayuda recibida en este trabajo y en otros más por Raúl. Asimismo, dar las gracias a todos los miembros del tribunal por acudir como expertos a la defensa de este trabajo y por sacar tiempo para su estudio.

Pero todo este trabajo no podría haber salido adelante sin mi familia, el mayor de los agradecimientos va para ellos, María, Víctor y Gema. Sin vuestra confianza en todo momento en mí y sin todo el trabajo que habéis hecho para que esto saliera adelante. Y como agradecerte Sara todo el apoyo que me has dado y todo el tiempo que hemos compartido y el que nos queda.

También me gustaría dar las gracias a todos los compañeros y amigos del Grupo de Informática Biomédica: Alberto, Diana, Miguel, Ana, Busta, Gema, Maxi, Juanma, Santi, Enrique, Guetón, por todo el tiempo que hemos vivido, por los grandes y los malos momentos, sin vosotros no hubiese sido lo mismo. Finalmente me gustaría agradecer por todo el tiempo que hemos trabajado codo con codo a Anca, Jasper, Brecht, Kristof, Njin Zu y Ahmed, y por todo lo que se puede aprender de grandes investigadores.

Madrid, Julio 2018.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

Capítulo 1

1. INTRODUCCIÓN Y OBJETIVOS

1.1. Introducción

Desde que la informática médica surgiera en la década de 1960 como una disciplina transversal entre la informática y la medicina [1] se han originado diversas líneas de investigación y aplicaciones dirigidas a distintos ámbitos y campos, que varían desde un nivel poblacional (nuevas tecnologías en la salud pública) hasta la informática clínica (centrada en el paciente) y la investigación biológica (bioinformática, centrada en el nivel molecular). Es en este último campo, bioinformática, donde se han producido mayores retos y sinergias con la informática médica, desde que en el año 1990 comenzara el proyecto del Genoma Humano [2].

El proyecto del Genoma Humano fue dirigido por los Institutos Nacionales de la Salud y por el Departamento de Energía de Estados Unidos para investigar la secuencia de bases químicas que definen y componen el ADN humano. Este proyecto publicó su primer borrador oficial en 2001 y culminó en 2003 con la publicación de la secuencia completa de un genoma humano [3]. Éste y otros proyectos relacionados han supuesto una revolución en el área ómica, marcando de esta forma las nuevas líneas a seguir en la

bioinformática, lo que requería nuevas metodologías y enfoques diferentes en las aplicaciones sanitarias [4][5] que comenzaron a crearse, particularmente en los que se ha llamado la medicina traslacional, que ha llevado a la medicina personalizada y la medicina de precisión [6][7].

Desde el año 2003 se produce una gran explosión en la aparición de información biológica, lo que incluye biomarcadores, microarrays, SNPs, etc., con importancia clínica creciente, y que se encuentran cada vez más a disposición de investigadores y médicos. Con esta información ha sido posible avanzar en la relación de la información genética, fisiológica y clínica. Esto puede ayudar a explicar los diversos procesos patológicos y moleculares subyacentes a las enfermedades, ampliando por tanto el espectro de preguntas que deben llevarse a cabo para ayudar en los diagnósticos, prognosis o terapias de distintas enfermedades.

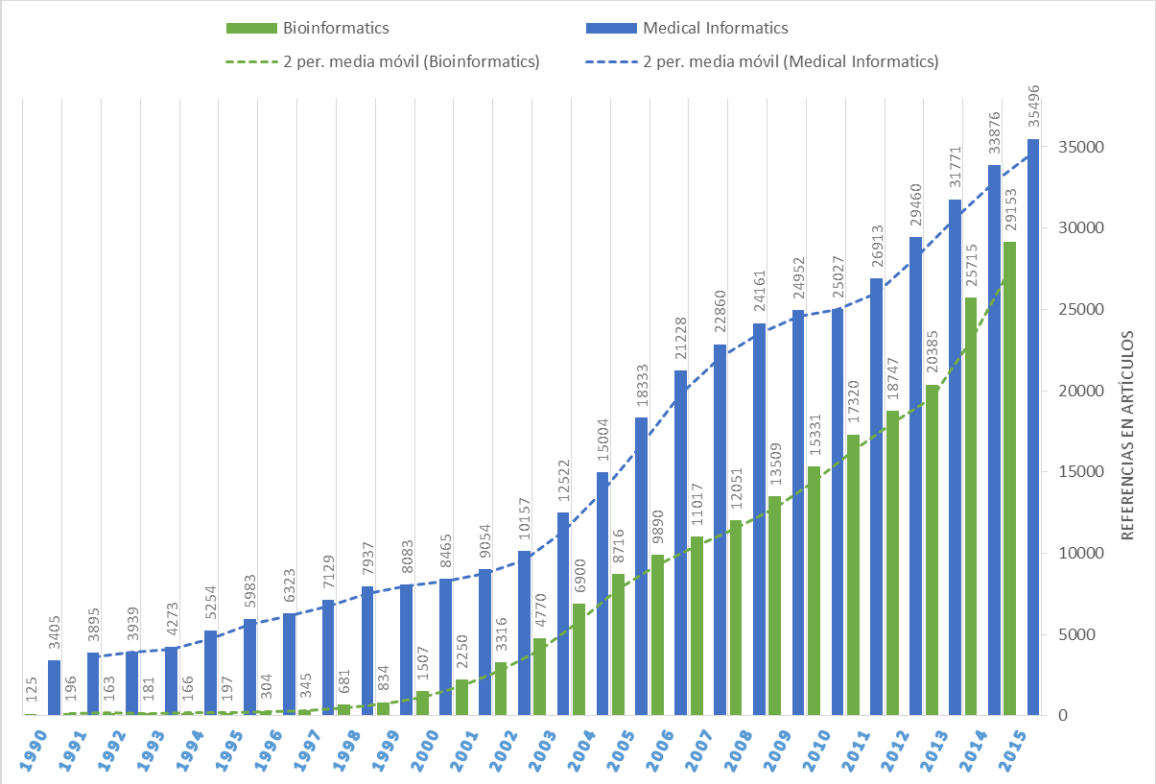


Ilustración 1: Diagrama que muestra la tendencia de los campos "Bioinformatics" y "Medical Informatics" en PubMed durante los últimos 25 años.

Esta explosión en la aparición de información biomédica ha tenido una profunda repercusión en las publicaciones científicas. En la Ilustración 1 se puede ver la tendencia asociada a las áreas de "*Bioinformatics*" y "*Medical Informatics*" en las publicaciones

científicas indexadas en *PubMed* durante los últimos 25 años. En esta gráfica se aprecia un salto generacional en las publicaciones científicas a raíz de los descubrimientos del Genoma Humano, dónde estas publicaciones se han multiplicado por cinco en el caso de la informática médica y por veinte en la bioinformática. Estos datos contrastan con la cantidad de referencias en *Medline* durante estos últimos años como se muestra en la siguiente gráfica. Donde, durante el mismo tiempo no se aprecia un incremento acorde en la cantidad de citas de publicaciones científicas en el área sanitaria, en general, por lo que se puede comprobar así el incremento en las áreas interdisciplinares citadas, la informática médica y la bioinformática.

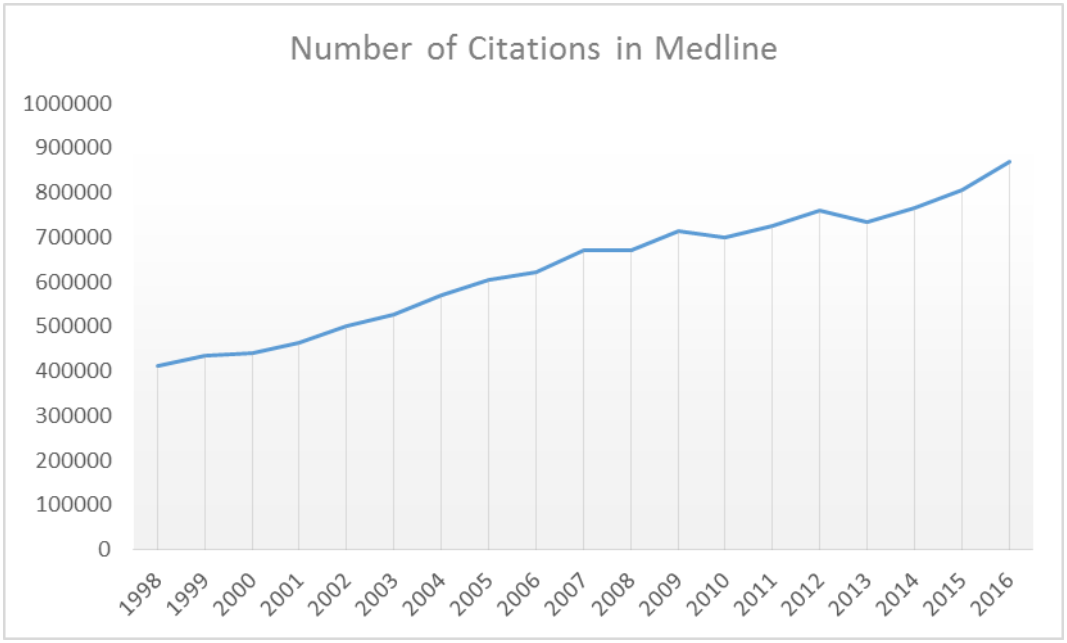


Ilustración 2: Número de citas de publicaciones científicas indexadas en Medline

Este aumento en la cantidad de información y variables introducidas por los últimos avances científicos se ha visto afectado también por la expansión de internet y las nuevas tecnologías durante los últimos años, favoreciendo la aparición de una gran cantidad de bases de datos biomédicas [8]. Estos avances también han supuesto cambios drásticos en la realización de nuevos ensayos clínicos, introduciendo nuevas pruebas moleculares que aumentan la especificidad de los ensayos. Esto ha provocado la necesidad de compartir información clínica entre distintos centros o institutos médicos para poder realizar ensayos clínicos multicéntricos, que aprovechen los avances biomédicos realizados en estas dos décadas [9][10].

Para favorecer este nuevo contexto de compartición de información e interacción de investigadores y clínicos, se introducen variaciones sustanciales en sistemas médicos y clínicos, como las Historias Clínicas Electrónicas (HCE) [11], que además de favorecer la investigación clínica, poblacional o incluso biológica, permitan realizar avances en la integración entre sistemas. Un ejemplo, de gran importancia, ha sido el tratamiento y gestión de heterogeneidades en la información clínica y la distribución de la misma [12]. Los sistemas de información clínica, además de poder ubicarse en distintas localizaciones, pueden presentar distintas heterogeneidades [13] debido a la procedencia o la complejidad de los datos, los vocabularios y modelos de representación, seguridad…etc. Esto requiere un proceso de homogeneización de datos que se ha hecho tradicionalmente —hasta ahora— de manera manual, pero debido al incremento de su complejidad se ha intentado automatizar lo máximo posible [14].

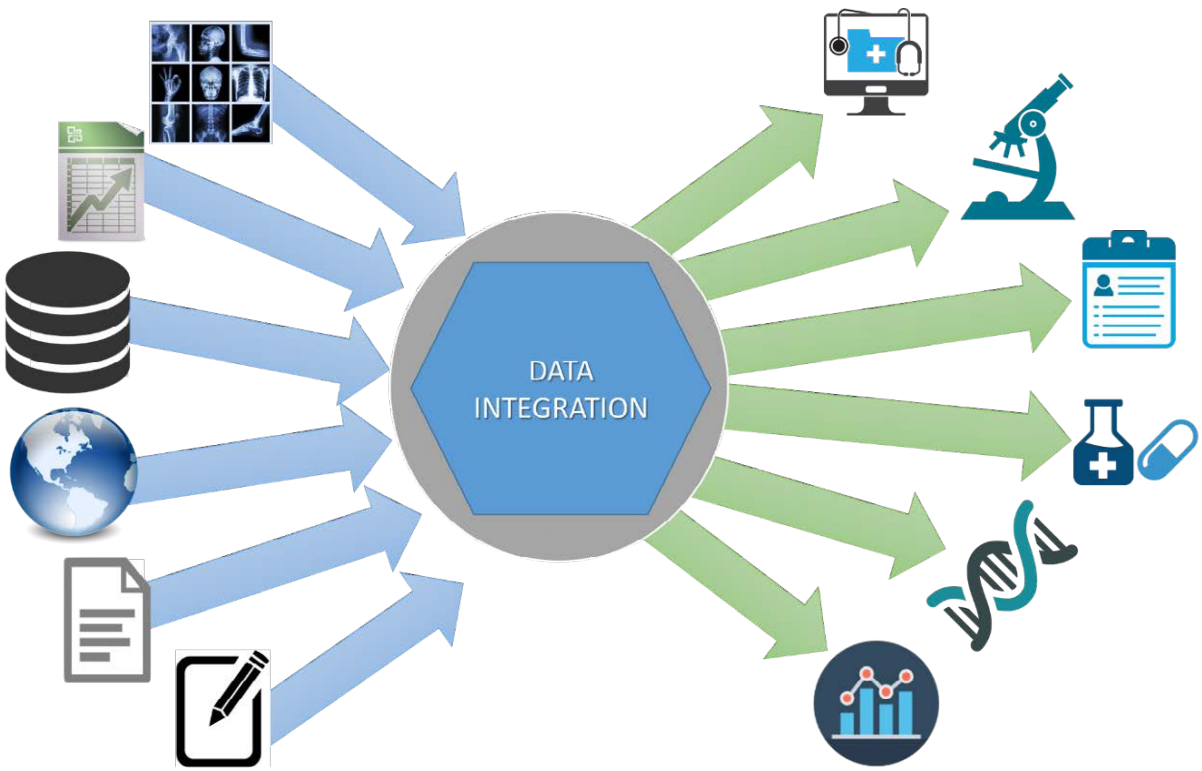


Ilustración 3: Acceso y aplicaciones de las diversas fuentes de datos

En este contexto, se pretende mejorar el acceso a la información clínica por parte de todos los actores implicados. Para abordar este problema, numerosos investigadores han desarrollado distintas metodologías y procesos de integración semántica e interoperabilidad de datos procedentes de multitud de fuentes heterogéneas durante las

dos últimas décadas [15]. A pesar de estos esfuerzos y los avances en la creación, mantenimiento y acceso a los numerosos recursos de información biomédica disponibles, en este momento todavía no existen mecanismos y canales establecidos internacionalmente que aseguren esta integración y que por tanto faciliten la interoperabilidad entre sistemas.

Durante los últimos años las numerosas necesidades en el campo de la bioinformática se han convertido en oportunidades bajo las cuales han surgido multitud de proyectos de investigación a escala europea e internacional. Curiosamente, mientras hay métodos y aplicaciones generales de integración de datos en distintos ámbitos profesionales, no pueden ser adaptados fácilmente al campo de la medicina. De esta forma aumenta la necesidad de desarrollar y adoptar estándares clínicos que faciliten el intercambio de información y su correcta utilización en distintos contextos y circunstancias. Así surgen distintos los tipos de estándares de interoperabilidad clínica, como se verá más detalladamente en los próximos capítulos: estándares de mensajería, terminologías y documentos, así como distintas iniciativas a nivel europeo e internacional que tratan de abordar estos problemas.

1.2. Objetivos e Hipótesis

Como se ha descrito en el apartado anterior, el aumento de la información biomédica disponible durante los últimos años ha representado un problema difícil de tratar, pero ha producido, a su vez, nuevas oportunidades en el mundo de la investigación. Comenzando por el desarrollo de nuevas tecnologías y arquitecturas que facilitaran el tratamiento e integración de estos datos, hasta enfocarse en la necesidad de adaptación de estándares que unifiquen la representación de la información para mejorar su consumo [16].

Surge una necesidad de gran importancia científica y práctica, como es la creación de nuevos métodos que faciliten y enriquezcan la integración de la información clínica en sistemas de información sanitarios [17], reduciendo los errores de representación —y por ende reduciendo los costes.

En este nuevo contexto científico e internacional relacionado con la integración e interoperabilidad entre sistemas clínicos, surgen varias preguntas de investigación, de las que podemos seleccionar las siguientes:

Pregunta científica 1: ¿Es posible homogeneizar la representación de datos clínicos existente mediante un proceso de transformaciones lógicas o normalización basado en las terminologías?

Pregunta científica 2: ¿Es posible abstraer el proceso de la búsqueda de información en un sistema en base del modelo de representación de datos elegido?

De esta forma, teniendo en cuenta lo expuesto anteriormente y tras una investigación a fondo realizada en al área de la informática médica se propone la siguiente hipótesis:

"Es posible enriquecer y corregir la representación clínica de datos mediante un proceso de normalización automática de conceptos médicos y otro de abstracción de consultas que facilite la integración, homogeneización y el acceso a estos sin conocimiento del modelo de datos utilizado"

Para contestar a esta hipótesis, se plantea la investigación de un sistema de interoperabilidad clínica que integre y en definitiva explote el uso de estándares biomédicos de facto de forma que facilite su inclusión en el entorno clínico real y que además que incluya las dos siguientes características:

1. Representación de la información clínica de manera homogénea, que incluya un proceso de normalización de datos que mejore la calidad de esta información.. Este proceso de normalización, además de evitar y corregir errores en la representación inicial de los datos, dotará a los sistemas que lo utilicen de mayor información para facilitar búsquedas y asegurar su trazabilidad.
2. Un método que permita acceder a los datos almacenados en el sistema abstrayendo al sistema del esquema de representación. Este método facilitará la búsqueda de información, encapsulando el esquema de datos y el lenguaje de consultas, para de esta forma y conjuntamente con el método de normalización, mejorar la representación del conocimiento.

Estos métodos partirán de un estudio analítico sobre los distintos sistemas y tecnologías de integración e interoperabilidad existentes haciendo un mayor énfasis en el uso de los estándares biomédicos más relevantes. Por tanto, es importante destacar en este aspecto la necesidad de un riguroso estudio del estado del arte en interoperabilidad clínica para la basar el estudio en las tecnologías más acordes a las necesidades del marco definido.

La hipótesis plantea la necesidad de mejorar el acceso y representación de la información clínica, aprovechando los mecanismos de gestión de conocimiento presentes en estándares biomédicos, además de facilitar y abstraer la búsqueda de esta información. Esta hipótesis se enmarcará dentro del trabajo de los proyectos de investigación europea, lo que facilitará una evaluación del método en distintos escenarios y entornos. La solución propuesta se centra principalmente en los problemas basados en la homogeneización semántica, desarrollando un método basado en procesos de normalización de vocabularios médicos, sincronizados con estándares de interoperabilidad, para facilitar la integración y recuperación de datos.

La metodología a seguir para verificar la hipótesis planteada anteriormente teniendo en cuenta la aplicación de un método hipotético-deductivo como método científico a realizar es el siguiente:

A. Definición del problema, la motivación y las preguntas científicas.
B. Análisis e investigación a fondo del estado de la cuestión.
    i. Estudio y análisis de los fundamentos científicos y así como las iniciativas y necesidades existentes en las áreas relacionadas.
    ii. Recopilación de recursos computacionales y herramientas de libre código destacadas en la integración semántica y vocabularios médicos, así como de la interoperabilidad entre los recursos seleccionados.
    iii. Estudio de la literatura científica e iniciativas relevantes en las áreas de integración semántica, modelos de datos, vocabularios y terminologías médicas. Estrategias de extracción de información y estudio de las distintas terminologías.

C. Diseño de un sistema de integración clínica que mejore la homogeneización de los datos y su interoperabilidad:

i. Integración de estándares y modelos de datos clínicos en el conjunto de terminologías seleccionado. Este diseño se debe basar en un modelo de datos y en un servidor de terminologías que estén basados en estándares clínicos y que cubran la totalidad de los datos clínicos en un área concreta y que nos permita validar la solución.
ii. Diseño de un método automático de enriquecimiento semántico en el servidor de terminologías basada en herramientas de normalización y enlazado de conceptos.
iii. Diseño de un modelo de abstracción de consultas sobre el modelo de datos, basada en la normalización del conjunto de terminologías médicas.

D. Evaluación y validación:

i. Análisis de impacto de la solución diseñada en el área.
ii. Evaluación del trabajo desarrollado en entornos y ecosistemas reales o cercanos a ésta.
iii. Comparativa del marco desarrollado con otros modelos existentes.

Esta metodología científica será seguida durante el trabajo realizado en la presente tesis y su solución será evaluada en el marco de investigación de proyectos europeos. Esta evaluación demostrará la capacidad de implantación de las soluciones diseñadas en un entorno clínico real definiendo a su vez las posibles líneas futuras o mejoras de la solución. Además, se comprobará la validez de la hipótesis expuesta mediante el análisis de los datos obtenidos durante estas validaciones, comparando cuantitativa y cualitativamente la representación de esta información. Se hará especial hincapié en indicadores que muestren la mejora en la forma de acceso de estos datos, la reducción y corrección de errores así como en el aumento de especificidad de los datos representados. El resultado de estas comparativas se medirá mediante la cantidad de recursos accedidos mediante estas metodologías en contraposición de otros modelos y metodologías.

Este trabajo, por tanto, será realizado y por tanto integrado como para de la solución entre las distintas herramientas de libre acceso y métodos diseñados en la ejecución de los proyectos europeos INTEGRATE (FP7-ICT-2009-6-270253) [18] y EURECA (FP7-ICT-2011-7-288048) [19] dentro de los cuales el Grupo de Informática Biomédica (GIB)[1]

[1] http://gib.fi.upm.es/

ha participado activamente y del que forma parte el autor de la presente tesis doctoral. El proyecto INTEGRATE comenzó en 2011 y tuvo como objetivo principal el estudio y desarrollo de una plataforma para el intercambio de datos y conocimiento entre investigadores y especialistas clínicos para fomentar la colaboración en el campo de la biomedicina, más concretamente sobre ensayos clínicos de cáncer. EURECA, por otro lado, comenzó en 2012 y tuvo como objetivo principal la búsqueda e implantación de una interoperabilidad semántica completa, segura y escalable de los datos médicos almacenados en distintos sistemas de registro electrónico mediante distintos casos de uso clínicos.

Durante los casi 5 años que duraron estos proyectos, gran parte de los esfuerzos estuvieron centrados en la investigación y desarrollo de métodos y herramientas que permitieran el acceso homogéneo de fuentes de datos heterogéneos que denominamos Capa de Interoperabilidad Semántica (CIS). Cada uno de estos proyectos tuvo un total de 3 revisiones logrando obtener ambos calificaciones excelentes, y destacando como producto obtenido la CIS como se puede apreciar en los anexos.

1.3. Organización del trabajo

Este apartado describe la estructura de la tesis doctoral. El primer capítulo introduce el contexto y la motivación del presente trabajo; presentando el concepto de informática médica y como la evolución en la investigación de este concepto en conjunto con la bioinformática ha supuesto un aumento en la cantidad de información clínica que es necesaria representar y acceder homogéneamente. En este capítulo también se describen los objetivos a realizar durante este trabajo, así como la hipótesis sobre la que se ha investigado en búsqueda de su verificación.

En el siguiente capítulo se realiza un análisis del estado de la cuestión en varios aspectos claves para este trabajo, como son la integración de datos biomédicos, la interoperabilidad clínica y un estudio sobre otros proyectos de investigación relevantes en estos aspectos, destacando el trabajo previo realizado durante la tesis de Master por el propio investigador [20]. Para la integración de datos biomédicos se analizarán los distintos tipos de heterogeneidad que pueden encontrarse en el área biomédica y los distintos enfoques

de integración de datos posibles según la procedencia de éstos. En el apartado de interoperabilidad clínica se hará una breve introducción sobre este concepto y su necesidad para, posteriormente, estudiar las distintas iniciativas presentes en los tres ejes de la interoperabilidad clínica: estándares de mensajería, terminologías y documentos.

En el tercer capítulo se explican el sistema diseñado como parte de la solución durante el presente trabajo: el método de normalización semántica basada en estándares terminológicos y el modelo de abstracción de consultas para mejorar la búsqueda de información. Para ambos métodos se detallará en profundidad las características y componentes involucrados durante el proceso. El siguiente capítulo, el cuarto, está dedicado a los experimentos realizados, detallando el marco experimental, dónde se han realizado así como los distintos escenarios clínicos donde han sido validados. En el quinto capítulo se discutirán los resultados obtenidos en los experimentos previamente definidos, así como la elaboración de unas comparativas con otros modelos de datos relevantes en el área. Finalmente, en el sexto capítulo se presentan las conclusiones obtenidas después del trabajo realizado en la tesis, además de unas posibles líneas futuras de cara a mantener activa esta investigación durante un futuro próximo

.

Capítulo 2

2. ESTADO DE LA CUESTIÓN

2.1. Introducción

En este capítulo, dedicado al estado de la cuestión, se analizará y describirá la investigación realizada sobre el estado actual de las principales áreas que abarca la presente tesis. Se comenzará analizando la necesidad de la integración de datos biomédicos y cómo han evolucionado las distintas soluciones, sus arquitecturas, cómo éstas tratan de resolver los distintos tipos y niveles de heterogeneidades presentes en la práctica clínica.

Para analizar la integración de datos biomédicos es necesario explicar y clasificar los distintos tipos de heterogeneidades de datos que pueden encontrarse y que dan lugar a distintos enfoques y arquitecturas de integración según la localización o el tipo de estas heterogeneidades.

Una vez descrita la integración de datos biomédicos nos encontramos ante la necesidad de utilizar e intercambiar dichos datos, como se explica en la sección de Interoperabilidad Clínica. Sección que presenta una clasificación de los distintos servicios de interoperabilidad clínica: mensajería, terminología y documentos. En estas subsecciones

se describirán los estándares más representativos de estos servicios explicando las diferencias básicas entre ellos.

Finalmente se realizará un análisis sobre los distintos proyectos de investigación y comerciales más importantes existentes en la actualidad en el área, explicando cómo aborda cada uno de ellos el problema de la interoperabilidad clínica y haciendo un especial hincapié en el trabajo realizado en los proyectos de investigación INTEGRATE y EURECA para abordar los problemas de heterogeneidad presentes en estudios sobre cáncer pecho a través de una capa de interoperabilidad semántica basada en estándares.

2.2. Integración de datos biomédicos

El éxito del Proyecto del Genoma Humano (PGH) generó expectativas para obtener un mayor o un completo conocimiento de las enfermedades humanas y, en general, del mundo de la salud [21]; sin embargo, un paso previo para completar este conocimiento en la práctica de la investigación clínica es la utilización del conocimiento resultante del PGH en otros proyectos y ámbitos. Paso que necesita de una previa gestión e integración de esta gran cantidad de información generada.

Hace más de 20 años, en relación con el PGH, los investigadores comenzaron a obtener y desarrollar grandes colecciones de datos biológicos y métodos para facilitar su gestión y acceso, con el fin de facilitar la investigación biológica y clínica relacionada y asegurar la disponibilidad de estos datos para otros investigadores [22]. En un área de vital importancia clínica y social, como es el cáncer, la integración de diversas fuentes de datos ha sido un paso esencial en proyectos colaborativos que han permitido aunar objetivos internacionalmente, como, por ejemplo, la búsqueda de nuevos biomarcadores para la mejora del diagnóstico y tratamiento en pacientes de cáncer [23][24][25].

Este nuevo paradigma de investigación, junto a los avances en nuevas tecnologías y el impacto de la *World Wide Web* (WWW), provocaron un incremento en la cantidad y disponibilidad de diversos recursos bioinformáticos y de información “ómica”. Recursos

que según la revista anual *Nucleic Acids Research* pasaron de poco más de 200 en el año 2000 [26] a los casi 1.700 actuales[2], con una gran variedad de contenidos y temas.

Al analizar la integración de datos para asegurar su acceso y disponibilidad, surge el problema de la falta de una metodología estándar de representación de las bases de datos [27] que facilite estas tareas. Además existe la dificultad añadida de dotar de semántica a la representación de estos datos [28] convirtiendo el proceso de integración en un problema complejo. En este aspecto, Wong [29] ha definido los cuatro aspectos básicos que cualquier proceso de integración debe cumplir, como son: i) tener un esquema de datos flexible, ii) adoptar un modelo de representación del conocimiento capaz de enlazar recursos externos, iii) tener en cuenta la futura integración de otras fuentes de datos y iv) la adopción de estándares de formatos de datos y consultas.

Estos procesos de integración se realizaban de manera manual hasta hace unos años. Como se ha comentado anteriormente, el auge de la WWW y de nuevas tecnologías ha incrementado el volumen de la información disponible. En este nuevo contexto, los esfuerzos realizados durante las últimas décadas se han centrado en la búsqueda de metodologías que guíen esta integración [30][31] y resuelvan sus dificultades. Éstas son generalmente clasificadas en base a la naturaleza semántica o sintáctica de las heterogeneidades que contienen [13] y a los enfoques y arquitecturas de integración que puedan tratarlas [32].

2.2.1. Heterogeneidades en integración

En la búsqueda de una representación homogénea de la información desde diversas fuentes heterogéneas debe resolverse una serie de problemas durante el proceso. Estos problemas pueden ser debidos a diferentes causas [33], como, por ejemplo, las distintas tecnologías utilizadas, esquemas de representación, estándares, lenguajes o incluso a errores humanos.

A continuación se detallan las posibles heterogeneidades que se pueden encontrar.

[2] http://www.oxfordjournals.org/nar/database/a

2.2.1.1. Sintácticas

Las heterogeneidades sintácticas son aquellas relacionadas con el modo de gestión y acceso a los datos almacenados en las bases de datos. Estas heterogeneidades son generalmente debidas a:

1. Diferentes esquemas de representación y/o tecnologías usadas. Estos problemas vienen derivados de las distintas formas de representar los datos obtenidos. Un ejemplo de esto puede ser una integración de datos estructurados que siguen un modelo de datos relacional con datos que no poseen una estructura en texto plano. Otro ejemplo de estas heterogeneidades puede ser debido a problemas más técnicos, como podrían ser la existencia de distintas arquitecturas software, así como la integración de datos provenientes de sistemas que utilicen distinta codificación de caracteres.
2. Diferentes lenguajes de consulta. Estos problemas surgen cuando se utilizan distintos motores de recuperación de la información en los sistemas, ya que aun estando almacenados los datos siguiendo un mismo esquema de representación, la recuperación de la información sería distinta.
3. Diferentes interfaces para recuperación de la información. Estos problemas son debidos a las distintas formas de acceder y visualizar los datos almacenados en la base de datos.

Las heterogeneidades sintácticas son problemas bastante comunes, a los cuales se han enfrentado los desarrolladores de sistemas desde los inicios de los procesos de integración. Este tipo de heterogeneidades suelen ser tratadas mediante el uso de '*wrappers*' (en castellano envoltorios) software que se desarrollan específicamente para tratar estos problemas a la hora de introducir los datos en la base de datos.

2.2.1.2. Semánticas

Otro tipo de problemas está relacionado con las heterogeneidades provocadas por las distintas formas de representación de datos y por su codificación. Las heterogeneidades semánticas pueden estar a su vez clasificadas en dos categorías, dependiendo de su procedencia.

Semántica de esquema. Son aquellas heterogeneidades debidas a las diferencias en la representación de datos en las distintas fuentes; es decir, debidas a los diferentes modelos de datos. Bergman [34] subdivide esta categoría como se expone a continuación, dependiendo del origen de este problema —sin que sean excluyentes unas de otras:

- Problemas de nombrado: debido al uso de sinónimos, abreviaturas o codificación. Es decir, que se nombre de distinta forma a campos o atributos del esquema de representación.
- Problemas de especificación: debido a la especialización y generalización del esquema de datos. Estos problemas se refieren a cuando en una fuente de datos un campo está dividido en varios campos (especialización) o viceversa (generalización). Un ejemplo frecuente podrían ser el almacenamiento de la fecha, dividida en los campos día, mes y año o todo en un campo fecha.
- Problemas de estructura: estos problemas engloban modificaciones mayores en la estructura del esquema de datos dentro de las distintas fuentes.

La solución a este tipo de problemas pasa siempre por la adopción de un modelo de representación de datos global para todas las fuentes [35]. De esta forma, las distintas fuentes de datos deberían ser transformadas a este esquema global de alguna manera. Esta solución evitaría este tipo de problemas, pero podría resultar en una pérdida de información o exactitud al transformar los datos.

Semántica de instancia. Este tipo de heterogeneidades surgen sobre todo debido las diferencias en la manera de representar datos equivalentes (en general codificación) y en errores en los datos.

En el área de la integración de datos biomédicos es muy común encontrar almacenados campos en texto libre o incluso codificado en algún lenguaje o terminología, cuando todos ellos se refieren al mismo término. Un ejemplo podría ser el caso que se produce cuando en distintas historias clínicas electrónicas se almacena un diagnóstico de cáncer de pecho de una paciente. Este diagnóstico podría estar siendo anotado en texto libre con distintos términos (cáncer de pecho, cáncer, o entrando en la especificidad del tumor, carcinoma o neoplasma). Además, este diagnóstico podría estar anotado con cualquiera de las múltiples terminologías médicas existentes [36].

Otro ejemplo muy común asociado a este tipo de problemas es el caso de las unidades de métrica o, en general, numéricos. Dependiendo de la ubicación, estos datos podrían estar almacenados en distintas unidades (centímetros, metros…etc.) o incluso en distinto orden, como podría ser el caso de las fechas.

Para este tipo de problemas Rahm y Do [37] han definido una metodología para resolverlos. Esta metodología se basa en el análisis y transformación de los datos para su correcta homogeneización. Además, es posible definir procesos de normalización de datos que transformen estos datos en formas de representación homogéneas [38].

2.2.2. Enfoques de integración

En esta sección se analizan los distintos enfoques y arquitecturas existentes para solucionar los problemas de heterogeneidad descritos anteriormente, facilitando un acceso homogéneo a los datos.

La expansión del conocimiento biomédico unido a la reducción de costes tecnológicos y el carácter individualista de cierta parte de la comunidad científica han provocado una gran brecha de información. Sujansky [39] analizó los principales problemas y definió el modelo e implementación de los Sistemas de Bases de Datos Heterogéneos (HDBS en inglés), útil para acceder de una manera homogénea a datos heterogéneos distribuidos. Estos sistemas han ido evolucionando, siendo capaces de mejorar el proceso de integración semántica de la información.

Por otro lado, a finales del siglo XX surgió en el mundo empresarial el modelo basado en repositorios o almacenes centrales (Data Warehouses) [40]. Estos modelos defendían distintos enfoques de integración, debido principalmente a la localización de los recursos a integrar.

El primero de estos enfoques se trata del **Enfoque Centralizado.** Este enfoque se basaría en la idea de Kimball de implementar un *data warehouse* que almacene la información proveniente de las distintas fuentes de datos, como muestra la Ilustración 4. Este enfoque se asocia con los procesos de extracción, transformación y carga (ETL de sus siglas en inglés, *Extract-Transform-Load*) y es el más usado en el mundo empresarial o en centros de una única institución que pueda mantener el control sobre sus propios datos.

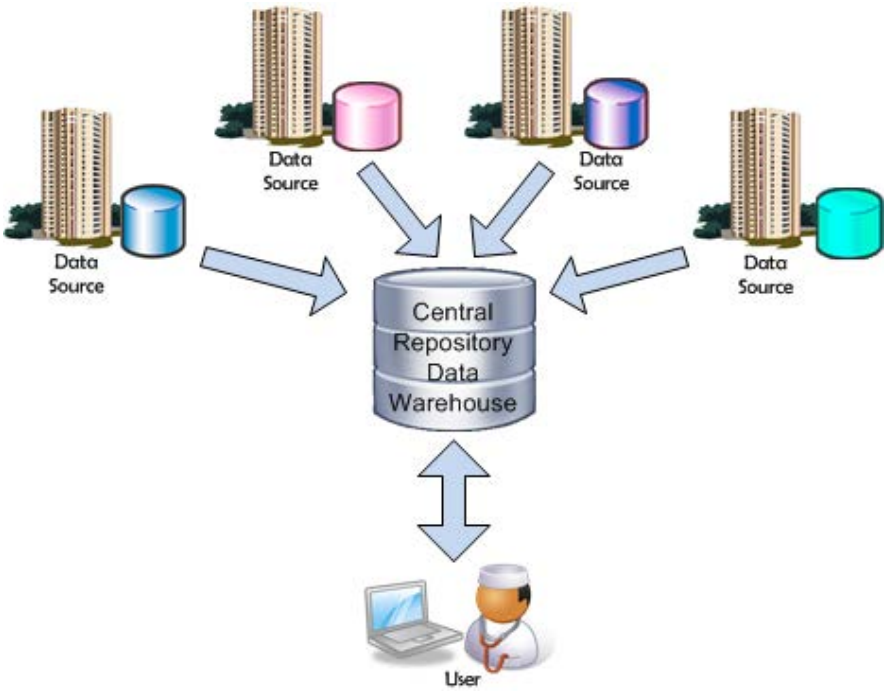


Ilustración 4: Diagrama de Enfoque Centralizado

Entre las ventajas de este enfoque se encontrarían; evitar la redundancia de datos en el repositorio central, eliminar inconsistencias o preservar la integridad de los datos. Por el contrario, entre sus desventajas se encontrarían la pérdida de control sobre los datos (privacidad y seguridad) y el incremento de los costes a la hora de gestionar grandes recursos. Otro aspecto a destacar en este enfoque es la gestión de actualización de datos. A la hora de actualizaciones o cambios en las distintas fuentes de datos el repositorio central se podría volver inconsistente.

El **Enfoque Distribuido** se basa en la idea de acceder a los datos distribuidos en distintas localizaciones. Este sistema distribuido ofrece una interfaz común de acceso a las fuentes de datos, fuentes que al permanecer sin modificar en sus ubicaciones de origen son controladas por las propias organizaciones, al contrario que en el enfoque centralizado. En este enfoque cada fuente de datos puede tener su propio esquema de representación de los datos, por lo que se necesita un proceso de traducción a un esquema virtual y '*mappings'*, como se puede apreciar en la Ilustración 5.

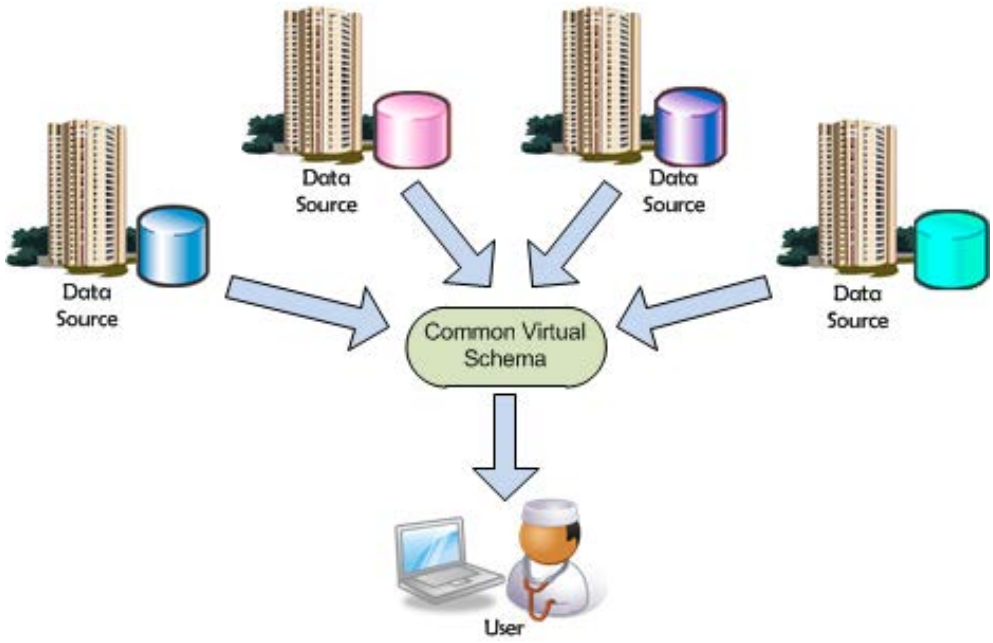


Ilustración 5: Diagrama de Enfoque Distribuido

La principal ventaja de este enfoque es el mantenimiento de la privacidad y la seguridad de los datos por cada organización, además de una cierta mejora en los costes relacionados con el mantenimiento. Este enfoque nos permitiría la división de la carga en distintas máquinas para almacenar los datos, huyendo de grandes estructuras. Otra ventaja importante con respecto al enfoque centralizado es el continuo acceso actualizado a los datos de cada institución, donde además, un fallo en uno de los nodos no supondría un fallo en el sistema. Por otro lado, las principales desventajas son la fiabilidad y la complejidad que conlleva tanto desarrollar el sistema como la incorporación de nuevas fuentes de datos al sistema.

Por tanto, mientras el enfoque centralizado es adecuado para compañías organizadas jerárquicamente, el enfoque distribuido sería más acertado para organizaciones que disponen de nodos en el mismo nivel. Durante varios años, los investigadores biomédicos centraron sus esfuerzos en el desarrollo de sistemas bajo este enfoque [41][42][43], pero la complejidad en la integración de algunas fuentes de datos supuso un cambio en este paradigma.

Tabla 1: Ventajas/desventajas de los enfoques distribuido y centralizado

	Ventajas	***Desventajas***
Enfoque Centralizado	***Complejidad*** - Facilidad de acceso y mantenimiento de los Datos	***Seguridad*** - Los propietarios sienten que pierden el control sobre sus datos
	Seguridad - Ventajas legales	***Concurrencia*** - Si se realizan continuas actualizaciones, el repositorio central podría estar desactualizado
	Complejidad - Mayor facilidad para controlar los datos a largo tiempo	***Eficiencia*** - Posibilidad de cuellos de botella en el acceso a datos
	Complejidad – Mayor rendimiento en la obtención de resultados en consultas sobre grandes volúmenes de datos.	***Falta de estándares*** - No existen metodologías o herramientas de facto.
	Económicos - Debido a la mejor complejidad se disminuyen los costes derivados	***Económicos*** - La disponibilidad de los datos depende de la financiación y mantenimiento de un repositorio central
Enfoque Distribuido	***Seguridad*** - Los propietarios mantienen el control sobre sus datos	***Complejidad*** - Requiere que todos los nodos se involucren en el mantenimiento
	Económicos - Los grupos no dependen de la financiación de una entidad central	***Económicos*** - Los técnicos necesitarán tiempo para especializarse en este sistema
	Concurrencia - Si se realizan actualizaciones no provocarán que la información no se encuentre actualizada	***Rendimiento*** - El rendimiento es mayor cuando los datos se encuentran de manera local
	Seguridad - Una caída de un servidor no supondría la caída del sistema	***Económicos*** - Incrementar la complejidad conlleva incrementar los gastos de desarrollo y mantenimiento
	Económicos - El coste de una red de ordenadores puede ser menor que el coste de un ordenador central	***Falta de estándares*** - Al igual que con el enfoque centralizado, no existen estándares o metodologías de facto

Esta tabla muestra una comparativa más extensa entre las ventajas y desventajas de ambos enfoques extraída de la literatura [13][20].

Por tanto, surge el **Enfoque Híbrido** como una mezcla de estos enfoques, donde se intenta subsanar las desventajas más importantes de los enfoques centralizados y distribuidos manteniendo sus mayores ventajas. De esta forma, el enfoque híbrido se basa

en la idea de transformar los datos de cada fuente de datos en un repositorio ubicado en cada localización. Repositorio que seguirían un modelo común de datos para el sistema siendo necesaria la intervención de procesos ETL desarrollados para cada fuente de datos como se puede apreciar en la Ilustración 6.

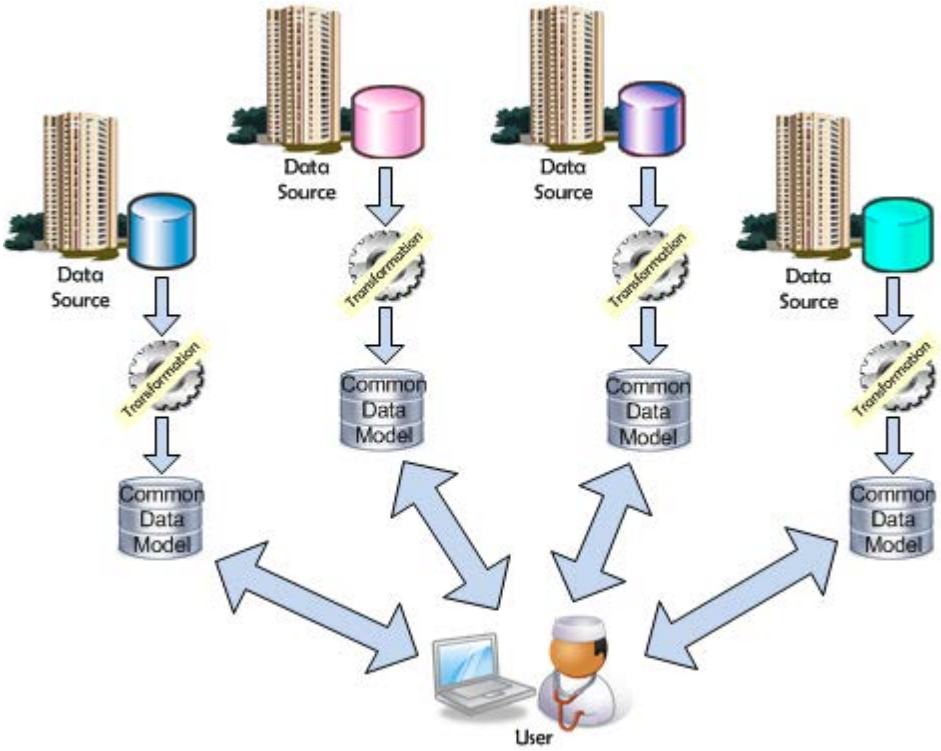


Ilustración 6: Diagrama de Enfoque Híbrido

Con este enfoque las organizaciones mantienen el control sobre sus propios datos, evitando los problemas de privacidad y seguridad que esta pérdida conllevaba. Además, facilita el acceso a los datos al pasar éstos por un proceso de transformación a un esquema común de datos. Por el contrario, las desventajas heredadas son principalmente debidas a los costes del desarrollo y mantenimiento de esta estructura. Debido a sus múltiples ventajas, este enfoque pasar a ser el más adecuado para la comunidad científica en los últimos años [44][45].

Pero también surge una clasificación distinta (pero de alguna forma conectada) sobre enfoques de integración debido al tratamiento de las heterogeneidades. Teniendo 3 categorías principales: i) enlazado de información, ii) traducción de datos y iii) traducción de consultas.

El enfoque de **enlazado de información** surge ante la proliferación de la web 2.0 y en concreto de la web semántica [46]. Este enfoque parte de la idea de enlazar los distintos recursos que existen a través de la web. Ejemplos de este enfoque pueden ser el acceso a recursos como *PubMed*[3] o *MedlinePlus*[4].

La gran ventaja de este enfoque es su facilidad de uso pero sin embargo carece de cierta flexibilidad, ya que los hipervínculos son enlaces unidireccionales. Este enfoque no puede ser adoptado en entornos en los que se necesita de interacción del usuario a la hora de componer búsquedas.

La **traducción de datos** estaría relacionada con las heterogeneidades semánticas y principalmente con los enfoques centralizados e híbrido, concretamente con todos los procesos ETL. Los sistemas que siguen estos enfoques, como se han descrito anteriormente, ofrecen en general un gran rendimiento a la hora de la recuperación de la información.

Si bien, el mayor inconveniente de estos sistemas es el desarrollo de las herramientas ETL que traduzcan los datos ad-hoc. Dependiendo siempre del formato y calidad de los datos de origen para conseguir un resultado óptimo. Esta técnica resulta útil en sistemas cerrados o que no sufren muchos cambios en sus estructuras, ya que estos cambios estructurales supondrían una gran carga a posteriori.

Finalmente, la **traducción de consultas** se relaciona con el enfoque distribuido o híbrido. Este proceso se basa en la traducción de una consulta realizada para un modelo en distintas sub-consultas que siguen esquemas globales o locales [33] para luego integrar los resultados.

Este enfoque se basa en la utilización de capas intermedias de software, '*wrappers*', que traduzcan las consultas a los distintos dominios y que a su vez hagan el proceso inverso en la traducción de resultados [47].

[3] https://www.ncbi.nlm.nih.gov/pubmed/
[4] https://medlineplus.gov/

2.3. Estándares de interoperabilidad

"I have applied everywhere for information, but in scarcely an instance have I been able to obtain hospital records fit for any purposes of comparison. If they could be obtained, they would enable us to decide many other questions besides the one alluded to. They would tell us (…) the exact sanitary state of every hospital and of every ward in it, where to seek for causes of insalubrity and their nature; and, (…) the relative value of particular operations and treatment than we have any means of ascertaining at present. They would enable us, besides, to ascertain the influence of the hospital with its numerous diseased inmates, (…) -or the reverse of all these - upon the general course of operations and diseases passing through its wards; and the truth thus ascertained would enable us to save life and suffering, and to improve the treatment and management of the sick and maimed poor".

Ilustración 7: Retrato de Florence Nightingale de la Universidad de Texas.

Florence Nightingale, pionera de la enfermería moderna y el manejo de información clínica, describió en su libro "*Notes on hospitals*" [48] la necesidad urgente de buscar y adoptar algún sistema de representación estadístico uniforme entre los distintos hospitales. Sistema que debe tener en consideración los registros hospitalarios para evaluar medidas y formas de tratamiento que puedan ayudar a gestionar mejor la información de los pacientes; analizar el curso general de las operaciones, mejorar tratamientos de pacientes o incluso gestionar la influencia de los distintos hospitales.

Estas necesidades se tradujeron en distintos trabajos que catalogaban y unificaban las causas de muerte y tratamientos de los pacientes [49], como veremos en las siguientes secciones, y más adelante en sistemas de gestión para poder guardar un registro electrónico de estas ocurrencias [50]. Si bien llegados a este punto, surge la necesidad no solamente de almacenar un registro de esta información, sino además cómo interpretarla para poder hacer que los sistemas puedan interconectarse entre ellos, dando lugar a la primera definición de interoperabilidad.

Dando lugar en 1990 a una definición del concepto de interoperabilidad [51], que se define como la habilidad de dos o más sistemas o componentes para intercambiar y utilizar la información.

Siguiendo esta definición, para integrar en el marco de la interoperabilidad clínica un sistema o tecnología clínica debe ser capaz de:

- Tener la capacidad de intercambiar información útil entre los sistemas de forma que todos sean capaces de interpretar el conocimiento.
- Tener la capacidad de hacer uso de esta información sin pérdida.

Estos dos conceptos guardan relación con la **interoperabilidad técnica** (capacidad de intercambiar información) y con la **interoperabilidad semántica** (usar la información) siguiendo la clasificación de Benson [52].

La interoperabilidad técnica se basa en la idea de trasladar información de un sistema informático a otro saltando la barrera de la distancia, en un proceso independiente del dominio. Por otro lado, la interoperabilidad semántica se define como la habilidad de los sistemas informáticos de establecer comunicación entre ellos para interpretar la información intercambiada [53].

Además de estos conceptos, aparecen otros niveles de interoperabilidad más relacionados con la parte de la interacción humana y entre las instituciones como son la **a) interoperabilidad de procesos y b) la clínica**. Se entiende como a) la capacidad de las organizaciones de coordinar procesos que puedan trabajar juntos y comunicarse entre ellos. En cuanto a b), la interoperabilidad clínica es la capacidad de transferir a pacientes entre distintas unidades y equipos médicos con el fin de proporcionar cualquier tipo de información [54].

La integración de aplicaciones cumpliendo estos niveles de interoperabilidad y estándares permite resolver problemas frecuentes en los sistemas informáticos como, por ejemplo:

A. **Evitan la fragmentación del conocimiento** y de esta forma proporcionar información útil para otros sistemas, ya que la pérdida de esta información podría afectar a la toma de decisiones.

B. **Evitan duplicidad de la información**. En un sistema incomunicado o no digital sería necesario replicar la información constantemente.

C. **Minimizan la entrada manual de datos**. La introducción manual supone grandes costes adicionales y un riesgo alto de cometer errores.

D. **Garantizan la calidad de los datos** ya que no solamente existen errores en la entrada manual de datos. Sin integración es casi imposible mantener la sincronización de la información entre sistemas.

E. **Reducen costes**.

La diferencia entre los distintos tipos de interoperabilidad es la que da sentido a las tres categorías de estándares clínicos:

- Modelos de datos clínicos, representación e intercambio de datos.
- Terminologías o vocabularios de referencia.
- Servicios de documentos.

En las siguientes subsecciones se describirán los principales estándares y tecnologías existentes en cada una de las tres categorías de estándares. En cada sección, además de realizar una descripción de cada estándar, se detallará las ventajas y desventajas en el ámbito de la presente tesis.

2.3.1. Modelos de datos clínicos e intercambio de datos

Este apartado describe los principales estándares relacionados con la representación de la información clínica en los sistemas electrónicos y por consecuencia en el intercambio de datos biomédicos. En esta sección se consideran los diferentes modelos de datos clínicos, que son aquellos que definen la estructura de representación de la información electrónica para la práctica biomédica y los estándares de mensajería encargados de definir cómo se va a realizar el intercambio de información entre los distintos sistemas.

Para almacenar e intercambiar la información, estos estándares definen una estructura, formato, lenguaje y tipos de datos que establecen unas reglas para poder compartir la información de una forma homogénea. Además, también definen cómo debe ser la comunicación e incluso qué información debe ser almacenada y cómo. La mayor parte de los estándares que se van a describir a continuación definen estas reglas para la práctica

clínica pero también para otros ámbitos como la gestión, cuidados asistenciales, evaluación clínica…etc. en el mundo de la salud, centrándonos sobre todo en los modelos de representación de información clínica más importantes.

Un estudio comparativo del autor analiza varios de los modelos que se van a describir a continuación, así como una primera versión de este enfoque [55].

2.3.1.1. IHE

La organización "*Integrating the Healthcare Enterprise*" (IHE)[5] es una plataforma internacional formada por profesionales, organismos de la salud y expertos en integración que trabajan para mejorar el intercambio de información sanitaria [56]. Para ello IHE desarrolla una serie de perfiles de integración en dominios clínicos basados en estándares ya existentes, con el fin de permitir la integración en los sistemas.

Estos perfiles de integración describen una solución de integración completa basada en un dominio clínico y su contexto. Es decir, específica para cada caso de uso clínico y los actores o usuarios involucrados en él, las distintas transacciones de información posible basada siempre en estándares de facto como DICOM [57], ISO[6] o HL7[7]. Por tanto, estos perfiles definen reglas para definir la interacción entre los sistemas involucrados, y no la implementación de los sistemas.

A fecha de 2018, en IHE se ha desarrollado 195 perfiles de integración divididos en 10 dominios clínicos:

- Cardiología
- Oftalmología
- Infraestructuras tecnológicas
- Patología y pruebas de laboratorio
- Coordinación del cuidado médico
- Dispositivos médicos
- Farmacia
- Investigación, calidad y salud pública
- Oncología
- Radiología

[5] https://www.ihe.net/
[6] https://www.iso.org
[7] http://www.hl7.org/

A pesar de la cantidad de perfiles desarrollados, muchos de ellos no se encuentran en versiones estables para su uso, lo que dificulta la adopción de IHE en la integración de sistemas. Además de este problema, los perfiles de integración son poco flexibles para su implementación técnica [58], debido a que la comunidad IHE se muestra aún poco eficiente a la hora de solucionar estos problemas. A pesar de esto, iniciativas como SemanticHealthNet [59] o epSOS [60] están realizando grandes contribuciones para enriquecer este enfoque [61].

2.3.1.2. OpenEHR

OpenEHR ("*An open domain-driven platform for developing flexible e-health systems*")[8] es una comunidad creada en 2002, en el marco de un proyecto financiado por la Comisión Europea, para investigar y desarrollar estándares con el fin de unificar el conocimiento clínico en formatos electrónicos que aseguren la interoperabilidad.

El objetivo principal es la gestión de historias clínicas electrónicas (o en inglés *Electronic Health Record*s de ahí el acrónimo EHR) mediante un sistema escalable basado en una arquitectura multinivel orientada a servicios como se puede apreciar en la Ilustración 8.

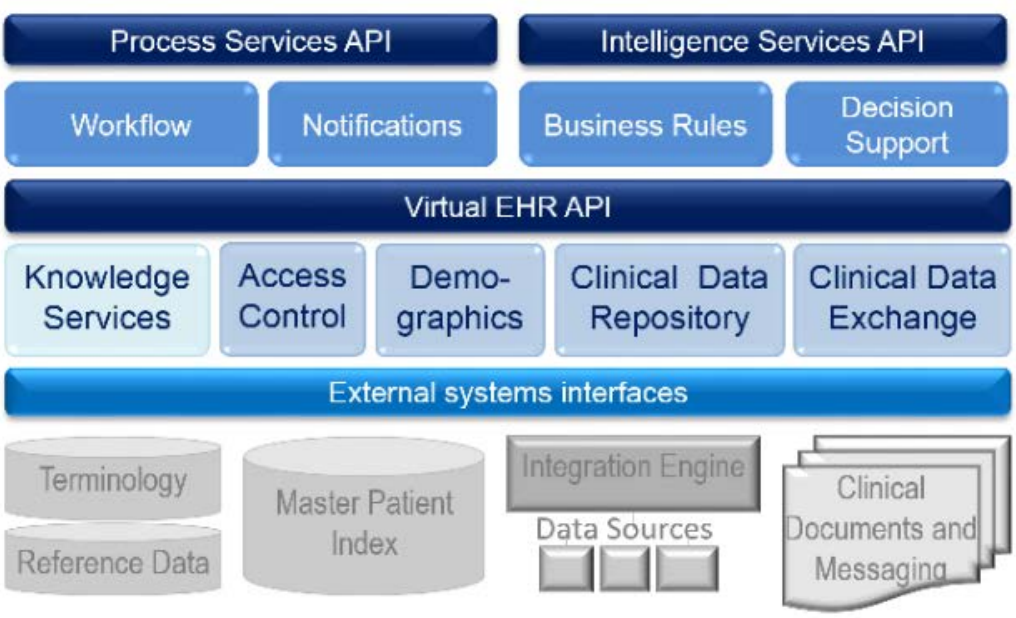


Ilustración 8: Arquitectura multinivel de OpenEHR

[8] https://www.openehr.org

La principal ventaja de esta arquitectura es su flexibilidad en gran cantidad de dominios biomédicos, debido al establecimiento de unos arquetipos [62] que definen un conjunto de reglas o estándares para un dominio clínico. Además de los arquetipos, openEHR se basa en otros 2 grandes pilares:

- **Modelo de Servicios**: Define los interfaces de los servicios y herramientas para el tratamiento de los datos médicos, basados sobre todo en la gestión de EHRs.
- **Modelo de Arquetipos**: Define el conjunto de reglas de representación y semántica de arquetipos basados en dominios clínicos para su posterior uso en los EHR.
- **Modelo de Referencia**: Se trata del esquema de representación del sistema. Este contiene el modelo de datos así como las librerías de arquetipos y las terminologías médicas.

Se trata de un caso parecido a IHE pero aplicado a otras tecnologías. Se trata de un modelo con buena acogida en cuanto a resultados [63][64], pero en ocasiones su dependencia y poca flexibilidad en la definición de los arquetipos dificulta su adaptación.

2.3.1.3. i2b2

La estructura "*Informatics for Integrating Biology and the Bedside*" (i2b2) [65] se inició en 2004 con el soporte de los *National Institutes of Health* (NIH) de los EE.UU. y un hospital de Boston (*Boston Children's Hospital*), entre otros. La primera versión se hizo pública en 2007 con el principal objetivo de crear un sistema escalable para su uso en investigaciones biomédicas. El sistema tiene como fin facilitar la identificación y manejo de datos de los pacientes en el contexto de la medicina traslacional.

i2b2 define un sistema que almacena los datos clínicos de distintas fuentes de manera homogénea siguiendo una arquitectura dividida en distintos componentes que juntos forma lo que sus desarrolladores llaman "la colmena" —i2b2 *hive*, en inglés. Cada componente de la colmena o celda es un servicio que comprende una unidad funcional del sistema, totalmente separada del resto.

Entre las distintas celdas que forman la colmena de i2b2 cabe destacar los componentes que formarían el núcleo del sistema:

- **Project Management** es la celda responsable de la gestión de usuarios, roles y grupos dentro del sistema. También es la encargada de gobernar la gestión y comunicación entre las distintas celdas.
- **File Repository** es la celda responsable del almacenamiento y gestión de los ficheros digitales externos o de gran tamaño como podrían ser resultados de pruebas genéticas o imágenes. Esta celda suele mantener una comunicación constante con el repositorio de datos.
- **Data Repository (CRC)** o modelo de datos, es la celda encargada de almacenar los fenotipos y genotipos de las pruebas y usuarios del sistema. De esta forma esta celda debería estar conectada con el repositorio de archivos y la celda de la ontología.
- **Ontology Management**. Esta celda es la responsable de la gestión y representación del conocimiento de la terminología médica del ecosistema.
- **Identity Management** es la celda responsable de la seguridad del sistema mediante la identificación de los niveles de accesos de los actores que intervienen en cada petición.
- **Workflow Framework** es la celda encargada de gestionar el flujo de trabajo entre las distintas celdas del sistema.

Además de éstas, existen otras celdas de carácter opcional o externo dependiendo de las aplicaciones a integrar. Entre estas celdas se encuentran algunas relevantes como pueden ser la celda de procesamiento de textos o de lenguaje natural, celdas de visualización de la información o la celda que representa el cliente web para gestionar el acceso web a todas las demás.

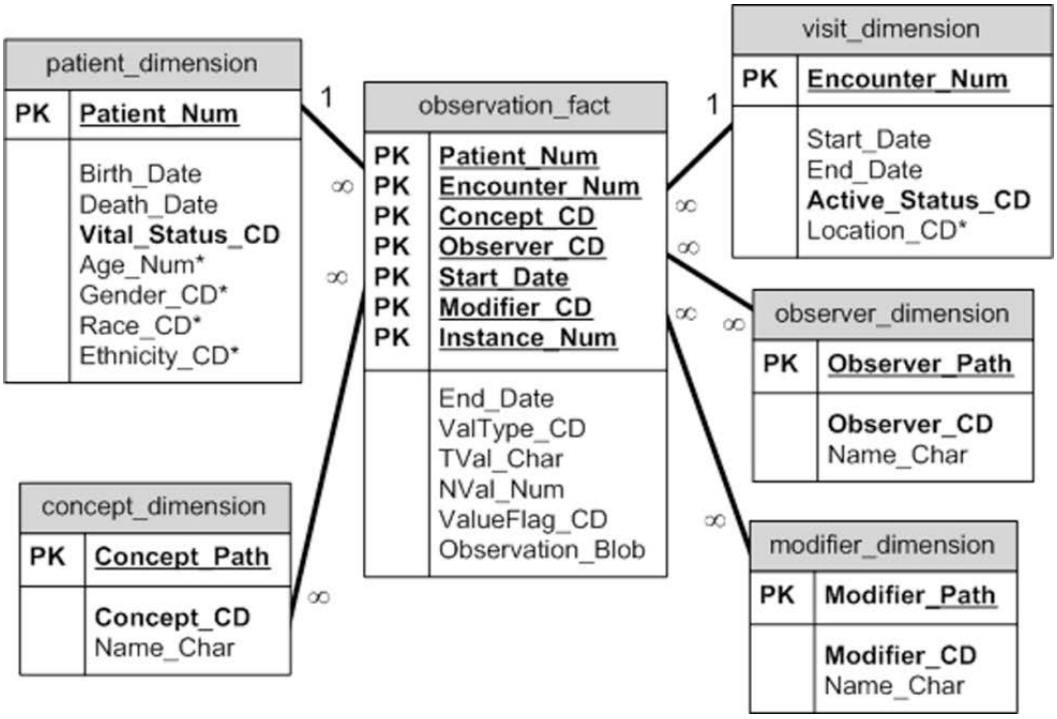


Ilustración 9: Modelo ER en forma de estrella del repositorio de datos de i2b2

Un aspecto a destacar entre sus distintos componentes es el modelo de datos representado en la celda de *Data Repository* en forma de estrella. Inicialmente el equipo de i2b2 desarrolló un modelo de datos basado en una estructura de estrella entre seis tablas de datos que representarían los datos. Estas tablas son las tablas referentes a la información de pacientes, visitas, observaciones, conceptos, modificadores y proveedores. Además de estas seis tablas existen la tabla de códigos (que no forma parte estrictamente del modelo de estrella) y las tablas de mappings de pacientes y visitas.

I2b2 continúa su desarrollo debido a su gran comunidad de usuarios y desarrolladores, en sus últimas versiones ha integrado el soporte al modelo de datos de OMOP en su sistema. Además facilitan la integración de sus soluciones mediante el acceso a máquinas virtuales con un entorno pre-configurado para su integración. Por otro lado, recientes estudios [66][67] han resaltado ciertas limitaciones en el rendimiento de sus consultas así como resultados incongruentes en dominios de diagnósticos y de procedimiento de su celda de ontología. Como ya se ha comentado anteriormente, las continuas versiones y mejoras desarrolladas en su entorno hacen suponer que i2b2 será una herramienta relevante en el futuro para los investigadores clínicos.

2.3.1.4. OMOP

Observational Medical Outcomes Partnership (OMOP) [68] fue un proyecto desarrollado por empresas públicas y privadas del mundo de la salud cuyo objetivo era estudiar los efectos de productos médicos en los pacientes. El proyecto concluyó en 2013 pero actualmente gran parte de su comunidad continúa el trabajo en la iniciativa *Observational Health Data Sciences and Informatics* (OHDSI)[69].

Después de cinco años de funcionamiento del proyecto, OMOP consiguió resultados que demostraron la viabilidad de su proyecto para desarrollar herramientas capaces de analizar y transformar diversas fuentes de datos en el ámbito clínico. Además, OMOP establece un recurso común y abierto para la investigación científica.

Entre sus componentes, cabe destacar el modelo de datos de OMOP. Se trata de una representación de una tabla que incluye todos los datos observacionales que son relevantes para la identificación de la información demográfica, intervenciones de la salud y resultados de pruebas en distintos dominios clínicos.

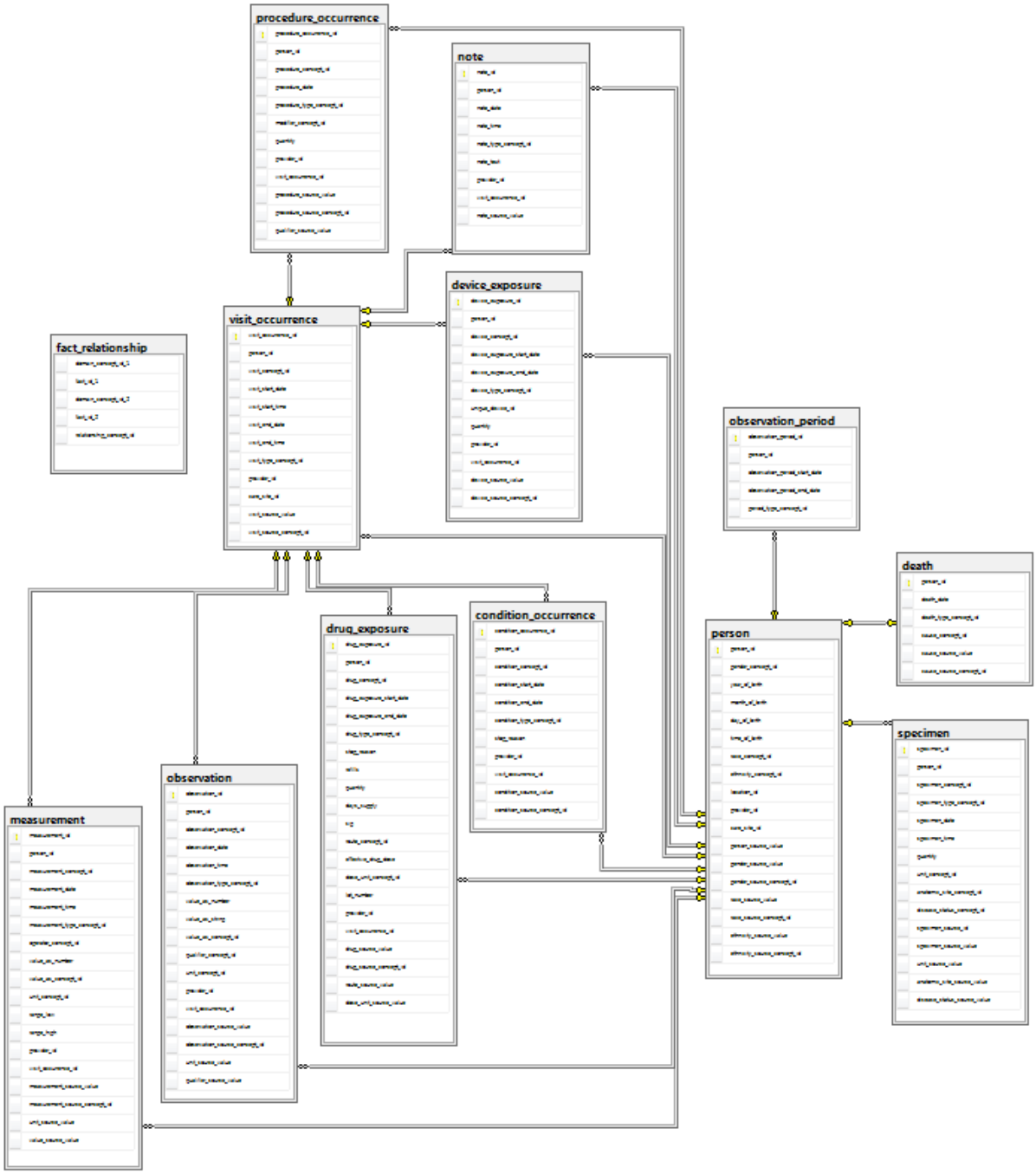


Ilustración 10: Modelo de datos conceptual de OMOP

Además, el modelo define la estructura para representar el vocabulario clínico del sistema, como se puede apreciar en la Ilustración 10. Este vocabulario actúa como una red semántica dentro del modelo de datos, que contiene todos los conceptos y sus relaciones (jerárquicas, sinónimos, atributos…etc.).

El modelo de datos de OMOP presenta resultados de relevancia [70] y actualmente es uno de los modelos de referencia en la práctica clínica. Además, tiene detrás una gran comunidad de usuarios y herramientas que facilitan su incorporación en la actualidad, a pesar de algunas dificultades encontradas anteriormente por investigadores [71] a la hora de la transformación y carga de datos en el modelo.

2.3.1.5. HL7 RIM

Health Level 7 (HL7) es una organización internacional especializada en el desarrollo de estándares en el área de la salud. Fundada hace más de 30 años, actualmente es la organización de referencia en la informática médica. Sus estándares han sido empleados en multitud de centros de distintas nacionalidades y entre sus desarrollos se encuentra:

- HL7 v2: Estándar internacional de mensajería para el intercambio de datos y comunicaciones entre sistemas de información. A pesar de que en sus inicios se presentaba en formato tabular, ahora incorpora esquemas en XML para aumentar la escasa semántica presente antes en sus mensajes. A pesar de estos problemas, HL7 v2 sigue siendo uno de los estándares más extendidos en la actualidad.
- HL7 v3: este estándar trata de cubrir la principal desventaja de la versión 2, la semántica. Para esto se definen estructuras XML de mensajes más restrictivos, lo que permite agregar un contexto más explícito los mensajes intercambiados.

Junto con la versión 3 surge el *Reference Information Model* (RIM). Esta estructura define un modelo orientado a objetos cuyo objetivo es contextualizar cualquier tipo de situación que pueda ocurrir en el ámbito sanitario, desde un diagnóstico de un paciente hasta la información del personal sanitario, contemplando el uso de intercambio de mensajes en XML.

El RIM define un modelo de datos basado en cuatro áreas principales que la forman tres clases principales (Acto, Entidad y Rol) y las respectivas relaciones entre éstas (Participación y las relaciones entre actos y entre roles).

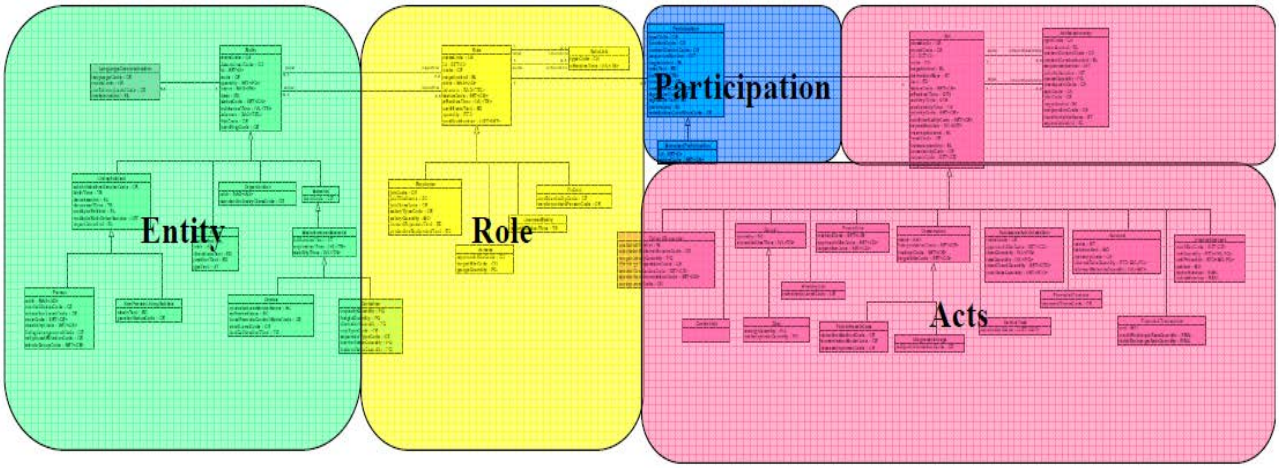


Ilustración 11: RIM y sus 4 áreas

- La clase **Acto** representa las acciones efectuadas. Es la clase principal del RIM ya que representa la mayor parte de registros del dominio. Esta clase está dividida a su vez en distintas especializaciones que heredan de la clase Acto: Observación, Procedimiento, Exposición, Administración de sustancias…etc. Los actos pueden relacionarse entre sí mediante las relaciones *ActRelationship* y con las clases Rol mediante las relaciones *RoleLink*.
- La clase **Entidad** define al objeto, persona, organización o similar que puede participar en un acto. De esta forma, una clase entidad participa en un acto con un rol definido por la clase Rol. Al igual que la clase de acto, esta clase está dividida a su vez en subclases que heredan jerárquicamente de ésta: Persona, material, organización, sustancia…etc.
- La clase **Rol** define el papel que desempeña una entidad al relacionar con un acto mediante la relación de participación. Este rol puede ser de paciente, empleado, acceso…etc.

Como se puede apreciar en la Ilustración 12 parece un modelo complejo debido a su extensión basada en la necesidad de representar todo tipo de información clínica. Sin embargo, muestra una gran flexibilidad ya que es fácilmente adaptable y acotable a las necesidades del momento.

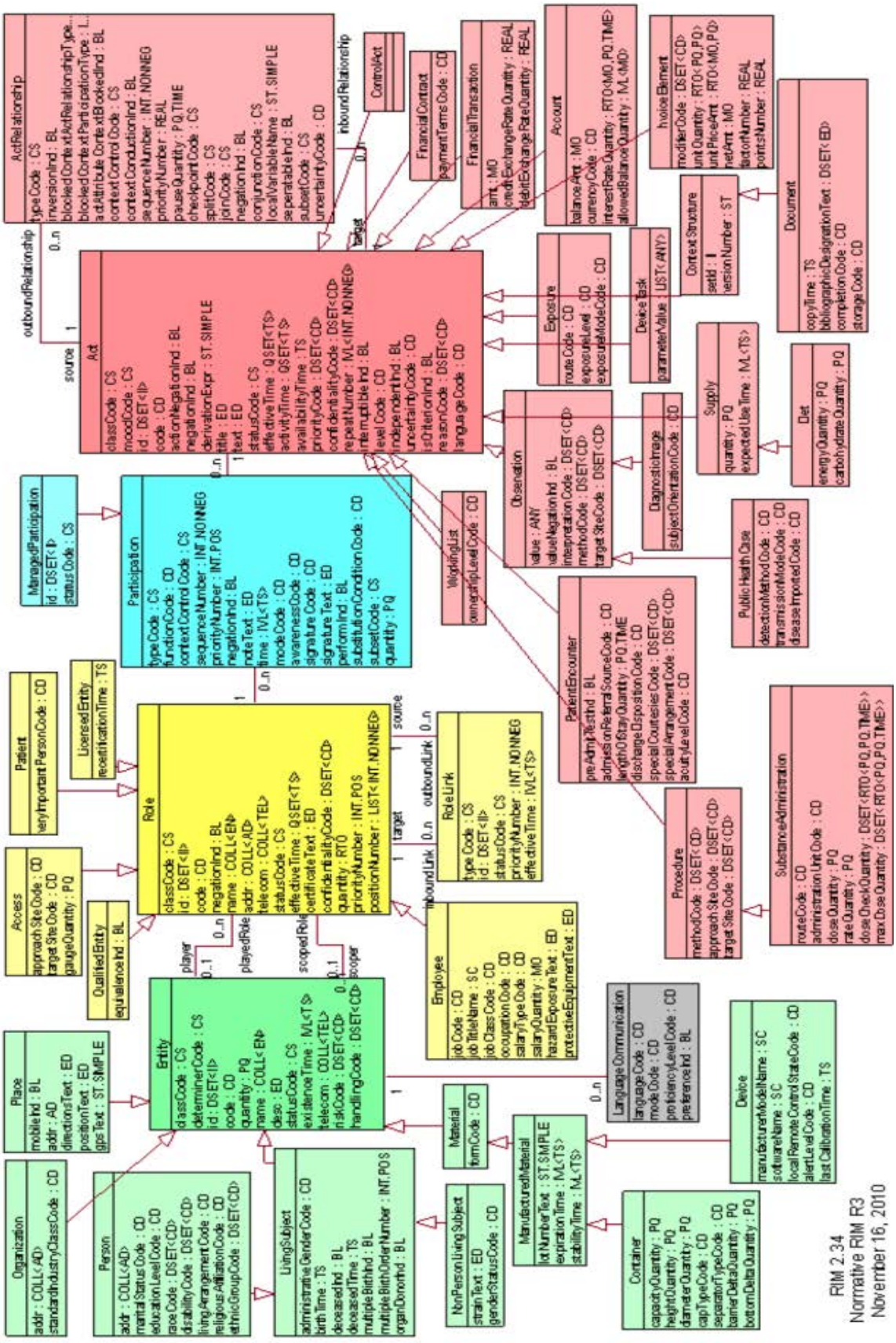


Ilustración 12: Diagrama de clases de HL7 RIM

HL7 RIM hace especial hincapié en la relación de estos modelos de representación con la necesidad de uso de terminologías médicas que permiten comprender la información almacenada. RIM está específicamente diseñado para ser usado conjuntamente con vocabularios o terminologías desarrolladas, sin necesidad de incorporar un vocabulario propio, como en otras iniciativas ya vistas. Adicionalmente, desde la versión 3 se propone la creación de listas de códigos propios para casos de uso más específicos.

Adicionalmente, existen distintos grupos de trabajo dentro de su extensa comunidad de desarrollo avanzando en distintos enfoques para estudiar cómo incorporar más información al modelo, como podría información genética [72] si bien existen soluciones que contemplan la inclusión de esta información de diversas formas [44][45].

La principal desventaja del RIM —que tiene que ver con su amplitud— es la ambigüedad [73], ya que la información puede estar representada correctamente de diversas formas. Para solventar estos problemas la comunidad de HL7 ha definido guías de recomendación de la representación, pero existen enfoques que corrigen estas ambigüedades haciendo uso del conocimiento adquirido por las terminologías médicas.

2.3.1.6. FHIR

"*Fast Healthcare Interoperability Resources*" (FHIR) [74] es el último desarrollo de HL7 para mensajería de datos. Este modelo combina la facilidad de uso de la versión 2 de HL7 con la riqueza semántica y estructural de la versión 3, con nuevos estándares y tecnologías web que facilitan su implementación e intercambio.

La comunidad de desarrolladores de HL7 está centrando sus recursos en la definición y desarrollo de recursos FHIR, como la unidad mínima de interoperabilidad. Estos recursos tratan de representar escenas concretas de conceptos del mundo sanitario: demográfico, observación, diagnóstico, medición, etc.

Estos recursos se basan en estructuras XML, como sus antecesores, y en estructuras JSON siguiendo la arquitectura de servicios REST para facilitar su intercambio en entornos web y móviles.

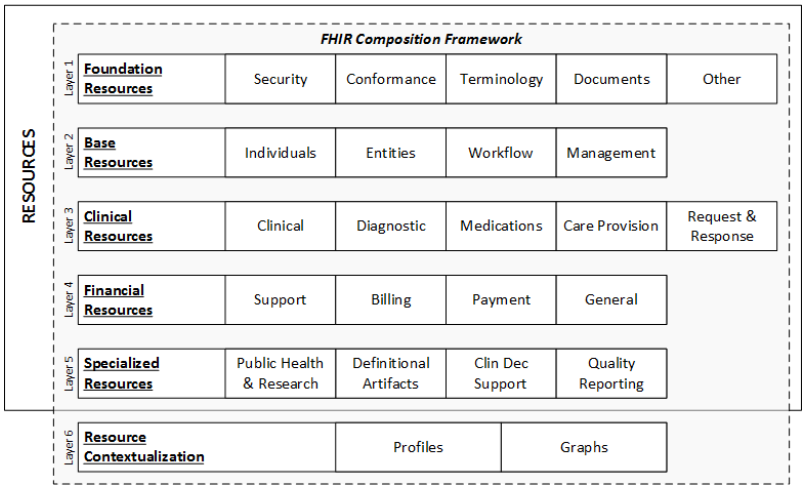


Ilustración 13: Recursos FHIR clasificados por su nivel de madurez

Durante 2018, HL7 presenta un total de 116 recursos FHIR divididos en seis categorías principales como muestra la Ilustración 13: recursos fundacionales, base, clínicos, financieros, específicos y de contexto.

Como se ha comentado anteriormente, se trata de un nuevo modelo de mensajería que dista mucho de estar en una versión completa o que pueda representar una amplia gama de conocimiento sin modificaciones. De los 116 recursos desarrollados, solamente un 5% se encuentra en un nivel de madurez máximo. Si bien la comunidad de HL7 mantiene una línea constante de trabajo en este estándar, en colaboración con otras comunidades y organizaciones como SNOMED, LOINC y los grupos de genómica.

2.3.2. Terminologías clínicas

Los estándares o servicios de terminologías son los encargados de asegurar que la información intercambiada pueda ser entendida por los distintos sistemas. Para ello, proporcionan vocabularios controlados de diversos conceptos clínicos según dominios.

Ilustración 14: William Farr

El origen de estas terminologías data de 1853, cuando William Farr (epidemiólogo y estadista inglés) desarrolló una clasificación de causas de muerte que pudiera ser aplicable de manera internacional. Su propuesta consistía en clasificar las enfermedades en cinco grandes grupos: enfermedades epidémicas, enfermedades constitucionales o generales, enfermedades locales ordenadas por localización

anatómica, enfermedades del desarrollo y enfermedades debidas a violencia. Esta clasificación fue evolucionando durante años hasta llegar a convertirse en la *International Classification of Diseases* (ICD) [75], detallada en la siguiente subsección.

La práctica habitual en estas terminologías o vocabularios es que se centren en un dominio clínico específico: enfermedades, alergias, reacciones adversas, etc., que facilite su desarrollo y por supuesto su mantenimiento. El principal objetivo de las terminologías es representar y proporcionar información sobre unidades de información clínica de cualquiera de estos dominios de manera precisa y fácilmente reconocible.

Estos vocabularios están generalmente formados por términos que identifican conceptos o ideas clínicas del dominio representados jerárquicamente, entiendo una idea clínica como un bloque de información personal de salud que puede comprender todo lo que creemos saber sobre salud, enfermedad, prevención, investigación y tratamiento. Cada término representa un concepto único pero puede haber sinonimias dentro del vocabulario (varios términos que representen o hacen referencia al mismo concepto).

A finales de 1990 se produjo un auge en la aparición de vocabularios médicos controlados que ofrecían el soporte necesario para el intercambio de información en aplicaciones informáticas. Ante esta proliferación incontrolable de terminologías que muchas veces se solapaban, Cimino [76] definió las bases o mínimos en los que se tiene que basar un vocabulario clínico para su correcta aplicación y futuro desarrollo.

Estos ejes se centran sobre todo en el contenido de la información y no en la estructura o en el lenguaje de representación (si bien sí hacía hincapié en la necesidad de usar identificadores únicos y no semánticos). Entre estos ejes se puede destacar la necesidad de que los conceptos de las terminologías representen de una manera clara y entendible las ideas clínicas, para que de esta forma, esta representación sea útil o fácilmente reproducible en cualquier aplicación que lo use. Además, estos vocabularios deben tener en cuenta dos necesidades para mejorar su difusión y dar el salto a otras aplicaciones y/o países; debe ser de fácil implantación en distintos entornos y aplicaciones, y se debe tener en cuenta la necesidad de representarlo en varios idiomas o facilitar su traducción. Desde la publicación de estos ejes se ha producido un cambio en la definición de terminologías al uso de ontologías [77]. Estos ejes tuvieron que ser actualizados hacia un enfoque más

adaptado al uso de ontologías [78], adaptando los ejes a la necesidad de soportar la búsqueda e intercambio de información y no solo el contenido clínico.

En las siguientes secciones se presentan los principales vocabularios médicos controlados utilizados en el ámbito médico, y que han sido usados en la experimentación de este trabajo., concretamente en el dominio de la investigación clínica en cáncer.

2.3.2.1. ICD

"*International Classification of Diseases*" (ICD) es el vocabulario desarrollado y mantenido por la Organización Mundial de la Salud (OMS) para proporcionar una clasificación completa de enfermedades mundiales. El principal propósito de este vocabulario desde sus primeras versiones es favorecer la clasificación y comparación de datos estadísticos de salud proporcionando una codificación completa de enfermedades, síntomas, hallazgos anormales, causas externas, etc. Multitud de centros clínicos lo utilizan en sus sistemas para favorecer la integración de datos [79] así como para comunicarla a Organismos Oficiales.

Actualmente se utilizan las versiones de ICD 9 y 10 (implantando en España en 1998 para seguimiento de mortalidad [80]), mientras que la versión 11 se encuentra en fase de desarrollo (previsto para el verano 2018). Las diferencias entre las distintas versiones se encuentran sobre todo en el sistema de codificación, debido a los cambios en la estructura de la clasificación y/o nuevas apariciones de conceptos. Todas las versiones son de libre acceso y están traducidos a 43 idiomas, entre ellos el castellano.

El número de conceptos codificados depende de la versión de ICD ya que varía su estructura de codificación, entre otros motivos. Así por ejemplo, ICD 10 puede contener más de 14.000 conceptos más otros 16.000 conceptos divididos en subcategorías.

ICD10 es una clasificación alfanumérica dividida en tres niveles o categorías que pueden llegar a tener hasta siete dígitos como máximo.

- Categoría; este nivel está formado por tres dígitos alfanuméricos, comenzando por una letra (A-Z) seguida de dos dígitos, reservando la letra U para enfermedades especiales, como se indica en la siguiente ilustración.

- A00-B99 Certain infectious and parasitic diseases
- C00-D49 Neoplasms
- D50-D89 Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
- E00-E89 Endocrine, nutritional and metabolic diseases
- F01-F99 Mental, Behavioral and Neurodevelopmental disorders
- G00-G99 Diseases of the nervous system
- H00-H59 Diseases of the eye and adnexa
- H60-H95 Diseases of the ear and mastoid process
- I00-I99 Diseases of the circulatory system
- J00-J99 Diseases of the respiratory system
- K00-K95 Diseases of the digestive system
- L00-L99 Diseases of the skin and subcutaneous tissue
- M00-M99 Diseases of the musculoskeletal system and connective tissue
- N00-N99 Diseases of the genitourinary system
- O00-O9A Pregnancy, childbirth and the puerperium
- P00-P96 Certain conditions originating in the perinatal period
- Q00-Q99 Congenital malformations, deformations and chromosomal abnormalities
- R00-R99 Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
- S00-T88 Injury, poisoning and certain other consequences of external causes
- V00-Y99 External causes of morbidity
- Z00-Z99 Factors influencing health status and contact with health services

Ilustración 15: Clasificación de enfermedades en ICD 10

- Detalles; este conjunto de hasta tres dígitos alfanuméricos puede indicar el origen, anatomía, localización, gravedad u otros detalles de la enfermedad del primer nivel.
- Extensión; dígito añadido en la versión 10 para indicar la gravedad o el estado de la enfermedad.

Además de la versión genérica existen otras clasificaciones derivadas de ICD que comprenden un dominio más concreto. Ejemplos de estas clasificaciones de ICD son la clasificación para oncología (ICD-O), neurología (ICD-NA) o modificaciones clínicas (ICD-CM).

En la actualidad ICD10 ha sido implantado en la gran mayoría de los más de 6.000 hospitales existentes en el territorio de los Estados Unidos con un coste superior a 10 mil millones de $. En España con más de 300 hospitales públicos y más de 400 hospitales privados (según el Ministerio de Sanidad [81]), existe desde 2015 un proyecto de implantación de ICD10 desde ICD9 que abarca desde la formación de profesionales en esta terminología al desarrollo e implantación de herramientas que faciliten su incorporación.

2.3.2.2. LOINC

"*Logical Observation Identifiers Names and Codes*" (LOINC) es la terminología desarrollada en 1994 por el *Regenstrief Institute* (RI) [82] para representar observaciones clínicas y/o de laboratorio. El propósito de LOINC es facilitar el intercambio de conocimiento y la gestión de resultados clínicos e investigación con el desarrollo y mantenimiento de un vocabulario que abarque primordialmente pruebas y resultados de laboratorio.

Actualmente, LOINC contiene 84.868 términos correspondientes a la versión de Junio de 2017, que representan una medición, una pregunta o una observación para estas pruebas. Cada término de LOINC está formado por un código numérico y un conjunto de nombres (nombre largo, corto y formal) para que pueda ser identificado por personas de manera gratuita.

Código LOINC	COMPONENT	PROPERTY	SYSTEM	SCALE	TIME	METHOD
2345-7	Glucose	MCnc	Ser/Plas	Qn	Pt	
3091-6	Urea	MCnc	Ser/Plas	Qn	Pt	
2160-0	Creatinine	MCnc	Ser/Plas	Qn	Pt	
3084-1	Urate	MCnc	Ser/Plas	Qn	Pt	
6470-9	Microscopic observation	Prid	Stool	Nom	Pt	We: preparation
29771-3	Hemoglobin.gastrointestinal	ACnc	Stool	Ord	Pt	Imm
13457-7	Cholesterol.in LDL	MCnc	Ser/Plas	Qn	Pt	Calculated

Ilustración 16: Varios términos de LOINC con su nombre formal

Este nombre formal está dividido a su vez en hasta seis componentes definidos por valores:

- Componente: valor de tipo texto que se refiere al nombre del componente al que afecta la medición
- Propiedad: valor de tipo texto que define la propiedad observada
- Tiempo o intervalo en el que se realizó la prueba.
- Sistema donde se realizó la prueba
- Escala en la que se mide la prueba
- Método con el que se ha realizado la prueba (opcional).

Además de identificar estos términos mediante el uso de estos seis componentes, LOINC permite utilizar su estructura jerárquica para ofrecer la información organizada según sus partes y códigos por sus niveles. Actualmente LOINC es una terminología de uso creciente debido a la traducción de la terminología a diversos idiomas (chino, portugués, francés, castellano, etc.) y el desarrollo de distintas aplicaciones. Destacan entre éstas la aplicación RELMA[9] que se facilita de manera gratuita para su explotación. En España existen diversos proyectos para implantar LOINC en la Historia Clínica Electrónica en distintas comunidades autónomas, donde se ha contribuido además al desarrollo de más de 1.000 nuevos términos LOINC.

La finalidad de LOINC es facilitar la integración de resultados clínicos procedentes de pruebas de laboratorio y observaciones clínicas; sin embargo, esta integración dejaría sin cubrir un amplio espectro de situaciones clínicas, de ahí que surjan colaboraciones para combinar terminologías médicas, como en el caso de LOINC y SNOMED [83] que comenzó en 2013. Esta colaboración asegura un acuerdo para soportar la consistencia entre ambas terminologías y evitar la pérdida de esfuerzo estableciendo enlaces entre las partes de LOINC y los términos de SNOMED.

2.3.2.3. HGNC

"*HUGO Gene Nomenclature Committee*" (HGNC) [84] desarrollada por la *Human Genome Organization* (HUGO) es una de las terminologías clínicas de referencia para la identificación de términos genéticos. Para ello, HGNC clasifica más de 40.000 términos genéticos con un identificado único, una abreviatura y un nombre largo. De esta forma se consigue representar de manera única los distintos genes facilitando el intercambio y la búsqueda de información en publicaciones y bases de datos públicas.

HGNC es de libre acceso y es revisada constantemente debido a las continuas contribuciones de investigadores. Para ello, HGNC colabora con los autores de investigaciones genéticas para la inclusión o modificación de conceptos en su base de datos, además de colaborar con el Comité Internacional de Nomenclatura y expertos en familias de genes específicos.

[9] https://loinc.org/relma/

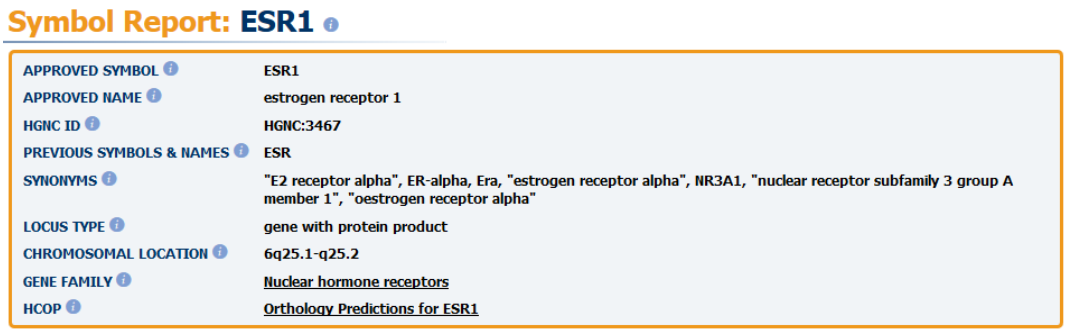


Ilustración 17: Gen ESR1 representado en HGNC

Como se ha indicado anteriormente, esta terminología actúa como un diccionario de genes y no como una clasificación jerárquica, como en el caso de las anteriores terminologías descritas. Si bien, como en el caso de LOINC, debido a la alta especificidad de HGNC en la práctica clínica, se puede proceder a su integración otras terminologías para mejorar la interoperabilidad.

2.3.2.4. SNOMED CT

Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) [85] es la terminología médica más extensa y más utilizada en el ámbito sanitario y clínico actualmente. Esta terminología de ámbito genérico fue desarrollada originalmente por el *College of Pathologists* (CAP) de los Estados Unidos pasando a pertenecer a la *International Health Terminology Standars Development Organisation* (IHTSDO) desde el año 2007 y que actualmente sigue siendo la encargada del mantenimiento de la terminología así como de su distribución.

Esta terminología abarca la totalidad de la información clínica facilitando de esta forma su integración en el entorno sanitario. Actualmente, SNOMED CT está formada por más de 330,000 conceptos clínicos organizados de una manera poli-jerárquica llegando a obtener una profundidad máxima de 32 niveles, en su rama más profunda. Para ello, además de contener esta información, SNOMED CT se encuentra traducido a varios idiomas, entre los que se encuentra el castellano.

SNOMED CT se distribuye bajo distintos tipos de licencias según el tipo de versión que se quiera utilizar (Internacional o nacional) o el fin con el que se vaya a utilizar la terminología (programa de afiliados, sublicencias, etc.). Hay varios países que participan como miembro de facto en la IHTSDO y que por tanto facilitan el acceso a sus

investigadores. Entre ellos se encuentra España. Desde el año 2008 el Ministerio de Sanidad entró a formar parte de este organismo internacional, dejando al Ministerio de Sanidad, Servicios Sociales e Igualdad (MSSSI) como centro nacional de referencia de SNOMED, facilitando su distribución internacional, así como la versión en castellano[10].

SNOMED CT se divide en 3 componentes básicos:

- **Conceptos**. Representan la idea clínica deseada con un identificador único para que sea entendible por las máquinas. Este identificador es una lista de entre 8 y 16 dígitos que no contienen información sobre el significado o la naturaleza del concepto. Facilitando de esta forma la granularidad del concepto, así como la codificación de todo tipo de extensiones, conjuntos o traducciones del propio SNOMED CT.
- **Términos y descriptores**. Cada concepto se representa por palabras o cadenas de texto que son entendibles por personas Cada término tiene que tener al menos un nombre completo ('*Fully Specified Name*' o FSN en inglés) que identifique claramente al concepto. Además los términos pueden tener varios sinónimos para referir al mismo concepto de distintas formas (lenguajes, tecnicismos, etc.) y nombres preferidos.
- **Relaciones**. Representan la conexión semántica entre dos conceptos, resultando una tripleta de la forma concepto-relación-concepto donde cada relación se representa con un identificador único. Existen dos tipos de relaciones en SNOMED; las relaciones de subtipo o *is-a* y las relaciones de atributo. Las primeras son las que definen la jerarquía dentro de la terminología, mientras que las relaciones de atributos son las que añaden significado al concepto. La combinación de las relaciones *is-a* y de atributo completan la definición lógica de un concepto.

[10] https://snomed-ct.msssi.es/snomed-ct/solicitudLicencia.do

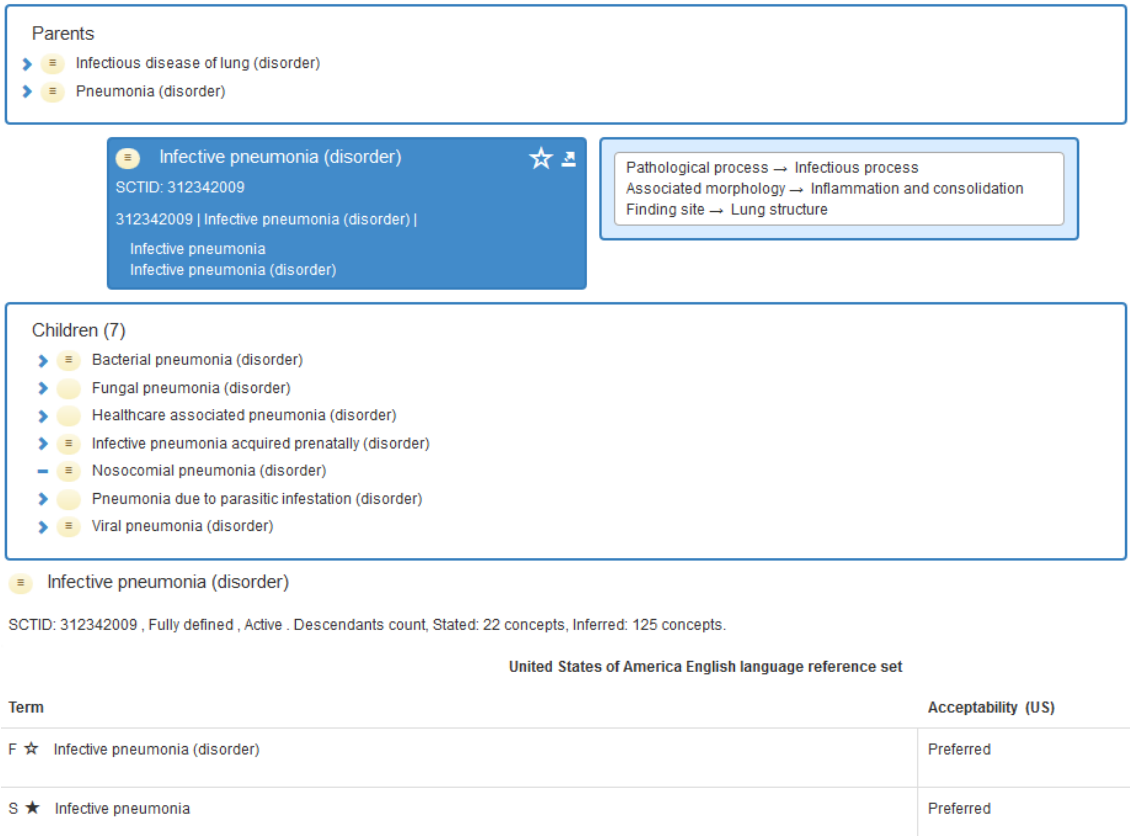


Ilustración 18: Vista de los componentes de "Infective pneumonia"

En la anterior ilustración se muestra un ejemplo de los componentes descritos anteriormente para el concepto "*Viral pneumonia*" en el buscador oficial de SNOMED[11]. En este caso, para este término, tenemos que su identificador de concepto sería el código 312342009. Este concepto se representa con el FSN "*Infective pneumonia (disorder)*" y además tendría como sinónimo "*Infective pneumonia*". Además es un subtipo (relación *is-a* o padres) de los conceptos "*Pneumonia (disorder)*" y "*Infectious disease of lung (disorder)*" y se relaciona con "*Lung structure (body structure)*" mediante "*Finding site (attribute)*", con "*Inflammation and consolidation (morphologic abnormality)*" mediante "*Associated morphology (attribute)*" y con "*Infectious process (qualifier value)*" mediante "*Pathological process (attribute)*". También se puede observar que este concepto es a su vez padre de otros siete conceptos.

Además de estos componentes, los conceptos de SNOMED pueden ser primitivos o completamente definidos según la exactitud de su definición lógica. Así por ejemplo, un concepto primitivo es aquel cuya definición no es suficiente para definir completamente

[11] http://browser.ihtsdotools.org

un concepto. Mientras que un concepto está completamente definido si su término define completamente y de manera única el concepto. En el ejemplo anterior, "*Infective pneumonia*" es completamente definitorio, mientras que un término como, por ejemplo, "*Lung structure*" sería primitivo.

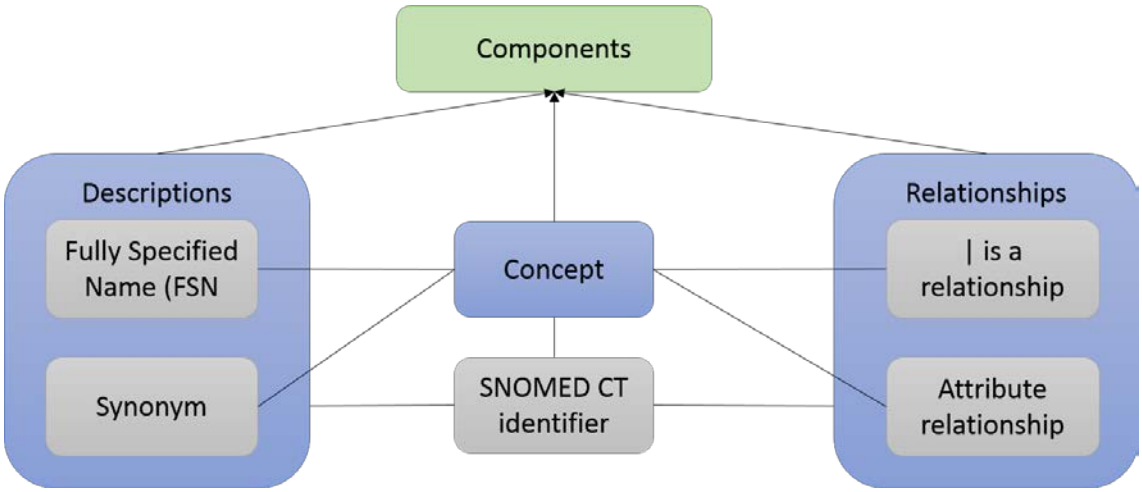


Ilustración 19: Modelo lógico de SNOMED

La representación de las ideas clínicas dentro de la terminología mediante la interacción de sus componentes está definida por el modelo lógico de SNOMED. Mientras que el modelo conceptual es el encargado de especificar cómo se definen los conceptos dentro de la propia jerarquía. Para ello se utiliza una combinación de reglas propias, así como el lenguaje de lógica descriptiva OWL2 en el que está desarrollado SNOMED.

Mediante la organización jerárquica de la terminología, se tiene un concepto raíz (138875005|"*SNOMED CT Concept*") del que descienden mediante la relación *is_a* (subtipos) las ramas principales.

Tabla 2: Ramas principales de SNOMED CT en la versión de Julio 2017

Término	***#Conceptos***	***Descripción***
"Body structure"	37.200	Son los conceptos que representan la estructura del cuerpo humano
"Clinical finding"	106.881	Esta rama abarca los conceptos que se refieren a observaciones clínicas, evaluaciones o juicios. Además incluye conceptos usados para estados clínicos normales y anormales
"Environments and geographical locations"	1.816	Representan ubicaciones
"Event"	3.624	Se trata de las ocurrencias que no sean ni procedimientos ni intervenciones.

"Observable entity"	8.702	Son los conceptos que representan una pregunta clínica que produzca una respuesta o resultado (ejemplos: género, color de los ojos…etc.)
"Organism"	33.819	Se trata de conceptos que representan organismos que interactúan en humanos o medicina animal.
"Pharmaceutical/ biologic product"	17.508	Representan los medicamentos
"Physical force"	170	Son los conceptos que representan la interacción con mecanismos médicos
"Physical object"	15.065	Se trata de los conceptos que representan objetos físicos en el contexto clínico
"Procedure"	56.545	Estos son los conceptos que representan actividades relacionadas durante el cuidado médico. Incluye métodos invasivos y no, además de administración de medicinas, imágenes, terapias…etc.
"Qualifier value"	10.138	Representan valores de atributos que no son conceptos de otros niveles de su jerarquía
"Record artefact"	264	Son los conceptos creados para proporcionar información sobre eventos o historial a terceros
"Situation with explicit context"	4.277	Se trata de conceptos en los que el contexto clínico especificado como parte de su definición. Un ejemplo de estos conceptos serían los que representan el historial familiar de…etc.
"SNOMED CT Model Component"	321.977	Contiene los metadatos de la terminología
"Social context"	4.718	Representan condiciones sociales relevantes en el ámbito médico
"Special concept"	648	Representan conceptos que no forman parte del modelo lógico ni conceptual.
"Specimen"	1.634	Son conceptos que identifican entidades obtenida de análisis médicos.
"Staging and scales"	1.420	Conceptos que identifican valores de escalas y tipos de tumores
"Substance"	25.911	Son los conceptos que representan sustancias en el contexto clínico.

Estas ramas definen por tanto el dominio de un concepto, que es definido como el tipo de concepto que puede ser representado en el origen de una relación. Por otro lado tenemos el rango de representación, que sería el conjunto de conceptos que pueden estar presentes como valor de una relación. Al ser una terminología polijerárquica, los conceptos de SNOMED pueden tener más de un padre, lo que permite que los conceptos pueden ser representados de distinta manera.

Además de estos modelos, SNOMED define un lenguaje de expresiones para representar los conceptos y que estas ideas clínicas sean entendibles por el lenguaje máquina y

humano. Para ello hace uso de una gramática composicional basado en la notación de Backus-Naur [86].

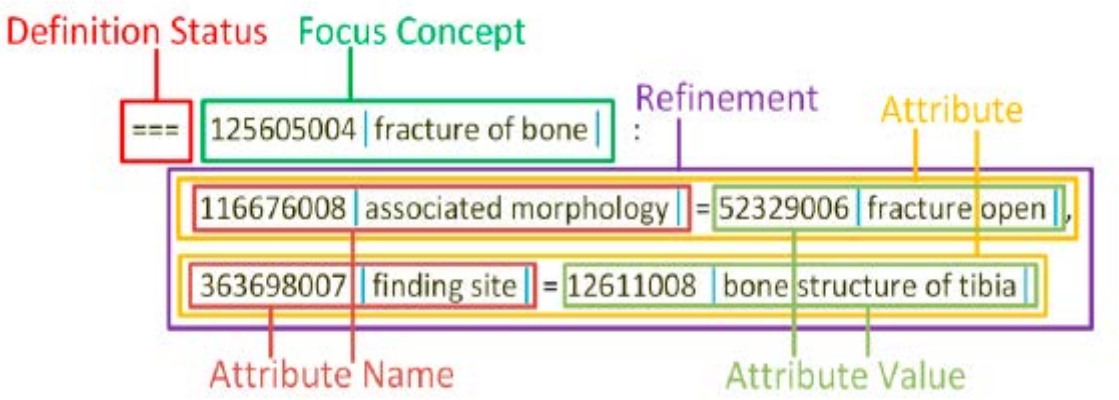


Ilustración 20: Expresión gramatical del concepto "31978002|fracture of tibia"

SNOMED CT publica 2 revisiones anuales de su terminología (Enero y Julio) en las que se añaden o modifican conceptos según las peticiones de cambio recibidas en su plataforma. Además, existen grupos de trabajo definidos para realizar traducciones a otros idiomas, así como enlaces con otras terminologías como ICD o LOINC.

La consistencia y la alta granularidad en la representación de ideas clínicas en SNOMED así como la extensa documentación y su continuo desarrollo son varios de sus puntos fuertes que contrastan con la dificultad que puede representar iniciarse en esta terminología. Si bien en este aspecto también la IHTSDO ha realizado avances ofreciendo distintos cursos de formación online enfocado a distintos tipos de usuarios dependiendo de su aplicación (generación de contenido o implementación de servicios).

Uno de los mayores potenciales que ofrece SNOMED es una serie de mecanismos propios de la terminología que aumentan el espectro de uso y representación de la misma. Entre estos destacaremos en las siguientes subsecciones el mecanismo de post-coordinación y un mecanismo de normalización de términos. Además, SNOMED dispone de un mecanismo de extensión de la terminología con una serie de guías para añadir nuevos conceptos a las distintas versiones de SNOMED y su posterior validación. O mecanismos para agrupar conceptos que representen un contexto común (Reference sets) o agrupaciones de atributos que definan conjuntamente un concepto.

2.3.2.4.1. Mecanismo de post-coordinación

Está relacionado con la formación de expresiones en SNOMED CT. Para ello, como se ha comentado anteriormente, SNOMED define una serie de reglas basadas en el lenguaje formal de composición para representar ideas clínicas con conceptos de la terminología. Estas expresiones son una combinación estructurada de conceptos de la terminología

En este contexto surgen tipos de expresiones: precoordinadas y postcoordinadas. Se entiende por expresiones precoordinadas aquellas que representan un único concepto definido en SNOMED. Siguiendo el ejemplo de la Ilustración 20 de "Fracture of tibia" sería un claro ejemplo de concepto precoordinado, a la información contenida en la imagen habría que añadir los conceptos padres del término.

Por otro lado, las expresiones postcoordinadas son aquellas que contienen dos o más conceptos de SNOMED. De esta forma se combinan conceptos para aumentar el grado de especificidad de la idea a representar. Esta expresión precoordinada no es únicamente una suma de conceptos, ya que se tienen que respetar la propia sintaxis y semántica de cada uno de los conceptos utilizados, así como una serie de reglas similares a las de creación de conceptos (extensiones).

Es decir, la idea de este mecanismo es facilitar la representación de ideas clínicas que en ese momento no están representados por un único concepto de la terminología. Si bien este mecanismo puede ser utilizado para representar conceptos precoordinados, aunque no refleja utilidad. Este mecanismo es de especial utilidad para incorporar en sistemas de procesamiento de lenguaje natural (NLP).

2.3.2.4.2. Forma normal

El mecanismo de forma normal —o normalización— es un mecanismo para generar una expresión válida con la cantidad de información necesaria mediante la aplicación de una serie de reglas de transformación [87]. Esta normalización es un proceso utilizado para homogeneizar la posible representación de ideas clínicas parecidas o con alguna relación.

Para ello, SNOMED ha definido un proceso recursivo de normalización basado en reglas en el que un concepto se transformaría en una expresión formada por conceptos primitivos. Es decir, partiendo de una expresión que represente un concepto, primero se comprueba si los conceptos que forman la expresión son primitivos. Si todos los

conceptos son primitivos, esta expresión sería la expresión normalizada. Si existen conceptos en su forma FSN, éstos se sustituyen por su expresión completa y se vuelve a comprobar si los nuevos conceptos en la expresión resultante son todos primitivos. En caso negativo se volvería a sustituir los conceptos no primitivos por sus expresiones hasta conseguir una representación completa de conceptos primitivos.

SNOMED define dos tipos de normalización, larga y corta. Mientras que la forma normal larga es el proceso recursivo explicado anteriormente, la forma normal corta sería la misma pero eliminando relaciones de atributos que puedan ser redundantes. Por tanto, ambas formas normales pueden llegar a ser la misma para un concepto. El propósito de la forma normal larga es la representación de expresiones clínicas en sistemas informáticos o en su defecto en el almacenamiento de estos sistemas. Mientras que por otro lado, la forma normal corta se utiliza únicamente para construir consultas en sistemas informáticos, ya que esta forma normal está contenida en la larga.

2.3.3. Estándares de documentos

En las anteriores secciones se han descrito modelos de datos y terminologías clínicas como la base de los estándares de interoperabilidad clínica, pero existen otros servicios o estándares clínicos de ámbito más concreto (documentos, imágenes, etc.) utilizados en la práctica de la informática médica: A continuación se describe diversos estándares sobre documentos clínicos.

Los estándares de documentos engloban aquellos estándares de intercambio de información clínica que tienen un contexto común como base. Estos estándares forman documentos completos legibles por el ser humano y que pueden ser firmados por el responsable de la petición o creación, donde estas características son las principales diferencias de los estándares de documentos y los estándares de representación de mensajería o modelos de datos clínicos. Ejemplos de estos documentos pueden ser un informe de alta, un cuaderno de recogida de datos de un ensayo clínico, un genograma genético, etc.

2.3.3.1. Ensayos clínicos

Dentro de la práctica clínica se define el ensayo clínico como la herramienta o estudio de investigación prospectiva que evalúa y compara el efecto de una intervención sanitaria

respecto a un control de pacientes [88]. Estas intervenciones pueden ser debidas a nuevos fármacos o combinaciones de éstos, células, procedimientos quirúrgicos o radiológicos, productos biológicos, etc. Por tanto, un ensayo clínico bien planificado y ejecutado es la principal herramienta experimental para evaluar la efectividad de las intervenciones médicas dentro de la medicina basada en la evidencia.

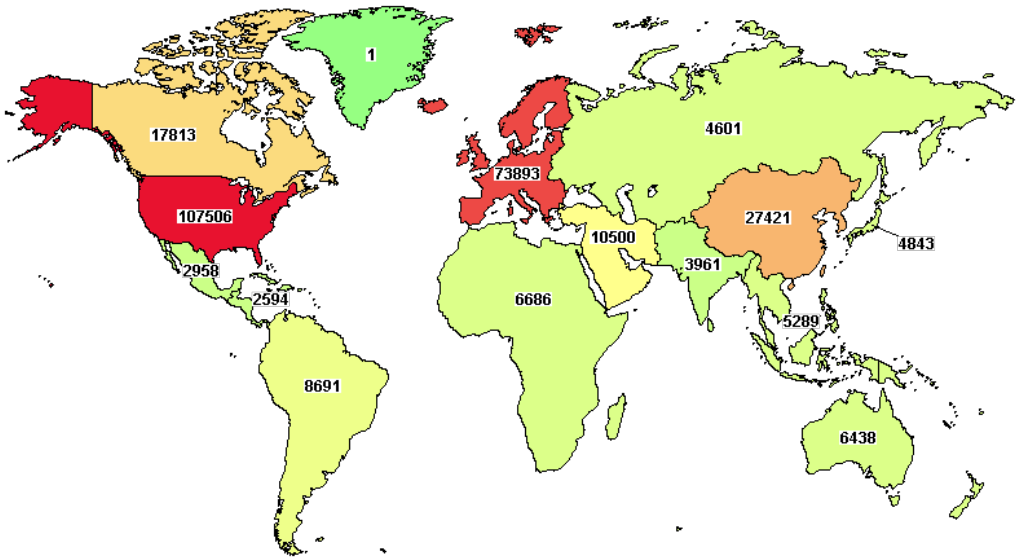


Ilustración 21: Número de ensayos clínicos en la base de datos de clinicaltrials.gov divididos por regiones

Existen distintas bases de datos públicas o registros sobre los ensayos clínicos, su evolución y resultados. Algunas de estas plataformas son de ámbito local y dependiendo del país de desarrollo del ensayo, es de carácter obligatorio la publicación del mismo. La más importante de estos registros es *ClinialTrials.gov*[12] gestionado por la *National Library of Medicine* (NLM) y los *National Institute of Health* (NIH) que a finales de 2017 contenía más de 300.000 ensayos clínicos registrados en todo el mundo. Estos ensayos clínicos pueden ser consultados en su página web, así como a través de una API de acceso, durante las distintas fases de desarrollo de un ensayo clínico, incluyendo resultados del mismo. Otros registros de ensayos clínicos relevantes son el *EudraCT*[13] mantenido por la Comisión Europea y la Agencia Europea de Medicamentos, que contiene más de 50.000 ensayos clínicos, así como la Plataforma de registros internacionales de ensayos clínicos de la OMS (ICTRP)[14].

[12] https://www.clinicaltrials.gov/
[13] https://eudract.ema.europa.eu/
[14] http://www.who.int/ictrp/es/

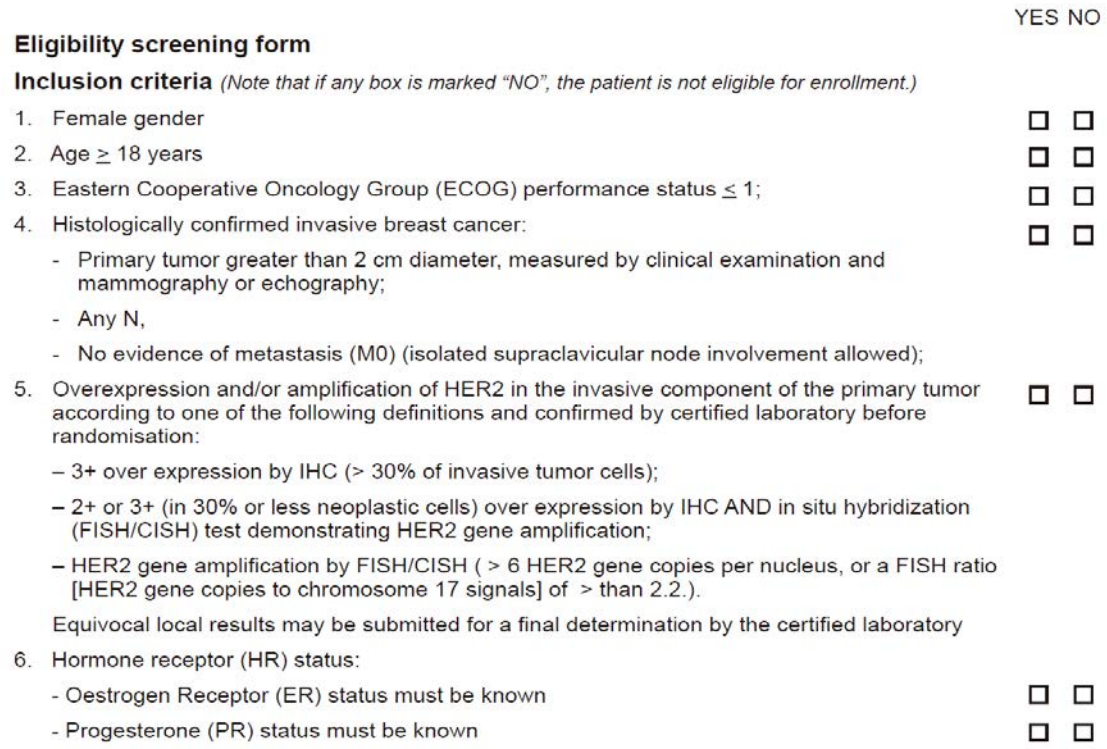
En las siguientes subsecciones describiremos dos aspectos destacados de los ensayos clínicos, como son los cuadernos de recogida de datos y los criterios de elegibilidad, de especial interés para este trabajo, concretamente para la recogida de datos clínicos y para la construcción de consultas que validen los criterios de elegibilidad.

2.3.3.1.1. Criterios de elegibilidad de ensayos clínicos

Cada ensayo clínico contiene información sobre el diseño del estudio, tipo de reclutamiento, localización, fases e intervenciones, pero uno de los aspectos más importantes está relacionado con los requisitos que deben cumplir los pacientes para poder entrar en el ensayo clínico. Cada ensayo clínico define una serie de guías o características, llamadas criterios de elegibilidad, que los pacientes deben cumplir para poder ser reclutados en el ensayo (género, sexo, edad, hábitos, antiguos tratamientos…etc.).

Estos criterios de elegibilidad se encuentran divididos en criterios de inclusión y exclusión; los primeros son los criterios que obligatoriamente debe cumplir para poder ser reclutado, mientras que los segundos son los criterios que no deben cumplir los pacientes. Estos criterios de elegibilidad se representan en texto libre como puede apreciarse en la siguiente ilustración, donde se aprecian algunos de los criterios de inclusión del estudio Neo ALTTO[15], uno de los estudios que ha sido considerado en este trabajo.

[15] https://clinicaltrials.gov/ct2/show/NCT00553358



YES NO

Eligibility screening form

Inclusion criteria *(Note that if any box is marked "NO", the patient is not eligible for enrollment.)*

1. Female gender ☐ ☐
2. Age ≥ 18 years ☐ ☐
3. Eastern Cooperative Oncology Group (ECOG) performance status ≤ 1; ☐ ☐
4. Histologically confirmed invasive breast cancer: ☐ ☐
   - Primary tumor greater than 2 cm diameter, measured by clinical examination and mammography or echography;
   - Any N,
   - No evidence of metastasis (M0) (isolated supraclavicular node involvement allowed);
5. Overexpression and/or amplification of HER2 in the invasive component of the primary tumor according to one of the following definitions and confirmed by certified laboratory before randomisation: ☐ ☐
   – 3+ over expression by IHC (> 30% of invasive tumor cells);
   – 2+ or 3+ (in 30% or less neoplastic cells) over expression by IHC AND in situ hybridization (FISH/CISH) test demonstrating HER2 gene amplification;
   – HER2 gene amplification by FISH/CISH (> 6 HER2 gene copies per nucleus, or a FISH ratio [HER2 gene copies to chromosome 17 signals] of > than 2.2.).

   Equivocal local results may be submitted for a final determination by the certified laboratory
6. Hormone receptor (HR) status:
   - Oestrogen Receptor (ER) status must be known ☐ ☐
   - Progesterone (PR) status must be known ☐ ☐

Ilustración 22: Criterios de inclusión del ensayo Neo ALTTO (*Neoadjuvant Lapatinib and/or Trastuzumab Treatment Optimisation*)

Como se ha comentado anteriormente, estos criterios se representan en texto libre, ya que no existe un estándar de representación, si bien, suelen existir guías prácticas de los estudios sobre cómo codificar los conceptos. Aunque no existe un estándar para representarlos, sí que existen diferentes lenguajes o expresiones de consulta en los que desarrollar los criterios de elegibilidad:

- Expresiones ad-hoc o conjuntos de reglas con valores booleanos, numéricos o categóricos junto con operadores lógicos o de comparación para la representación de criterios de elegibilidad.
- Sintaxis Arden, estándar HL7 para la codificación de términos médicos con condiciones en forma de reglas. Este lenguaje admite la definición de funciones, razonamiento y codificación de idiomas.
- Lenguajes basados en lógica que permiten la codificación de reglas lógicas más complejas que las expresiones ad-hoc (Ej: SQL).
- Lenguajes de consulta orientados a objetos que operan sobre modelos basados en objetos similares a lenguajes de programación. (ej: GELLO)

- Lenguajes de consulta temporal en los que prima la representación del tiempo en los criterios de elegibilidad. Ofrecen funciones de razonamiento temporal (fechas relativas, margen de error, patrones cíclicos…etc.).

Hay distintos estándares de representación de conocimiento en los ensayos clínicos, entre los que destacan: ERGO y CRFQ. ERGO es una ontología de representación de criterios de elegibilidad basada en un modelo de información orientados a objetos. CRFQ se basa en el uso de parámetros semánticos para la estandarización de criterios. La complejidad y la heterogeneidad en la representación de los criterios de elegibilidad han demostrado la necesidad de combinar varios de estos estándares, así como la necesidad de procesos manuales para llevar a cabo esta tarea [89], demostrando la necesidad de adopción de un estándar de representación de criterios de elegibilidad.

2.3.3.2. Cuaderno de recogida de datos

El Cuaderno de Recogida de Datos (CRD) es un documento establecido para la recogida y registro de datos de participantes en ensayos clínicos. Se trata de cuestionarios que hasta hace unos años se realizaban en formato papel pero que en la actualidad ya se realizan en formato electrónico, y que contienen preguntas basadas o relacionadas con uno o varios ensayos clínicos. Estos documentos son los que facilitan el reclutamiento y la evaluación de los ensayos clínicos.

Estos CRD actúan como un registro robusto de datos que permiten no solamente su almacenamiento, sino también su correcta trazabilidad. Si bien no existe un estándar de facto para los CRD, sí que existen algunas iniciativas como *Clinical Data Interchange Standards Consortium* (CDISC) [90] que desarrollan propuestas de estándares para CRDs. Estos desarrollos se centran en áreas o formularios muy genéricos que no incluyen aspectos más específicos de los ensayos clínicos.

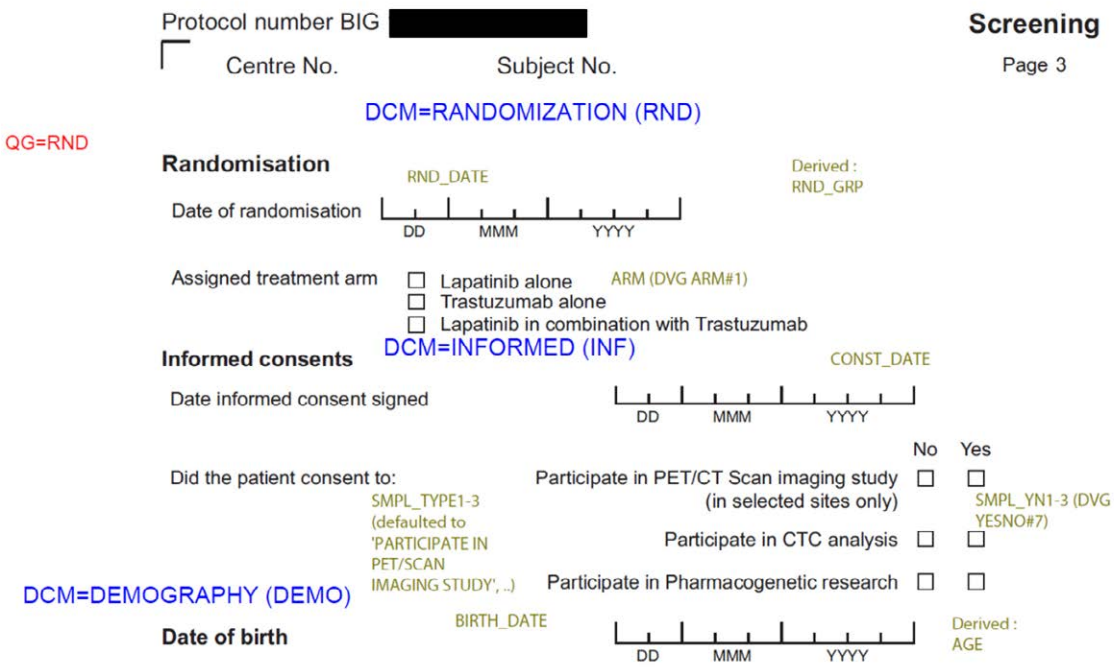


Ilustración 23: Ejemplo de guía de CRD en el proyecto EURECA

En cuanto a los CRD en formato papel (e incluso en formato electrónico), generalmente se acompañan de una manual de uso que indican cómo rellenar el formulario, tipos de unidades, conversión o corrección de errores, ya que generalmente los usuarios de estos CRD no son "encuestadores" habituados a rellenar estos formularios. Debido a la falta de un diseño estándar, estos manuales suelen además incluir una guía de estilo que aconseja valores terminológicos o conjuntos de terminologías para la anotación de nombres de campos. Generalmente estos CRD son posteriormente anotados y almacenados de manera permanente en sistemas de bases de datos electrónicas de manera que puedan ser consultados fácilmente para su posterior análisis.

2.3.3.3. Genograma genético familiar

Un genograma médico es una representación gráfica de las relaciones entre las generaciones familiares. En el ámbito clínico se utilizan como otra forma de representación además la historia clínica, con el objetivo de encontrar una causa común o a una probable sintomatología familiar. Es un documento que viene utilizándose durante el último siglo en distintos contextos sanitarios [91][92] y que se ha explorado en la presente tesis en su vertiente genética.

El genograma genético familiar definido por la *National Society of Genetic* y su grupo de estandarización en los años 90 [93], tiene como propósito definir cómo tiene que ser intercambiada la información genética en una representación visual en forma de árbol genealógico. Es decir, qué información tiene que contener este genograma, cómo se tiene que representar gráficamente y cómo se tiene que anotar.

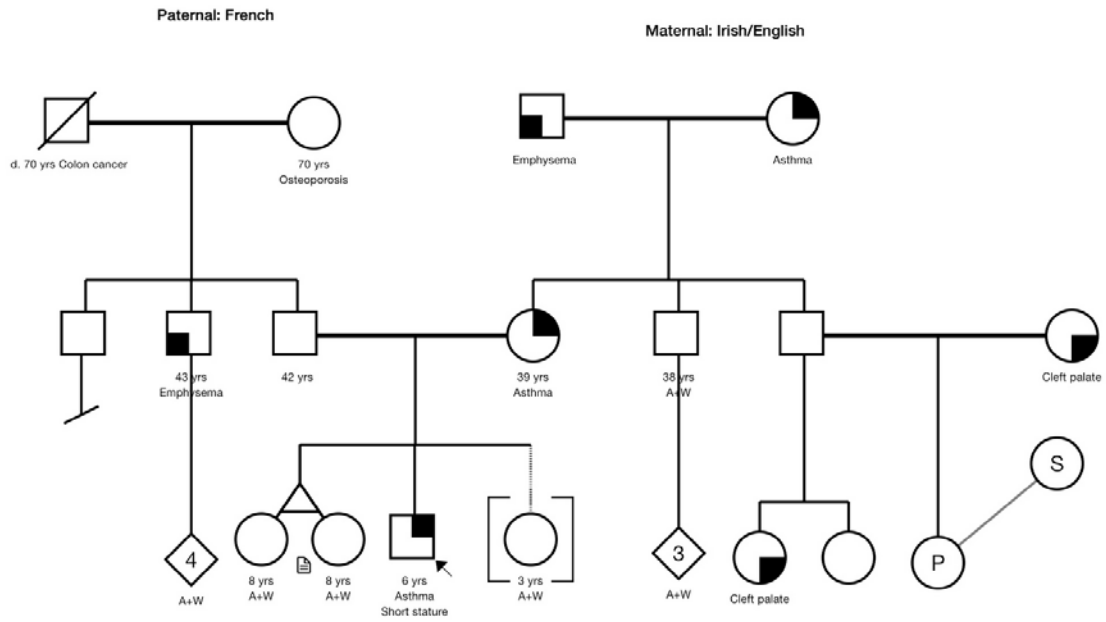


Ilustración 24: Genograma genético de una familia de 3 niveles

Existen vayas revisiones o actualizaciones sobre de este estándar [94][95] que definen estos aspectos en el ámbito genético para la integración de este documento en la Historia Clínica Electrónica. Aún no existe un consenso sobre las terminologías clínicas que deben ser utilizadas para la anotación de las patologías y fenotipos, aunque existen herramientas como *Phenomyzer*[16] de la *Human Phenotype Ontology* [96] que ayudan a hacer diagnósticos basados en cálculos probabilísticos, usando genogramas como referencia.

2.4. Proyectos e iniciativas relevantes

Durante los últimos años, se han venido realizando multitud de esfuerzos a diferentes niveles, como el científico y clínicos, dirigidos a investigar modelos de integración que

[16] http://compbio.charite.de/phenomizer/

sean transferibles a la práctica clínica. Como se ha comentado en la introducción, el incremento de los recursos accesibles a través de bases de datos públicas, el incremento de variables e investigaciones clínicas con respecto a otras épocas y la urgente necesidad de transferir estos avances a la práctica clínica diaria han convertido esta necesidad en una oportunidad para grupos de investigación. Los proyectos e iniciativas relevantes, se han centrado sobre todo en la reutilización de datos recogidos para otros objetivos (EHR CRDs, etc.) para la investigación clínica entre distintos centros.

Debido a la amplitud y la complejidad de las tecnologías y las posibles implantaciones clínicas, muchos de estos proyectos se han centrado en uno o varios aspectos de los descritos anteriormente. Un ejemplo de estas iniciativas se ha visto dentro de los estándares de intercambio y modelos clínicos de datos, Sección 2.3.1 Modelos de datos clínicos e intercambio de datos, donde los modelos OMOP e i2b2 forman parte de iniciativas con un espectro mayor.

A continuación se detallan varias iniciativas relevantes en la investigación clínica del campo de la integración de datos biomédicos a nivel internacional. Las dos primeras iniciativas se pueden incluir entre las más relevantes a nivel internacional, mientras que a continuación se detallan otros proyectos de investigación europeos más relevantes para el presente trabajo, incluyendo los dos proyectos en los que se ha desarrollado el trabajo.

2.4.1. caBIG

Cancer Biomedical Grid (caBIG) [97] fue una red financiada por el *US National Cancer Institute* (NCI) cuyo principal objetivo era el intercambio de forma segura de investigaciones clínicas sobre cáncer. Para ello, desarrollaron una red formada por más de 60 centros del NCI y 16 hospitales, mediante computación *grid*, desde sus inicios en 2004. En este año hubo un apoyo decidido al el desarrollo de guías y estándares, así como en la definición de lenguajes comunes para facilitar el intercambio de datos y comunicación.

caBIG se centró sobre todo en el intercambio de datos en el contexto de gestión de ensayos clínicos, imágenes clínicas y en la gestión de pruebas y bio-especímenes. Entre sus desarrollos cabe destacar el uso de software libre, lo que facilitó definir una gran comunidad de usuarios en Estados Unidos. Su estructura básica, a nivel semántico, se

basó en 2 componentes principales; caCORE y caGRID, encargados de definir las funcionalidades y herramientas para gestionar los vocabularios y facilitar el intercambio de información.

Tras varios años y más de 350 millones de dólares invertidos en la red, el NCI decidió la finalización del proyecto [98]. Al considerable gasto asumido se unieron multitud de críticas y problemas [99], así como la dependencia en una tecnología (grid) que empezaba a ser desbancada por otras tecnologías más modernas.

2.4.2. epSOS

El proyecto epSOS es un proyecto financiado por la Comisión Europea (2008 – 2014) en el que participaron 45 entidades de 25 países europeos [60], entre los que destacaban los ministerios de sanidad de varios países. El principal objetivo de epSOS era asegurar un marco de interoperabilidad europeo en dos contextos digitales, como son el de las prescripciones médicas electrónicas (*ePrescription*) y el de los resúmenes de pacientes (*Patient summary*).

El proyecto se dividió en distintas fases. Comenzando por el diferente estado en el que se encontraba cada miembro de la comunidad europea se examinaban las barreras tecnológicas, legales y éticas para implantar las bases para un modelo común de intercambio. En las fases finales se realizaron las pruebas y validaciones necesarias para asegurar que estas soluciones podrán ser implantadas en los próximos años.

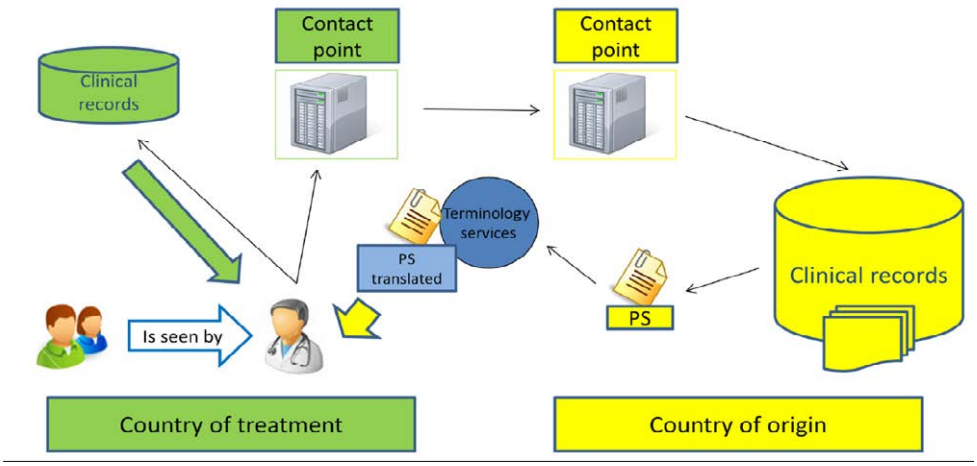


Ilustración 25: Diagrama de funcionamiento de Patient Summary en Europa

Para ello, definieron un formato de resumen (*Patient summary*) que el propio usuario podría acceder y trasladar a los países donde se desplazase para facilitar y mejorar su asistencia sanitaria, a través de entidades nacionales que aseguraran este proceso. Además, se definieron los estándares y guías a utilizar también en la *ePrescription* en Europa, así como sus posibles correlaciones con las utilizadas en Estados Unidos (*Blue Button* [100]) y con los tiempos de implantación en Europa.

2.4.3. INTEGRATE y EURECA

El proyecto **INTEGRATE** (*Driving Excellence in Integrative Cancer Research through Innovative Biomedical infrastructures*) fue un proyecto financiado por la Comisión europea (2011-2014) [18] formado por 6 entidades entre las que se encontraba el Grupo de Informática Biomédica de la Universidad Politécnica de Madrid. El principal objetivo de este proyecto de investigación era el desarrollo de estructuras tecnológicas para promover la colaboración en la investigación sobre cáncer.

Para esto se definió un modelo de capa semántica basada en estándares biomédicos que sirviera como núcleo central del proyecto. Esta capa semántica se basada en un componente llamado *Common Information Model* que a su vez estaba formado por un modelo común de datos (basado en HL7 RIM), un conjunto de vocabularios clínicos y los enlaces entre ellos, además de una serie de servicios y tecnologías alrededor de ellos como se puede apreciar en la imagen.

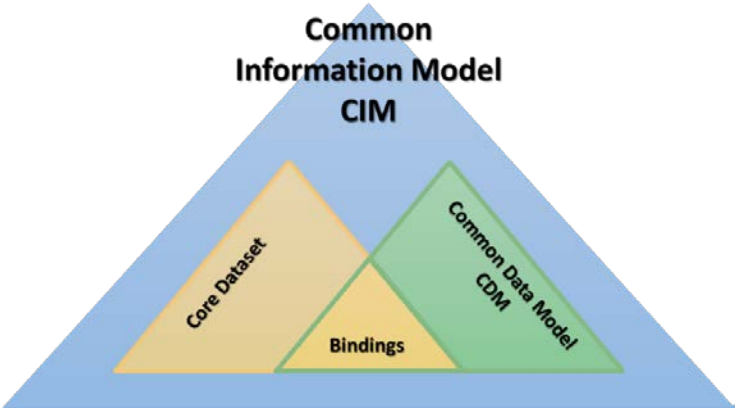


Ilustración 26: Common Information Model de la plataforma INTEGRATE

El proyecto tuvo una duración de 4 años y el modelo de capa semántica desarrollado fue evaluado por expertos europeos en los pilotos de reclutamiento y selección de cohortes de pacientes para ensayos clínicos de cáncer de pecho, con buenos resultados.

Enabling information re-Use by linking clinical REsearch and Care (**EURECA**) es el Proyecto de investigación financiado por la Comisión Europea [19] en el que se ha realizado la mayor parte de la presente tesis. Este proyecto liderado por *Philips Research* contó con la colaboración de hasta 18 entidades europeas y canadienses en el dominio de la oncología clínica. El principal objetivo de este proyecto era el establecimiento de enlaces semánticos las Historias Clínicas Electrónicas y ensayos clínicos desarrollando una solución escalable e interoperable.

Al igual que en el proyecto INTEGRATE, se definió un *Common Information Model* en el que se expandieron tanto los vocabularios clínicos como el modelo de datos usado, y por consiguiente los posibles enlaces entre ellos. De esta forma se consiguió gestionar datos de diversas fuentes heterogéneas en un único sistema que tuvo que tratar con distintas barreras:

- Documentos estructurados, semiestructurados y de texto libre.
- Diferentes leyes y políticas.
- Diversos sistemas de atención clínica y de investigación.
- Multitud de estándares clínicos y terminologías, si bien estas tenían en la práctica una baja adopción frente a recursos propios.

Este proyecto concluyó 2016. La capa de interoperabilidad semántica tuvo una gran acogida por parte de los expertos de la Comisión Europea en los pilotos definidos en el proyecto, como eran: reclutamiento de pacientes, aplicación de generación de hipótesis, clasificador de diagnósticos, seguimiento de pacientes, selección de cohortes, etc. Parte de este trabajo está accesible en una versión abierta de la solución[17] para facilitar la implantación del sistema y su posible evolución.

2.4.3.1. Capa de Interoperabilidad Semántica

La capa de interoperabilidad semántica desarrollada para estos proyectos europeos, y en las que desarrolló su trabajo el autor de este trabajo, actuó como uno de los ejes centrales de las plataformas desarrolladas. Esta capa trata de garantizar un acceso homogéneo a la información procedente de los distintos ensayos clínicos, pruebas de laboratorios e

[17] https://bitbucket.org/dperezdelrey/hl7-snomed-semantic-solution

historias clínicas electrónicas. Por tanto, la CIS es la encargada de dar acceso homogéneo a los datos presentes en cada uno de los centros clínicos que formaron parte de los proyectos para todas las aplicaciones desarrolladas, como se muestra en la siguiente ilustración.

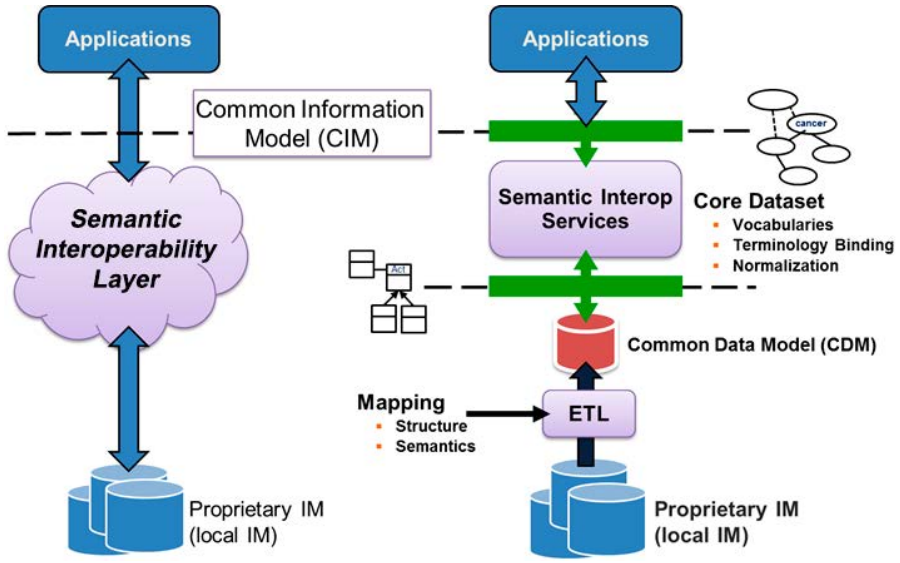


Ilustración 27: Capa de Interoperabilidad Semántica en relación con las aplicaciones y fuentes

Como se puede apreciar en la ilustración, la capa se encuentra entre las aplicaciones y las bases de datos o fuentes de datos de cada centro. Abarcaba de esta forma, desde el proceso de acceso de los datos por las distintas aplicaciones desarrolladas, hasta la carga y gestión de las distintas fuentes de datos. A la derecha se puede apreciar con más detalle el conjunto de componentes y servicios que forman la capa de interoperabilidad semántica y su interacción con las distintas fuentes de datos (*Propietary IM – local IM*) y las aplicaciones.

Para asegurar una correcta representación homogénea de los datos, la CIS se basa en un Modelo de Información Común (*Common Information Model*, CIM, en inglés) que está formado por dos componentes principales: Modelo Común de Datos (CDM del inglés, *Common Data Model*) y el vocabulario central (o *Core Dataset*), además de los enlaces entre los dos componentes (*Bindings*), véase Ilustración 26 [20][101]. El CDM es la estructura que almacena la información de las distintas fuentes de datos siguiendo un esquema de representación de datos basado en un estándar. El CDM guarda una relación directa con el *Core Dataset*, ya que éste lo forman los vocabularios clínicos que se utilizan en la plataforma, conteniendo, por tanto, los conceptos que representan la información

clínica del proyecto, así como las relaciones entre los conceptos y los mecanismos y herramientas semánticas para la inferencia de conocimiento en la plataforma. Estos componentes se han desarrollado dentro de una arquitectura orientada a servicios, por lo que estos componentes interactúan con las aplicaciones y las fuentes de datos a través de una serie de servicios desarrollados para ese fin.

2.4.3.1.1. Modelo común de datos

El modelo común de datos o CDM actúa como el esquema del repositorio de datos de la CIS garantizando la representación y el almacenamiento de la información clínica de una forma homogénea en la plataforma. De los distintos estándares de modelos de datos clínicos vistos anteriormente (Sección 2.3.1) se realizó una implementación de este CDM basándose en HL7 RIM debido a su alta capacidad de acoplamiento con vocabularios clínicos y a su cohesión con las fuentes de datos disponibles en los proyectos donde se realizaba la investigación [102]. Sin embargo, debido a la amplitud y complejidad de HL7 RIM, se escogió en un subconjunto de éste, como figura en la siguiente imagen.

Dónde, se realizó una simplificación de algunas de las áreas de HL7 RIM en comparación con los modelos vistos en las Ilustración 11 y 12. En especial se han acotado las clases y atributos relativos a entidades y roles, ya que la información que se va a codificar proviene mayoritariamente de ensayos clínicos por tanto la cantidad de roles y entidades involucradas es limitada o fuera del rango de valores en estos casos.

Este diseño se realizó sobre una base de datos relacional SQL que además incorporó un *wrapper* (o envoltorio) para permitir realizar consultas bajo otros lenguajes de consulta. En la primera versión este *wrapper* utilizó la herramienta D2R Server [103] permitiendo realizar consultas en SPARQL, pero problemas de eficiencia en ciertas consultas que relacionaban distintas tablas [104] hicieron que este componente se actualizara con la herramienta Morph-RDF [105]. De esta forma, se facilita la recuperación de información en varios lenguajes de consulta evitando en cierta forma un gran conocimiento del esquema elegido.

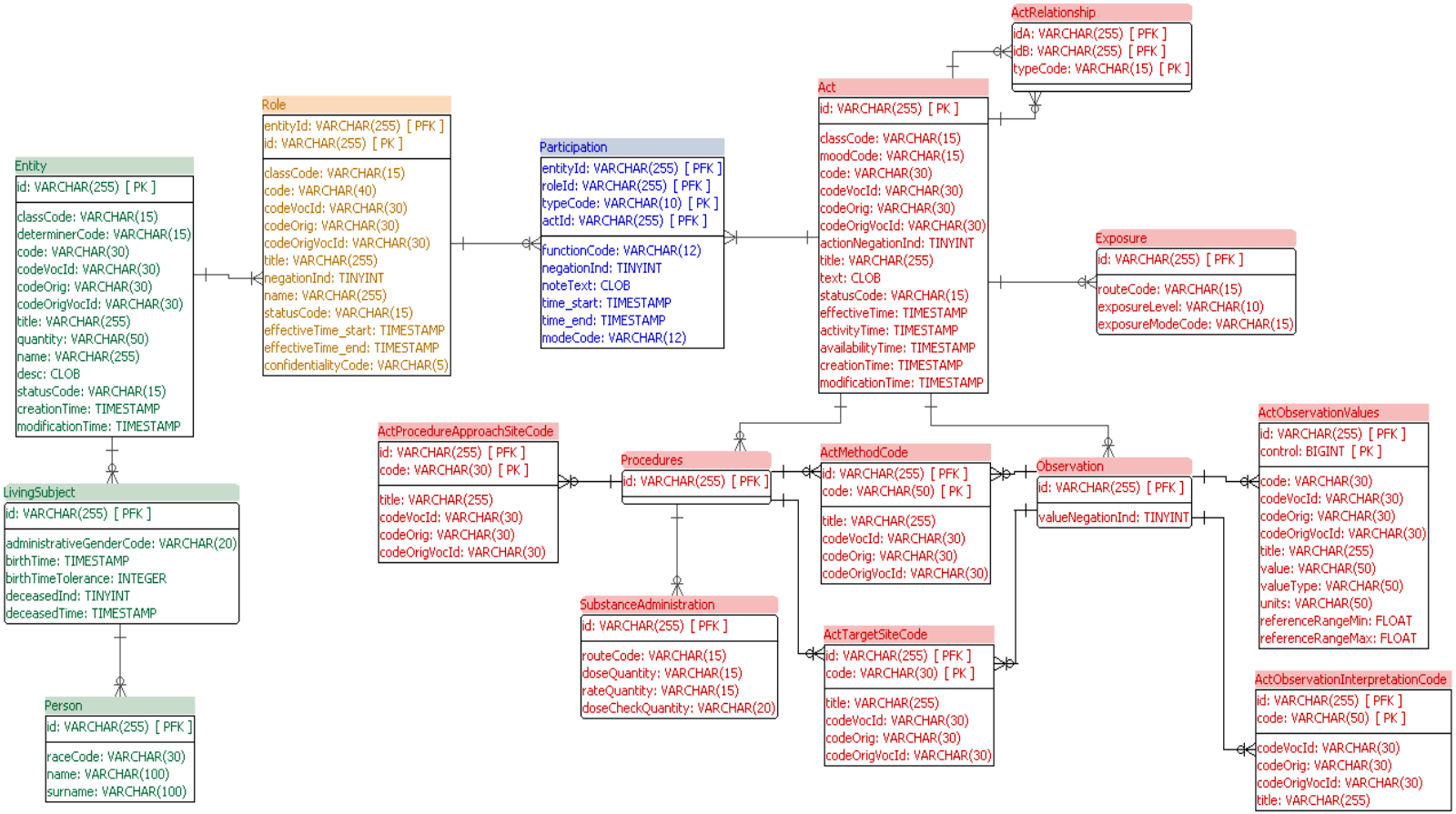


Ilustración 28: Subconjunto de HL7 RIM que compone el CDM

2.4.3.1.2. Core Dataset

El *Core Dataset* es el componente que contiene la base central de conocimientos médicos de la CIS, es decir, contiene los conceptos clínicos presentes en el CDM (extraídos de varias terminologías biomédicas) y las relaciones entre éstos. El principal objetivo de este componente es el de facilitar la extracción e intercambio de la información clínica presente en el CDM, dotándola de un significado homogéneo.

Con este propósito, se realizó un *Core Dataset* formado por varios de los vocabularios descritos en la sección 2.3.2 - Terminologías clínicas, en concreto:

- SNOMED CT: que cubre la mayoría de conceptos clínicos a representar debido a la generalidad de su ámbito.
- LOINC: para la anotación de pruebas y resultados de laboratorios practicados a los pacientes de ensayos clínicos.
- HGNC: para la anotación de observaciones genéticas de los pacientes.

El *Core Dataset* contiene además de la información de estos vocabularios (códigos, términos, relaciones, sinónimos, etc.) los métodos necesarios para inferir este conocimiento en la CIS y los métodos para enlazar estos conceptos con el CDM, así como métodos para mantener una trazabilidad de la información. Todo esto se gestiona bajo el lenguaje estructurado OWL 2.0, que contiene las versiones actualizadas de las terminologías biomédicas y que además contiene información para representar estos conceptos en el CDM. Estos mapeos surgen de iniciativas como el *Term Info Project* [106] que definieron una serie de mapeos y guías para enlazar conceptos a HL7 RIM.

Toda esta gran cantidad de información está representada en un archivo OWL pero para asegurar la gestión de este componente, se implementó este componente utilizando el repositorio semántico *Sesame Server* [107], que facilita la gestión y el acceso a este recurso.

Capítulo 3

3. METODOS

En este apartado se presentan los métodos diseñados para dar respuesta a las preguntas científicas identificadas al comienzo del trabajo en la sección 1.2 y que buscan validar la hipótesis planteada. Estos métodos surgen tras la realización de un estudio centrado en la integración y la interoperabilidad en el área de estándares y proyectos de investigación clínicos. De este estudio se observa como punto en común entre los proyectos la utilización de procesos de interoperabilidad que reutilizan otras tecnologías y estándares para explotar nuevas oportunidades y avances de otras comunidades. Sin embargo también se concluye de este estudio que el diseño o utilización de vocabularios y esquemas de datos no consigue representar de una manera homogénea la información clínica, lo que justifica la necesidad de una solución diferente.

De este análisis y del trabajo realizado en el marco proyectos de investigación surgen los métodos que dan título a la presente tesis doctoral y que, siguiendo otros procesos similares en otros ámbitos [108][109] buscan diseñar un método de normalización y abstracción de consultas que puedan ser integrados y replicados en otros contextos de la práctica clínica. Se ha dividido este capítulo en los dos métodos que lo componen detallando en la sección correspondiente a cada uno, los procesos y pasos que forman

parte de su diseño, acabando cada apartado con la aplicación de cada método dentro del entorno de investigación.

Esta sección comenzará con una introducción en el apartado 3.1 donde se explicará cómo interactúan los métodos diseñados en el presente trabajo entre ellos y su entorno, a modo de visión global de la solución planteada. A continuación se describirá cada uno de los métodos y los componentes que lo forman.

3.1. Visión global de la solución

Los métodos propuestos en la presente tesis forman parte de la solución diseñada para enriquecer y corregir la representación de datos clínicos y facilitar su posterior recuperación. Ambos métodos se diseñan para dar respuesta a unas preguntas científicas que buscan mejorar la representación de datos clínicos y su posterior acceso.El diseño de estos métodos surge tras el análisis realizado en profundidad en proyectos de integración de datos clínicos, donde el punto en común es el uso de interfaces o mediadores de interoperabilidad que utilizando distintos estándares clínicos forman un sistema para la representación de información clínica. Estas soluciones han conseguido avances en la práctica clínica [110] pero la complejidad de sus sistemas y la falta de una forma común de representación de la información son los puntos que buscan ser abordados en el presente trabajo.

Los procesos de normalización en tecnologías como bases de datos surgen a comienzos de los 80, cuando Codd realiza distintas formulaciones y normalizaciones que buscan mejorar los diseños de bases de datos relacionales mediante la realización de transformaciones de datos. Simulando estos procesos de normalización se piensa en una normalización semántica basada en las terminologías, que mejore la representación de los bloques de información clínica (o ideas clínicas) e indirectamente su representación en el modelo de datos, pero sin modificar éste. Además, se busca mejorar la complejidad de los sistemas estudiados mediante la integración de un método que abstraiga de la generación de consultas a usuarios y aplicaciones, y que por tanto, abstraiga de la representación de los datos. Para ello se integran estos métodos en la Capa de Interoperabilidad Semántica (CIS) que comprobará la validez de los métodos y a su vez,

demostrará la posible generalización y adaptación de los métodos a otras CIS de otros ámbitos y sistemas.

Si bien cada método está diseñado para un propósito concreto, es la unión de ambos métodos el que dota de una gran estabilidad al acceso a la representación homogénea de los datos. Mientras el método de normalización semántica está diseñado para interactuar en el proceso de almacenamiento de los datos en el CDM, el método de abstracción de consultas busca mejorar el proceso de la búsqueda y recuperación de información. De esta forma se consigue una homogeneización basa en dos pasos; donde en un primer momento se produce una normalización para mejorar y corregir posibles errores a la hora de almacenar los datos en el CDM, y posteriormente se facilita la creación de consultas utilizando conceptos ya normalizados para consultar los datos.

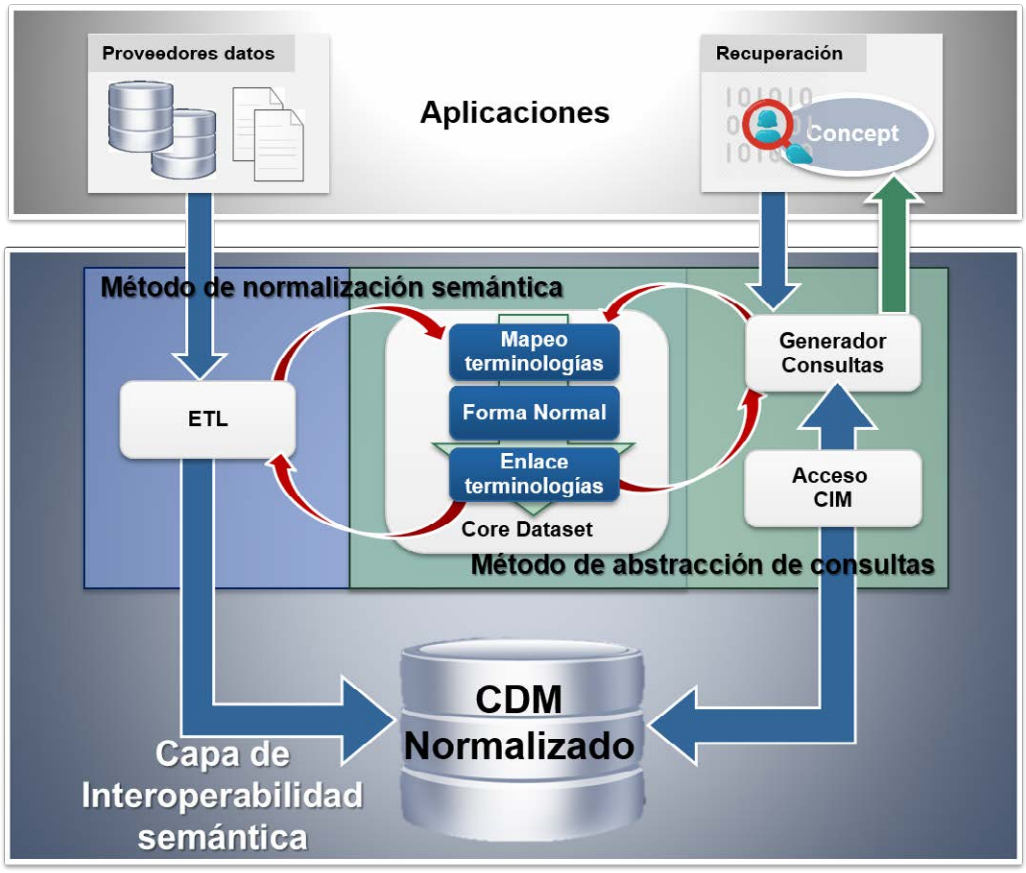


Ilustración 29: Interacción de los métodos diseñados en la CIS

En la anterior ilustración se puede comprobar cómo ambos métodos se integran dentro de la CIS y cómo interactúan con el nivel de aplicaciones, ya sea con las fuentes de datos o

para la recuperación de información a partir de términos clínicos. Donde el componente central, además del CDM, sería el *Core Dataset* que aglutinaría el conocimiento de las terminologías clínicas presentes en el sistema y que por tanto se utilizaría para facilitar la homogeneización de los datos constantemente.

3.2. Método de normalización semántica

La normalización semántica basada en estándares biomédicos es el método diseñado para mejorar la eficiencia y representatividad de los datos en el CDM. Este método surge debido a la complejidad asociada a la integración de datos clínicos y la necesidad de su representación dentro de los esquemas de datos. De estas necesidades y sus posibilidades de representación, en conjunción con estándares terminológicos surge una pregunta científica que es la que trata de abordar este método. ¿Sería posible homogeneizar la representación de datos clínicos existente mediante un proceso de transformaciones lógicas o normalización basado en las terminologías?

Aun escogiendo uno o varios vocabularios clínicos y un único modelo de datos para representar la capa de interoperabilidad semántica, podrían encontrarse distintos tipos de heterogeneidades semánticas debido sobre todo a las distintas posibilidades de representación de un mismo bloque de información clínica (idea clínica) o incluso a errores. Para resolver estos conflictos y problemas de ambigüedad se diseña un método de normalización semántica que transforma los datos desde los orígenes siguiendo las guías y la lógica descriptiva de los estándares seleccionados en la CIS.

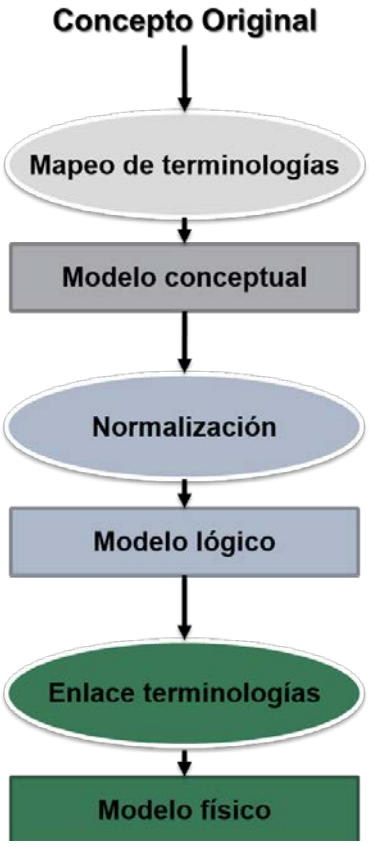


Ilustración 30: Proceso de normalización semántica diseñado

Este método se basa en la interacción de 3 distintos modelos o fases que guardan relación con la transformación de la representación de ideas clínicas hasta su almacenamiento en el modelo de datos. Como se puede apreciar en la Ilustración 30, la idea es que un concepto

pase por tres estados, correspondientes al: i) modelo conceptual, ii) modelo lógico y iii) modelo físico. El primero de ellos (modelo conceptual) hace referencia a la traducción de conceptos a términos presentes en el *Core Dataset* (o las terminologías del sistema). El modelo lógico es el que trata la conversión de estos términos del *Core Dataset* a su forma normalizada, mientras que el modelo físico trata la parte de la traducción de esta expresión a su representación en el modelo de datos (CDM).

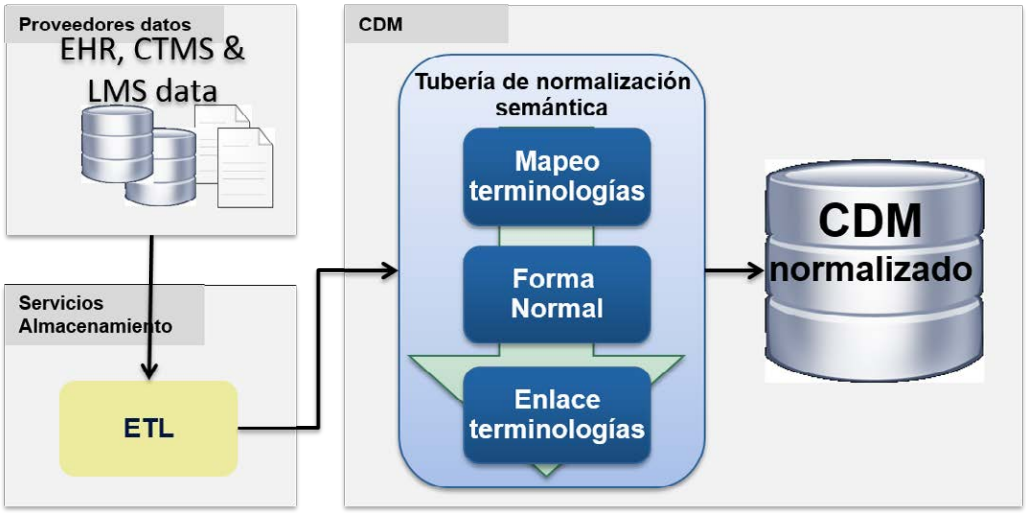


Ilustración 31: Diagrama de componentes durante la normalización semántica

En esta ilustración se puede apreciar el diseño de interacción con el CDM, el método de normalización semántica y las fuentes de datos de cada proveedor. El método de normalización semántica diseñado estaría formado a su vez por tres procesos diferenciados: i) Mapeo de terminologías, ii) normalización y iii) enlace de terminologías con el modelo de datos. Este método es lanzado por las ETLs o los servicios de almacenamiento cuando los proveedores de datos desean introducir nueva información en el sistema.

El primer proceso que se lleva a cabo dentro del método de normalización semántica es la traducción de los conceptos clínicos presentes en las fuentes de datos a sus conceptos más representativos en términos de los vocabularios presentes en el *Core Dataset*. Es decir, este enlazado de terminologías se produce cuando las fuentes de datos están anotadas con terminologías no contempladas en el *Core Dataset* para su conversión a términos "conocidos" por el sistema, pero siempre conservando su concepto original para mantener la trazabilidad y asegurar que se conserva el significado.

Después de que los conceptos sean traducidos a conceptos del *Core Dataset*, éstos se ven envueltos en el proceso de normalización que transforma los conceptos a una forma de representación más informativa y homogénea basándonos en el conocimiento extraído de las terminologías [87][111]. De esta forma, los conceptos se representarán con su forma normal evitando la redundancia y la duplicación en su representación.

Finalmente, se realiza el proceso de unión de las terminologías del *Core Dataset* en el CDM. Este proceso es el encargado de asignar el contexto o ubicación a los conceptos ya normalizados dentro del CDM. Esta unión tiene una fuerte dependencia del modelo de datos elegido y por tanto se debe trabajar con las comunidades de estos desarrolladores y clínicos para poder definir una unión exacta entre terminologías y modelos.

El método de normalización semántica diseñado genera información implícita de las distintas fuentes de datos. Pero se asegura una trazabilidad completa de estos cambios debido al almacenado por un lado de las distintas fuentes; en una versión del CDM sin mediar este método de normalización y en otra versión con normalización. Donde esta última, además almacena los códigos de conceptos que han sido transformados en el paso de procesado a la forma normal, así como las versiones que las distintas terminologías y enlaces utilizados. En las siguientes secciones subsecciones se describirá más detalladamente cada uno de los componentes del método resumido anteriormente.

3.2.1. Mapeo de terminologías

El proceso de mapeo de terminologías es el componente encargado de traducir y enlazar conceptos de terminologías no presentes en el *Core Dataset* a terminologías que forman de él para incrementar la interoperabilidad entre las distintas aplicaciones y las fuentes de datos presentes. Este componente, como el resto de componentes del método de normalización semántica y la CIS, está diseñado para ser integrado en sistemas hospitalarios. Por tanto un requisito importante es que no dependan del acceso a sistemas externos. Es por esto, pensando en la posible integración en otros sistemas o entornos, que se plantea en una solución modular que contenga información sobre posibles mapeos y sinónimos entre distintas terminologías ya conocidas. Utilizando para estos casos

listados ya existentes como UMLS[18] o Bioportal[19], que facilitarían la interacción de este componente en distintos ámbitos.

La principal funcionalidad de este componente es la de traducir un concepto codificado de una terminología existente a un concepto de una de las terminologías del *Core Dataset*, utilizando el *mapping* que sugiere UMLS para ese concepto. En caso de no encontrar un concepto del *Core Dataset* que pueda representar el concepto original, se almacenará únicamente el concepto original.

Además de esta funcionalidad se permite la búsqueda de términos no codificados (cadenas de texto) para su codificación en un concepto presente en el *Core Dataset*, si bien esta funcionalidad requiere la validación de los conceptos por expertos clínicos o por los proveedores de los datos.

3.2.2. Normalización basada en el *Core Dataset*

Una vez los conceptos son representados con términos del *Core Dataset*, se ejecuta el proceso de normalización semántica de conceptos para homogeneizar la representación de la información en el CDM. Para conseguir esto, los datos son transformados a una forma de representación normal que está basada en la transformación de conceptos en su mínima forma de expresión.

Esta normalización aprovecha el conocimiento de los estándares terminológicos presentes en el *Core Dataset* para transformar los conceptos originales en una representación "base" enriquecida con expresiones que la dotan de la especificidad del significado original. Mediante un proceso de transformación basado en reglas lógicas se comprueba si el concepto representado está en su forma de representación más generalista; en caso negativo se transforma éste en su concepto padre y se añaden además los pares relación-valor que representan la especificidad que se pierde al subir un nivel en su jerarquía, para después volver a comprobar si el nuevo concepto se encuentra en su forma normal o hay que volver a repetir el proceso.

[18] https://www.nlm.nih.gov/research/umls/
[19] https://bioportal.bioontology.org/

Es decir, el proceso de normalización de un concepto es un proceso recursivo que surge al normalizar todos los conceptos que forman parte de la expresión definitoria de un concepto hasta conseguir una expresión formada por sus conceptos definitorios. Por tanto, esta normalización de un concepto resulta en una expresión formada por uno o más conceptos del *Core Dataset* que representan el mismo significado de su concepto original pero formado por una composición de conceptos más generalistas que ayudan a mantener una representación homogénea del sistema. En la siguiente ilustración se puede apreciar la formalización de este proceso de normalización.

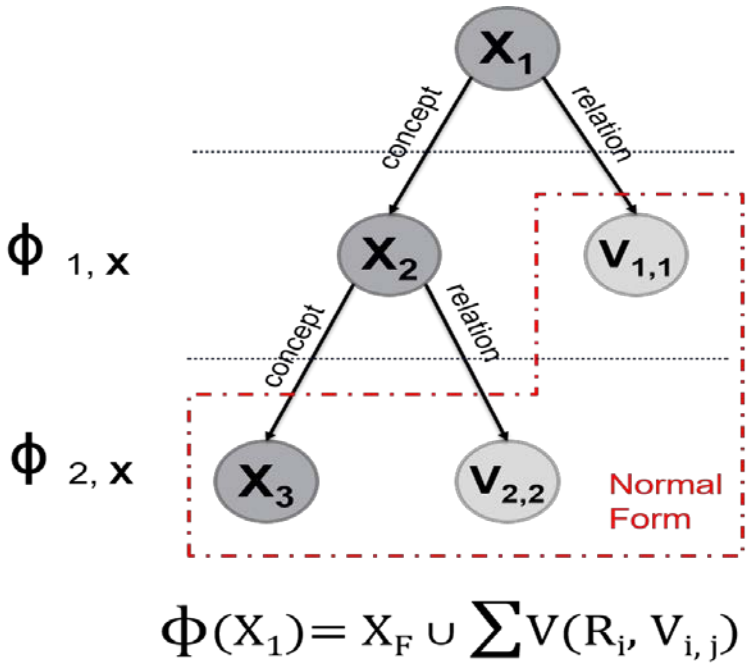


Ilustración 32: Formalización del proceso de normalización semántica

En el caso de no encontrar una representación que mejore el concepto original, se mantendría el mismo concepto como su expresión más generalista. Este proceso de normalización explora y expande la semántica de los conceptos con sus relaciones definitorias; pero, como en el caso del enlazado de terminologías, se mantienen también los conceptos originales para conseguir una trazabilidad completa de los datos originales o para realizar búsquedas por sus conceptos originales en caso de que fuera necesario.

3.2.3. Enlace de terminologías y modelos de datos

Una vez que los conceptos han sido traducidos a conceptos del *Core Dataset* y éstos han sido normalizados semánticamente, se tiene que representar la información relacionada con estos conceptos en el CDM, es decir, la representación física en el modelo de datos de los términos y expresiones ya normalizadas. Por tanto, este proceso es el encargado de determinar en qué clases y atributos del CDM se representan y almacenan cada uno de los términos presentes en el *Core Dataset*.

Para saber en qué clases y atributos del CDM se pueden encontrar los conceptos del CDM, se diseña un sistema basado en anotaciones donde a cada concepto del *Core Dataset* se le indican posibles ubicaciones (ubicación primaria y alternativas) dentro del CDM, debido a la semántica y tipología del término. Estas anotaciones, por tanto, incluyen la ubicación exacta (tabla y atributo) del CDM donde se puede almacenar el concepto en cuestión y se añade al archivo que forma el *Core Dataset* durante la creación de la ontología que integra las distintas terminología. Los conceptos, en este punto se encuentran, ya normalizados, por tanto, se pasarían a clasificar cada uno de los conceptos que forman la expresión normalizada del concepto.

De esta forma, en el caso de una expresión normalizada de más de un concepto, cada uno de ellos tendría una ubicación en el CDM y guardarían una relación directa ya que hacen referencia a una única expresión clínica. Obteniendo finalmente una ubicación preferida del CDM para cada concepto que forma una expresión normalizada, además de una serie de ubicaciones alternativas para evitar solapamientos dentro de las terminologías.

3.2.4. Aplicaciones en el CIM

Estos tres procesos descritos anteriormente forman parte del método de normalización semántica y son diseñados para mejorar la representación de los datos clínicos y para su posterior evaluación, éstos son integrados en la CIS, interactuando por tanto con el CDM y el *Core Dataset*.

Su ejecución viene precedida por los servicios de almacenamiento a la hora de introducir nuevos datos en el CDM. Una vez el servicio recibe nuevos datos para almacenar en el CDM se producen dos líneas de ejecución para almacenar la información en dos CDM;

uno con los datos normalizados y otro sin normalizar. Será únicamente el primero de estos CDM el que involucre al método de normalización semántica ya que el almacenamiento de los datos sin normalizar no requiere de ningún proceso de normalización ni corrección de la información original.

Por tanto de esta forma se asegura una completa trazabilidad de la información original y se reduce el riesgo de pérdida de información en cualquiera de los casos. Ya que además de presentar la información original sin transformar, en cada uno de los componentes del método de normalización semántica se registran las transformaciones realizadas si las hubiera y las versiones de los componentes involucrados en los cambios.

Además, como se detallará en el método de abstracción de consultas, la normalización semántica será utilizada a la hora de generar posibles consultas en el CDM normalizado. Para, de esta forma, asegurar una completa recuperación de la información deseada.

3.3. Método de abstracción de consultas

La abstracción de consultas es el método diseñado para facilitar la recuperación de información clínica presente en un modelo de datos. Para esto, se diseña un método basado en estándares terminológicos que faciliten la creación de consultas en base a la representación de términos o expresiones clínicas presentes en el sistema. Este método como en el caso anterior, surge de a la complejidad asociada al campo de la integración de datos clínicos y la gran extensión de modelos de datos. Al igual que en el apartado anterior se explicaba un método que mejoraba la homogeneización de los datos en el CDM, en este caso se trata de la necesidad de acceder y recuperar la información ya almacenada. El presente método, debido a la modularidad y necesidad de utilizar diferentes terminologías y modelos de datos, se diseña para dar respuesta a la siguiente pregunta científica, ¿es posible abstraer a las aplicaciones y/o investigadores del modelo de representación de datos elegido?

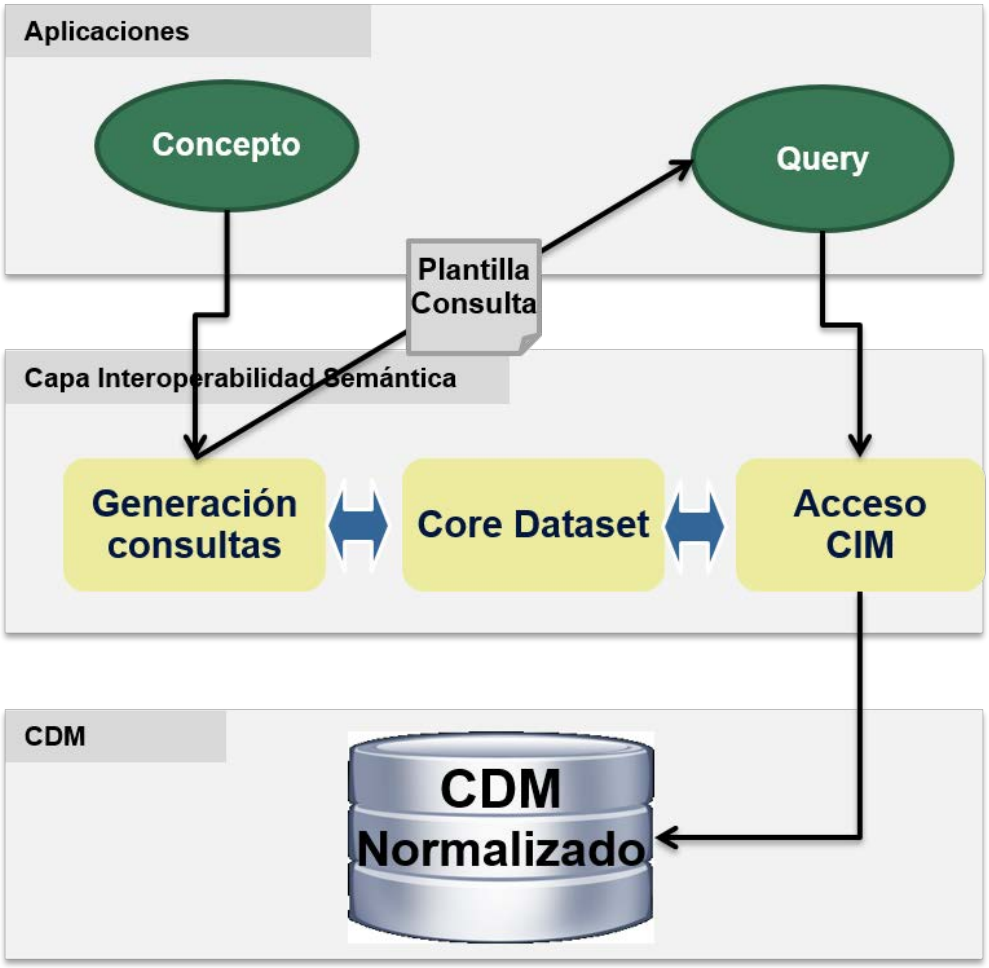


Ilustración 33: Diagrama de componentes en la abstracción de consultas

El principal objetivo de este método de abstracción es facilitar un acceso homogéneo para recuperar la información normalizada del CDM. Para este fin, se ha diseñado un método como el mostrado en la Ilustración 33 formado por la interacción de las aplicaciones con tres servicios de la CIS y el CDM para encapsular el esquema del CDM y el lenguaje de consulta utilizado.

En la imagen se aprecian tres niveles (aplicaciones, CIS y CDM), donde el método se iniciaría con la petición de una plantilla de consulta para un término de los vocabularios del sistema. Este proceso está diseñado para ser accedido de dos maneras distintas dependiendo de la utilización del contexto del término a buscar (se describirá en las siguientes subsecciones), devolviendo una plantilla con la consulta mínima y necesaria para obtener información presente en el CDM sobre ese concepto y una serie de filtros

adicionales para restringir aún más la consulta final. Para finalmente lanzar esta consulta en el CDM a través de los puntos de acceso al modelo de datos.

Para ello, este método de abstracción de consultas se diseña con dos componentes principales que aseguren la modularidad del sistema y su fácil integración en otros sistemas con otros modelos de datos y lenguajes de consulta: i) la **librería de consultas** y el ii) **envoltorio** para la traducción de consultas.

La **librería de consultas** forma parte de la generación de consultas y es una estructura compuesta por un conjunto de plantillas que contienen consultas genéricas y sus posibles filtros o añadidos basadas en el CDM. Estas plantillas cubren todas las posibles expresiones clínicas (conceptos del *Core Dataset*) presentes en el CDM, y están divididas por los bloques de información disponibles en el CDM. Es decir, estas plantillas cubren la información presente en el CDM dividiendo éste en bloques temáticos y

Una vez se tiene la plantilla de consulta deseada, la aplicación puede construir una consulta más o menos específica siguiendo la información contenida la plantilla. A partir de la consulta, se podrá usarla para acceder a la información del CDM. Además se utilizará un **envoltorio** para realizar la traducción de la consulta de la plantilla al modelo de datos utilizado para, de esta forma, facilitar la traducción entre distintos modelos de datos y lenguajes de consultas utilizados para su representación.

La modularidad de la solución diseñada facilita la actualización o migración de cualquiera de los componentes envueltos en el método de abstracción de consultas: CDM, servicios, lenguaje de consulta, plantillas, etc. En la siguiente subsecciones se describen las dos formas de recuperar las plantillas de consultas dependiendo del contexto de los conceptos o ideas clínicas que se quieran consultas o del CDM del que se quiera recuperar la información.

3.3.1. Búsqueda de información sin contexto o no guiada

El componente central o principal del método de abstracción de consultas es el generador de consultas. La función de este componente, como se ha descrito anteriormente, es la de facilitar una plantilla de consulta para un concepto del *Core Dataset* del que se quiera recuperar información en el CDM. Pero estos argumentos varían dependiendo de dos

situaciones o interfaces dependientes de si el contexto del concepto está implícito o no en la búsqueda.

Por tanto, el generador de consultas tiene dos interfaces para obtener una plantilla de consulta sobre un concepto del *Core Dataset*: i) en base a un concepto descontextualizado y normalizado en el CDM y ii) en base a conceptos contextualizados. El primero de estos interfaces se refiere al concepto del *Core Dataset* del cual se quiere recuperar información en el CDM.

Esta interfaz es la que nos ocupa en esta subsección e involucra al método de normalización semántica, concretamente al componente de normalización semántica y a su enlace con el CDM. Por tanto, esta interfaz recibe como único argumento el concepto del *Core Dataset* del que se quiere buscar información en el CDM. Este concepto es procesado de manera análoga a como lo hacen en el método de normalización semántica. Primero se busca un sinónimo que pueda representar este concepto con términos del *Core Dataset*. Después se normaliza este concepto obteniendo su expresión normalizada para finalmente obtener su posible ubicación en el CDM. Dependiendo de esta ubicación se obtendrá una de las plantillas que forman la librería de consultas asegurándonos mediante este proceso que el concepto del que se desea recuperar información del CDM sufre las mismas transformaciones que a la hora de almacenarlo. Finalmente se rellena esta plantilla con la expresión normalizada además de la meta-información que indica con que versiones de que componente se ha producido cada proceso hasta llegar a esta plantilla.

Esta solución está ideada para obtener consultas de manera rápida y abstrayendo a la aplicaciones totalmente del modelo de datos utilizado en el CDM, de los conceptos presentes en el CDM y de las terminologías del *Core Dataset*. Además esta interfaz está pensada para ser utilizada en el CDM que contiene la información normalizada, completando de esta forma un círculo a la hora de normalizar los datos al almacenar y al consultar la información almacenada.

3.3.2. Búsqueda de información contextualizada o guiada

Además de la interfaz ideada para un conjunto de datos normalizados, se ha diseñado una interfaz para obtener plantillas de consultas más específicas de un contexto que puedan ser utilizadas en datos sin normalizar o simplemente con otro tipo de búsquedas. A este

interfaz lo llamamos contextualizado o guiado, ya que además del concepto del *Core Dataset* del que se quiere recuperar información en el CDM, se utilizan como parámetros una lista de clases y atributos del modelo de datos donde se pretende realizar esta búsqueda.

Por tanto, en esta interfaz no se realiza una normalización de los conceptos argumento, ya que se indican las ubicaciones del CDM (contexto) donde se desea realizar la búsqueda. Es la propia aplicación o investigador, los encargados de enviar esta información al generador de consultas para obtener la plantilla. Este componente, por tanto, únicamente selecciona una de las cinco plantillas que más se asemeje al contexto enviado y añade el concepto a la plantilla para poder realizar la consulta.

El principal hándicap de esta solución viene dado por las diferencias que puede haber entre las posibles ubicaciones de los elementos (redundancia), interpretaciones de relaciones existentes entre conceptos del *Core Dataset* y el CDM por parte de especialistas clínicos e incluso errores humanos a la hora de representar la información en el CDM o anotarla. Si bien este método permite a las aplicaciones cierta libertad para poder construir de una manera guiada consultas relacionadas con el CDM.

Capítulo 4

4. PRUEBAS Y EXPERIMENTOS

En el capítulo anterior se describieron los métodos de normalización semántica y abstracción de consulta basada en estándares. Mientras que la Capa de Interoperabilidad Semántica (CIS), se ha descrito como el marco de trabajo que integra estos métodos. En el presente capítulo se describen las pruebas y experimentos llevados a cabo para demostrar la hipótesis planteada en este trabajo.

"Es posible enriquecer y corregir la representación clínica de datos mediante un proceso de normalización automática de conceptos médicos y abstracción de consultas que facilite la integración, homogeneización y el acceso a estos sin conocimiento del modelo de datos utilizado"

Para verificarla, se ha implementado una solución programática siguiendo el diseño explicado en el capítulo 3 mediante servicios SOAP (axis2 1.6.2 y java 1.7). Para validar ambos métodos, esta implementación ha sido integrada en la CIS descrita en la sección 2.4.3 dentro de los proyectos INTEGRATE y EURECA con el fin de facilitar la integración con sus aplicaciones. Dado que esta solución se trata de un prototipo de un sistema completo de integración, resulta recomendable realizar sus experimentos a través

de requisitos concretos y casos de uso reales para validar su posible incorporación a un entorno real. Por tanto, se utilizará como caso de uso su validación en estos proyectos mediante la validación de los distintos componentes en los diferentes escenarios en los que hayan sido utilizados. Concretamente, INTEGRATE se desarrolló desde el 1 de Febrero de 2011 a 31 de Octubre de 2014 y EURECA desde el 1 de Febrero de 2012 a 31 Noviembre de 2015. El autor de la tesis ha trabajado en ellos como investigador del Grupo de Informática Biomédica.

De esta forma, en las siguientes secciones del presente capítulo se va a describir en primer lugar la implementación realizada para probar los distintos métodos estudiados (Sección 4.1). También se van a describir los conjuntos de datos y fuentes con datos reales donde se han realizado las pruebas (Sección 4.2) y como se han anotado cada uno de estas fuentes de datos (Sección 4.3). Se describen en la sección 4.4 los experimentos y pruebas básicas que se han realizado para comprobar el correcto funcionamiento de los métodos diseñados con respecto a la hipótesis. Además como escenario de validación se tienen en cuentan dos de los escenarios clínicos de los proyectos europeos donde se ha evaluado la utilización de estos métodos en la Sección 4.5 y los conjuntos de pruebas para validar los resultados de los escenarios mediante el método de abstracción de consultas (Sección 4.6). Finalmente se adjunta la evaluación obtenida en la revisión final de los proyectos europeos, desarrollada por expertos de la Comisión Europea.

4.1. Aplicación de los métodos integrados en la CIS

El diseño de los métodos de normalización semántica y de abstracción de consultas se realizó utilizando el modelo presentado en el anterior capítulo. Además, estos métodos fueron integrados en la CIS desarrollada para mejorar la representación de los datos y facilitar su acceso. Esta CIS fue desarrollada [20] basándose en una arquitectura SOAP para asegurar una comunicación estable y permanente entre los distintos componentes, donde además se utilizaban tecnologías como *Sesame Server* para la gestión del *Core Dataset* en memoria y MySQL 5.7 como sistema generador de las bases de datos.

La CIS fue desplegada como una solución completa en cada institución proveedora de los datos para cada uno de los escenarios que se validarían en los proyectos europeos, emulando una instalación real en su propio entorno, siguiendo de esta forma el enfoque híbrido de distribución de los datos. Esta CIS fue instalada en servidores Ubuntu server con alrededor de 8 GB de memoria RAM donde además se aseguraba el acceso a estos servicios mediante un servidor proxy de autenticación desarrollado por otra institución para cumplir con el protocolo de seguridad.

A continuación se detallará la implementación e integración de cada uno de los métodos diseñados en la presente tesis para realizar las pruebas y experimentos necesarios para validar la tesis en el entorno de los proyectos INTEGRATE y EURECA.

4.1.1. Normalización semántica basada en SNOMED CT

La implementación del método de normalización semántica sigue el diseño detallado en la sección 3.2, donde se presentaba un método basado en tres componentes principales: i) mapeo de terminologías, ii) normalización y iii) enlace de terminologías con el modelo de datos. Cada uno de estos componentes se presenta como una serie de procesos modulares y unitarios, facilitando la actualización o modificación de uno o varios de los componentes.

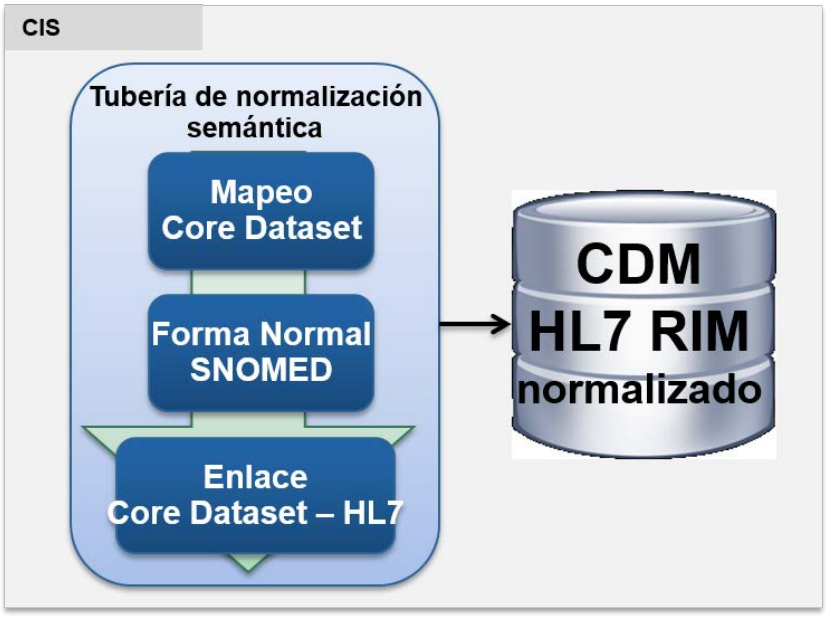


Ilustración 34: Componentes que forman el método de normalización semántica implementado

4.1.1.1. Implementación del enlazado de terminologías

El proceso de enlazado de terminologías sigue el diseño presentado en la sección 3.2.1 donde se realiza una implementación a modo de un método del servicio *Core Dataset*. Este método utiliza la base de datos de UMLS Methateshaurus para conseguir la mejor traducción posible de cada concepto, dando prioridad a su traducción a conceptos de SNOMED. Si bien es cierto que la cobertura de SNOMED varía entre un 60-80% en los ensayos clínicos utilizados [112], se complementa este porcentaje con los otros vocabularios que forman el *Core Dataset* y con la posibilidad de introducir traducciones a mano siguiendo el consejo de expertos en la anotación, aunque siempre se conserva el concepto original para mantener su trazabilidad.

Si bien en UMLS existen más de 200 terminologías indexadas [113], realizar búsquedas en todas las terminologías podría suponer una pérdida de eficiencia que podría llegar a sobrecargar la CIS, por tanto, se ha seleccionado un subconjunto de terminologías. Este subconjunto se ha seleccionado en base a la descripción de los datos que se van a utilizar en los distintos pilotos de EURECA e INTEGRATE dando prioridad a las terminologías más reconocidas en la práctica clínica y que no están presentes en el *Core Dataset* como ICD 9 y 10, CTCAE, NCIt o MedDRA.

Estas terminologías junto con las que están incluidas en el *Core Dataset* (SNOMED, LOINC y HGNC) cubren la mayoría de los datos presentes en estos proyectos de investigación, si bien la modularidad de este componente permite fácilmente incorporar otras terminologías presentes en UMLS.

4.1.1.2. Forma normal de SNOMED

Para enriquecer la representación de los conceptos en el CDM se diseña el proceso de normalización detallado en la sección 3.2.2, en el que se basa el siguiente desarrollo para su validación. Este proceso se basa en la transformación de los conceptos a una forma de representación normal que está basada principalmente en el mecanismo de normalización definido por SNOMED CT, ya que este actúa como el vocabulario central del *Core Dataset*.

La forma normal de SNOMED está definida como la representación de un concepto formada por conceptos pre-coordinados en la cual todos los conceptos son primitivos (un concepto primitivo es aquel cuya composición no es suficiente para definir únicamente a la expresión clínica) y que ocurre a raíz de aplicar un conjunto de reglas de transformaciones lógicas en todos sus conceptos. El proceso de normalización es un proceso recursivo donde se comprueba que el concepto a estudiar se encuentre en su forma primitiva de SNOMED y si tiene relaciones definitorias que lo complementen. En caso de encontrarse relaciones definitorias, se sustituye este concepto por la expresión de SNOMED que lo define (conjunto de pares atributo-valor que no son relaciones jerárquicas). Una vez se sustituye el concepto original por los nuevos conceptos, se vuelve a comprobar que los nuevos conceptos sean primitivos, en cuyo caso ya habríamos llegado a su forma normal, o en caso negativo se volverían a sustituir el concepto o los conceptos que no sean primitivos por su expresión definitoria de SNOMED y se volvería a comprobar como en el paso anterior. Si además el concepto original es no primitivo (completamente definido) se sustituye el concepto original por su super-tipo primitivo más próximo como concepto foco. En caso de que el concepto original sea primitivo, se mantiene este. Queda por tanto su forma normal formada por su concepto foco o el concepto original (en caso de que sea primitivo) más las relaciones definitorias en forma normal y primitiva.

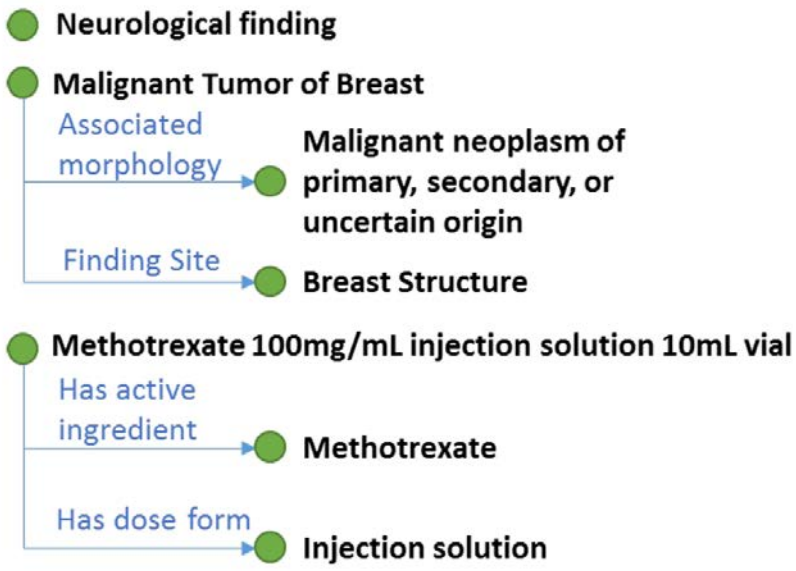


Ilustración 35: Ejemplos de forma normal de SNOMED

En la imagen anterior se pueden apreciar 3 ejemplos de conceptos de SNOMED CT:

- *Neurological finding:* Este es un concepto primitivo que no tiene relaciones definitorias que no sean jerárquicas (is_a), por tanto este concepto ya se encuentra en su forma normal.
- *Malignant Tumor of Breast:* Este es un concepto definido completamente (no primitivo) por sus relaciones definitorias. Por tanto, su forma normal estará formada por sus relaciones definitorias (que en este caso son primitivas) además de un concepto foco o primario que sustituye al original y que en este caso sería *Disease* (concepto raíz de la rama de *Malignant Tumor of Breast*) como concepto supertivo primitivo más próximo al original.
- *Methotrexate 100mg/mL injection solution 10mL vial:* Este es un concepto primitivo cuya expresión además se forma con las relaciones definitorias de dosis e ingrediente activo. Por tanto, su forma normal estará formada por sus relaciones definitorias (en este caso también son primitivas) y como concepto primario el mismo concepto ya que éste es primitivo.

La diferencia en la normalización de estos dos últimos conceptos es la que da lugar a la forma normal corta y larga definidas por SNOMED. Ya la forma normal corta de un concepto primitivo debería ser el propio concepto únicamente, mientras que la forma normal larga de un concepto siempre contiene además sus relaciones definitorias. En la CIS se ha decidido utilizar la forma normal corta de SNOMED en la mayoría de casos ya que la utilización de la forma normal larga podría suponer un aumento considerable de la información en la base de datos. Las excepciones en las que se utiliza la forma normal larga de SNOMED es con conceptos que pertenecen a las ramas y unidades semánticas de:

- *Organism*
- *Pharmaceutical / biologic product*
- *Physical object*
- *Specimen*
- *Substance*
- *Environment or geographical location*

Por tanto, los conceptos que pertenecen a cualquiera de estas ramas de SNOMED son transformados a su forma normal larga tras consultar con expertos clínicos la necesidad de añadir información de al menos estas ramas al CDM. Es decir, en la CIS se realiza un proceso de normalización mixto basado en el mecanismo de normalización de SNOMED dependiendo de la semántica de los conceptos a estudiar. En caso de conceptos que no pertenezcan a SNOMED estos se mantienen intactos en este proceso de normalización.

4.1.1.3. Desarrollo del enlace de terminologías con el modelo de datos

El proceso de unión entre las terminologías y el CDM diseñado en la sección 3.2.3 ayuda en el proceso de ubicar los conceptos del *Core Dataset* en el modelo de datos para su almacenamiento y la posterior recuperación de información. Este proceso se basa en la anotación de las posibles ubicaciones de cada concepto del *Core Dataset* en la estructura que lo gestiona [114]. Para esto se realiza un script que recorre los conceptos y los anota en las guías de interoperabilidad y recomendaciones de HL7 [52] y sus grupos de trabajo, así como del proyecto TermInfo[106].

Estas recomendaciones se basan en la ubicación de los conceptos de SNOMED dentro de su propia jerarquía para elegir la ubicación dentro de HL7 RIM. Por tanto, si un concepto del *Core Dataset* se encuentra en la jerarquía de *Clinical Finding*, estos conceptos se almacenarán como observaciones en el CDM, y serán anotados como tal en el propio *Core Dataset*. En caso de conceptos de terminologías como LOINC y HGNC su ubicación en el CDM es bastante evidente debido a la especificidad de estas terminologías. Por tanto, los conceptos de LOINC representan pruebas de laboratorio y por tanto serán anotados como presentes en la clase *Act* como Observaciones en el CDM. Mientras que los conceptos genéticos de HGNC serán representados y anotados en el *Core Dataset* como pertenecientes a la clase *Entity* del CDM. Estos últimos serán relacionados con los pacientes en el CDM mediante la observación anotada en SNOMED como *Genetic Finding*.

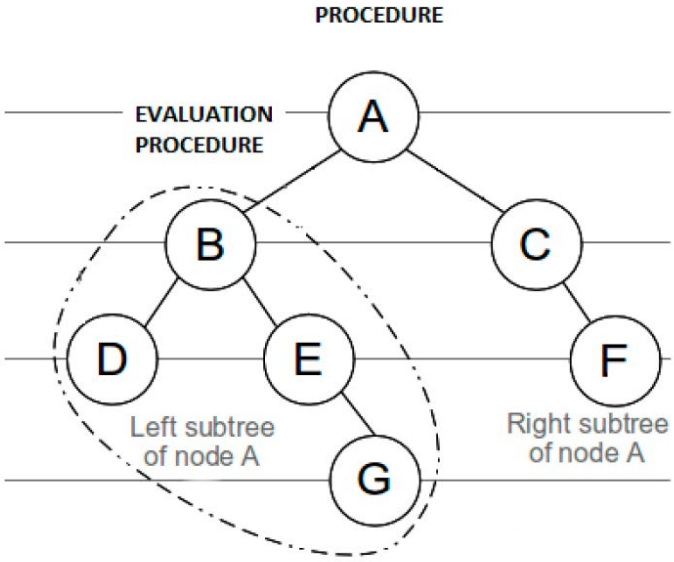


Ilustración 36: Ejemplo de solapamiento de conceptos en SNOMED

Al realizar una categorización de los conceptos del *Core Dataset* para saber en qué clases y atributos del CDM se pueden almacenar estos conceptos y al ser SNOMED CT una terminología polijerárquica, nos podemos encontrar con solapamientos en la posible ubicación de conceptos que pueden guardar relación con varias ramas. Un ejemplo de este solapamiento se muestra en la Ilustración 36, donde nos encontraríamos con conceptos de la rama de *Procedure* que se anotarían como procedimientos en el CDM y su subrama *Evaluation Procedure* que deberían ser por su contexto observaciones del CDM. Estos solapamientos no tienen recomendaciones claras en los proyectos y grupos de trabajo anteriores, por tanto, se han decidido anotar distintas las distintas alternativas, dando prioridad a la específica en cada caso, es decir, se anotará como primera alternativa la que corresponda a la menor sub-rama (en el ejemplo anterior, para uno de los conceptos B-D-E-G se anotaría como primera alternativa observación).

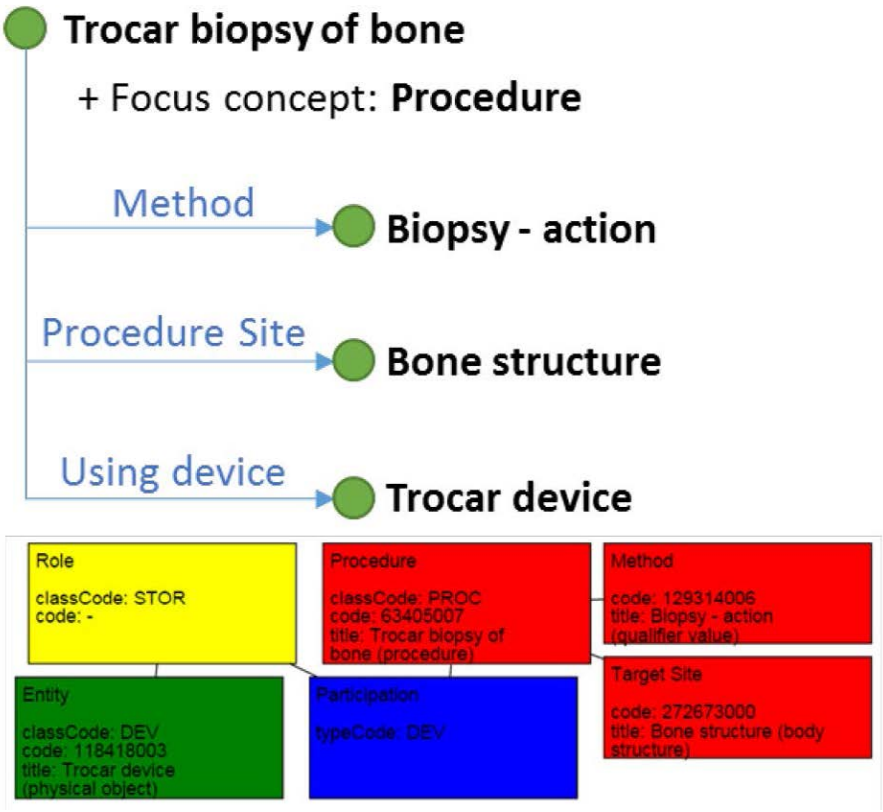


Ilustración 37: Ejemplo de enlazado de un concepto normalizado en el CDM

Los conceptos en este punto se encuentran ya normalizados, por lo que se pasaría a clasificar cada uno de los conceptos que forman la expresión normalizada del concepto. En el ejemplo de la Ilustración 37 se puede apreciar el concepto *Trocar biopsy of bone* y como se normaliza con los conceptos *Procedure* (concepto foco), *Biopsy – action* (a través de su relación *Method*), *Bone structure* (a través de su relación *Procedure Site*) y el concepto *Trocar device* (a través de su relación *Using device*). Los cuatro conceptos tendrían una ubicación en el CDM y lógicamente guardarían una relación directa ya que hacen referencia a una única expresión clínica (*Trocar biopsy of bone*).

4.1.2. Abstracción de consultas basadas en HL7 RIM

El método de abstracción de consultas detallado en la sección 3.3 es el método ideado para mejorar la recuperación de información del CDM en base a conceptos presentes en el *Core Dataset*. Por tanto, para esta implementación se parte de un método que tiene que estar dirigido para ser usado en un CDM basado en HL7 RIM y en un *Core Dataset* y un proceso de normalización semántica basados principalmente en SNOMED CT.

El componente principal de este método es el servicio *Query Builder* que contiene la librería de consultas que definen las plantillas para la recuperación de información en el CDM. Esta librería de consultas es una estructura compuesta por un conjunto de consultas genéricas y sus posibles filtros basados en bloques de información del CDM. Estas plantillas cubren todas las posibles expresiones clínicas (conceptos del *Core Dataset*) presentes en el CDM. Para esto se han definido 5 plantillas genéricas de consulta en base al tipo de información a recuperar y al concepto del *Core Dataset* asociado:

- Observación o diagnosis: esta plantilla contiene consultas para recuperar información asociada a conceptos del *Core Dataset* (todos los conceptos de LOINC y parte de conceptos de SNOMED CT) que después de ser normalizados su enlace con el CDM es la subclase de *Act, Observation* del CDM.
- Procedimientos: esta plantilla contiene consultas y filtros para recuperar información asociada a conceptos del *Core Dataset* que después de ser normalizados, su enlace con el CDM es la sublcase de *Act, Procedure* del CDM.
- Administración de substancia: esta plantilla contiene consultas y filtros para recuperar información asociada a conceptos del *Core Dataset* que después de ser normalizados, su enlace con el CDM es la sublcase de *Act, Substance Administration* del CDM.
- Entidades: esta plantilla contiene consultas y filtros para recuperar información asociada a conceptos del *Core Dataset* que después de ser normalizados, su enlace con el CDM es la clase *Entity* del CDM. Esta clase hace referencia sobre todo a dispositivos, productos, sustancias o genes, por tanto todos los conceptos de HGNC se obtendrán con esta plantilla.
- Demográfica: esta plantilla contiene consultas y filtros para recuperar toda la información asociada a pacientes (id, género, fecha de nacimiento…etc.).

```
<template version="0.5" id ="001" description="Template for querying observations">
    <templateClass>Observation</templateClass>
    <rimClasses>
    <inputConcept>%concept%</inputConcept>
    <normalizedOutput>%normalform%</normalizedOutput>
    <sqparqlQuery>
        <![CDATA[
            SELECT DISTINCT ?id ?code ?patientId ?birthTime ?effectiveTime $$optionalAttributes$$
            WHERE {
                ?instPerson             hl7rim:person_id ?patientId.
                ?instPerson             hl7rim:person_code '337915000'.
                OPTIONAL{?instPerson    hl7rim:person_birthTime ?birthTime}
                ?instPerson             hl7rim:person_role ?instRole2.
                ?instRole2              hl7rim:role_entityId ?patientId.
                ?instRole2              hl7rim:role_classCode "PAT".
                ?instRole2              hl7rim:role_participation ?instPart2.
                ?instPart2              hl7rim:participation_entityId ?patientId.
                ?instPart2              hl7rim:participation_act ?instAct.
                ?instAct                hl7rim:act_code ?code;
                                        hl7rim:act_id   ?id.
                OPTIONAL {?instAct      hl7rim:act_effectiveTime ?effectiveTime}
                $$optionalQueries$$
                FILTER (?code IN (isAnySubclassOf(%Observation_code%)))
                $$optionalFilters$$
            }
        ]]>
    </sqparqlQuery>
    <optionals>
        <optional id="classCode">
        <optional id="moodCode">
        <optional id="codeVocId">
        <optional id="codeOrig">
        <optional id="codeOrigVocId">
        <optional id="actionNegationInd">
        <optional id="title">
        <optional id="text">
        <optional id="statusCode">
        <optional id="effectiveTime">
        <optional id="activityTime">
        <optional id="availabilityTime">
        <optional id="creationTime">
        <optional id="modificationTime">
        <optional id="value">
        <optional id="interpretationCode">
        <optional id="methodCode">
        <optional id="targetSiteCode">
        <optional id="entity">
        <optional id="relationShip">
    </optionals>
</template>
```

Ilustración 38: Plantilla genérica para los diagnósticos u observaciones.

Todas estas plantillas están diseñadas en el lenguaje estructurado XML y como ya se ha indicado anteriormente, contiene información como la consulta con los parámetros y filtros mínimos para recuperar información asociada al concepto argumento. También contiene una serie de parámetros y filtros opcionales para realizar consultas más específicas. Además, como se puede apreciar en la Ilustración 38, se incluye información estructurada sobre la representación del concepto original en su forma normalizada y su enlace correspondiente en el CDM.

Una vez se tiene la plantilla de consulta deseada, la aplicación puede construir una consulta más o menos específica siguiendo la información contenida en el XML. Y a partir de la consulta, utilizar el servicio de acceso para recuperar la información del CDM. Este servicio está formado además por un **envoltorio o *wrapper* de Morph-RB [105]** para realizar la traducción de la consulta de la plantilla al CDM, ya que éste se encuentra en un sistema relacional.

4.2. Descripción de los datos utilizados

Los dos métodos presentados en la presente tesis doctoral, han sido integrados dentro de la CIS, explicada en la sección 2.4.3.1, para posibilitar el acceso homogéneo a datos clínicos. Si bien estos métodos pueden utilizarse en distintos ámbitos de la práctica clínica, se ha integrado esta solución en los proyectos europeos INTEGRATE y EURECA, que están principalmente enfocados a la integración de datos provenientes de ensayos clínicos en el estudio de cáncer.

En esta sección se describen los datos facilitados por cinco instituciones clínicas que formaron parte de estos proyectos (Universität des Saarlandes, Institute Jules Bordet, Maastro Clinic, University of Oxford y German Breast Group Forschungs GmBH), y que facilitaron datos reales de pacientes para la validación de las distintas herramientas en los escenarios clínicos que se definieron en los proyectos. Estos datos por tanto, se representaron utilizando los métodos propuestos por la CIS, siendo almacenados dentro del CDM a través del método de normalización semántica basado en estándares. Posteriormente se realizó una validación mediante las herramientas que harían uso del método de abstracción de consultas.

Cada una de las fuentes de datos han sido almacenadas en el CDM como se ha indicado, si bien, cada uno de los proveedores facilitaba el accesos a estos datos de una manera distinta, dependiendo de diversos factores: idioma, tecnología de exportación, anotación, modelo de datos…etc. Por tanto, se necesitaba de un primer paso de entendimiento y preparación para su almacenaje dentro del CDM que principalmente era realizado manualmente con la ayuda de la especificación de los datos facilitada por los proveedores de datos y los propios expertos clínicos de cada institución.

Cabe destacar que cada uno de las fuentes de datos fue representada en un CDM distinto compuesto por dos bases de datos (una con los conceptos normalizados y otra sin normalizar) para una correcta simulación de los pilotos en un entorno multicéntrico. También hay que destacar que la información clínica era previamente anonimizada por los proveedores y que por tanto se trata de información privada, simulando un enfoque distribuido en cada uno de los pilotos clínicos, donde primaba la protección y trazabilidad de los datos por cada institución poseedora de éstos.

A continuación se detallan los distintos datos integrados haciendo hincapié en el origen de éstos y en los procesos de transformación necesarios para su incorporación en la CIS.

4.2.1. Datos del Institute Jules Bordet

Institute Jules Bordet (IJB)[20] es uno de los socios clínicos que formaron parte en los proyectos de investigación europeos. Este instituto trabaja en la realización de ensayos clínicos y tratamientos especialmente enfocados en cáncer de pecho.

Los datos facilitados por IJB provienen de su sistema de Historia Clínica Electrónica (Oribase) y de uno de sus ensayos clínicos, TOP TRIAL[21] Los datos procedentes de la HCE se encuentran almacenados en su sistema en distintas ramas-especialidades: registro de cáncer, datos multidisciplinares, laboratorio, datos anatomopatológicos e informes. Estos datos por tantos son facilitados electrónicamente por cada uno de los sistemas en un formato estructurado (XML, HL7…etc.) como indica se indica en la tabla 3. Además estos datos, dependiendo de la especialidad de donde provienen, se encuentran anotados con distintas terminologías.

Tabla 3: Datos procedentes de IJB

	Área	Formato	Terminología
HCE	Registro Cáncer	• Propio • HL7 v3 CDA	• Propio • ICD-O • SNOMED
	Multidisciplinar	• XML	• Propio

[20] https://www.bordet.be/en
[21] https://clinicaltrials.gov/ct2/show/NCT00162812

		• HL7 v3 CDA	• ICD-O • SNOMED
	Laboratorio	• HL7 v2	• LOINC • Propio
	Anatomopatología	• HL7 v3 CDA	• SNOMED
	Informes	• HL7 v3 CDA	• LOINC
Ensayo Clínico	TOP TRIAL	• CRD	• NCI CTACE • MedDRA

Además de los datos de las HCE de pacientes, IJB facilitó datos procedentes de un ensayo clínico realizado entre 2003 y 2009 [115]. Este ensayo, realizado en su institución, que contó con 338 participantes, tenía como fin la evaluación prospectiva de la amplificación del gen alfa de la topoisomerasa II y la sobreexpresión de proteínas como marcadores que predicen la eficacia de la epirubicin en el tratamiento primario de pacientes con cáncer de pecho. Por tanto, se incorporaron datos de los Cuadernos de Recogida de Datos (CRD) de pacientes que habían participado en este ensayo clínico y que habían dado su consentimiento expreso. Los datos de IJB pertenecían a datos anonimizado de 101 pacientes de estos sistemas.

4.2.2. Datos de la Universität des Saarlandes

La Universidad de Saarland (UdS)[22] ha participado en el proyecto de investigación EURECA como uno de los socios clínicos. UdS además de participar en la elaboración de los escenarios clínicos y requisitos, ha facilitado datos anonimizados provenientes de su sistema de información hospitalario (SIH) (Registro de cáncer de Saarland y HCE del Hospital universitario de Saarland) y datos del ensayo clínico de nefroblastoma de SIOP-GPOH [116][117].

[22] https://www.uni-saarland.de/nc/startseite.html

Tabla 4: Datos procedentes de UdS

	Área	Formato	Terminología
HIS	Registro Cáncer Saarland	• XML	• Propio • ICD
	HCE	• XML	• Propio • ICD
Ensayo Clínico	SIOP Nefroblastoma	• CRD • Base de datos propia	• Propia

Estos datos representan a un total de 3120 pacientes anotados en terminologías propias y que por tanto, han sido traducidos al vocabulario del *Core Dataset* con la supervisión y colaboración de los expertos clínicos de UdS.

4.2.3. Datos de Maastro Clinic

La Maastro Clinic (Maastro)[23] de Holanda participó también en el proyecto de investigación EURECA como uno de los socios clínicos. Maastro además de participar en la elaboración de los escenarios clínicos y requisitos, ha facilitado datos provenientes de sus HCE.

Tabla 5: Datos procedentes de Maastro

	Área	Formato	Terminología
HCE	Dossier médico	• MySQL	• ICD9 • NCI Thesaurus
	Informes clínico	• Texto • XML	
Ensayo Clínico	SUPREMO[24]	• CRD	• NCI Thesaurus

[23] https://www.maastro.nl/
[24] **http://clinicaltrials.gov/ct2/show/NCT00966888**

	Varios [25] [26]	• Base de datos propia	• Propia

Principalmente estos son datos anonimizados de pacientes relacionados con cáncer de pecho o de pulmón. Estos datos provienen del sistema de Maastro de dossier médico, que contiene los datos anotados con ICD 9 y NCIt y que contiene también datos procedentes de varios ensayos clínicos sobre cáncer de pecho. Además también facilita datos de informes clínicos en formato texto. Estos datos son tratados por otros socios del proyecto de investigación para leer mediante procesamiento de texto los datos relevantes de cada informe y proceder a su anonimización antes de ser trasladados a la Capa de Interoperabilidad Semántica. Maastro ha facilitado datos de un total de 8685 pacientes anonimizados para su procesamiento en la CIS.

4.2.4. Datos de la University of Oxford

La universidad de Oxford (UOxf)[27] formó parte también del proyecto EURECA como parte del hospital universitario de Oxford. Como el resto de socios clínicos participó en distintas tareas dentro del proyecto y facilitó además acceso a datos anonimizados. Estos datos procedían de registros de pacientes del propio hospital: base de datos de patologías, informes, notas clínicas, genómica…etc.

Tabla 6: Datos procedentes de UOxf

	Área	Formato	Terminología
HCE	Notas clínicas e informes	• HL7 v3 • XML	• NCI • ICD
	Genómica	• RAW	• Texto
	Patologías	• Base de datos propia	• Propia • ICD

[25] **http://clinicaltrials.gov/ct2/show/NCT00470236**
[26] **http://clinicaltrials.gov/ct2/show/NCT01803958**
[27] https://www.ouh.nhs.uk/

Principalmente estos datos procedían de la historia clínica del hospital, donde se encuentran los datos estructurados y desestructurados. Se consiguieron datos anonimizados de un total de 249 pacientes relacionados con distintos tipos de sarcoma. A estos datos se les agregaron datos de genómica en texto plano para almacenar todo esto en la CIS [44] y comprobar la validez de la integración de datos clínicos y genómicos en el mismo sistema.

4.2.5. Datos de German Breast Group (GBG)

El último de los socios clínicos de EURECA es el *German Breast Group* (GBG)[28]. GBG participó en las distintas tareas clínicas del proyecto de investigación y facilitó acceso a datos procedentes de distintos ensayos clínicos sobre cáncer de pecho realizados por su organización:

- TBP (Metastatic study)[29] : ensayo clínico sobre el uso de trastuzumab en pacientes con cáncer de pecho
- GAIN (Adjuvant study)[30]: ensayo clínico para la observación y seguimiento de pacientes con cáncer de pecho y su tratamiento con epirubicin.
- GeparQuattro (NeoAdjuvant study)[31]: ensayo clínico que compara el tratamiento con y sin trastuzumab y capacitabine en pacientes con cáncer de pecho y sin operaciones de cirugía previas.

Tabla 7: Datos procedentes de GBG

	Área	Formato	Terminología
Ensayos Clínicos	TBP (Metastatic study)	• CRD • CSV	• Texto • Propia
	GAIN (Adjuvant study)	• CRD • CSV	• Texto • Propia

[28] https://www.gbg.de/
[29] https://www.clinicaltrials.gov/ct2/show/NCT00148876
[30] https://www.clinicaltrials.gov/ct2/show/NCT00196872
[31] https://www.clinicaltrials.gov/ct2/show/NCT00288002

	GeparQuattro (NeoAdjuvant study)	• CRD • CSV	• Texto • Propia

Los datos de estos ensayos clínicos provienen en su mayoría de los CRD y de las bases de datos propias de GBG. Estos datos son anonimizados por herramientas facilitadas dentro del proyecto de investigación EURECA para su posterior incorporación en la CIS. Previamente hay que estructurar y anotar los conceptos mediante terminologías presentes en el *Core Dataset*. Este paso se realiza bajo la supervisión de los propios clínicos del GBG para su validación, obteniendo finalmente un conjunto de datos anonimizados con 4929 pacientes.

4.3. Codificación de las fuentes de datos

La codificación o representación de los conceptos procedente de cada uno de las fuentes de datos es un proceso que se realizó manualmente con la ayuda de herramientas y buscadores de terminologías [114][113]. Si bien este proceso de codificación de las fuentes de datos no es un proceso realizado expresamente para la presente tesis doctoral, se cree oportuno realizar una descripción de cómo han sido codificados las distintas fuentes de datos con términos pertenecientes al *Core Dataset* para su posterior evaluación con los métodos desarrollados. Este proceso debía ser validado siempre por los propios socios clínicos que cedían sus datos, para asegurar la coherencia en la anotación de estas ideas clínicas, así como asegurar que se mantenía el significado de estos.

Por tanto, partiendo de las guías facilitadas por cada uno de los proveedores de datos, se estudian los conceptos o ideas clínicas presentes. De ese estudio una primera representación de cada uno de los conceptos de las guías en términos presentes del *Core Dataset* (SNOMED, LOINC o HGNC). Esta representación ha de ser refrendada por los proveedores de los datos para asegurar el correcto seguimiento de su información y el mantenimiento de su representación.

En este paso se encontraron diversas diferencias dependiendo de cada uno de las fuentes de datos, ya que cada una de éstas tenían un ámbito y un formato distinto, así como unas terminologías distintas, llegando a estar representando en alguno casos con cadenas de texto o identificadores propios del país y del idioma. A continuación se detallan las diferencias básicas entre cada uno de los conjuntos de datos:

- **IJB:** Los datos de IJB procedían en su mayoría de CRD de varios de sus ensayos clínicos y de su propio sistema hospitalario. Por tanto, la mayor parte de sus datos se encontraban representados en una codificación propia del centro y en cadenas de texto en francés. Por tanto, se realizó una representación de estos conceptos haciendo uso de términos del *Core Dataset*, utilizando un total de 253 conceptos del *Core Dataset*, como puede comprobarse en el ANEXO C.
- **UdS:** En el caso de UdS, sus datos procedían mayormente del CRD del ensayo clínico SIOP, del que ya han realizado varios experimentos y aplicaciones [118] y que por tanto ya tienen representados. Por tanto, como puede comprobarse en el ANEXO C.2, sus términos surgen de preguntas y variables de su representación en el sistema de los CRD. De este conjunto de datos de más de 3.000 pacientes, se obtiene una representación de únicamente 23 términos del *Core Dataset*.
- **Maastro:** Los datos de Maastro proceden de su sistema electrónico y representan más de 8.000 pacientes. Estos conceptos se encontraban diferenciados de dos formas: representados mediante códigos de la terminología NCI y mediante cadenas de texto en holandés. Para el primero de los caso se hizo uso de la herramienta "Enlazado de terminologías" para su traducción a conceptos del *Core Dataset*. Para el segundo caso se necesitó de la ayudar de los proveedores de datos para asegurar el significado de los términos holandeses para su posterior anotación mediante terminologías del CD. Estos datos fueron representados por un total de 580 términos de del *Core Dataset*
- **UOXF:** En el caso de los datos del Hospital Universitario de Oxford, se trata de datos procedentes de su sistema sobre pacientes con distintos tipos de sarcoma. Estos ya datos ya se encontraban estructurados y anotados mediante terminologías extendidas en la práctica clínica. Por tanto, la búsqueda de términos sinónimos o similares en el *Core Dataset* fue bastante directa, siempre bajo la supervisión de

los proveedores de los datos. En estos datos se obtuvo una representación mediante el uso de 137 términos del *Core Dataset*.

- **GBG:** Los datos del GBG proceden de los CRD de 3 ensayos clínicos realizado por la institución. Estos datos se encontraban almacenados en su sistema en su propia estructura y anotados con sus propias terminologías, por tanto, como en los otros casos, se realizó una representación de estos términos mediante vocabularios presentes en el *Core Dataset* bajo la supervisión de los proveedores de los datos. Obteniendo que para los casi 5.000 pacientes presentes en estos datos, un total de 75 términos del *Core Dataset* de la CIS.

Una vez se conseguía una representación homogénea para todas las fuentes de datos mediante los vocabularios presentes en el *Core Dataset*, se pasaba a la generación de mensajes HL7 con esta información para almacenar los datos en el CDM. Esta generación se realizaba mediante el desarrollo de *scripts ad-hoc* que generaban mensajes genéricos en base al contexto para su posterior carga. Finalmente la carga de estos mensajes se hacía mediante le servicio *Data Push* que se encargaba de almacenar esta información en su correspondiente CDM y en hacer uso del método de normalización semántica para almacenar la representación normalizada de esta información en un CDM normalizado sobre el que posteriormente se harán las consultas.

4.4. Pruebas básicas de los métodos diseñados

4.4.1. Fuentes de datos normalizadas y no normalizadas

El método de normalización semántica implementado además de transformar los datos originales genera una réplica de cada CDM con la información sin normalizar, por tanto, se puede realizar un análisis sobre cómo ha afectado el proceso de normalización semántica en cada uno de los conjuntos de datos. Este análisis buscará comprobar la validez de la hipótesis sobre la posibilidad de enriquecer y corregir la representación clínica mediante el proceso de normalización planteado. Para esta comprobación se realizan experimentos con los distintos conjuntos de datos descritos, teniendo en cuenta la presencia de diferencias en la anotación de los conceptos dependiendo de la institución, principalmente estos casos se refieren a la especificidad de los términos usados para representar mismas situaciones clínicas o la representación de valores asignados a

observaciones (por ejemplo un tipo de tumor debido a su tamaño). Además también se considera la aparición de errores de codificación (etiquetas o textos que no coinciden con su término asociado) y representación (términos del *Core Dataset* representados originalmente en ubicaciones incorrectas del CDM).

Las diferentes formas de anotación de las fuentes de datos dan lugar a heterogeneidades semánticas que son abordadas por el método de normalización semántica. En general, estas heterogeneidades surgen de la amplitud de los vocabularios presentes en el *Core Dataset* y de su facilidad para representar conceptos post-coordinados. A pesar de que el proceso de anotación es realizado por parte del equipo de desarrollo de la CIS, estas anotaciones se basan en los datos facilitados y en la supervisión de los proveedores de los datos. Por tanto, es común que éstos decidan utilizar conceptos de distinta especificidad para representar las mismas situaciones clínicas y distintos sinónimos. El método de normalización semántica está diseñado para solucionar estas heterogeneidades representando de una manera común y generalista estos casos, y enriqueciéndolos con nuevas relaciones que completen el significado del concepto original.

También es posible encontrar errores de distinta tipología en la anotación de las fuentes de datos. Estos pueden varias desde simple errores humanos en la anotación textual o codificada de los conceptos o incluso en la representación de estas anotaciones en la generación de los mensajes. El proceso de normalización semántica debe encargarse también de corregir estos errores y de asegurar la trazabilidad de estos cambios como en el caso anterior.

Estas pruebas, por tanto, serán realizadas con cada conjunto de datos, donde se procederá a comparar los datos normalizados y sin normalizar de cada fuente, analizando los cambios ocurridos en un CDM normalizado y como han mejorado su representación. Ya que el punto de partida son conjuntos de datos reales, no se valora la práctica de “ensuciar” o introducir errores en los CDM para añadir ruido, ya que se comprende que los errores ya presentes son los suficientemente representativos en la práctica real Partiendo de estos resultados, se puede obtener una lista de casos comunes en las fuentes de datos que faciliten la recuperación de información conjunta en todos los CDM.

4.4.2. Implementación y comparación con otros sistemas relacionados

Finalmente, se decide realizar una implementación de otros sistemas similares para realizar una comparación cuantitativa y cualitativa sobre los mismos conjuntos de datos y las consultas realizadas dentro del mismo entorno. Estas implementaciones cubrirán los distintos estándares de modelos de datos más utilizados; OMOP, i2b2 y HL7, como parte del CDM. Por tanto, se realizará una comparativa en la que se analizará las diferencias en la CIS desarrollada al comparar estos 3 modelos de datos con la CIS implementada.

Para probar su validez, se selecciona un subconjunto de los datos del IJB (procedentes de TOP TRIAL) que contienen 80 pacientes con más de 500 observaciones, incluyendo expresiones genéticas siguiendo las recomendaciones sobre cáncer de pecho de St. Gallen [119]. Estos datos son, por tanto, almacenados para su posterior recuperación de la información en cuatro sistemas distintos: (i) CIS basada en HL7 RIM sin normalización semántica, (ii) OMOP, (iii) i2b2 y (iv) CIS basada en HL7 RIM con normalización semántica. Donde el primero de éstos se trata de una implementación de la CIS actual que no contaría con la integración del método de normalización semántica, mientras que los sistemas basados en i2b2 y OMOP han sido implementados siguiendo las recomendaciones oficiales de cada proyecto.

Las consultas a realizar para recuperar la información de estos sistemas están basadas en criterios de elegibilidad del ensayo TOP trial simulando un escenario de reclutamiento de pacientes para este ensayo. Estos criterios varían desde la búsqueda de información general de un paciente sobre diagnósticos previos, tratamientos con medicamentos y expresiones genéticas[32].

4.5. Escenarios para la validación

La forma de comprobar el correcto seguimiento y consecución de un proyecto es mediante la validación de las soluciones realizadas durante el proyecto en escenarios que cubran las necesidades de las partes interesadas en el proyecto. Dentro de INTEGRATE y EURECA se definieron hasta un total de 15 escenarios clínicos, que fueron definidos durante el primer año de ambos proyectos tras realizar cuestionarios personales entre los

[32] https://bitbucket.org/sparaiso/semantic-normalization-and-query-abstraction-based-on-snomed

distintos socios (clínicos, técnicos…etc.) y potenciales usuarios de las distintas herramientas o soluciones realizadas, validando de esta forma las soluciones desde una perspectiva de salud básica orientada al paciente hasta una perspectiva más orienta a la investigación específica en la realización de ensayos clínicos.

Si bien la CIS fue utilizada hasta en un total de 11 de los 15 escenarios definidos en estos proyectos, los experimentos realizados para comprobar los métodos realizados en la presente tesis doctoral se centraron en los escenarios que guardaban una relación directa con la realización de ensayos clínicos: i) reclutamiento y ii) factibilidad de ensayos clínicos. En las siguientes subsecciones se describirá en qué consistía cada uno de estos escenarios, su relación con los conjuntos de datos y como se utilizan los métodos diseñados.

4.5.1. Reclutamiento para ensayos clínicos

El escenario de reclutamiento de pacientes busca mejorar el tiempo y la eficacia en la realización de los ensayos clínicos. El proceso de reclutamiento de los ensayos clínicos es una de las labores más costosas en cuanto a tiempo debido a la, cada vez más, alta especificidad de los ensayos. El principal objetivo de este escenario es reducir ese tiempo mediante una aplicación que procese los criterios de elegibilidad de los ensayos clínicos y realice la búsqueda en las HCE de posibles pacientes.

Este escenario involucra, por tanto, distintos componentes realizados por varios grupos participantes de los proyectos para dotar a la solución de una alta fiabilidad, seguridad y facilidad de uso para los investigadores o reclutadores que tengan que utilizar la solución. Para validar correctamente este escenario y asegurarse la existencia se pacientes que cumplan los criterios de elegibilidad se utilizan datos anonimizados de pacientes procedentes de los ensayos clínicos de IJB, GBG y Maastro. Además se modelarán los criterios de elegibilidad de estos ensayos clínicos (y otros) para demostrar su correcto funcionamiento.

Este escenario hace uso de datos de pacientes anonimizados que son normalizados por el método de normalización semántica del presente trabajo. Finalmente se recuperan los datos almacenados a través del método de búsqueda información des-contextualizada

contenido en el método de abstracción de consultas, simulando mediante este método los distintos criterios de elegibilidad de cada uno de los ensayos clínicos definidos.

4.5.2. Factibilidad de ensayos clínicos

Este escenario describe si un nuevo ensayo clínico es factible que para ser ejecutado en un conjunto de pacientes. Por tanto, como en el anterior escenario, este piloto busca reducir el tiempo de ejecución de un ensayo clínico mediante el diseño de una solución que simplifique la viabilidad de éstos.

La viabilidad de esta solución será realizada por los investigadores y gestores de ensayos clínicos y por tanto se realizará en un entorno muy similar al reclutamiento de pacientes, donde se integran todos los componentes presentes en la CIS junto con componentes de seguridad, autenticación y gestores de ensayos clínicos que procesen los estudios.

Para validar este escenario se hará uso también de datos de pacientes anonimizados procedentes de ensayos clínicos de GBG, UOxF, UdS e IJB que se cruzarán con distintos ensayos clínicos obtenidos de bases de datos públicas como ClinicalTrials.gov. Por tanto, se hace un uso similar de los métodos investigados en la presente tesis, donde los datos ya se encuentran normalizados semánticamente en la CIS y se recupera la información mediante la construcción de consultas haciendo uso de la búsqueda de información contextualizada del método de abstracción de consultas.

4.6. Consultas a generar mediante la abstracción

Estos escenarios recuperan la información presente en el CDM para su posterior análisis en las distintas aplicaciones realizadas. Para ello obtienen las consultas a realizar gracias al método no contextualizado (o no guiado) de la abstracción de consultas sobre los conceptos presentes en cada CDM y otros conceptos sobre les que se desee buscar (según el escenario). En estos casos, como ya se ha indicado, se hace un uso prácticamente general de este método en la totalidad de estos escenarios debido a su facilidad de uso e integración en los distintos sistemas.

Por otro lado, los escenarios que tienen un objetivo con un mayor componente de investigación están relacionados con la mejora en la práctica en las distintas fases de los

ensayos clínicos. Por tanto, en estos escenarios las pruebas se realizan simulando ensayos clínicos existentes relacionados con la procedencia de los datos:

4.6.1. Ensayo 1 – TOPTRIAL

Este ensayo[33] fue realizado entre 2003 y 2008 por el Instituto Jules Bordet para realizar una evaluación prospectiva de la amplificación de la topoisomerasa II y otros biomarcadores para predecir y evaluar la eficacia de tratamientos con epirubicin en pacientes con cáncer de pecho. TOP Trial se modela en el escenario de factibilidad de ensayos clínicos, para demostrar si sería posible efectuar este ensayo clínico en el conjunto de datos existentes en el CDM.

Para que sea factible la realización de un ensayo clínico se deben tener un número necesario de personas que cumplan los criterios de elegibilidad (CE) del ensayo clínico. Esto significa que deben cumplir los criterios de inclusión y no cumplir los criterios de exclusión. Por tanto, lo que buscan las aplicaciones es obtener respuesta para cada uno de los criterios de elegibilidad modelados.

Tabla 8: Ejemplo de formalización de CE de TOPTRIAL

	#CE	Texto	Anotación de CD	Abstracción de consultas
Inclusión	1	Diagnóstico de cáncer de pecho confirmado	254837009 | Malignant neoplasm of breast (disorder)	Contextualized(254837009) YES/NO
	4	Tamaño de tumor de al menos 2 cm.	263605001 | Tumor size (observable entity) |	Contextualized(263605001) Values>2
	5	Negativo en tumores ESR	106221001 | Genetic finding (finding) ESR | HGNC	Contextualized (106221001, entity(ESR)) YES/NO
	9	Escala 0 o 1 en estado ECOG	423740007 | ECOG performance status	Contextualized (423740007)

[33] https://clinicaltrials.gov/ct2/show/NCT00162812

				Value = 1 or 2
Exclusión	1	Metástasis en cáncer de pecho	94649002 | Secondary malignant neoplasm of trunk (disorder) |	Contextualized(94649002) YES/NO
	2.3	Infección activa	40733004 | Infectious disease (disorder)	Contextualized(40733004, effectiveTime(NOW)) YES/NO
	6	Tratamiento previo de antracyclines para cáncer de pecho	432102000 | Administration of substance (procedure) | 372540003 | Anthracycline (substance) |	Contextualized(432102000, entity(372540003)) YES/NO

Donde este modelado de los criterios de elegibilidad, se basa en codificar estos criterios con términos del *Core Dataset* para obtener consultas que nos permitan recuperar esta información del CDM. Como el modelado de los CE se trata de un proceso manual, ya que requiere de su codificación, se decide utilizar en estos casos el método contextualizado de la abstracción de consultas y añadir los filtros necesarios en cada caso. En la anterior tabla se muestra un ejemplo de la anotación y generación de consultas de algunos criterios de elegibilidad de TOP Trial.

4.6.2. Ensayo 2 – TBP

Ensayo[34] prospectivo y controlado realizado por el grupo GBG entre 2003 y 2011 en distintos centros para comparar la eficacia del tratamiento de pacientes con positivo en HER2 con trastuzumab. TBP se modela en el escenario de factibilidad de ensayos clínicos y en el escenario de reclutamiento de pacientes para ensayos. Este escenario busca responder si un paciente seleccionado es elegible para su participación en un ensayo.

[34] https://www.clinicaltrials.gov/ct2/show/NCT00148876

De una manera similar al anterior, para que un paciente puede ser reclutado para su participación en un ensayo clínico debe cumplir con los CE del mismo. Por tanto, en este caso, las aplicaciones buscan responder si un paciente seleccionado cumple todos los criterios de inclusión de un ensayo y no cumple los criterios de exclusión.

Tabla 9: Ejemplo de formalización de CE de TBP

	#CE	Texto	Anotación de CD	Abstracción de consultas
Inclusión	2	Carcinoma de pecho patológicamente confirmado	254838004 | Carcinoma of breast (disorder) |	Contextualized(254838004) YES/NO
	4	Expresión de tumor con metástasis via gen HER2-neu	106221001 | Genetic finding (finding) ERBB2 | HGNC	Contextualized (106221001, entity(ERBB2)) YES/NO
	11	Prueba de Karnofsky con resultado mayor de 60	423740007 | ECOG performance status	Contextualized (423740007) Value = 0 or 1
	13	Número de neutrofitos mayor de 1500 micro por células	30630007 | Neutrophil count (procedure) |	Contextualized (30630007) Value > 1500 Units: m/cells
	16	LVEF mayor de 50%	250908004 | Left ventricular ejection fraction (observable entity) |	Contextualized (250908004) Value > 50
Exclusión	2	Inmunoterapia concurrente o terapia hormonal	309542002 | Endocrine therapy (procedure) | 277132007 | Therapeutic procedure (procedure) |	Contextualized(309542002) or Contextualized(277132007) YES/NO

En la tabla anterior se muestra una selección del modelado de varios CE del ensayo clínico TBP. Donde se puede apreciar la descripción del CE, su codificación realizada manualmente con términos del *Core Dataset* y finalmente la forma en que se obtiene la consulta para recuperar la información del CDM. En estos casos también se decide utilizar el método contextualizado de la abstracción de consultas e ir añadiendo los filtros necesarios en cada caso, dependiendo del resultado a buscar en el CE.

4.6.3. Ensayo 3 – GAIN

GAIN[35] fue un ensayo clínico intervencional aleatorio realizado por el grupo GBG entre los años 2004 y 2014 con más de 3.000 pacientes enrolados. El objetivo de este ensayo era analizar la evolución del tratamiento de pacientes de cáncer de pecho con la administración de epirubicin. Como el anterior ensayo, GAIN participa en los escenarios de reclutamiento y factibilidad de ensayos clínicos.

Tabla 10: Ejemplo de formalización de CE de GAIN

	#CE	Texto	Anotación de CD	Abstracción de consultas
Inclusión	5	Al menos un ganglio linfático mamario o axilar involucrado	385381005 | pN category finding (finding) |	Contextualized (250908004) Value: YES/NO
	6	No hay evidencia de metástasis después del diagnóstico	385380006 | Metastasis category finding (finding) |	Contextualized (385380006) Value: YES/NO
Exclusión	2	Disfunciones orgánicas: - ANC < 1.5 G/l - Plaquetas <100 G/l - Billirubina >1.25UNL	30630007 | Neutrophil count (procedure) | 61928009 | Platelet count (procedure) | 302787001 | Bilirubin measurement (procedure) |	Contextualized (30630007) Value > 1500 m/cells Contextualized (61928009) Value > 100 g/l Contextualized (302787001)

[35] https://www.clinicaltrials.gov/ct2/show/NCT00196872

				Value > 2 UNL

En la tabla anterior se muestran una selección de la creación de CE distintos a los presentes en TBP, desde la descripción del ensayo clínico hasta la obtención de consultas para recuperar la información del CDM anotada con los conceptos del CD y los valores deseados. En estos casos también se decide utilizar el método contextualizado de la abstracción de consultas e ir añadiendo los filtros necesarios en cada caso, dependiendo del resultado a buscar en el CE.

4.6.4. Ensayo 4 – Geppar Quattro

Geppar Quattro[36] fue un ensayo clínico intervencional y aleatorio realizado por el grupo GBG entre los años 2005 y 2013. Este ensayo analiza y compara la evolución en pacientes de cáncer de pecho primario con el tratamiento con y sin de sustancias como trastuzumbab y capacitabine. Como los dos anteriores ensayos, Geppar Quattro participa en los escenarios de reclutamiento y factibilidad de ensayos clínicos.

Tabla 11: Ejemplo de formalización de CE de Geppar Quattro

	#CE	Texto	Anotación de CD	Abstracción de consultas
Inclusión	5	Tumores avanzados cT4 o ct3 ER o PgR tumores negativos	14410001 | T3 category (finding) | 65565005 | T4 category (finding) | 06221001 | Genetic finding (finding) ER | HGNC PGR | HGNC	Contextualized (14410001) Contextualized (65565005) Contextualized (106221001, entity(ER)) Contextualized (106221001, entity(PGR)) YES/NO

[36] https://www.clinicaltrials.gov/ct2/show/NCT00288002

	6	HER2/neu localizado en prueba	106221001 | Genetic finding (finding) ERBB2 | HGNC	Contextualized (106221001, entity(ERBB2)) YES/NO
	11	Requisitos de Laboratorio: ANC >= 2.0 Hemoglobina >= 10 ASAT (SGOT) & ALAT (SGPT) <= 2.5 x UNL	30630007 | Neutrophil count (procedure) | 441689006 | Measurement of total hemoglobin concentration (procedure) | 45896001 | Aspartate aminotransferase measurement (procedure) | 34608000 | Alanine aminotransferase measurement (procedure) |	Contextualized (30630007) Value > 2000 m/cells Contextualized (441689006) Value > 10 g/l Contextualized (45896001) Value > 2,5 UNL Contextualized (34608000) Value > 2,5 UNL
Exclusión	3	Cualquier quimioterapia	367336001 | Chemotherapy (procedure) |	Contextualized (367336001) Value: YES/NO
	4	Cualquier radioterapia	108290001 | Radiation oncology AND/OR radiotherapy (procedure) |	Contextualized (108290001) Value: YES/NO
	10	Tratamiento concurrente con hormonas sexuales	432102000 | Administration of substance (procedure) | 312263009 | Sex hormone (substance) |	Contextualized(432102000, entity(312263009)) Value: YES/NO

En la tabla anterior se muestran una selección de la generación de CE distintos a los presentes en los ensayos anteriores, desde la descripción del ensayo clínico hasta la

obtención de consultas para recuperar la información del CDM anotada con los conceptos del CD y los valores

4.7. Evaluación final

La ejecución de los escenarios de validación descritos en la sección 4.3 demostró la viabilidad de los proyectos INTEGRATE y EURECA, y por tanto de los componentes que formaron parte de sus evaluaciones. Los escenarios sirvieron para comprobar el correcto funcionamiento de las distintas soluciones y servicios en condiciones reales. En todos los casos se obtuvieron resultados positivos por expertos y revisores de la Comisión Europea.

Los métodos presentados en la presente tesis fueron integrados en de la Capa de Interoperabilidad Semántica (CIS) para comprobar la validez y viabilidad de la hipótesis inicial. Por tanto, éstos fueron evaluados en los escenarios clínicos definidos, demostrando el potencial de los métodos para mejorar la homogeneización de los distintos conjuntos de datos y su facilidad de integración en las aplicaciones utilizadas. Todos los escenarios fueron demostrados en vivo en las distintas revisiones anuales de los proyectos a expertos de la Comisión Europea.

Los evaluadores destacaron la capacidad técnica de las aplicaciones que formaron parte de las demostraciones, si bien destacan sobre todo, la capa de seguridad y la CIS como los servicios que aconsejan ser explotados y tener en cuenta para futuros proyectos.

Cabe destacar finalmente, que ambos proyectos concluyeron obteniendo una evaluación positiva por la Comisión Europea, animando a futuras sinergias de estos proyectos con otras agrupaciones para su posible continuación. Estas evaluaciones pueden consultarse en el ANEXO C.

Capítulo 5

5. RESULTADOS Y DISCUSIÓN

El trabajo realizado en la presente tesis doctoral ha consistido en el diseño de un método de normalización semántica y un método de abstracción de consultas basadas en estándares biomédicos para mejorar el acceso a fuentes de datos heterogéneas. Estos métodos surgen de la colaboración y el trabajo del doctorando en los proyectos de investigación europea INTEGRATE y EURECA, donde el investigador ha implementado e integrado su solución en la Capa de Interoperabilidad Semántica realizada en estos proyectos, para verificar las características y posibilidades de ambos métodos y probar las posibles generalizaciones y adaptabilidad de los métodos a otros sistemas y soluciones.

Dentro de este marco de trabajo definido las secciones 3 y 4 se describían a fondo los métodos desarrollados así como los experimentos diseñados para poner a prueba el sistema realizado. En este capítulo se ofrece una argumentación sobre la relevancia de los métodos propuestos y la utilidad de la CIS en su conjunto como un sistema de integración de datos que facilita un acceso a homogéneo a éstos. Para ello, se analizan las transformaciones de conceptos en bases de datos normalizadas y no normalizadas que demuestran cómo han evolucionado en su representación estos métodos en cada uno de

los conjuntos de datos y de manera global. Asimismo, se realiza una comparativa analizando en su conjunto los métodos y la CIS con otras soluciones existentes y basadas en otros modelos de datos clínicos basados en los estándares más relevantes (HL7 RIM, i2b2 y OMOP).

5.1. Importancia y marco de los modelos propuestos

Los grandes avances en la investigación clínica han venido provocado un aumento en la cantidad de información y en la aparición de nuevas variables clínicas. Estos avances han provocado un gran cambio en los ensayos clínicos, ya que surgen nuevos estudios y pruebas moleculares que antes no se contemplaban. Pruebas que a su vez añaden una especificidad que imposibilita en muchas ocasiones la realización de ensayos clínicos en poblaciones o conjuntos localizados, pasando a tener que ser realizados en distintas ubicaciones mediante la colaboración de diversos centros. Ante esta diversidad en la representación de los datos se investigan distintos métodos de integración que faciliten el acceso a los datos, tomando mayor relevancia el uso de estándares clínicos para mejorar la interoperabilidad entre los distintos sistemas y aplicaciones.

Siguiendo la línea de investigación sobre interoperabilidad clínica destacan distintas iniciativas que tratan de integrar esta información dependiendo de los distintos enfoques de integración (debidos a la ubicación de los datos) y los tipos de heterogeneidades que afecten al sistema. Es en este contexto donde se inician proyectos de investigación como INTEGRATE y EURECA, donde la integración de los datos y la interoperabilidad entre sus soluciones es un punto crucial para la validación de sus escenarios clínicos enfocados en el estudio de tratamientos sobre cáncer de pecho.

El trabajo realizado en la presente tesis trata de enriquecer y corregir la representación clínica de datos mediante el uso de un método de normalización semántica y un método de abstracción que combinados explotan las ventajas y el conocimiento de estándares terminológicos para asegurar la homogeneización. Para comprobar la validez y la viabilidad de estos métodos, se integra una implementación de estas soluciones en la Capa de Interoperabilidad Semántica (CIS) desarrollada como parte de los proyectos europeos

INTEGRATE y EURECA para dotar a sus aplicaciones de un punto de acceso a la información clínica.

El método de normalización semántica de datos diseñado se presenta como una solución para la práctica de la integración de datos utilizando estándares terminológicos y su conocimiento implícito. El método de abstracción de consultas mejora la interoperabilidad de los sistemas para acceder a esta información para su posterior análisis. Estos dos métodos, combinados e integrados en la CIS, dotan a los sistemas de una solución capaz de homogeneizar la información clínica procedente de distintas fuentes de datos y facilitar la recuperación de los datos.

Los métodos presentados en esta tesis doctoral reúnen las características comunes a los sistemas utilizados en la práctica real, de hecho han sido verificados con datos de pacientes en unos escenarios clínicos definidos por expertos clínicos. Se ha realizado una simulación de un entorno federado donde la CIS representa un ente que contiene un esquema global de los datos almacenados basado en HL7 y un servidor de terminologías (con las terminologías SNOMED, HGNC y LOINC), además de los métodos diseñados, proporcionado una solución a problemas de interoperabilidad reales de la práctica clínica. De esta forma, se proporcionan unos métodos genéricos y fácilmente adaptables a otras soluciones y componentes, que permiten mejorar el acceso y la integración de datos mediante el uso de estándares clínicos.

5.2. Análisis de los resultados

El desarrollo e integración de los métodos diseñados en la capa de interoperabilidad semántica utilizada como sistema de integración de datos clínicos, es un proceso complejo que requiere la aplicación de técnicas de ingenierías del software, pero que dado su carácter mediador entre aplicaciones y fuentes de datos, depende en gran medida de éstas. El diseño, implementación e integración de todos los componentes requiere un mayor esfuerzo debido a la multitud de actores involucrados.

En el capítulo anterior se detallaba y explicaba el trabajo realizado para la implementación e integración de los métodos diseñados para formar parte de la CIS. Además se detallaban las distintas fuentes de datos de pacientes que iban a servir como

conjunto de pruebas y validación de las aplicaciones que se demostrarían en los escenarios clínicos. La mayor parte del trabajó se realizó en la integración del componente de normalización semántica, debido a su especial sensibilidad a los datos de origen y a la constante comunicación con los proveedores de datos para asegurar el mantenimiento del significado en sus datos.

El trabajo expuesto evidencia la necesidad de utilización de estándares de representación comunes entre los sistemas para facilitar el intercambio de información entre distintos centros y de esta forma facilitar la realización de estudios clínicos y acortar los tiempos de éstos. Estos estándares facilitan herramientas para la mejora de la representación de los datos y para su posterior utilización, lo que conlleva una mejora en el acceso a la información clínica que se traduciría en mejores prestaciones para todos los sistemas que accedan a estos datos y por tanto para los investigadores y pacientes.

Los métodos propuestos mejoran la representación de los datos mediante un método de normalización semántica basada en estándares terminológicos que transforma los datos en una representación homogénea y sin ambigüedades permitiendo el enriquecimiento de los datos y la corrección de posibles errores. Además, el trabajo realizado permite un punto de acceso común que abstrae de la representación de los datos y o tecnologías utilizadas que permitiría la replicación y adaptación de estos métodos en otros sistemas.

En las siguientes subsecciones se presentarán los análisis obtenidos al representar los datos utilizando el método de normalización semántica en cada uno de las fuentes de datos, donde se analizarán los conceptos obtenidos y su aplicación en el CDM que enlaza con el punto de acceso de esta representación. Asimismo, se presenta una comparativa cuantitativa y cualitativa de la CIS y los métodos diseñados con otros modelos e iniciativas relevantes.

5.2.1. Análisis del impacto de los métodos en las fuentes de datos

Esta sección presenta un análisis de cómo han sido representadas las distintas fuentes de datos descritas en la sección 4.2. Para ello se han realizado experimentos utilizando el método de normalización semántica con cada fuente de datos para así poder obtener una representación de los datos en un CDM normalizado y en otro CDM sin normalizar (original).

5.2.1.1. Datos normalizados de IJB

Los datos del *Institute Jules Border* (IJB) procedían de pacientes anonimizados que se involucraron en varios ensayos clínicos y en su sistema, donde la mayor parte de sus datos se encontraban representado en una anotación propia e incluso en otro idioma.

Como se puede comprobar en la siguiente tabla, en este caso, los datos originales fueron anotados con un total de 253 términos, siendo en su mayoría estos representados en el CDM con la clase *Observation* de HL7, algo bastante frecuente en esta clase de datos ya que la mayoría de los datos son diagnósticos y observaciones. Con la columna de instancias nos referimos al número de ocurrencias relacionadas con esta observación que tienen lugar en el CDM (estos datos representan a un total de 101 pacientes).

Tabla 12: Conceptos originales vs normalizados por atributo en el CDM de IJB

Clases CDM	**CDM Original**	**Instancias**	**CDM Normalizado**	**Instancias**
Observation	166	2127	157	2118
Procedure	30	469	31	483
Entity	20	32	73	215
SBADM	6	319	4	305
Target site	13	342	27	632
Interpretation	2	119	2	119
Method	2	206	12	625
Values	14	78	15	89
TOTAL	**253**	**3692**	**321**	**4586**

El método de normalización semántica efectuado a través del servicio *Data Push* para almacenar los datos al CDM, ha pasado a homogeneizar los términos obteniendo un total de 321 términos en total, lo que representa un aumento de un 26% en el total de conceptos. La mayor parte de este aumento de términos se produce sobre todo en las clases de HL7 correspondientes a *TargetSite, Method* y a *Entity*, mientras que el caso de la clase *Observation* se produce un descenso importante en el número de términos, pero no tan relevante en el número de instancias. Esto se explica debido al método de normalización basado en la forma normal de SNOMED (sección 4.1.1.2), ya que al normalizar conceptos menos generalistas como pueden ser conceptos referentes a neoplasmas o tumores como pueden ser: "*Malignant tumor of colon", "Malignant tumor of pancreas", "Malignant*

tumor of ovary" y "*Malignant tumor of thyroid gland"*, se obtiene una expresión normalizada formada por el concepto "*367651003| Malignant neoplasm of primary, secondary, or uncertain origin (morphologic abnormality)*" y la relación *target site* con el término que hace referencia a su ubicación. Esta transformación mantiene por tanto el número de instancias para la misma clase, ya que se cambia el término original por uno más generalista, pero añade especificidad en otra clase del CDM (en este caso target site), lo que explica la diferencia entre el mayor descenso en el número de instancias al número de términos para la clase *Observation*. Algo parecido pasa con el aumento de términos e instancias para términos de las clases de *Entity y Method*, donde además ocurre le caso de la corrección de errores al representar términos de tratamientos de sustancias como si fueran observaciones.

5.2.1.2. Datos normalizados de UdS

Los datos de la *Universitat de Saarland* (UdS) procedían de pacientes anonimizados que se involucraron en el ensayo clínico SIOP. Como se puede comprobar en la siguiente tabla, en este caso, los datos originales fueron anotados con un total de 23 términos, siendo en su mayoría estos representados en el CDM con la clase *Observation* y *TargetSite* de HL7, algo bastante frecuente en esta clase de datos ya que la mayoría de los datos son diagnósticos y observaciones (estos datos representan a un total de 3120 pacientes).

Tabla 13: Conceptos originales vs normalizados por atributo en el CDM de UdS

Clases CDM	**CDM Original**	**Instancias**	**CDM Normalizado**	**Instancias**
Observation	8	11204	8	11204
Procedure	2	3571	2	3571
Entity	1	1	1	1
SBADM	2	27748	2	27748
Target site	7	8386	9	11400
Interpretation	0	0	0	0
Method	0	0	1	2947
Values	3	39	3	39
TOTAL	**23**	**50949**	**26**	**56910**

Al contrario que en el anterior conjunto de datos, en este las modificaciones son mínimas, debido a la escasa diversidad en el número de términos utilizados para su representación. En este conjunto no hubo ninguna corrección de error, y el único caso representable es la generación de nuevos términos e instancias para las clases de *Method y TargetSite* debido a la normalización de SNOMED como en el caso anterior.

5.2.1.3. Datos normalizados de Maastro

Los datos de la clínica de Maastro procedían de pacientes anonimizados que se involucraron en varios ensayos clínicos y en su sistema electrónico. Donde la mayor parte de sus datos se encontraban representados en otras terminologías no presentes en el *Core Dataset*, por lo que hubo que hacer un primer proceso de enlazado de terminologías.

Como se puede comprobar en la siguiente tabla, en este caso, los datos originales fueron anotados con un total de 580 términos, siendo en su mayoría estos representados en el CDM con la clase *Observation y Entity* de HL7, algo frecuente en esta clase de datos ya que la mayoría de los datos son diagnósticos y observaciones, donde además al utilizar el enlazado de terminologías previo a la normalización consigues conceptos más cercanos a su forma normal (estos datos representan a un total de 8685 pacientes).

Tabla 14: Conceptos originales vs normalizados por atributo en el CDM de Maastro

Clases CDM	**CDM Original**	**Instancias**	**CDM Normalizado**	**Instancias**
Observation	251	69558	198	86545
Procedure	30	15615	29	7244
Entity	271	284	279	297
SBADM	2	5864	2	5864
Target site	14	346	66	5819
Interpretation	0	0	0	0
Method	1	2	11	6814
Values	13	40348	7	23762
TOTAL	**582**	**132017**	**592**	**136345**

En este caso, se obtiene un descenso de 53 (casi un 20%) en el número de términos pertenecientes a la clase Observation, dando lugar a prácticamente el mismo aumento en el número de términos clasificados como *TargetSite*. Este es un caso muy similar al de los datos de IJB, ya que estos datos provienen de pacientes clasificados con distintos tipos

de cáncer, con lo cual se normalizan a un mismo concepto y se añade la especificidad debida a su ubicación en la clase *targetSite*.

También se corrigen errores debidos a heterogeneidades a nivel de instancia en el caso de *Values*. Es un error muy común debido a las diferentes formas de anotar en este caso los valores positivos y negativos (*Pos, Positive, Negative, Neg*), donde el método de normalización los representa de una única forma utilizando el concepto de SNOMED correspondiente. Además, nos encontramos con un error manual debido a la incorrecta ubicación de conceptos relativos al *ECOG Performance Status*, que deberían ser conceptos de la clase *Observation*, lo que explica a su vez el aumento en el número de instancias en la clase *Observation* entre el CDM original y el normalizado.

5.2.1.4. Datos normalizados de UOXF

Los datos de la Universidad de Oxford procedían de pacientes anonimizados del sistema electrónico hospitalario. Donde la mayor parte de sus datos se encontraban representados en modelos estructurados en terminologías presentes en el *Core Dataset* (estos datos representan a un total de 249 pacientes).

Como se puede comprobar en la siguiente tabla, en este caso, los datos originales fueron anotados con un total de 137 términos, siendo en su mayoría estos representados en el CDM con la clase *Observation* de HL7, algo frecuente en esta clase de datos ya que la mayoría de los datos son diagnósticos y observaciones.

Tabla 15: Conceptos originales vs normalizados por atributo en el CDM de UOXF

Clases CDM	**CDM Original**	**Instancias**	**CDM Normalizado**	**Instancias**
Observation	80	3492	78	3762
Procedure	11	763	10	706
Entity	6	7	7	8
SBADM	2	84	2	84
Target site	25	483	39	1290
Interpretation	2	508	2	508
Method	4	452	9	878
Values	8	167	6	167
TOTAL	**138**	**5956**	**153**	**7403**

En el caso de esta fuente de datos, se reduce mínimamente los términos anotados con la clase *Observation*, si bien la normalización de términos de esta clase es la que da lugar al aumento en la clase *targetSite*. Algo similar ocurre con el aumento de términos de la clase *Method*. También se corrigen errores debidos a heterogeneidades a nivel de instancia en el caso de *Values*, las diferentes formas de anotar en este caso los valores positivos y negativos (*Pos, Positive, Negative, Neg*), donde el método de normalización los representa de una única forma utilizando el concepto de SNOMED correspondiente.

5.2.1.5. Datos normalizados de GBG

Los datos de la clínica de GBG procedían de pacientes anonimizados que se involucraron en varios ensayos clínicos de cáncer de pecho. La mayor parte de sus datos se encontraban representado en una anotación propia e incluso en otro idioma.

Como se puede comprobar en la siguiente tabla, en este caso, los datos originales fueron anotados con un total de solamente 75 términos para un total de 4.929 pacientes, siendo en su mayoría estos representados en el CDM con la clase *Observation* de HL7, debido a que la mayoría de los datos proceden de observaciones y diagnósticos relacionados con ensayos de cáncer de pecho.

Tabla 16: Conceptos originales vs normalizados por atributo en el CDM de GBG

Clases CDM	**CDM Original**	**Instancias**	**CDM Normalizado**	**Instancias**
Observation	53	64162	49	64162
Procedure	3	188	3	188
Entity	7	7	15	16
SBADM	2	582	2	582
Target site	1	14094	10	32373
Interpretation	2	13695	2	13695
Method	0	0	5	14681
Values	7	22403	7	22403
TOTAL	**75**	**115131**	**93**	**148100**

Este es un ejemplo donde se puede apreciar perfectamente cómo se mantienen el número de instancias para los términos etiquetados con la clase *Observation* mientras disminuye el número de términos de esta clase debido a su normalización. Como resultado de la normalización y generalización de estos términos se produce un aumento relacionado en

los términos e instancias etiquetados como *TargetSite y Method*. El aumento en los términos e instancias de la clase *Entity* están relacionados con la anotación inicial de los datos de administración de sustancia como un único concepto, y que, al normalizar, se transforma en un término de la clase SBADM y la entidad o sustancia relacionada.

5.2.1.6. Comparativa de datos normalizados y no normalizados

Analizados todos los conjuntos de datos en sus versiones normalizadas y sin normalizar se procede a realizar una comparativa que nos permita sacar conclusiones más amplias sobre los métodos diseñados y su aplicación en los distintos conjuntos de datos así como su uso en las diferentes aplicaciones durante la validación en los escenarios clínicos de EURECA e INTEGRATE.

De esta comparativa se desprende en términos generales como el proceso de normalización semántica basado en SNOMED afecta sobre todo a la heterogeneidad presente en las anotaciones de la clase *Observation* en primer lugar. Fruto de la normalización de estos términos, se consigue representar de una forma más homogénea estos términos y se aumenta su especificidad con términos de otra clase del CDM, como es el caso de *Entity, TargetSite y Method*. Además de errores que pueden ocurrir debido a malas ubicaciones de términos en los orígenes, se puede apreciar como la normalización semántica ha abordado errores semánticos a nivel de estancia en la clase de *Values*.

Se produce finalmente un aumento de casi un 10% sobre el total de los términos originales, lo que permite que el método de abstracción de consultas pueda recuperar información sobre conceptos que no estaban presentes en su origen y que además, puede integrar datos de otros términos normalizados en una expresión más homogénea.

Tabla 17: Comparativa de conceptos normalizados vs originales en todas las fuentes de datos clasificados por clases del CDM

	CDM	***Observation***	***Procedure***	***Entity***	***SBADM***	***Target Site***	***Method***	***Interpretation***	***Values***	***Total***
IJB	Original	166	30	20	6	13	2	2	14	***253***
	Normalizada	157	31	73	4	27	2	12	15	***321***
	Varianza	***-5,73%***	***3,23%***	***72,60%***	***-50,00%***	***51,85%***	***0,00%***	***83,33%***	***6,67%***	***21,18%***
UdS	Original	8	2	1	2	7	0	0	3	***23***
	Normalizada	8	2	1	2	9	0	1	3	***26***
	Varianza	***0,00%***	***0,00%***	***0,00%***	***0,00%***	***22,22%***	***0,00%***	***100,00%***	***0,00%***	***11,54%***
Maastro	Original	251	30	271	2	14	0	1	13	***582***
	Normalizada	198	29	279	2	66	0	11	7	***592***
	Varianza	***-26,77%***	***-3,45%***	***2,87%***	***0,00%***	***78,79%***	***0,00%***	***90,91%***	***-85,71%***	***1,69%***
UOXF	Original	80	11	6	2	25	2	4	8	***138***
	Normalizada	78	10	7	2	39	2	9	6	***153***
	Varianza	***-2,56%***	***-10,00%***	***14,29%***	***0,00%***	***35,90%***	***0,00%***	***55,56%***	***-33,33%***	***9,80%***
GBG	Original	53	3	7	2	1	2	0	7	***75***
	Normalizada	49	3	15	2	10	2	5	7	***93***
	Varianza	***-8,16%***	***0,00%***	***53,33%***	***0,00%***	***90,00%***	***0,00%***	***100,00%***	***0,00%***	***19,35%***
TOTAL	***Original***	***558***	***76***	***305***	***14***	***60***	***6***	***7***	***45***	***1071***
	Normalizada	***490***	***75***	***375***	***12***	***151***	***6***	***38***	***38***	***1185***
	Varianza	***-13,88%***	***-1,33%***	***18,67%***	***-16,67%***	***60,26%***	***0,00%***	***81,58%***	***-18,42%***	***9,62%***

5.2.2. Comparativa con otros modelos de datos

Esta sección ofrece una comparativa del sistema realizado, incluyendo los métodos realizados en la presente tesis como parte de la CIS, con los sistemas y modelos de datos más relevantes que aparecen en la literatura. Para ello se realiza una evaluación cualitativa y cuantitativa de la solución propuesta en la CIS, la CIS basada en HL7 RIM sin normalización y los sistemas i2b2 y OMOP. Para ello se selecciona un subconjunto de los datos de IJB (TOP Trial) que incluyen expresiones genéticas como se detalló en la sección 4.4.2.

La evaluación cuantitativa incluye una serie de características generales que hacen referencia a los requisitos funcionales de cada sistema:

a) **Ambigüedad de datos**: diferentes representaciones de la misma información en el mismo modelo de datos.

b) **Seguridad**: capacidad del sistema para filtrar datos según el permiso del usuario.

c) **Lenguajes de consulta**: utilizado para acceder a los datos.

d) **Consultas temporales**: capacidad de recuperar información con restricciones temporales.

e) **Trazabilidad**: al almacenamiento de la información original de las fuentes de datos y al mantenimiento de un log que permita consultar todas las transformaciones ocurridas.

f) **Inferencia de conocimiento de vocabularios**: explotación de información jerárquica y de sinónimos de vocabularios de dominio.

g) **Abstracción de consultas del modelo de datos**: posibilidad de generar consultas para el esquema del modelo de datos sin conocimiento previo.

h) **Información multimedia**: capacidad de almacenar información relacionada con otras observaciones.

i) **Información genética**: capacidad de almacenar información genética en el modelo de datos.

Además de la evaluación cualitativa, se han definido un pequeño conjunto de características que nos permitan realizar una evaluación cuantitativa sobre el mismo conjunto de datos y las mismas consultas.

j) **Redundancia de datos** en el mismo conjunto de datos**.**

k) Tiempo de ejecución.

Estas características han sido evaluadas en un subconjunto de datos procedentes de IJB con los pacientes del ensayo clínico TOP trial, utilizando además consultas extraídas de los criterios de elegibilidad del mismo ensayo clínico, simulando un escenario de reclutamiento de pacientes para un ensayos clínico. En la siguientes subsecciones se describen los resultados obtenidos con cada implementación del sistema.

5.2.2.1. HL7 RIM sin normalización

Esta solución utiliza la implementación de la CIS previa al diseño de los métodos descritos en la presente tesis doctoral, detallada en la sección 3.4. En lo referente a la parte cualitativa de la evaluación, este sistema tiene problemas debido a posibles ambigüedades en la representación de la información, ya que es posible encontrar los mismos datos representados en distintas ubicaciones del HL7 RIM, dependiendo del proceso ETL definido. Un ejemplo de esto serían datos originales que incluyen expresiones SNOMED-CT normalizadas y no normalizadas, se almacenan en diferentes clases HL7 RIM.

Con respecto a otras características de la evaluación, como puede verse en la siguiente tabla, esta implementación de una CIS basada en HL7 RIM sin normalizar admite la incorporación de seguridad. Además HL7 RIM está diseñado para ser utilizado con terminologías médicas y por tanto facilita la inferencia de conocimiento de estos en el sistema. Además de proporcionar la posibilidad de almacenar imágenes y relaciones entre distintas observaciones y un mantenimiento de la trazabilidad de cambios a niveles de terminologías.

Con respecto a la evaluación cuantitativa, es importante tener en cuenta que cada EC se ha traducido como una consulta SPARQL. Estos procesos tienen un impacto directo en el rendimiento al aumentar el tiempo de ejecución, como se puede observarse en la tabla.

Si bien la flexibilidad es una de las principales características importantes de HL7 RIM, también es una de sus principales desventajas, ya que consultar un modelo basado en HL7 RIM requiere que los usuarios conozcan tanto el esquema del modelo como la representación de datos. Aunque se puede proporcionar un servicio de abstracción de consultas, no cubre todos los resultados posibles debido a la falta de normalización de datos.

Tabla 18: Comparativa de los métodos propuestos con otros sistemas

	CIS con HL7	***OMOP***	***i2b2***	***CIS (HL7) con métodos diseñados***
Ambigüedad de datos	Sí, dependiendo de las ETL y las fuentes de datos	Sí, dependiendo de las ETL y las fuentes de datos	Sí, dependiendo de la estructura de árbol definida para mostrar los datos	No, el proceso de normalización semántica no permite ambigüedades
Seguridad	Sí, en el nivel de aplicaciones	Sí, en el nivel de aplicaciones	Sí, en el nivel de aplicaciones como otra celda	Sí, en el nivel de aplicaciones
Lenguaje de consulta	SPARQL	SQL	SQL	SPARQL
Consultas temporales	Sí, el modelo almacena distintos valores temporales	Sí, se almacenan fechas en el modelo	Sí, cada datos tiene su fecha asociada	Sí, el modelo almacena distintos valores temporales
Trazabilidad	Sí, la información no es transformada	No, dependerá de la ETL	No, dependerá de la ETL	Sí, la información es transformada pero se almacenan ambas en el CDM
Inferencia de conocimiento de vocabularios	Sí, gracias al Servidor de Terminologías	Manualmente, usando la tabla de vocabularios como un diccionario	Es un proceso manual sin conocimiento semántico, se busca por cadenas de texto	Sí, gracias al Servidor de Terminologías
Abstracción de consultas	No, es necesario conocimiento del modelo para realizar consultas	No, es necesario conocimiento del modelo para realizar consultas	No, es necesario conocimiento del modelo para realizar consultas	Sí, la abstracción de consultas permite crearlas en base al

				servidor de terminologías
Relación entre observaciones y multimedia	Sí, metadatos o imágenes pueden ser almacenadas mediante relaciones entre actos	No existe relación entre observaciones y procedimientos	No es posible relacionar actos entre sí, solos los definidos en el árbol de datos	Sí, metadatos o imágenes pueden ser almacenadas mediante relaciones entre actos
Datos genéticos	Sí, dependiendo del vocabulario utilizado para representarlo	No, sería necesario modificar el modelo con más información	Sí, este modelo soporta datos clínicos y genómicos	Sí, dependiendo del vocabulario utilizado para representarlo
Redundancia de datos	Sí (ej: Breast cancer versus Infiltrating Duct Carcinoma located on a Breast finding)	No, los conceptos son anotados en términos de su tabla vocabulario	Sí, ya que i2b2 depende de su árbol vocabulario teniendo contemplando únicamente cadenas de texto	No, el proceso de normalización semántica elimina datos redundantes
Tiempo de ejecución	300-600 ms	20-250 ms	300-650 ms	400-750 ms

5.2.2.2. OMOP

Siguiendo las recomendaciones facilitadas por OMOP se creó una base de datos relacional que se rellenó con el conjunto de datos de la prueba TOP incluyendo las expresiones génicas como en el sistema anterior. Las consultas de los criterios de elegibilidad se tradujeron al lenguaje SQL para probar el modelo de datos de OMOP. En comparación con el enfoque anterior, se obtuvieron resultados similares en diferentes puntos: ambigüedad de los datos, seguridad y consultas temporales, donde las principales diferencias aparecen con respecto al vocabulario médico. Usando este sistema, la inferencia usando el conocimiento del vocabulario implícito se debe realizar manualmente en las consultas, es decir, que no existe. La búsqueda de sinónimos está presente en las tablas de vocabulario del modelo OMOP, pero debe incluirse manualmente en las consultas.

La generación de consultar para recuperar información del modelo OMOP requiere conocer el esquema de representación de los datos. La información genética sí que se

puede almacenar parcialmente en el modelo para pruebas de laboratorio; sin embargo, esto solo es posible si los conceptos de vocabulario están presentes en las *mappings* facilitados por OMOP.

En cuanto a la evaluación cuantitativa, es importante señalar que produce mejores resultados con respecto al tiempo, debido a que las consultas ejecutadas utilizando SQL se realizan directamente contra una base de datos con elementos ya indexados y sin ningún mediador. El principal inconveniente de este enfoque es que es necesario extender manualmente las consultas con tablas de vocabulario para utilizar el conocimiento semántico del dominio médico. De hecho, todos los conceptos de vocabularios médicos como SNOMED-CT o LOINC están mapeados en estas tablas de vocabulario.

5.2.2.3. I2b2

Se desarrolló una instancia de un repositorio de datos i2b2 implementado como una base de datos relacional para comparar éste con el resto de los sistemas. La principal dificultad para poblar el repositorio es que el modelo de datos i2b2 es muy genérico. Para esto, se han seguido las guías del CRC de i2b2 para modelar y representar el conjunto de datos de investigación para el estudio actual.

Los valores de diagnóstico y pruebas de laboratorio se codificaron utilizando vocabularios estándar como SNOMED-CT. La estructura del vocabulario definido solamente incluye etiquetas de cadena por lo que la mayor parte de la información semántica se pierde. Por tanto, las búsquedas dependerán de las definiciones de esta estructura de vocabulario que generalmente es un árbol jerárquico que tiene que ser construido *ad-hoc* para los datos.

5.2.2.4. HL7 RIM con normalización

Este es el sistema desarrollado siguiendo el diseño descrito en la presente tesis doctoral. Donde las principales diferencias con los otros sistemas son, como puede apreciarse en la Tabla 18, el método de normalización semántica y el método de abstracción de consultas para obtener las plantillas de consultas para recuperar información del CDM.

Este sistema comparte ventajas con el sistema de HL7 sin normalización, coincidiendo por tanto en varias medidas (seguridad, lenguaje, consultas temporales, inferencia de conocimiento y relaciones multimedia). Con respecto a la medida sobre la ambigüedad

de los datos, este sistema no permite datos representados de manera ambigua en el CDM ya que el método de normalización semántica se encarga de transformar y corregir estos casos, muy frecuentes en conjuntos reales como vimos en la sección 5.2.1. El método de abstracción de consultas permite la generación de consultas para la recuperación de esta información normalizada sin ninguna ambigüedad ni conocimiento necesario en el esquema de representación utilizado.

Finalmente, los resultados cuantitativos son también similares al primer sistema. El rendimiento es ligeramente peor debido a la inferencia de conocimiento realizada en estos casos que requieren de la normalización semántica en el método de abstracción de consultas, pero estos métodos no surgen para mejorar el rendimiento del sistema. En este enfoque, se evitan datos redundantes como resultado del proceso de normalización.

Capítulo 6

6. CONCLUSIONES Y LÍNEAS DE FUTURO

6.1. Conclusiones

La interoperabilidad entre los sistemas implicados en la práctica clínica y en los ensayos clínicos modernos sigue siendo uno de los principales cuellos de botella en la gestión de la información para la investigación clínica. Hecho que se ha ido incrementado en las últimas décadas debido al aumento en la cantidad de información clínica presente en la red y en la aparición de nuevas variables clínicas y que han provocado un gran cambio en el diseño de los ensayos clínicos. Estos ensayos deben estudiar el impacto de nuevas pruebas moleculares que aumentan la especificidad de los ensayos y por tanto, complica la realización de los mismos debido a la cada vez más difícil tarea de encontrar pacientes que cumplan estos requisitos.

La mayoría de estas tareas se realizan actualmente de forma manual, por lo que la sostenibilidad de los estudios, incluida la información económica, es especialmente dependiente de los métodos avanzados para automatizar ciertos procedimientos. Aún más

importante, sin embargo, es explotar el conocimiento semántico inferido de los vocabularios clínicos que se ha desarrollado en los últimos años.

En esta tesis doctoral se ha abordado el diseño de dos métodos que tratan de mejorar la representación de la información clínica mediante un proceso de normalización semántica y abstracción de consultas basadas en estándares biomédicos. En concreto, la presente tesis en el capítulo de introducción plantea la posibilidad de enriquecer y corregir la representación clínica de datos mediante un proceso de normalización automática de conceptos médicos y abstracción de consultas que facilite la integración, homogeneización y el acceso a estos sin conocimiento del modelo de datos utilizado.

Asimismo, tras un análisis del estado de la cuestión donde se estudian desde los distintos tipos de heterogeneidades y enfoques de integración a los diferentes tipos niveles de interoperabilidad clínica, se aborda la integración de los métodos como parte de la Capa de Interoperabilidad Semántica de proyectos de investigación europeos formando el núcleo de central de su sistema en el capítulo 3.

El proceso de diseño e implementación descrito en los capítulos 3 y 4, comenzó con un estudio exhaustivo de los distintos estándares biomédicos y cuáles eran los más acordes a la tipología de los datos disponibles en los proyectos. Este análisis es el que permitió obtener el conjunto de tecnologías y estándares más apropiado para la solución implementada para las pruebas. Como resultado de esta implementación, se obtiene una CIS basada en dos componentes principales que además integran los métodos de normalización semántica y abstracción de consultas como parte de su solución para dotar al sistema de una solución capaz de homogeneizar la información clínica procedente de distintas fuentes de datos y facilitar la recuperación de los datos.

El método de normalización semántica diseñado para actuar en el proceso de integración de los datos en el CDM, se basa en un proceso de 3 pasos que comienza buscando los posibles términos del *Core Dataset* que representan cada idea clínica. Acto seguido se efectúa la normalización semántica, basada en la forma normal de SNOMED, como la mínima forma de expresión de un término y sus relaciones. Y finalmente se busca el lugar más idóneo para representar estos datos en el CDM.

Por otro lado, el método de abstracción de consultas es el encargado de mejorar la interoperabilidad del sistema con las aplicaciones o tecnologías encargadas de recuperar la información del CDM para su posterior consumo. Este método se caracteriza por el uso de dos componentes: librería de consultas y un envoltorio del lenguaje de consultas, que son los que permiten completar la CIS facilitando un punto de acceso homogéneo para los datos almacenados en su forma normal.

Uno de las principales ventajas de los métodos diseñados es su alto grado de modularidad, ya que si bien se ha realizado una implementación basada en unos estándares y modelos de datos muy concretos, su división en distintos componentes hacen factible su adaptación a otros estándares. Además, se puede destacar cómo la combinación de ambos métodos en la propia CIS facilita una integración de los datos procedentes de distintas fuentes de datos en un entorno homogéneo que asegura la trazabilidad de los datos y su posterior recuperación en cualquier entorno.

Para validar el trabajo diseñado se han realizado varias pruebas y experimentos con los datos de pacientes reales procedentes de los socios clínicos de proyectos europeos. Estas pruebas comenzaron analizando el impacto surgido al normalizar las fuentes de datos descritas en el capítulo 4, mediante el método de normalización semántica, y su evolución o comparación con las mismas fuentes de datos sin normalizar. Además, se ha realizado una comparativa del sistema obtenido al integrar estos métodos con otros sistemas representativos en la práctica clínica (Hl7, i2b2 y OMOP) obteniendo una evaluación positiva con respecto a unas métricas cualitativas, mostrando de esta forma la validez de la solución aportada.

Asimismo, como parte fundamental de los proyectos europeos donde se realizó la investigación (INTEGRATE y EURECA), se dispuso de la evaluación de los métodos como parte de la CIS en la validación de los escenarios clínicos definidos durante su ejecución y descritos en la sección 4.5. Escenarios que fueron validados por expertos de varias áreas de la Comisión Europea durante la ejecución de los proyectos europeos. Estos expertos, además de evaluar positivamente ambos proyectos, destacaron la CIS desarrollada como uno de los componentes más relevantes del proyecto, de la que además

aconsejaron su explotación y su seguimiento en futuros proyectos debido a su adaptabilidad a otras especialidades de la práctica clínica.

Todos estos resultados detallados en el capítulo 5 y los resultados científicos obtenidos durante esta investigación que se detallarán a continuación, sirven para corroborar la hipótesis planteada inicialmente, convirtiéndose de esta forma en tesis. Donde, los métodos diseñados pueden suponer una base para futuras herramientas de integración de datos clínicos, debido sobre todo a su sencillez y a su adaptabilidad a otros estándares.

6.2. Publicaciones obtenidas en este trabajo

6.2.1. Artículos de Revista

1. Paraiso-Medina, S., Perez-Rey, D., Bucur, A., Claerhout, B., & Alonso-Calvo, R. (2015). ***Semantic normalization and query abstraction based on SNOMED-CT and HL7: supporting multicentric clinical Trials***. Biomedical and Health Informatics, IEEE Journal Biomed Health Informatics, 19(3), 1061-1067.

   Publicación JCR. Factor de Impacto: 2,093 (Q1)

2. Perez-Rey, D., Alonso-Calvo, R., Paraiso-Medina, S., Munteanu, C.R., Garcia-Remesal, M., ***SNOMED2HL7: A tool to normalize and bind SNOMED CT concepts to the HL7 Reference Information Model,*** Computer Methods and Programs in Biomedicine, 2017, ISSN 0169-2607, http://dx.doi.org/10.1016/j.cmpb.2017.06.020.

   Publicación JCR. Factor de Impacto: 2,503 (Q1)

3. Priyatna, F., Alonso-Calvo, R., Paraiso-Medina, S., & Corcho, O. (2017). ***Querying clinical data in HL7 RIM based relational model with morph-RDB***. *Journal of Biomedical Semantics, 8*.

   Publicación JCR. Factor de Impacto: 1,845 (Q1)

4. Alonso-Calvo, R., Perez-Rey, D., Paraiso-Medina, S., Claerhout, B., Hennebert, P., & Bucur, A. (2015). ***Enabling semantic interoperability in multi-centric***

clinical trials on breast cancer. Computer methods and programs in biomedicine, 118(3), 322-329.

Publicación JCR. Factor de Impacto: 1,862 (Q1)

5. Alonso-Calvo, R., Paraiso-Medina, S., Perez-Rey, D., Alonso-Oset, E., van Stiphout, R., Yu, S., ... & Maojo, V. (2017). ***A semantic interoperability approach to support integration of gene expression and clinical data in breast cancer***. Computers in Biology and Medicine, Volume 87, 1 August 2017, Pages 179-186, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2017.06.005.

   Publicación JCR. Factor de Impacto: 1,836 (Q2)

6.2.2. Capítulos de libros

1. Paraiso-Medina, S, Perez-Rey, D, Alonso-Calvo, R, Munteanu CR., Pazos A, Kulikowski CA and Maojo V, ***Translational Bioinformatics: Informatics, Medicine, and -Omics***, In *Reference Module in Biomedical Sciences*, Elsevier, 2017, ISBN 9780128012383, https://doi.org/10.1016/B978-0-12-801238-3.10879-7

6.2.3. Ponencias en conferencias

1. Bucur, A., van Leeuwen, J., Chen, N.-Z., Claerhout, B., de Schepper, K., Perez-Rey, D., Paraiso-Medina, S., Alonso-Calvo, R., Krykwinski, C. (2016). **Cohort *Selection and Management Application Leveraging Standards-based Semantic Interoperability and a Groovy DSL***. *AMIA Summits on Translational Science Proceedings*, *2016*, 25–32.

2. Priyatna, F., Alonso-Calvo, R., Paraiso-Medina, S., Padron-Sanchez, G., & Corcho, O. (2015). ***R2RML-based Access and Querying to Relational Clinical Data with Morph-RDB***. In *SWAT4LS* (pp. 142-151).

3. Bucur, A.I., van Leeuwen, J., Chen, N.Z., Claerhout, B., de Schepper, K., Pérez-Rey, D., Paraiso-Medina, S. and Mehta, K., 2015. ***Cohort Selection Tool for Efficient Exploration of Patient Data***. In *AMIA*.

4. Paraiso-Medina, S., Pérez-Rey, D., Alonso-Calvo, R., Claerhout, B., de Schepper, K., Hennebert, P., & Bucur, A. I. (2013). ***Semantic Interoperability Solution for Multicentric Breast Cancer Trials at the Integrate EU Project***. In HEALTHINF (pp. 34-41).

5. Moratilla, J. M., Alonso-Calvo, R., Molina-Vaquero, G., Paraiso-Medina, S., Perez-Rey, D., & Maojo, V. (2012). ***A data model based on semantically enhanced HL7 RIM for sharing patient data of breast cancer clinical trials***. Studies in health technology and informatics, 192, 971-971.

6.3. Líneas futuras

Si bien esta tesis presentaba unos métodos que buscaban la mejora en la integración de los datos clínicos mediante el uso de un proceso de normalización y abstracción de consultas basados en estándares terminológicos, la implementación se basa en la utilización de un subconjunto de varios estándares para su evaluación. Una línea futura evidente sería la extensión de los métodos con otros modelos. A continuación, se listan las posibles líneas futuras:

- Extensión de los métodos a otros estándares. El trabajo realizado en la presente tesis doctoral se ha evaluado en una implementación concreta de los métodos y su integración en la CIS bajo el modelo de datos HL7 RIM y las terminologías SNOMED CT, LOINC y HGNC. Si bien en el diseño de los métodos se hizo especial hincapié en su modularidad para facilitar su adaptación en otros sistemas, uno de los procesos centrales se basa en la normalización mediante el mecanismo de forma normal de SNOMED. Por tanto, la adaptación de este método pasaría por definir en primer lugar un proceso de normalización semántica basada en otras terminologías o en una terminología que actuara de *lingua franca*. En este caso se podría pensar en utilizar el conocimiento de SNOMED CT como *lingua franca* y así poder utilizar este conocimiento en futuras normalizaciones para posteriormente trabajar con los *mappings* o enlaces existentes en otras terminologías. Por otro lado, la extensión a otros modelos de datos sería un proceso bastante sencillo ya que en su adaptación se vería involucrada únicamente

la librería de consultas definida en el método de abstracción de consultas. Prioridad: alta, dificultad: baja.

- Fruto de esta línea surge la idea o necesidad de extender otras terminologías clínicas con el conocimiento de SNOMED CT. Si bien existen terminologías y metatesauros como UMLS que facilitan los mapeos entre terminologías, estás únicamente facilitan el correspondiente sinónimo para su traducción. SNOMED CT es un vocabulario que no solo contiene un abanico más grande de términos, si no que contiene unos mecanismos que favorecen la generación de conocimiento implícito en él. De esta forma, se plantea la opción de enriquecer otras terminologías médicas con el conocimiento adquirido en SNOMED CT. Prioridad: baja, dificultad: media.
- Evaluación de los métodos en otras áreas de conocimiento. El trabajo realizado en la presente tesis doctoral se ha centrado específicamente en el dominio biomédico, concretamente en casos cercanos a tratamientos e investigaciones sobre cáncer. Los métodos han sido diseñados para su adaptación a otras áreas, si bien el campo de evaluación han sido proyectos de investigación europea centrados en esta área. Por tanto, sería interesante evaluar su fiabilidad en otros ámbitos o estudios clínicos, si bien, lo que parece afectar más en estos cambios sería las anotaciones con términos del *Core Dataset* y la generación de consultar que simulen los criterios de elegibilidad de ensayos. Prioridad: alta, dificultad: media.
- Integración de un método de extracción de información de criterios de elegibilidad textuales. De las anteriores líneas también surge la idea de una posible integración de un método automático para anotación y representación de criterios de elegibilidad mediante términos del *Core Dataset*. De esta forma se facilitaría la creación automática de consultas desde bases de datos públicas como ClinicalTrials.gov y además se podría realizar un proceso de normalización semántica de estos criterios para analizar y mejorar la composición de estos criterios. Prioridad: baja, dificultad: alta.

- Por último, se ha abierto una nueva línea de investigación relacionado con la estandarización e integración del genograma genético en la historia clínica electrónica. Si bien actualmente existe un estándar internacional para su representación, en la práctica clínica no se viene utilizando este estándar, por lo que sería relevante su adopción en el propio sistema clínico. Con una integración total de genogramas genéticos en las historias clínicas, se podría enriquecer la historia clínica de pacientes y familiares debido a la representación de patologías o fenotipos de toda una familia. De esta forma se podría tener una representación bastante más amplia de los antecedentes familiares de los pacientes. Prioridad: media, dificultad: media.

Capítulo 7

7. REFERENCIAS

[1]. Dahlen, R. W. (1997). A History of Medical Informatics in the United States: 1950 to 1990. Bulletin of the Medical Library Association, 85(4), 443.

[2]. Sawicki, M. P., Samara, G., Hurwitz, M., & Passaro, E. (1993). Human genome project. The American journal of surgery, 165(2), 258-264.

[3]. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., ... & Funke, R. (2001). Initial sequencing and analysis of the human genome. Nature, 409, 860-921.

[4]. Feero, W. G., Guttmacher, A. E., & Collins, F. S. (2010). Genomic medicine—an updated primer. New England Journal of Medicine, 362(21), 2001-2011.

[5]. Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. New England Journal of Medicine, 372(9), 793-795.

[6]. Hamburg, M. A., & Collins, F. S. (2010). The path to personalized medicine. *New England Journal of Medicine*, *363*(4), 301-304.

[7]. Jameson, J. L., & Longo, D. L. (2015). Precision medicine—personalized, problematic, and promising. *Obstetrical & Gynecological Survey*, *70*(10), 612-614.

[8]. Kanehisa, M., & Bork, P. (2003). Bioinformatics in the post-sequence era. *Nature genetics*, *33*, 305.

[9]. Mello, M. M., Francer, J. K., Wilenzick, M., Teden, P., Bierer, B. E., & Barnes, M. (2013). Preparing for responsible sharing of clinical trial data. The New England journal of medicine, 369(17), 1651-1658.

[10]. Ross, J. S., Lehman, R., & Gross, C. P. (2012). The importance of clinical trial data sharing: toward more open science. *Circulation: Cardiovascular Quality and Outcomes*, *5*(2), 238-240.

[11]. Häyrinen, K., Saranto, K., & Nykänen, P. (2008). Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International journal of medical informatics*, *77*(5), 291-304.

[12]. Lenzerini, M. (2002, June). Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 233-246). ACM.

[13]. Anguita, A., Martin, L., Perez-Rey, D., & Maojo, V. (2010). A review of methods and tools for database integration in biomedicine. *Current Bioinformatics*, *5*(4), 253-269.

[14]. Suter, E., Oelke, N. D., Adair, C. E., & Armitage, G. D. (2009). Ten key principles for successful health systems integration. *Healthcare quarterly (Toronto, Ont.)*, *13*(Spec No), 16.

[15]. Vest, J. R., & Gamm, L. D. (2010). Health information exchange: persistent challenges and new strategies. *Journal of the American Medical Informatics Association*, *17*(3), 288-294.

[16]. Barbarito, F., Pinciroli, F., Mason, J., Marceglia, S., Mazzola, L., & Bonacina, S. (2012). Implementing standards for the interoperability among healthcare providers in the public regionalized Healthcare Information System of the Lombardy Region. *Journal of biomedical informatics*, *45*(4), 736-745.

[17]. Hanauer, D. A., Rhodes, D. R., & Chinnaiyan, A. M. (2009). Exploring clinical associations using '-omics' based enrichment analyses. *PloS one*, *4*(4), e5203.

[18]. European Comission (2011). INTEGRATE, Driving Excellence in Integrative Cancer Research through Innovative Biomedical Infrastructures. Retrieved from https://cordis.europa.eu/project/rcn/97843_es.html

[19]. European Comission (2012). EURECA, Enabling information re-Use by linking clinical REsearch and CAre. Retrieved from https://cordis.europa.eu/project/rcn/102156_es.html

[20]. Paraíso-Medina, S. (2013). "Interoperabilidad Semántica en Ensayos Clínicos Multicéntricos sobre Cáncer de Mama". Tesis Fin de Máster, Máster Universitario en Inteligencia Artificial, Universidad Politécnica de Madrid.

[21]. Gurwitz, D., Lunshof, J. E., & Altman, R. B. (2006). A call for the creation of personalized medicine databases. *Nature Reviews Drug Discovery*, *5*(1), 23.

[22]. Philippi, S., & Köhler, J. (2006). Addressing the problems with life-science databases for traditional uses and systems biology. *Nature Reviews Genetics*, *7*(6), 482.

[23]. Ludwig, J. A., & Weinstein, J. N. (2005). Biomarkers in cancer staging, prognosis and treatment selection. *Nature Reviews Cancer*, *5*(11), 845.

[24]. Allegra, A., Alonci, A., Campo, S., Penna, G., Petrungaro, A., Gerace, D., & Musolino, C. (2012). Circulating microRNAs: new biomarkers in diagnosis, prognosis and treatment of cancer. *International journal of oncology*, *41*(6), 1897-1912.

[25]. Bertoli, G., Cava, C., & Castiglioni, I. (2015). MicroRNAs: new biomarkers for diagnosis, prognosis, therapy prediction and therapeutic tools for breast cancer. *Theranostics*, *5*(10), 1122.

[26]. Baxevanis, A. D. (2001). The Molecular Biology Database Collection: an updated compilation of biological database resources. *Nucleic Acids Research*, *29*(1), 1-10.

[27]. Malterud, K. (2001). Qualitative research: standards, challenges, and guidelines. *The lancet*, *358*(9280), 483-488.

[28]. Cimino, J. J. (1996). Coding systems in health care. *Methods of information in medicine*, *35*, 273-284.

[29]. Wong, L. (2002). Technologies for integrating biological data. *Briefings in bioinformatics*, *3*(4), 389-404.

[30]. Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., ... & Tegnér, J. (2014). Data integration in the era of omics: current and future challenges.

[31]. Mate, S., Köpcke, F., Toddenroth, D., Martin, M., Prokosch, H. U., Bürkle, T., & Ganslandt, T. (2015). Ontology-based data integration between clinical and research systems. *PloS one*, *10*(1), e0116656.

[32]. Martín, L., Bonsma, E., Anguita, A., Vrijnsen, J., García-Remesal, M., Crespo, J., ... & Maojo, V. (2007, November). Data Access and Management in ACGT: Tools to solve syntactic and semantic heterogeneities between clinical and image databases. In *International Conference on Conceptual Modeling* (pp. 24-33). Springer, Berlin, Heidelberg.

[33]. Anguita Sanchez, A. (2012). *Modelo de mediación semántica para la integración de fuentes de datos heterogéneas* (Doctoral dissertation, Informatica).

[34]. Bergman, M. (2006). Sources and classification of semantic heterogeneities. *Web Blog: AI3-Adaptive Information, Adaptive Innovation, Adaptive Infrastructure*.

[35]. Hakimpour, F., & Geppert, A. (2001, October). Resolving semantic heterogeneity in schema integration. In *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001* (pp. 297-308). ACM.

[36]. Cimino, J. J. (1998). Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of information in medicine*, *37*(4-5), 394.

[37]. Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, *23*(4), 3-13.

[38]. Tsuruoka, Y., McNaught, J., & Ananiadou, S. (2008, April). Normalizing biomedical terms by minimizing ambiguity and variability. In *BMC bioinformatics* (Vol. 9, No. 3, p. S2). BioMed Central.

[39]. Sujansky, W. (2001). Heterogeneous database integration in biomedicine. *Journal of biomedical informatics*, *34*(4), 285-298.

[40]. Kimball, R., & Ross, M. (2011). *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons.

[41]. Maojo, V., García-Remesal, M., Billhardt, H., Alonso-Calvo, R., Pérez-Rey, D., & Martín-Sánchez, F. (2006). Designing new methodologies for integrating biomedical information in clinical trials. *Methods of information in medicine*, *45*(02), 180-185.

[42]. Tsiknakis, M., Kafetzopoulos, D., Potamias, G., Analyti, A., Marias, K., & Manganas, A. (2006). Building a European biomedical grid on cancer: the ACGT Integrated Project. *Studies in health technology and informatics*, *120*, 247.

[43]. Newman, A., Hunter, J., Li, Y. F., Bouton, C., & Davis, M. (2008). A scale-out RDF molecule store for distributed processing of biomedical data. In *Semantic Web for Health Care and Life Sciences Workshop*.

[44]. Alonso-Calvo, R., Paraiso-Medina, S., Perez-Rey, D., Alonso-Oset, E., van Stiphout, R., Yu, S., ... & Maojo, V. (2017). A semantic interoperability approach to support integration of gene expression and clinical data in breast cancer. *Computers in biology and medicine*, *87*, 179-186.

[45]. McCarty, C. A., Chisholm, R. L., Chute, C. G., Kullo, I. J., Jarvik, G. P., Larson, E. B., ... & Struewing, J. P. (2011). The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC medical genomics*, *4*(1), 13.

[46]. Shadbolt, N., Berners-Lee, T., & Hall, W. (2006). The semantic web revisited. *IEEE intelligent systems*, *21*(3), 96-101.

[47]. Brochhausen, M., Spear, A. D., Cocos, C., Weiler, G., Martín, L., Anguita, A., ... & Sfakianakis, S. (2011). The ACGT Master Ontology and its applications–Towards an ontology-driven cancer research and management system. *Journal of biomedical informatics*, *44*(1), 8-25

[48]. Nightingale, F. (1863). *Notes on hospitals*. Longman, Green, Longman, Roberts, and Green.

[49]. Farr, W. (1885). *Vital statistics: a memorial volume of selections from the reports and writings of William Farr*. sanitary institute.

[50]. Haux, R. (2006). Health information systems–past, present, future. *International journal of medical informatics*, *75*(3-4), 268-281.

[51]. Geraci, A., Katki, F., McMonegal, L., Meyer, B., Lane, J., Wilson, P., ... & Springsteel, F. (1991). *IEEE standard computer dictionary: Compilation of IEEE standard computer glossaries*. IEEE Press.

[52]. Benson, T. (2010). *Principles of health interoperability HL7 and SNOMED* (p. 263). London:: Springer.

[53]. Dolin, R. H., & Alschuler, L. (2010). Approaching semantic interoperability in health level seven. *Journal of the American Medical Informatics Association*, *18*(1), 99-103.

[54]. Benson, T., & Grieve, G. (2016). Principles of FHIR. In *Principles of Health Interoperability* (pp. 329-348). Springer, Cham.

[55]. Paraiso-Medina, S., Perez-Rey, D., Bucur, A., Claerhout, B., & Alonso-Calvo, R. (2015). Semantic normalization and query abstraction based on SNOMED-CT and HL7: supporting multicentric clinical Trials. *IEEE journal of biomedical and health informatics*, *19*(3), 1061-1067.

[56]. Rhoads, J. G., Cooper, T., Fuchs, K., Schluter, P., & Zambuto, R. P. (2010). Medical device interoperability and the Integrating the Healthcare Enterprise (IHE) initiative. *Biomed Instrum Technol*, (Suppl), 21-27.

[57]. Mildenberger, P., Eichelberg, M., & Martin, E. (2002). Introduction to the DICOM standard. *European radiology*, *12*(4), 920-927.

[58]. Wozak, F., Ammenwerth, E., Hörbst, A., Sögner, P., Mair, R., & Schabetsberger, T. (2008). IHE based interoperability-benefits and challenges. In *MIE* (Vol. 136, pp. 771-776).

[59]. Kalra, D., Stroetmann, V., Sundgren, M., Dupont, D., Schlünder, I., Thienpont, G., ... & De Moor, G. (2017). The European Institute for Innovation through health data. *Learning Health Systems*, *1*(1), e10008.

[60]. Lindén, F. (2009). epsos, smart open services for European patients from strategies to services health as the enabler for cross-border healthcare. *Infrastructures for Health Care*, *23*.

[61]. European Commission. eHealth action plan 2012–2020. http://ec.europa.eu/health/ehealth/docs/com_2012_736_en.pdf (accessed on July 2018).

[62]. Beale, T., & Heard, S. (2008). openEHR architecture overview. openEHR Foundation. *London, UK*.

[63]. González-Ferrer, A., Peleg, M., Verhees, B., Verlinden, J. M., & Marcos, C. (2013). Data integration for clinical decision support based on openEHR archetypes and HL7 virtual medical record. In *Process Support and Knowledge Representation in Health Care* (pp. 71-84). Springer, Berlin, Heidelberg.

[64]. Dentler, K., ten Teije, A., Cornet, R., & de Keizer, N. (2012). Semantic integration of patient data and quality indicators based on openEHR archetypes. In *Process Support and Knowledge Representation in Health Care* (pp. 85-97). Springer, Berlin, Heidelberg.

[65]. Murphy, S. N., Weber, G., Mendis, M., Gainer, V., Chueh, H. C., Churchill, S., & Kohane, I. (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, *17*(2), 124-130.

[66]. Ganslandt, T., Mate, S., Helbing, K., Sax, U., & Prokosch, H. U. (2011). Unlocking data for clinical research–the German i2b2 experience. *Applied clinical informatics*, *2*(01), 116-117.

[67]. Johnson, E. K., Broder-Fingert, S., Tanpowpong, P., Bickel, J., Lightdale, J. R., & Nelson, C. P. (2014). Use of the i2b2 research query tool to conduct a matched case–control clinical research study: advantages, disadvantages and methodological considerations. *BMC medical research methodology*, *14*(1), 16.

[68]. Stang, P. E., Ryan, P. B., Racoosin, J. A., Overhage, J. M., Hartzema, A. G., Reich, C., ... & Woodcock, J. (2010). Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Annals of internal medicine*, *153*(9), 600-606.

[69]. Hripcsak, G., Duke, J. D., Shah, N. H., Reich, C. G., Huser, V., Schuemie, M. J., ... & Van Der Lei, J. (2015). Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Studies in health technology and informatics*, *216*, 574.

[70]. Madigan, D., Ryan, P. B., Schuemie, M., Stang, P. E., Overhage, J. M., Hartzema, A. G., ... & Berlin, J. A. (2013). Evaluating the impact of database heterogeneity on observational study results. *American journal of epidemiology*, *178*(4), 645-651.

[71]. Fitzhenry, F., Resnic, F. S., Robbins, S. L., Denton, J., Nookala, L., Meeker, D., ... & Matheny, M. E. (2015). Creating a common data model for comparative effectiveness with the observational medical outcomes partnership. *Applied clinical informatics*, *6*(03), 536-547.

[72]. HL7 Clinical genomics working group. Retrieved from http://www.hl7.org/Special/committees/clingenomics.

[73]. Hasman, A. (2006). HL7 RIM: an incoherent standard. In *Ubiquity: Technologies for Better Health in Aging Societies, Proceedings of Mie2006* (Vol. 124, p. 133).

[74]. Bender, D., & Sartipi, K. (2013, June). HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. In *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on* (pp. 326-331). IEEE.

[75]. Trott, P. A. (1977). International classification of diseases for oncology. *Journal of clinical pathology*, *30*(8), 782.

[76]. Cimino, J. J. (1998). Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of information in medicine*, *37*(4-5), 394.

[77]. Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., ... & Rosse, C. (2005). Relations in biomedical ontologies. *Genome biology*, *6*(5), R46.

[78]. Cimino, J. J., & Zhu, X. (2006). The practical impact of ontologies on biomedical informatics. *Yearbook of medical informatics*, *15*(01), 124-135.

[79]. de Lusignan, S., Minmagh, C., Kennedy, J., Zeimet, M., Bommezijn, H., & Bryant, J. (2001). A survey to identify the clinical coding and classification systems currently in use across Europe. *Studies in health technology and informatics*, (1), 86-89.

[80]. Cirera, L., & Vázquez, E. (1998). La implantación en España de la Clasificación Internacional de Enfermedades 10ª-Revisión (CIE-10). *Santiago de Compostela: Sociedad Española de Epidemiología*.

[81]. Ministerio de Sanidad, Consumo y Bienestar Social. https://www.msssi.gob.es/estadEstudios/estadisticas/normalizacion/CIE10/home.htm. Retrieved on July 2018.

[82]. Forrey, A. W., Mcdonald, C. J., DeMoor, G., Huff, S. M., Leavelle, D., Leland, D., ... & Tullis, A. (1996). Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clinical Chemistry*, *42*(1), 81-90.

[83]. LOINC & SNOMED International agreement: https://loinc.org/collaboration/snomed-international/. Accessed on July 2018.

[84]. Gray, K. A., Yates, B., Seal, R. L., Wright, M. W., & Bruford, E. A. (2014). Genenames. org: the HGNC resources in 2015. *Nucleic acids research*, *43*(D1), D1079-D1085.

[85]. Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, *121*, 279.

[86]. Benson, T. (2012). SNOMED CT Concept Model. In *Principles of Health Interoperability HL7 and SNOMED* (pp. 253-266). Springer, London.

[87]. Spackman, K. A. (2001). Normal forms for description logic expressions of clinical concepts in SNOMED RT. In *Proceedings of the AMIA Symposium* (p. 627). American Medical Informatics Association.

[88]. Organización Mundial de la Salud. Retrieved from: http://www.who.int/ictrp/es/

[89]. Weng, C., Tu, S. W., Sim, I., & Richesson, R. (2010). Formal representation of eligibility criteria: a literature review. *Journal of biomedical informatics*, *43*(3), 451-467.

[90]. Kuchinke, W., Aerts, J., Semler, S. C., & Ohmann, C. (2009). CDISC standard-based electronic archiving of clinical trials. *Methods of information in medicine*, *48*(05), 408-413.

[91]. Galton, F. (1889). Natural inheritance (1889). *SERIES E: PHYSIOLOGICAL PSYCHOLOGY*.

[92]. Resta, R. G. (1993). The crane's foot: The rise of the pedigree in human genetics. *Journal of Genetic Counseling*, *2*(4), 235-260.

[93]. Bennett, R. L., Steinhaus, K. A., Uhrich, S. B., O'Sullivan, C. K., Resta, R. G., Lochner-Doyle, D., ... & Hamanishi, J. (1995). Recommendations for standardized human pedigree nomenclature. *Journal of Genetic Counseling*, *4*(4), 267-279.

[94]. Bennett, R. L., French, K. S., Resta, R. G., & Doyle, D. L. (2008). Standardized human pedigree nomenclature: update and assessment of the recommendations of the National Society of Genetic Counselors. *Journal of genetic counseling*, *17*(5), 424-433.

[95]. Bennett, R. L. (2011). *The practical guide to the genetic family history*. John Wiley & Sons.

[96]. Robinson, P. N., & Mundlos, S. (2010). The human phenotype ontology. *Clinical genetics*, *77*(6), 525-534.

[97]. Saltz, J., Oster, S., Hastings, S., Langella, S., Kurc, T., Sanchez, W., ... & Covitz, P. (2006). caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid. *Bioinformatics*, *22*(15), 1910-1916.

[98]. George A. Komatsoulis. "Program Announcement". National Cancer Institute. Archived on : https://web.archive.org/web/20120730234757/https://cabig.nci.nih.gov/program_announcement

[99]. Foley, J. (2011). Report blasts problem-plagued cancer research grid. *InformationWeek. April*, *8*, 12.

[100]. Turvey, C., Klein, D., Fix, G., Hogan, T. P., Woods, S., Simon, S. R., ... & Wakefield, B. (2014). Blue Button use by patients to access and share health record information using the Department of Veterans Affairs' online patient portal. *Journal of the American Medical Informatics Association*, *21*(4), 657-663.

[101]. Paraiso-Medina, S., Pérez-Rey, D., Alonso-Calvo, R., Claerhout, B., de Schepper, K., Hennebert, P., ... & Bucur, A. I. (2013). Semantic Interoperability Solution for Multicentric Breast Cancer Trials at the Integrate EU Project. *HEALTHINF*, *34*, 41.

[102]. Aso, S., Perez-Rey, D., Alonso-Calvo, R., Rico-Diez, A., Bucur, A., Claerhout, B., & Maojo, V. (2013). Analyzing SNOMED CT and HL7 terminology binding for semantic interoperability on post-genomic clinical trials. *Studies in health technology and informatics*, *192*, 980-980.

[103]. Moratilla, J. M., Alonso-Calvo, R., Molina-Vaquero, G., Paraiso-Medina, S., Perez-Rey, D., & Maojo, V. (2013). A data model based on semantically enhanced HL7 RIM for sharing patient data of breast cancer clinical trials. *Studies in health technology and informatics*, *192*, 971-971.

[104]. Priyatna, F., Alonso-Calvo, R., Paraiso-Medina, S., Padron-Sanchez, G., & Corcho, O. (2015). R2RML-based Access and Querying to Relational Clinical Data with Morph-RDB. In *SWAT4LS* (pp. 142-151).

[105]. Priyatna, F., Alonso-Calvo, R., Paraiso-Medina, S., & Corcho, O. (2017). Querying clinical data in HL7 RIM based relational model with morph-RDB. *Journal of biomedical semantics*, *8*(1), 49.

[106]. Markwell, D., Sato, L., & Cheetham, E. (2008). Representing clinical information using SNOMED Clinical Terms with different structural information models. In *KR-MED* (Vol. 2008, pp. 72-79).

[107]. Broekstra, J., Kampman, A., & Van Harmelen, F. (2003). Sesame: An architecture for storing and querying RDF data and schema information. *Spinning the semantic web: Bringing the world wide web to its full potential*, *197*.

[108]. Codd, E. F. (1971, November). Normalized data base structure: A brief tutorial. In *Proceedings of the 1971 ACM SIGFIDET (now SIGMOD) Workshop on Data Description, Access and Control* (pp. 1-17). ACM.

[109]. Beeri, C., Bernstein, P. A., & Goodman, N. (1988). A sophisticate's introduction to database normalization theory. In *Readings in Artificial Intelligence and Databases* (pp. 468-479).

[110]. Benson, T., & Grieve, G. (2016). Why interoperability is hard. In *Principles of Health Interoperability* (pp. 19-35). Springer, Cham.

[111]. Kiers, H. A. (2000). Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *14*(3), 105-122.

[112]. Milian, K., Bucur, A., & Ten Teije, A. (2012, October). Formalization of clinical trial eligibility criteria: Evaluation of a pattern-based approach. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on* (pp. 1-4). IEEE.

[113]. Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, *32*(suppl_1), D267-D270.

[114]. Perez-Rey, D., Alonso-Calvo, R., Paraiso-Medina, S., Munteanu, C. R., & Garcia-Remesal, M. (2017). SNOMED2HL7: A tool to normalize and bind SNOMED CT concepts to the HL7 Reference Information Model. *Computer methods and programs in biomedicine*, *149*, 1-9.

[115]. Desmedt, Christine, et al. "Multifactorial approach to predicting resistance to anthracyclines." *J Clin Oncol* 29.12 (2011): 1578-1586.

[116]. Reinhard, Harald, et al. "Wilms' tumor in adults: results of the Society of Pediatric Oncology (SIOP) 93-01/Society for Pediatric Oncology and Hematology (GPOH) Study." *Journal of Clinical Oncology* 22.22 (2004): 4500-4506.

[117]. Graf, Norbert, et al. "Characteristics and outcome of stage II and III non-anaplastic Wilms' tumour treated according to the SIOP trial and study 93-01." *European Journal of Cancer* 48.17 (2012): 3240-3248.

[118]. Graf, N., Anguita, A., Bucur, A., Burke, D., Claerhout, B., Coveney, P., ... & Jefferys, B. (2012). P-medicine: A solution for translational research. *Paed Blood Cancer*, *59*(6), 1101.

[119]. Goldhirsch, A. 2011, et al. "Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011." *Annals of oncology* 22.8 (2011): 1736-1747.

ANEXOS

ANEXO A. Acrónimos

API	Application Programming Interface
caBIG	cancer Biomedical Informatics Grid
CAP	College of Pathologists
CD	Core Dataset
CDISC	Clinical Data Interchange Standards Consortium
CDM	Common Data Model
CE	Criterio de Elegibilidad
CIM	Common Information Model
CIS	Capa de Interoperabilidad Semántica
CRC	Clinical Data Repository
CRD	Cuaderno de Recogida de Datos
CRFQ	Clinical Research Filtered Query
CT	Clinical Trial

CTCAE	Common Terminology Criteria for Adverse Events
DICOM	Digital Imaging and Communication in Medicine
EHR	Electronic Health Records
epSOS	Smart Open Services for European Patients
ERGO	Eligibility Rule Grammar and Ontology
ETL	Extract, Transform and Load
EudraCT	European Clinical Trials Database
EURECA	Enabling information re-Use by linking clinical Research
FHIR	Fast Healthcare Interoperability Resources
FSN	Fully Specified Name
GBG	German Breast Group
GIB	Grupo de Informática Biomédica
HCE	Historia Clínica Electrónica
HDBS	Sistemas de Bases de Datos Heterogéneos
HGNC	HUGO Gene Nomenclature Committee
HL7	Health Level 7
HL7 RIM	Health Level 7 Reference Information Model
i2b2	Informatics for Integrating Biology and the Bedside
ICD 9 - 10	International Classification of Diseases 9 - 10
ICD-CM	International Classification of Diseases, Clinical
ICD-NA	International Classification of Diseases to Neurology
ICD-O	International Classification of Diseases to Oncology
ICTRP	Plataforma de registros internacionales de ensayos clínicos
IHE	Integrating the Healthcare Enterprise
IHTSDO	International Health Terminology Standars Development
IJB	Institute Jules Bordet
INTEGRATE	Driving excellence in integrative cancer research
ISO	International Organization for Standardization

JSON	JavaScript Object Notation
LOINC	Logical Observation Identifiers Names and Codes
MedDRA	Medical Dictionary for Regulatory Activities
MSSSI	Ministerio de Sanidad, Servicios Sociales e Igualdad
NCI	National Cancer Institute
NCIt	National Cancer Institute thesaurus
NIH	National Institutes of Health
NLM	National Library of Medicine
NLP	Natural Language Processing
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
OMS	Organización Mundial de la Salud
OpenEHR	An open domain-driven platform for developing flexible e-
OWL2	Web Ontology Language
PGH	Proyecto del Genoma Humano
REST	REpresentational State Transfer
RI	Regenstrief Institute
SBADM	Substance Administration
SIH	Sistema de Información Hospitalario
SNOMED CT	Systematized Nomenclature of Medicine and Clinical Terms
SNP	*Single Nucleotide Polymorphism*
SOAP	Simple Object Access Protocol
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
UdS	Universität des Saarlandes
UMLS	Unified Medical Language System
Uoxf	University of Oxford
XML	Extensible Markup Language

ANEXO B. Anotaciones de los datos

B.1. Anotación de datos del Institute Jules Bordet

Tabla 19: Conceptos anotados del CD en el Institute Jules Bordet

Etiqueta original	Traducción a *Core Dataset*	Código (CD)	Terminología
DEBHOSP	Hospital admission	32485007	SNOMED CT
FINHOSP	Patient discharge	58000006	SNOMED CT
DATECONS	HISTORY OF PAST ILLNESS	11348-0	LOINC
ATCSEIN	Malignant tumor of breast	254837009	SNOMED CT
ATCOVAIR	Malignant tumor of ovary	363443007	SNOMED CT
ATCCANC	Malignant neoplastic disease	363346000	SNOMED CT
ATCHTA	Hypertensive disorder	38341003	SNOMED CT
ATCCAR	Heart disease	56265001	SNOMED CT
ATCVASC	Vascular disorder	27550009	SNOMED CT
ATCDIAB	Diabetes mellitus	73211009	SNOMED CT
ATCDEPRESS	Depressive disorder	35489007	SNOMED CT
ATCCATAR	Cataract	193570009	SNOMED CT
ATCAUTR	Disease	64572001	SNOMED CT
ATCANNEXUNIT	Unilateral excision of adnexa of uterus	176907004	SNOMED CT
ATCANNEXBILAT	Bilateral excision of adnexa of uterus	176901003	SNOMED CT
ATCHYSTER	Hysterectomy	236886002	SNOMED CT

MENARCH	Age at menarche	398700009	SNOMED CT
MENSTA & AGMENOP 1	Premenopausal state	22636003	SNOMED CT
MENSTA & AGMENOP 2	Postmenopausal state	76498008	SNOMED CT
TRTHORM	Endocrine therapy	309542002	SNOMED CT
FSH	Follicle stimulating hormone measurement	31003009	SNOMED CT
LSH	Luteinizing hormone measurement	69527006	SNOMED CT
GESTITE	Gravida – finding	366321006	SNOMED CT
PARITE	Number of live deliveries	248991006	SNOMED CT
LATER=1	Left	7771000	SNOMED CT
LATER=2	Right	24028007	SNOMED CT
LATER=3	Right and left	51440002	SNOMED CT
DIAGHISTO	Tru-cut biopsy of breast	303689004	SNOMED CT
DIAGBIOPS	Biopsy of breast	122548005	SNOMED CT
CT=0	T0 category	58790005	SNOMED CT
CT=1b	Tumor stage T1b	261649005	SNOMED CT
CT=1c	Tumor stage T1c	261650005	SNOMED CT
CT=2	T2 category	67673008	SNOMED CT
CT=3:	T3 category	14410001	SNOMED CT
CT=is	Tis category	44401000	SNOMED CT

CN=0	N0 category	62455006	SNOMED CT
CN=1	N1 category	53623008	SNOMED CT
CM=0	M0 category	30893008	SNOMED CT
CM=X	MX category	27167007	SNOMED CT
CA153	CA 15-3 measurement	113058009	SNOMED CT
TYPECH=101	Lumpectomy of breast	392021009	SNOMED CT
TYPECH=102	Lumpectomy of breast	392021009	SNOMED CT
	related with Excision of sentinel lymph node	443497002	
TYPECH=103	Lumpectomy of breast	392021009	SNOMED CT
	related with Excision of axillary lymph node	234262008	
TYPECH=104	Simple mastectomy	172043006	SNOMED CT
TYPECH=105	Modified radical mastectomy	406505007	SNOMED CT
TYPECH=108	Excision of axillary lymph node	234262008	SNOMED CT
TYPECH=110	Simple mastectomy	172043006	SNOMED CT
	related with Excision of axillary lymph node	234262008	
TYPECH=111	Lumpectomy of breast	392021009	SNOMED CT
	related with Excision of sentinel lymph node	443497002	
	and Excision of axillary lymph node	234262008	

TYPECH=116	Simple mastectomy related with Reconstruction of breast with immediate insertion of breast prosthesis	172043006 443611007	SNOMED CT
TYPECH=118	Simple mastectomy Related with Level 1 axillary clearance of lymph nodes	172043006 408484000	SNOMED CT
TYPECH=134	Lumpectomy of breast with Right and left	392021009 51440002	SNOMED CT
RECONS	Mammoplasty	33496007	SNOMED CT
VERIDEX	Polymerase chain reaction observation with a related observation for Secondary malignant neoplastic disease (128462008)	315010001 128462008	SNOMED CT
TOPOCH=50,1	Structure of central portion of breast	49058007	SNOMED CT
TOPOCH=50,2	Structure of upper inner quadrant of breast	77831004	SNOMED CT
TOPOCH=50,3	Structure of lower inner quadrant of breast	19100000	SNOMED CT
TOPOCH=50,4	Structure of upper outer quadrant of breast	76365002	SNOMED CT
TOPOCH=50,5	Structure of lower outer quadrant of breast	33564002	SNOMED CT

TOPOCH=50,82	6 o'clock position	260337008	SNOMED CT
TOPOCH=50,83	9 o'clock position	260343005	SNOMED CT
TOPOCH=99,5	Multifocal tumor	399506006	SNOMED CT
HISTCH= 8140,3	Carcinoma of breast	254838004	SNOMED CT
HISTCH= 8500,2	Intraductal carcinoma in situ of breast	109889007	SNOMED CT
HISTCH= 8500,3	Infiltrating duct carcinoma of breast	408643008	SNOMED CT
HISTCH= 8520,3	Infiltrating lobular carcinoma of breast	278054005	SNOMED CT
HISTCH= 8522,3	Mixed ductal and lobular carcinoma of breast	444604002	SNOMED CT
HISTCH= 0	Malignant tumor of breast	254837009	SNOMED CT
HISTCH= 99,6	Residual tumor stage R0	258254000	SNOMED CT
TAIL	Diameter of lump	248530000	SNOMED CT
SBR	Nottingham Combined Grade finding	373378009	SNOMED CT
GRAD=1	G1 grade	61026006	SNOMED CT
GRAD=2	G2 grade	1663004	SNOMED CT
GRAD=3	G3 grade	61026006	SNOMED CT
IPVN	Determination of prognosis	20481000	SNOMED CT
EMBV	Status of vascular invasion by tumor	371512006	SNOMED CT
EMBL	Status of invasión by tumor	370052007	SNOMED CT

PCINSITU	Percentage of carcinoma in situ in neoplasm	444916005	SNOMED CT
NGGSPOS	Number of lymph nodes with isolated metastatic neoplastic cells	444510006	SNOMED CT
NGGSPREL	Number of sentinel lymph nodes examined	444411008	SNOMED CT
NGGNSPOS	Number of lymph nodes involved by malignant neoplasm	443527007	SNOMED CT
NGGNSPREL	Number of lymph nodes examined	444025001	SNOMED CT
ER	Molecular genetic test with a related Entity for estrogen receptor 1	405825005 ESR1	SNOMED CT
PgR	Molecular genetic test with a related Entity for progesterone receptor	405825005 PGR	SNOMED CT
CERB	Molecular genetic test with a related Entity for v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2	405825005 ERBB2	SNOMED CT
KI67	Molecular genetic test with a related Entity for marker of proliferation Ki-67	405825005 MKI67	SNOMED CT
pT= 1b	pT1b category	443506002	SNOMED CT
pT= 1C	pT1c category	443642002	SNOMED CT

pT= 2	pT2 category	80898003	SNOMED CT
pT= 3	pT3 category	90402004	SNOMED CT
pT= is	pTis category	84921008	SNOMED CT
pN= 0	pN0 category	21917009	SNOMED CT
pN= 1	pN1 category	45552005	SNOMED CT
pN= 1mi	pN1mi category	443824009	SNOMED CT
pN= 1a	pN1a category	443716008	SNOMED CT
pN= 3	pN3 category	49182004	SNOMED CT
pN= x	pNX category	54452005	SNOMED CT
pM= 0	pM0 category	19408000	SNOMED CT
pM= x	pMx category	17076002	SNOMED CT
REFISH	Human epidermal growth factor receptor 2 gene detection by fluorescence in situ hybridization	434363004	SNOMED CT
TTTPL1	Chemotherapy Fluorouracil	367336001 387172005	SNOMED CT
TTTPL2	Chemotherapy Epirubicin	367336001 417916005	SNOMED CT
TTTPL3	Chemotherapy Cyclophosphamide	367336001 387420009	SNOMED CT
TTTPL4	Chemotherapy Docetaxel	367336001 372817009	SNOMED CT

TTTPL4	Chemotherapy	367336001	SNOMED CT
	Doxorubicin	386918005	
TTTPL4	Chemotherapy	367336001	SNOMED CT
	Paclitaxel	387374002	

B.2. Anotación de datos del Universität des Saarlandes

Tabla 20: Conceptos anotados del CD en la Universidad de Saarland

Etiqueta Original	**Comentario**	**Traducción a *Core Dataset***	**Código (CD)**	**Terminología**
public_patient_anonymised(pnr)	Patient id	---	---	---
public_patient_anonymised(sex)	Patient gender	Sex="1" → Male (248153007) Sex="2" → Female (248152002) Sex="" → Unknown (394743007)		SNOMED
public_patient_anonymised(gebdat)	Patient birthdate	---	---	---
public_f2_anonymised(gewicht)	Patient weight	Body weight measure	363808001	SNOMED
public_f6_anonymised(prim_feld)	Whether (and how much) radiotherapy was applied on the right side	Radiation oncology	419815003	SNOMED
	Site of radiotherapy	prim_feld="1" → Tumor site(371480007)		SNOMED

		prim_feld=”2 or 3” → Para-aortic lymph node group (245285004) prim_feld=”4” → Abdomen(302553009)		
public_f3a_anonymised(lokal)	Where the tumor is located	Tumor finding	395557000	SNOMED
		Kidney	64033007	
		with laterality	78615007	
		Right	24028007	
		Left	7771000	
public_f3a_anonymised	Surgery (is always Nephrectomy)	Nephrectomy	175905003	SNOMED
public_chemo7_anonymised	Whether (and how much) the patient got actynomycin	Pre-operative chemotherapy	394894008	SNOMED
		Dactinomycin	387353003	
public_f2_anonymised(vod) public_f7_anonymised(vod)	Whether and when the patient developed vod (veno occlusive disease)	Obstruction of vein	307221007	SNOMED
public_f8b_anonymised(verlauf)		Death	419620001	
		Life threatening severity	442452003	
		Active	55561003	
		Inpatient stay	308540004	

public_f9_anonymised(l_unt)	When was the last followup	Last follow-up hospital	22023-6	LOINC
public_f9_anonymised(totdat)	Date of death	Death	419620001	SNOMED
public_f9_anonymised(rezdat)	Date of relapse	Relapse	263855007	SNOMED

B.3. Anotación de datos del Maastro Clinic Lung

Tabla 21: Conceptos anotados del CD en los datos de pulmón de la Clínica Maastro

Etiqueta original	**Código (NCI)**	**Traducción a *Core Dataset***	**Código (CD)**	**Terminología**
Forced Expiratory Volume in 1 Second	C38084	Forced expired volume in 1 second	59328004	SNOMED CT
Percentage	C25613	Percentage unit	415067009	SNOMED CT
Zubrod Performance Status 0	C19998	ECOG performance status - grade 0	425389002	SNOMED CT
Zubrod Performance Status 1	C19999	ECOG performance status - grade 1	422512005	SNOMED CT
Zubrod Performance Status 2	C17846	ECOG performance status - grade 2	422894000	SNOMED CT
Zubrod Performance Status 3	C17847	ECOG performance status - grade 3	423053003	SNOMED CT
Tumor Burden	C28384	Tumor volume	258261001	SNOMED CT
Male gender	-	Male	248153007	SNOMED CT
Female gender	-	Female	248152002	SNOMED CT

CountPet5g	-	Positron emission tomography	82918005	SNOMED CT
CountPet5g, value=0	-	N0 category	62455006	SNOMED CT
CountPet5g, value=1	-	N1 category	53623008	SNOMED CT
CountPet5g, value=2	-	N2 category	46059003	SNOMED CT
CountPet5g, value=3	-	N3 category	5856006	SNOMED CT
CountPet5g, value=4	-	N4 category	22079002	SNOMED CT
Survival Time	-	Survival time	445320007	SNOMED CT
-	-	Neoplasm of lung	126713003	SNOMED CT

B.4. Anotación de datos del Maastro Clinic

Tabla 22: Conceptos anotados del CD en la clínica de Maastro

Etiqueta original	Traducción a *Core Dataset*	Código (CD)	Terminología
Levermetastasen	Secondary malignant neoplasm of liver	94381002	SNOMED CT
Hersenmetastasering	Secondary malignant neoplasm of brain	94225005	SNOMED CT
Longmetastasen	Secondary malignant neoplasm of lung	94391008	SNOMED CT
Botmetastase	Secondary malignant neoplasm of bone	94222008	SNOMED CT
cervicale wervelmetastase	Secondary malignant neoplasm of cervical vertebral column	94250008	SNOMED CT
invasief ductaal adenocarcinoom	Invasive ductal carcinoma of breast	408643008	SNOMED CT
ductaalcelcarcinoom	Intraductal carcinoma of breast	278053004	SNOMED CT

B.5. Anotación de datos del University of Oxford

Tabla 23: Conceptos anotados del CD en los datos del Hospital de Oxford

Etiqueta original	Traducción a *Core Dataset*	Código (CD)	Terminología
Discomfort in Right Breast	Chest discomfort	279084009	SNOMED CT
dull pain in right breast	Dull chest pain	3368006	SNOMED CT
Uncomfortable right breast site	Chest discomfort	279084009	SNOMED CT
dull pain breast	Dull chest pain	3368006	SNOMED CT

Sleeplessness in night	Insomnia	193462001	SNOMED CT
Dullache pain at disease site	Dull pain	83644001	SNOMED CT
Dullache in left breast	Dull chest pain	3368006	SNOMED CT
Discomfort on Breast	Chest discomfort	279084009	SNOMED CT
Bruise at site of cancer due to core biopsy	Core needle biopsy	9911007	SNOMED CT
Dullache at lump left breast	Dull pain	83644001	SNOMED CT
Dull Pain in Right hip joint	Dull pain	83644001	SNOMED CT
Bruise on right arm	Contusion	125667009	SNOMED CT
Left breast lump with overlying skin erythema	Breast lump	89164003	SNOMED CT
Lower back pain	Low back pain	279039007	SNOMED CT
Swollen Right Breast	Swelling of breast	300885006	SNOMED CT
Swollen Right Arm	Swelling of upper arm	449619004	SNOMED CT
Pain Right Arm	Pain in right arm	287046004	SNOMED CT
Hypertension	Hypertensive disorder	38341003	SNOMED CT
Pain Right upper Arm	Pain Right upper arm structure	22253000 368209003	SNOMED CT
Pain at breast lump	Pain Breast structure	22253000 76752008	SNOMED CT
Dull pain & lump Right Breast	Dull pain Breast lump	83644001 89164003	SNOMED CT

	Right	24028007	
Lump in Left breast	Breast lump	89164003	SNOMED CT
	Left	7771000	
Redness Left Breast	Erythema	247441003	SNOMED CT
	Left breast structure	80248007	
Pain in right breast	Pain of breast	53430007	SNOMED CT
	Right	24028007	
Pain in breast lump	Pain	22253000	SNOMED CT
	Breast structure	76752008	
Age	Age	102518004	SNOMED CT
Menopause	Menopause finding	276477006	SNOMED CT
TSize	Tumor size	263605001	SNOMED CT
Nstatus	Generic lymph node tumour invasion status stage	258309004	SNOMED CT
Nnumb	Number of lymph nodes involved by malignant neoplasm	443527007	SNOMED CT
Nsampled	Lymph node tissue sample	309078004	SNOMED CT
ERElisa	Enzyme-linked immunosorbent assay	76978006	SNOMED CT
EGFR	Assay technique	272392009	SNOMED CT
Grade1	Histological grading system	277457005	SNOMED CT
ReGrade	Histological grading systems	277457005	SNOMED CT
Histology	Neoplasm (morphologic abnormality)	108369006	SNOMED CT

RT	Radiation oncology AND/OR radiotherapy	108290001	SNOMED CT
Chemo	Chemotherapy (procedure)	367336001	SNOMED CT
Side	Laterality (attribute)	272741003	SNOMED CT
Operation	Excision of breast tissue	69031006	SNOMED CT
Axilla	Neoplasm of axilla (disorder)	126639006	SNOMED CT
eRFSsteep	Surviving free of recurrence of neoplastic disease	445150007	SNOMED CT
tRFSsteep	Duration of recurrence-free survival	445397003	SNOMED CT

B.6. Anotación de datos del German Breast Group

Tabla 24: Conceptos anotados del CD en los datos del GBG

Etiqueta original	Comentarios	Traducción a *Core Dataset*	Código (CD)	Terminología
Menopausal status		Finding related to menstrual cycle	373412001	SNOMED CT
Premenopausal	value=premenopausal	Premenopausal state	22636003	SNOMED CT
Postmenopausal	value=postmenopausa	Postmenopausal state	76498008	SNOMED CT
Height		Body height	50373000	SNOMED CT
Weight		Body weight	27113001	SNOMED CT
Karnofsky index, %		Karnofsky index	273546003	SNOMED CT
pT at 1st diagnosis	value=1	pT1 category	53786006	SNOMED CT
pT at 1st diagnosis	value=2	pT2 category	80898003	SNOMED CT
pT at 1st diagnosis	value=3	pT3 category	90402004	SNOMED CT
pT at 1st diagnosis	value=4	pT4 category	6123003	SNOMED CT
pN at 1st diagnosis	value=0	pN0 category	21917009	SNOMED CT
pN at 1st diagnosis	value=1	pN1 category	45552005	SNOMED CT
pN at 1st diagnosis	value=2	pN2 category	15076001	SNOMED CT
pN at 1st diagnosis	value=3	pN3 category	49182004	SNOMED CT
M at 1st diagnosis	value=0	M0 category	30893008	SNOMED CT
M at 1st diagnosis	value=1	M1 category	55440008	SNOMED CT
Grade	value=2	Tumor grade G2	1663004	SNOMED CT

Grade	value=3	Tumor grade G3	61026006	SNOMED CT
Hormone receptor status	value=positive	Hormone receptor positive tumor	417742002	SNOMED CT
Hormone receptor status	value=negative	Hormone receptor negative neoplasm	438628005	SNOMED CT
Her2	Genetic finding (106221001)	erbb2 interacting protein	HGNC:3430	HGNC
locoregional_BL		Neoplasm by body site	127331007	SNOMED CT
liver_BL		Hepatoblastoma	109843000	SNOMED CT
lung_BL		Neoplasm of lung	126713003	SNOMED CT
bone_BL		Secondary malignant neoplasm of bone	94222008	SNOMED CT
CNS_BL		Neoplasm of central nervous system	126951006	SNOMED CT
other_BL		Neoplasm and/or hamartoma	399981008	SNOMED CT
Haemoglobin, baseline		Anemia	271737000	SNOMED CT
Haemoglobin, baseline	value=1	Mild	255604002	SNOMED CT
Haemoglobin, baseline	value=2	Moderate (severity modifier)	6736007	SNOMED CT
Haemoglobin, baseline	value=3	Severe (severity modifier)	24484000	SNOMED CT

Haemoglobin, value=4 baseline	Life threatening severity	442452003	SNOMED CT
Leucocytes, baseline	White blood cell count	767002	SNOMED CT
Neutrophils, baseline	Neutrophil count	30630007	SNOMED CT
Thrombocytes, baseline	Platelet count	61928009	SNOMED CT
AP, baseline	Alkaline phosphatase measurement	88810008	SNOMED CT
SGOT, baseline	Aspartate aminotransferase measurement	45896001	SNOMED
SGPT, baseline	Alanine aminotransferase measurement	34608000	SNOMED
Bilirubin, baseline	Bilirubin measurement	302787001	SNOMED
Serum Creatinine, baseline	Creatinine measurement, serum	113075003	SNOMED
ECG	Electrocardiographic procedure	29303009	SNOMED
Echocardiography	Echocardiography	40701008	SNOMED
LVEF	LVEF - Left ventricular ejection fraction	250908004	SNOMED
Any chemotherapy	Chemotherapy	367336001	SNOMED

Chemotherapy adjuvant or neo-adjuvant	Chemotherapy	367336001	SNOMED
Chemotherapy palliative	Chemotherapy	367336001	SNOMED
Anthracycline-containing chemotherapy	Chemotherapy	367336001	SNOMED
Anthracycline-containing chemotherapy adjuvant or neo-adjuvant	Chemotherapy	367336001	SNOMED
Anthracycline-containing chemotherapy palliative	Chemotherapy	367336001	SNOMED
Taxane-containing chemotherapy	Chemotherapy	367336001	SNOMED
Taxane-containing chemotherapy adjuvant or neo-adjuvant	Chemotherapy	367336001	SNOMED
Taxane-containing chemotherapy palliative	Chemotherapy	367336001	SNOMED
Endocrine therapy	Endocrine therapy	309542002	SNOMED

Endocrine therapy adjuvant	Endocrine therapy	309542002	SNOMED
Endocrine therapy palliative	Endocrine therapy	309542002	SNOMED
Radiotherapy	Radiation oncology AND/OR radiotherapy	108290001	SNOMED
Radiotherapy adjuvant	Radiation oncology AND/OR radiotherapy	108290001	SNOMED
Radiotherapy palliative	Radiation oncology AND/OR radiotherapy	108290001	SNOMED
Trastuzumab	Trastuzumab	327397006	SNOMED
Trastuzumab adjuvant	Trastuzumab	327397006	SNOMED
Trastuzumab palliative	Trastuzumab	327397006	SNOMED
Bisphosphonate treatment	Bisphosphonates	96281001	SNOMED
Bisphosphonate treatment adjuvant	Bisphosphonates	96281001	SNOMED
Bisphosphonate treatment palliative	Bisphosphonates	96281001	SNOMED
Other treatments	Therapeutic procedure	277132007	SNOMED
Other treatments adjuvant	Therapeutic procedure	277132007	SNOMED

Other treatments palliative	Therapeutic procedure	277132007	SNOMED
Duration of trastuzumab	Trastuzumab	327397006	SNOMED
Nausea	Nausea	422587007	SNOMED
Nausea, grade 3-4	Nausea	422587007	SNOMED
Vomiting	Vomiting	422400008	SNOMED
Vomiting, grade 3-4	Vomiting	422400008	SNOMED
Diarrhea	Diarrheal disorder	128333008	SNOMED
Diarrhoea, grade 3-4	Diarrheal disorder	128333008	SNOMED
Mucositis	Mucositis	95361005	SNOMED
Mucositis, grade 3-4	Mucositis	95361005	SNOMED
Constipation	Constipation	14760008	SNOMED
Constipation, grade 3-4	Constipation	14760008	SNOMED
Other gastrointestinal disorders	Disorder of gastrointestinal tract	119292006	SNOMED
Other gastrointestinal disorders, grade 3-4	Disorder of gastrointestinal tract	119292006	SNOMED

Anorexia, loss of appetite	Anorexia	79890006	SNOMED
Anorexia, loss of appetite, grade 3-4	Anorexia	79890006	SNOMED
Allergic reactions	Allergic reaction	418925002	SNOMED
Allergic reactions, grade 3-4	Allergic reaction	418925002	SNOMED
Oedema	Oedema	423666004	SNOMED
Oedema, grade 3-4	Oedema	423666004	SNOMED
Asthenia (fatigue)	Asthenia	13791008	SNOMED
Asthenia (fatigue), grade 3-4	Asthenia	13791008	SNOMED
Alopecia	Alopecia	56317004	SNOMED
Skin changes (including HFS)	Traumatic skin changes	238535007	SNOMED
Skin changes (including HFS), grade 3-4	Traumatic skin changes	238535007	SNOMED
Hand-foot-syndrome	Hand-foot syndrome	371104006	SNOMED
handfoot_34	Hand-foot syndrome	371104006	SNOMED
Nail changes	Nail changes	416596008	SNOMED
Nail changes, grade 3-4	Nail changes	416596008	SNOMED

Sensory neuropathy	Sensory neuropathy	95662005	SNOMED
Sensory neuropathy, grade 3-4	Sensory neuropathy	95662005	SNOMED
Other neurological disorders	Neurological disorder	118940003	SNOMED
Other neurological disorders, grade 3-4	Neurological disorder	118940003	SNOMED
Pain	Pain	22253000	SNOMED
Pain, grade 3-4	Pain	22253000	SNOMED
Infection (including pneumonia)	Infectious disease	40733004	SNOMED
Infection, grade 3-4	Infectious disease	40733004	SNOMED
Fever	Fever	386661006	SNOMED
Fever, grade 3-4	Fever	386661006	SNOMED
Thromboembolic events	Thromboembolic disease	371039008	SNOMED
Thromboembolic events, grade 3-4	Thromboembolic disease	371039008	SNOMED
Dyspnoea	Dyspnoea	267036007	SNOMED
Dyspnoea, grade 3-4	Dyspnoea	267036007	SNOMED
Other respiratory or pulmonary disorder	Disorder of respiratory system	50043002	SNOMED

Other respiratory or pulmonary disorders, grade 3-4	Disorder of respiratory system	50043002	SNOMED
Cardiac events	Cardiac disorder	56265001	SNOMED
Cardiac events, grade 3-4	Cardiac disorder	56265001	SNOMED
Renal and urinary disorders	Disorder of kidney and/or ureter	443820000	SNOMED
Renal and urinary disorders, grade 3-4	Disorder of kidney and/or ureter	443820000	SNOMED
Eye disorders	Eye disorder	371405004	SNOMED
Eye disorders, grade 3-4	Eye disorder	371405004	SNOMED
Hot flushes and sweating	Hot flushes	198436008	SNOMED
Hot flushes and sweating, grade 3	Hot flushes	198436008	SNOMED
Musculosceletal disorders	Disorder of musculoskeletal system	928000	SNOMED
Musculosceletal disorders, grade 3-4	Disorder of musculoskeletal system	928000	SNOMED
Hepatobiliary disorders	Disorder of liver	235856003	SNOMED
Hepatobiliary disorders, grade 3-4	Disorder of liver	235856003	SNOMED

other	Disease	64572001	SNOMED
Other, grade 3-4	Disease	64572001	SNOMED
Febrile neutropenia	Febrile neutropenia	409089005	SNOMED
Anaemia, any grade	Anaemia	271737000	SNOMED
Anaemia, grade 3-4	Anaemia	271737000	SNOMED
Leukopenia, any grade	Leukopenia	84828003	SNOMED
Leukopenia, grade 3-4	Leukopenia	84828003	SNOMED
Neutropenia, any grade	Neutropenia	165517008	SNOMED
Neutropenia, grade 3-4	Neutropenia	165517008	SNOMED
Thrombopenia, any grade	Thrombocytopenia	302215000	SNOMED
Thrombopenia, grade 3-4	Thrombocytopenia	302215000	SNOMED
Alkiline phosphatase, any grade	Alkaline phosphatase measurement	88810008	SNOMED
Alkiline phosphatase, grade 3-4	Alkaline phosphatase measurement	88810008	SNOMED

ASAT, any grade		Aspartate aminotransferase measurement	45896001	SNOMED
ASAT, grade 3-4		Aspartate aminotransferase measurement	45896001	SNOMED
ALAT, any grade		Alanine aminotransferase measurement	34608000	SNOMED
ALAT, grade 3-4		Alanine aminotransferase measurement	34608000	SNOMED
Bilirubin, any grade		Bilirubin measurement	302787001	SNOMED
bilirubin, grade 3-4		Bilirubin measurement	302787001	SNOMED
Serum creatinine, any grade		Creatinine measurement, serum	113075003	SNOMED
Serum creatinine, grade 3-4		Creatinine measurement, serum	113075003	SNOMED
133-164 repeated concepts	May differ on the grade, the grade does not need any additional concept.			
Any toxicity		Poisoning	75478009	SNOMED
Any toxicity, grade 3-4		Poisoning	75478009	SNOMED

Any haematological toxicity		Agents affecting blood constituents, causing poisoning	90687000	SNOMED
Any haematological toxicity, grade 3-4		Agents affecting blood constituents, causing poisoning	90687000	SNOMED
Any biochemistry toxicity	No	Poisoning due to chemical substance	441952005	SNOMED
Any biochemistry toxicity, grade 3-4	No	Poisoning due to chemical substance	441952005	SNOMED
Any other toxicity	No	Poisoning	75478009	SNOMED
Any other toxicity, grade 3-4	No	Poisoning	75478009	SNOMED

ANEXO C. Informe final de los evaluadores

C.1 Resultados Proyecto INTEGRATE

1. OVERALL ASSESSMENT

a. Executive summary

Please give your overall assessment of the project, commenting on the following:
- *main scientific/technological achievements of the project*
- *quality of the results*
- *attainment of the objectives and milestones for the period*
- *adherence to the workplan, any deviations (whether justified) and remedies (whether acceptable)*
- *take-up of the recommendations from the previous review (if applicable)*
- *contribution to the state of the art*
- *use of resources*
- *impact*

The INTEGRATE project aims were to develop a proof-of-concept demonstrator of a series of tools to promote more effective enrolment into clinical trials, facilitate consensus annotation of histological slides and search for new biomarkers in existing clinical trial databases. Innovative infrastructures developed were supposed to enable large-scale data and knowledge sharing, and to foster multi-data collaboration in the biomedical research.
It was expected to bring together the heterogeneous multi-scale biomedical data generated through standard and novel technologies within post-genomic clinical trials by various research teams including clinical investigators in local hospitals, pathologists, basic researchers in molecular labs, statisticians and trial administrators. At the same time it aimed at adoption of a more personalized treatment approach allowing to match automatically breast cancer patients with various possible clinical trials according to patients clinical and biological eligibility. This would significantly reduce the physicians' burden of work required for trial selection and improve the knowledge about currently run studies, leading to an ultimate improvement of recruitment.

At the final review, the DECIMA tool for checking trial eligibility was largely complete and was the subject of interest from end users in particular from the German Breast Group. The patient recruitment tool NONA also completed its technical development and Philips expressed potential interest in integrating this tool in an internal platform. Both tools underwent successful validation by the target user groups, although with a necessarily limited sample size since the number of suitable end-users in the consortium is limited. The validation results confirmed that the functionality in both tools is appropriate and desirable, but also indicated that further development is required before the tools can be put to routine use. Thus, the first aim was substantially achieved.

The tool to share histopathology slides for consensus annotation addresses a very important issue of relevance to clinical practice. This tool was also successfully evaluated identifying the need for further incremental improvements of practical value for routine use, so a qualified level of success has been achieved, albeit not at the originally intended level of user readiness.

The analytical tool for biomarker discovery enables clinicians to quantify differences between defined samples and calculate suitable significant tests. The revised final report indicated that adjustments were made for multiple testing using standard methodologies for false detection rate correction.

INTEGRATE intended to provide tools to streamline the screening phase of breast cancer clinical trials. Before a patient is enrolled in a clinical trial, she must meet a certain number of eligibility criteria such as age, cancer type and stage, or previous or concomitant treatments.

INTEGRATE is supposed to facilitate this by managing lists of eligibility criteria for registered trials, and by automating electronic data capture and the evaluation of the criteria. It also provides interfaces to allow linking and extracting of clinical data for eligibility from electronic health records, the acquisition of molecular testing data from central laboratories, and the tracking of biological samples.

Tools for statistical and bioinformatics descriptive analyses of data were incorporated in INTEGRATE, too, and a collaborative environment was provided where researchers could share and annotate statistical models built from the data. The modular nature of the architecture will make it possible to easily plug in analytical components on top of the data querying component.

In the fourth and final year of the project a crucial objective was to complete building the tools which task was previously significantly delayed in several areas. This was successfully achieved as stated above, however little opportunities and time were left for proper and extensive tools validation. The consortium had tried to overcome these shortcomings, also using the project's extension which was granted until October 2014. Definitely the project's evaluation and validation was enhanced by the recruitment of a new partner, the German Breast Group, as actually expected during the previous review. Some plans for the project's sustainability and exploitation were developed, as advised during the previous review. However, plans for sustainability and exploitation are still largely unfulfilled.

Overall, the project overall made acceptable progress in relation to its original objectives. Final project's report, as well as 10 deliverables were submitted for this review.

b. Recommendations concerning the period under review

Please give your recommendations on the acceptance or rejection of resources, work done and required corrective actions – e.g., resubmission of reports or deliverables, further justifications, etc.

All required tool demonstrators were built and presented during the fourth final review. They included:

- Patient screening tool (Decima),
- Cohort selection tool (Nona),
- Analysis tool (Collaboratory) which included also Predictive models tool (which was supposed, at least up to my understanding, to be developed as a separate tool and which became somewhat the weakest part of the project)
- Central Pathology Review tool (CRP).

Although in general the project has achieved further progress from the last annual review, the development of final tools was delayed, which left limited time for their evaluation, improvement and validation, even though the project was extended by 9 months.
Nevertheless, the consortium addressed the recommendations made at the last review, including the greater involvement of end-users especially during final evaluation. The limited number of subjects available meant that the validation was mostly qualitative.
It is of concern, however, that former recommendations of reviewers, particularly R1, R2 and R3, were not fully addressed.
In general, several of the provided deliverables were not adequate with respect to structure, content and quality. Some of the recommendations were at least in the deliverables neglected.

Moreover, some aspects concerning sustainability, exploitation and dissemination of tools developed remained weak. This seemed to result from the fact that project has from the beginning been driven by engineering and IT companies rather than by clinicians. Overall, ultimate involvement of clinicians in the project was insufficient, despite steady improvement observed throughout the course of the project. Some of the problems resulted from the fact that functional tools requirements were not adequately specified by the project partners and in particular end-users at the very beginning. Although improved, completion of automatic execution of some processes was not fully achieved which was actually pinpointed by the internal evaluators.

A detailed pathway to exploitation remains aspirational, in part because the tools need further work to meet end-user requirements. Detailed tutorials to illustrate good practice would also need to be developed before the tools can be released for commercial or routine use. Hence, the project's exploitation remained the weakest part of the project with only partly defined interests from its partners.

The functional level of the analysis tool was not present in the deliverables, although some more advanced functionality was reported at the consensus meeting. A re-submission of the final report noting the required functionality was provided.

Each demonstrator tool must include good practice guidelines so that non-expert users can profit from their use without falling into known pitfalls of poor practice.

Nevertheless, with a generous view on the project, one could conclude that the project was successfully finished.

c. Recommendations concerning future work

Since the project has been finalized this is section is strictly speaking non-applicable.

It seems to be worthwhile, however, to issue some recommendations to assist in future commercial project's execution:

1. There needs to be yet greater emphasis on end-users needs, especially regarding project's validation.

2. A detailed pathway to exploitation needs to be still much improved, detailing potential involvement and intentions from current partners.

3. Each demonstrator tool must include good practice guidelines, a sort of cookbook, so that non-expert users can learn them easily and profit from their use.

4. Although and automatic patients recruitment and trial matching tool is an excellent idea, which definitely has its niche in the medical market, it is of concern that some aspects seemed not to be sufficiently taken into account, i.e. consenting the patients regarding the use of their electronic health records for research purposes, as well as inclusion of eCRFs (electronic Case Report Forms).

C.2 Resultados Proyecto EURECA

1. OVERALL ASSESSMENT

a. Executive summary

Please give your overall assessment of the project, commenting on the following:
- *main scientific/technological achievements of the project*
- *quality of the results*
- *attainment of the objectives and milestones for the period*
- *adherence to the workplan, any deviations (whether justified) and remedies (whether acceptable)*
- *take-up of the recommendations from the previous review (if applicable)*
- *contribution to the state of the art*
- *use of resources*
- *impact*

EURECA aims to "build an advanced, standards-based and scalable semantic integration environment enabling seamless, secure and consistent bi-directional linking of clinical research and clinical care systems" (Dow). No less than 7 services were proposed to be used to test this "environment". In a nutshell, its promise was to capitalize on previous projects and tools and re-use as much as possible all existing developments/standards in both the scientific and business community in the field of semantic interoperability.

A more secure secondary use of care data for research will be easier after this project, at least in the scenarios where the instrument has been proved.

EURECA is focused on semantic interoperability and has described specific clinical areas, starting from disease and treatment-related sets of concepts in oncology.

The overall EURECA content cover 14 webapplications, 16 webservices and 2 API. **The main development** is related to the SIL (Semantic Interoperability Layer) which uses existing standards, but also some specific ontological development. The security framework provides a technological solution that covers all identified security requirements and guarantees compliance of the complete EURECA platform to the legal framework governing the project.

As such, it presented an important number of objective challenges (and risks), among which the following ones are the most important:
- This was a huge consortium with consequently potential important coordination problems.
- As in most IT research projects, a strong involvement of the clinical partners is an absolute condition of success. For many reasons, this is usually difficult to obtain. Use cases in EURECA have been originally selected more on the basis of existing tools of the partners than on objectively identified and focused clinical objectives. The weakness of the methodology used in D.1.1 was originally perceived as an indicator that the key focus was not on users and clinicians, with direct consequences on the project outputs.
- Use of the available (and subsequent) resources can only be justified by the capacity to integrate them in real life setting environment and document the process of implementation in such a way that it can allow its generalisation.
- The clinical sites selected are not homogeneous be it at clinical, organizational or legal levels.

288048 EURECA 3rd Review

5. USE AND DISSEMINATION OF FOREGROUND

a. Impact

Is there evidence that the project has so far had, and is it likely to have, significant scientific, technical, commercial, social or environmental impact (where applicable)?

There is evidence that this project may have scientific impact on clinical and research, especially if effort to coordinate with other projects (i.e., TransForm, Ponte,…) working in the same fields are done in the near future.

However, the project results do not really go beyond the pure R&D field but at the very minimum it will have contributed to the improvement of the tools developed by the partners and should enrich the current debate on semantic interoperability strategy. Before the project formally closes, a formal description of EURECA findings related to this debate (and specially use of SNOMED-CT) should ideally still be realized.

b. Use of results

Comment on whether the plan for the use of foreground, including any updates, is still appropriate. Comment also on the plan for the exploitation and use of foreground for the consortium as a whole, or for individual beneficiaries or groups of beneficiaries, and its progress to date.

The tools developed are technically ok, but no products near to the market are still foreseen. The final exploitation plan is acceptable and in line with the project final outputs.
Best chances of (re)use are related to services which were already associated with existing described needs at the clinical site level, future EU projects or direct support to the overall investment in the field of SIL by the technological partners.

c. Dissemination

Assess whether the dissemination of project results and information (via the project website, publications, conferences, etc.) has been adequate and appropriate.

The activities developed to support dissemination have been "standard" and are thus acceptable although for a project of that size and ambition one could have expected something more original and ambitious. This is partially due to the fact that the project has not reached the TLA originally foreseen.
One would have appreciated more efforts dedicated to activities which would help people to have a better understanding of the global picture and the solutions at stake, focusing thus also on policy makers rather than mainly the R&D milieu.
The video about EuroREC is a bit too promotional, and do not provide a real idea of all the work done by EURECA.