Normas internacionales de información de HCE

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN
Instituto de Salud Carlos III

IMPaCT
Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología

Programa	IMPacT: Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología		
Nombre Proyecto	IMPaCT-Data: Programa de Ciencia de Datos de IMPaCT		
Expediente	IMP/00019		
Duración	Enero 2021 – Diciembre 2023		
Paquete Trabajo	WP4 – Datos Médicos e Imagen		
Tarea	Tarea 4.1 – Adaptación, instalación y uso de software de código abierto para la extracción de variables a partir de HCE.		
Entregable	E4.1 Normas internacionales de información de HCE. Revisión de las normas internacionales para la anotación de la información extraída como mecanismo para minimizar la generación de nuevos estándares de interoperabilidad.		
Versión	1.1.1		
Fecha Entrega	31/03/2022	***Fecha Aprobación***	17/03/2022
Responsable	IACS		
Nivel Diseminación	X	PU	Público
		CO-IMP	Confidencial, sólo participantes de los pilares de IMPaCT, incluyendo la comisión de evaluación de IMPaCT.
		CO-DATA	Confidencial, sólo participantes de IMPaCT-Data, incluyendo la comisión de evaluación de IMPaCT.

Autores		
Organización	**Nombre**	**Rol**
IACS	Javier Gómez-Arrue Azpiazu	Autor
IACS	Carlos Tellería Orriols	Autor
HCB	Xavier Pastor Duran	Autor
HCB	Santiago Andrés Frid	Autor
ISCIII-UITeS	Adolfo Muñoz Carrero	Autor
IMIM	Miguel Angel Mayer	Autor
H12O	Miguel Pedrera Jiménez	Autor
H12O	Pablo Serrano Balazote	Autor
FISABIO-UMIB	María de la Iglesia Vayá	Autor
FISABIO-UMIB	Silvia Nadal Almela	Autor
SAS-HUVR	Carlos Luis Parra-Calderón	Autor
SAS-HUVR	Sara Gonzalez García	Autor
Navarrabiomed	Javier Gorricho	Revisor
FPS	Joaquin Dopazo	Revisor

Historial de versiones			
Nro.	**Fecha**	**Descripción**	**Autor**
v 0.1	10/01/2022	Borrador Índice	J Gómez-Arrue (IACS) C Tellería (IACS)
v 0.2	14/02/2022	Índice revisado	J Dopazo (FPS) J. Gorricho (Navarrabiomed)
v 0.3	28/02/2022	Borrador 1	Todos los autores
v 1.0	17/03/2022	Versión para enviar a coordinación	Todos los autores
v 1.1	17/05/2023	Cambio visibilidad a público y aprobado	Comité Dirección
v 1.1.1	14/06/2023	Cambio de formato para publicar en la web de IMPaCT	David Velasco (ISCIII)

Contenido

Figuras

Resumen Ejecutivo

Este documento trata de recopilar los distintos estándares para la descripción normalizada de datos de salud para su uso secundario, tanto a nivel de mapeo conceptual de los datos, como de los metadatos y modelos de información, de manera que puedan ser interoperables, y por tanto que permitan la realización fácil y coherente de proyectos de investigación utilizando conjuntos de datos procedentes de múltiples fuentes de información. El documento no pretende ser una recopilación exhaustiva de todas las normas y estándares existentes en el ámbito biomédico, lo que sería una tarea realmente titánica. El presente documento se focaliza en aquellos estándares y normas que son de uso generalizado en el contexto de los sistemas de información asistenciales y en la investigación biomédica en Europa, y más particularmente en España, como punto de partida para una propuesta de estándares, normas, herramientas y buenas prácticas para el desarrollo de una infraestructura de investigación en salud, en el marco del proyecto de Innovación en Medicina de Precisión a través de la Ciencia de Datos, IMPaCT.

Se complementa el documento con una referencia a los principios FAIR, que deben estar en la base de cualquier desarrollo de interoperabilidad de datos para la investigación, y con una breve exposición de los principios y técnicas básicas para garantizar el cumplimiento de los criterios éticos y legales en el uso compartido de datos de salud.

Introducción

Audiencia

Este documento está destinado a todos los participantes del proyecto IMPaCT.Data, como punto de referencia para el conocimiento y selección de las tecnologías de interoperabilidad más adecuadas a utilizar en las distintas fases del paquete de trabajo 4, así como en la integración de dato clínico con dato genómico (paquete 5), y su utilización en los casos de uso globales del proyecto (paquete 6).

Ámbito

El presente documento se empleará como referencia para la selección y uso de distintas normas y tecnologías de interoperabilidad en los demostradores a desarrollar en el paquete 4, así como en los procesos de integración global del paquete 5, y en los casos de uso propuestos por el paquete 6.

Relación con otros Entregables

Este entregable guarda relación con el entregable 4.4 relativo a normas de anotación en imagen médica, ya que muchos de los conceptos desarrollados en ambos documentos son similares o directamente coincidentes. El presente entregable guarda también relación con el entregable E5.1.Técnicas de Integración de Datos Biomédicos, con el E4.2. Comparación de Técnicas de Gestión de Información de HCE, y con el E6.4. Aspectos de Seguridad en el Manejo de Datos Sensibles, todos ellos previstos para el mes 18.

Estructura Entregable

El documento se abre con una sección en la que se motiva la necesidad del documento. En la sección 22, se introducen a modo de glosario los conceptos principales que son objeto del análisis realizado en el documento, estableciendo las diferencias entre normas de metadatación, estándares terminológicos y de clasificación, modelos comunes de datos, y estándares de interoperabilidad.

A continuación, en secciones diferenciadas, se van desglosando y analizando los distintos estándares y normas estudiados para cada una de las categorías:

1. Normas de metadatación
2. Estándares de terminológicos y de clasificación
3. Modelos comunes de datos
4. Estándares de interoperabilidad

La sección 7 se dedica a proponer un marco de implementación de los principios FAIR basado en estándares, que pueda ser de aplicación en las subsiguientes fases del paquete de trabajo en el contexto de IMPaCT-Data.

La sección 8 enumera brevemente aspectos relacionados con la anonimización / pseudonimización de la información, y otros aspectos éticos y legales que deben estar presentes en todos los procesos de análisis masivo de dato sanitario, y de forma especial cuando estos datos son compartidos o intercambiados entre distintos nodos de una red de investigación.

En la sección 9 se han recogido algunas implementaciones de referencia y casos de éxito de integración de datos clínicos para investigación, tanto en España como en el ámbito internacional, que pueden utilizarse como modelo para el futuro desarrollo de la integración de datos clínicos en IMPaCT-Data

Por último, se exponen las conclusiones finales del análisis desarrollado para la elaboración del presente documento, así como recomendaciones para las siguientes fases del proyecto.

1. Motivación

El proyecto IMPaCT-Data tiene como principal objetivo el desarrollo inicial de la Infraestructura y los protocolos necesarios para coordinar, integrar, gestionar y analizar datos clínicos, de imagen médica y genómica, con el objetivo de proporcionar herramientas validadas para facilitar la implementación, eficaz y coordinada, de la Medicina Personalizada en los Sistemas de Atención Sanitaria, siempre a través del marco legislativo vigente y de las estructuras de alto nivel establecidas al efecto. (Plan estratégico de IMPaCT-Data).

Dentro de ese objetivo general, la tarea 4.1 tiene como objetivo específico la adaptación, instalación y uso de software de código abierto para la extracción de variables a partir del HCE, y como primera tarea dentro de ello, la identificación de aquellas Normas Internacionales de Anotación de Información clínica que puedan ser útiles para la interoperabilidad sintáctica, y sobre todo semántica de información clínica procedente de diversas fuentes de información primaria, y que de forma conjunta deban ser procesadas y analizadas con fines de investigación (objetivos técnicos 1 y 5 del Plan estratégico de IMPaCT-Data).

El presente documento es el resultado del estudio y análisis realizado sobre los estándares y normas de metadatación, ontologías y terminologías, estándares de interoperabilidad y herramientas útiles para el abordaje de estas tareas en el contexto actual del uso secundario del dato sanitario de vida real para la investigación biomédica. En el mismo, se enumeran, explican, detallan y comparan los distintos elementos, y dicho análisis será el punto de partida para la propuesta de normas, ontologías y herramientas a utilizar en los demostradores y casos de uso a desarrollar dentro del alcance de IMPaCT-Data.

A lo largo de las últimas décadas, se han venido desarrollando múltiples iniciativas tecnológicas y organizativas orientadas al intercambio e interoperabilidad de datos médicos. En un primer momento, las iniciativas, herramientas y estándares propuestos iban especialmente orientados a la interoperabilidad entre sistemas de información dentro de una misma institución. Así, por ejemplo, se buscaba integrar los sistemas de admisión de pacientes con los sistemas de gestión de servicios centrales hospitalarios, tales como laboratorio, farmacia o diagnóstico por imagen. En paralelo, se fueron desarrollando normas de clasificación de información clínica y ontologías clínicas que permitían mejorar la calidad de la documentación clínica, a la vez que facilitar la computabilidad de la actividad y de los resultados clínicos.

A medida que las redes de comunicaciones fueron permitiendo la integración de los sistemas de información entre distintos nodos, y que el desarrollo de los sistemas de Historia Clínica Electrónica pusieron sobre la mesa la necesidad del acceso remoto a la información clínica de los pacientes, los estándares y normas de interoperabilidad se fueron haciendo más imprescindibles en el contexto del uso primario de la información clínica, y se hizo necesaria la evolución de los mismos, y el desarrollo de nuevas herramientas que facilitaran esa labor.

Más recientemente, con el desarrollo de las técnicas de análisis masivo de datos (Big Data) y de las técnicas de aprendizaje automático, ha empezado a tomar cuerpo el concepto de "uso secundario de los datos de salud", entendiendo como tal el conjunto de técnicas y desarrollos conducentes a la extracción de los datos de salud recogidos en la práctica médica habitual, y su uso -con las debidas medidas de seguridad y privacidad- para una finalidad distinta a la atención médica de los pacientes: la investigación, la adopción de políticas sanitarias, la regulación farmacoterapéutica o la monitorización de la salud pública, entre otras. Estos procesos exigen la extracción de la información clínica desde los sistemas primarios, su preprocesamiento y adaptación a las necesidades específicas del uso que se le va a dar a esa información. En este contexto, aspectos como la calidad de la información recogida -información que, no olvidemos, no se ha recogido para su uso directo en investigación o análisis, sino para la asistencia clínica del paciente-, la privacidad de los pacientes, o el correcto mapeo conceptual -los datos son datos, pero expresan conceptos que deben ser correctamente interpretados cuando esos datos se sacan de su contexto inicial-, cobran especial relevancia.

Sólo en época muy reciente se ha empezado a plantear y trabajar seriamente en entornos de investigación con dato sanitario de vida real multinodo, o incluso infraestructuras federadas, en las que distintos nodos realizan análisis equivalentes contra subconjuntos distintos de datos, para luego inferir conclusiones a partir de la agregación de resultados parciales, o bien agregando desde el inicio los datos parciales para su análisis conjunto en un nodo central. En cualquiera de estos escenarios, la utilización de modelos comunes de datos, y el correcto mapeo conceptual entre los datos de origen (que pueden tener muy diversa procedencia) y los datos analizados es una exigencia incuestionable.

Para conseguir esa interoperabilidad sintáctica y semántica se han propuesto a lo largo de los años distintas normas, modelos y herramientas, algunas estandarizadas formalmente, otras "de facto". Algunas de ellas proceden de los distintos proyectos de interoperabilidad para el uso primario de la información clínica (Historia Clínica Electrónica interoperable), y han sido reutilizados y adaptados al uso secundario, con más o menos éxito. Otras normas y modelos se han desarrollado específicamente para resolver casos de uso no asistenciales, como pueden ser ensayos clínicos, ensayos pragmáticos o estudios post-autorización. Estos casos de uso de investigación, aun no siendo estrictamente "uso secundario de dato de salud", están mucho más cerca de los casos de uso planteables en el contexto de IMPaCT-Data, y su adaptación será mucho más sencilla, o incluso inmediata.

A la hora de plantearse el desarrollo de una infraestructura para la integración y uso compartido de información de salud -clínica, genómica, de imagen-, con toda seguridad no tiene sentido proponer un esquema de normalización e interoperabilidad nuevo, existiendo numerosos estándares y modelos que pueden resolver en gran medida esas necesidades.

El presente documento es un compendio de los análisis realizados sobre un gran número de estándares, normas y modelos existentes en el ámbito de la información clínica en sus distintos usos -primario, secundario, investigación clínica...-, y su análisis como posible

propuesta para la infraestructura IMPaCT-Data. En ese análisis se han evaluado aspectos como la cobertura de los estándares y normas -casos de uso que pueden ser realizados con los mismos, y cantidad de información que puede ser normalizada con los mismos-, facilidad de implementación o adaptación, posibilidad de integrar información no estructurada como imágenes o texto en lenguaje natural, así como las organizaciones que las usan o que las soportan y mantienen.

En el alcance de este documento se han tenido en cuenta los aspectos relacionados con la interoperabilidad sintáctica y semántica. Otros niveles de interoperabilidad, como los técnicos, legales y organizativos, quedan fuera del alcance de este documento, pero deben ser tenidos en cuenta en el contexto del paquete 4 y de IMPaCT-Data en conjunto.

2. Conceptos previos

2.1. Principios FAIR en datos médicos

La iniciativa IMPaCT en su eje estratégico 2, desarrolla su estrategia de Ciencia de Datos. En dicho eje pretende desarrollar y validar un entorno de integración y análisis conjunto de datos, para el uso secundario de los datos clínicos, moleculares y genéticos. Y como objetivo técnico 3, se pretende el "Desarrollo e implementación de protocolos, métodos y sistemas integrados de análisis de datos basados en la infraestructura de bases de datos, métodos, sistemas de evaluación y mecanismos de FAIRificación".

En el marco de la Ciencia Abierta, los principios FAIR[1], formalmente publicados en 2016 por la comunidad Force11 [2] , proveen guías para mejorar los datos de modo que sean Encontrables, Accesibles, Interoperables y Reutilizables, por sus siglas en inglés (Findable, Accessible, Interoperable and Reusable), con el objetivo de garantizar que los datos puedan localizarse, accederse, interoperar y reutilizarse por parte de personas y máquinas. Aunque FAIR surgió de un taller para la comunidad de las ciencias de la vida, estos principios fueron diseñados para ser aplicados a datos y metadatos de forma transversal a todas las disciplinas científicas. A continuación, se incluye un resumen de cada dimensión:

- Datos encontrables: Los datos deben describirse con metadatos ricos y relacionados entre sí. Los (meta)datos deben tener un identificador único y estar registrados o indexados en un recurso que permita realizar búsquedas.
- Datos accesibles: Los (meta)datos deben poder recuperarse utilizando su identificador, utilizando un protocolo de comunicación abierto, gratuito y de aplicación universal, que permita la autenticación y la autorización si es necesario. Los metadatos deben ser accesibles incluso cuando los datos ya no estén disponibles.
- Interoperabilidad de los datos: Los (meta)datos deben utilizar un lenguaje específico para la representación del conocimiento, y deben utilizar vocabularios conformes a los principios FAIR, incluyendo referencias a otros (meta)datos.
- Reutilización de los datos: Los (meta)datos deben estar descritos con atributos, incluyendo información de licencia y procedencia entre otros.

Las directrices de gestión de datos de la Comisión Europea se actualizaron en 2017 para introducir la noción de FAIR. Dichos principios FAIR han sido asumidos por la iniciativa de la Comisión Europea en el marco del Mercado Digital Único: la European Open Science Cloud (EOSC) [3], y su reciente Strategic Research and Innovation Agenda [4] .

Es fundamental el informe publicado por la Unión Europea (UE) sobre los costes de NO tener datos de investigación FAIR [5] , cuyas principales conclusiones son que, para la UE: i) el coste de NO tener datos FAIR es de aproximadamente 10.200 millones de euros al año, ii) el impacto en innovación del uso de los principios FAIR podría añadir otros 16.000 millones de

euros a la economía de datos abiertos; todo ello generaría un total de al menos 26.200 millones de euros al año.

En un contexto sanitario, el uso de los principios FAIR está en aumento, hecho demostrable teniendo en cuenta la multitud de iniciativas que se pueden localizar tanto del dominio académico como de investigación [6-12]. En el marco del proyecto FAIR4Health, tras analizar diferentes perspectivas (técnica, ética, de seguridad, jurídica, cultural, conductual y económica), y basándose en el flujo de FAIRificación de GO FAIR [13], se definió el flujo de FAIRificación para datos sanitarios [14] .

La adopción de las prácticas FAIR se están extendiendo entre un amplio sector de interesados bajo un enfoque de ciencia abierta. En el caso de los datos biomédicos, al igual que en otros dominios de la ciencia, este proceso de adopción debe tener en cuenta numerosas consideraciones relativas a la utilización de los recursos específicos del propio dominio y de la infraestructura disponible. Estas consideraciones deben hacerse para cada uno de los Principios Rectores FAIR y deben incluir objetivos supra-dominiales como la máxima reutilización de los recursos existentes (es decir, minimizar la reinvención de la rueda) o la máxima interoperabilidad con los datos y servicios FAIR existentes. Muchas decisiones son las mismas que en otras comunidades científicas y éstas puedan acelerar el propio proceso de adopción de FAIR reutilizando juiciosamente las decisiones de implementación ya tomadas por otros. Para aprovechar estas redundancias y acelerar la convergencia en la reutilización generalizada de las implementaciones FAIR, en la anotación de datos de la HCE en IMPaCT data proponemos tener en cuenta el concepto de Perfil de Implementación FAIR (FIP) que capta un amplio conjunto de opciones de implementación realizadas por las comunidades de práctica individuales [15].[15]

En todo caso, el plan estratégico de IMPaCT Data establece el siguiente compromiso respecto al cumplimiento de los principios FAIR en la acción nº 2 de la LET1 de Integridad científica, en la que se debe establecer los procedimientos para garantizar los principios FAIR en la organización y gestión de la información en IMPaCT. Esto se materializará en un entregable E1 “Sistema de Evaluación de principios FAIR”.

La propuesta de anotación de información de la HCE que proponemos aquí debe permitir un nivel profundo de cumplimiento de los principios FAIR en IMPaCT Data.

2.2. Normas de metadatación

Al abordar proyectos de intercambio y análisis de datos de salud, hay una serie de pasos que debe realizar el usuario -investigador- antes de poder trabajar directamente con los datos. El primero de esos pasos es el descubrimiento de los datos o data discovery, consistente en interrogar a los distintos nodos o fuentes de datos acerca de la disponibilidad de los datos necesarios para hacer el estudio pretendido. Se podría realizar este paso consultando directamente el conjunto completo de datos, pero esto no es siempre posible, y casi siempre es ineficiente. Resulta mucho más adecuado poder hacer la consulta sobre un conjunto de

metadatos que describa con suficiente nivel de detalle el tipo, contenido y nivel de detalle de los datos recogidos en un conjunto de datos determinado. Pero para que esta tarea pueda realizarse adecuadamente, y de una forma semi-automatizada, es necesario que los metadatos consultados se ajusten a determinada normalización.

Entendemos, por tanto, como normas de metadatación aquellos estándares, normas, buenas prácticas, herramientas e interfaces que establezcan la forma en la que los conjuntos y fuentes de datos que se integran en un espacio de datos deben describir el detalle de sus conjuntos de datos, y permiten el acceso y la consulta remota de esta información. Aquí encontraremos normas que establezcan la denominación, y tipo de dato de cada uno de los metadatos que describen un conjunto de datos, su obligatoriedad y opcionalidad, las distintas ontologías que sustentan la organización de los metadatos, los estándares de publicación de dichos metadatos, y las herramientas que permiten la publicación de catálogos de datos y su consulta semi-automatizada.

También incluimos en este apartado aquellas normas y herramientas que permiten la definición de estructuras de datos ad hoc para un determinado proyecto de investigación con dato de salud para su uso compartido por distintos nodos y/o proveedores de datos, de manera que hagan factible la agregación de dichos datos, o su análisis distribuido homogéneo (mismo código analizando datos homólogos en distintos nodos).

2.3. Estándares terminológicos y de clasificación

El vocabulario propio de las ciencias de la salud se caracteriza por su riqueza, ambigüedad, dependencia contextual y uso de jerga, acrónimos y términos altamente especializados. Estos atributos determinan la necesidad de elaborar estrategias de representación del conocimiento que optimicen el rendimiento de los sistemas de información en salud, tanto para fines asistenciales como estadísticos y de investigación. La codificación emerge entonces como una solución para representar el significado de estos conceptos del dominio de la salud, y que los sistemas puedan “comprender” el lenguaje. Esta vinculación semántica se realiza a dos niveles:

- Enlace semántico, para representar el significado de los componentes de los modelos de información. Por ejemplo, el concepto “Diagnóstico principal” o “Test de antígenos para COVID-19”.
- Enlace de valores, para representar el conjunto de valores que un concepto codificable puede tomar. Por ejemplo, “Diabetes Mellitus Tipo 2” o “Resultado positivo”.

Dentro de los recursos de información diseñados para este propósito, podemos distinguir entre terminologías y clasificaciones.

Las terminologías son conjuntos limitados de términos que tienen por objetivo representar, de manera granular, conceptos del dominio para así desambiguar la información transmitida en un ámbito particular. Una vez que una terminología de un dominio específico es revisada,

depurada y validada por alguna organización, se convierte en una nomenclatura. Ejemplos de terminologías ampliamente utilizadas en el dominio sanitario son SNOMED CT y LOINC.

Por su parte, las clasificaciones son sistemas ordenados y agrupados de conceptos pertenecientes a un dominio y diseñadas con un fin particular. Se conforman por listas de categorías, cada una de las cuales alberga conceptos clínicos vinculados por alguna característica común. Ejemplos de clasificaciones ampliamente empleadas en el dominio de la salud son las propuestas por la Organización Mundial de la Salud (OMS) como las Clasificaciones Internacionales de Enfermedades (CIE) o la Clasificación Internacional de Atención Primaria (CIAP-2).

En este aspecto, es importante diferenciar las terminologías de las clasificaciones. Una clasificación, al constituir una organización de conceptos clínicos en clases agrupadas para propósitos concretos de explotación, no permite representar de manera granular la información. Es por ello que utilizar clasificaciones de manera nativa en los sistemas de registro reduce la calidad y utilidad de los datos, tanto en asistencia como para usos secundarios. En cambio, una terminología sí pretende representar, de manera fina estos los conceptos que aplican al dominio clínico, por lo que son el recurso ídoneo del que partir y posteriormente mapear a clasificaciones si algún caso de uso lo requiere.

A pesar de ello, hay diversos motivos por los cuales las clasificaciones tienen una mayor adopción a nivel asistencial que las terminologías. En primer lugar, los organismos estatales encargados de gestionar información en salud a nivel poblacional utilizan clasificaciones construidas para estos fines (como la CIE). Por ello, los efectores en salud habitualmente se ven obligados a reportar información utilizándolas, de lo cual suele depender también la contraprestación que reciben, como ocurre con los conjuntos mínimos básicos de datos (CMBD).

Otro punto a tener en cuenta es la facilidad en la adopción e implantación. Aunque se pierda granularidad o no se permita representar correctamente el concepto clínico que el profesional de la salud desea en el punto de cuidado, las clasificaciones son más sencillas y permiten una codificación primaria por el usuario sin mucha dificultad.

Debe también tenerse en cuenta que clasificaciones como las desarrolladas por la OMS pueden utilizarse de manera gratuita, mientras que algunas terminologías (por ejemplo, SNOMED-CT) requieren el pago de membresías ya sea a nivel estatal o a nivel institucional.

2.4. Modelos comunes de datos

Los modelos de datos en salud nos permiten representar y persistir, de manera estructurada, información relativa a conceptos sanitarios y su contexto. Existen multitud de modelos de datos, algunos considerados estándares y otros, casi tantos como casos de uso, propietarios. Por ello, es útil clasificarlos, atendiendo a criterios de interés, para comprender cómo y cuándo utilizar cada uno de ellos. No es extraño que se atribuyan errores de diseño a un

modelo de datos que vienen originados de un uso para el cual no ha sido diseñado. En este análisis se toman como criterios de clasificación su arquitectura y su usabilidad.

Al hablar de arquitectura, distinguimos dos tipos de modelos:

- Modelo de arquitectura no-dual: no existe distinción entre el modelo de datos y el modelo conceptual. La semántica del dato va implícita, por diseño, en la propia estructura del modelo (funciona "por columnas"). Este tipo de modelos son de fácil diseño, pero pierden toda escalabilidad: para añadir datos relativos a un nuevo concepto es necesario modificar la arquitectura del modelo.

- Modelos de arquitectura dual: modelo de datos y modelo conceptual son independientes. La persistencia de los datos sigue una estructura clave – valor, de manera que, para cada concepto, se define su significado explícitamente en un campo del registro. Este tipo de modelos requieren un diseño más avanzado que el no-dual, pero son mucho más escalables y sostenibles: para añadir datos relativos a nuevos conceptos no es necesario modificar la estructura del modelo de datos, únicamente definirlo en su modelo conceptual.

Del mismo modo, en cuanto a la usabilidad, se distinguen dos tipos de modelos:

- Modelos de uso primario: los datos se utilizan en la atención sanitaria individual del paciente que los ha generado. Esto implica que el modelo debe incluir una serie de información de contexto que cumpla con los requisitos ético-legales establecidos por la organización y su entorno para la asistencia sanitaria.

- Modelos de uso secundario: los datos se utilizan en actividades que difieren de la asistencia individual del paciente que los ha generado. Dentro de estos no primarios, o secundarios, se incluyen actividades como la investigación sanitaria, la evaluación de resultados en salud, auditorias, contabilidad analítica o facturación.

Así, atendiendo a estos dos criterios, es posible clasificar distintos tipos de sistemas de información en función del modelo de datos que implementan:

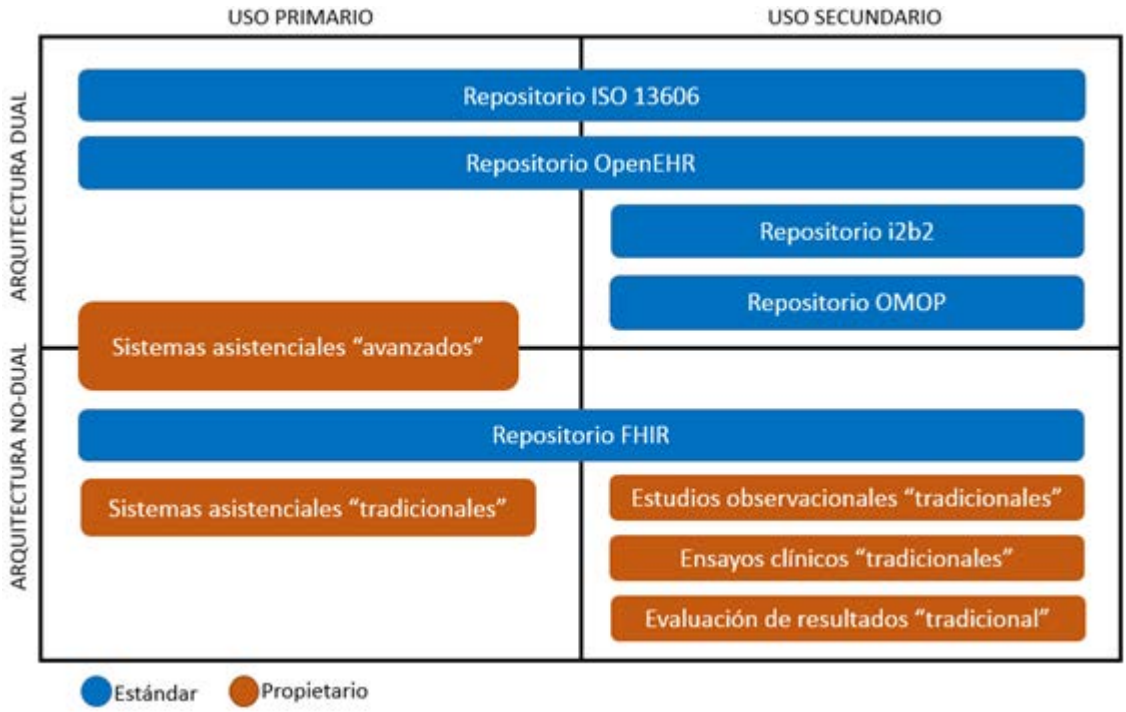

Ilustración 1 Modelos de datos y estándares de interoperabilidad

Fuente: Hospital Universitario 12 de Octubre

2.5. Estándares de interoperabilidad

La interoperabilidad se puede definir como la capacidad entre diferentes sistemas para procesar, de manera automática, la información intercambiada entre ellos, o combinada a partir de los mismos. Se pueden definir cuatro niveles de interoperabilidad de la información:

- Técnica o fundamental (nivel 1): Establece los requisitos de interconectividad necesarios para que un sistema o aplicación comunique datos a otro y los reciba de forma segura.
- Sintáctica o estructural (nivel 2): Define el formato, la sintaxis y la organización del intercambio de datos, incluso a nivel de campo de datos para su interpretación.
- Semántica (nivel 3): Proporciona modelos subyacentes comunes y codificación de los datos, incluyendo el uso de elementos de datos con definiciones estandarizadas de conjuntos de valores y vocabularios de codificación disponibles públicamente, proporcionando una comprensión y un significado compartidos para el usuario.
- Organizativo (nivel 4): Incluye consideraciones de gobernanza, políticas, sociales, legales y organizativas para facilitar la comunicación y el uso seguros, fluidos y

oportunos de los datos tanto dentro como entre organizaciones, entidades y personas. Estos componentes permiten el consentimiento compartido, la confianza y los procesos y flujos de trabajo integrados de los usuarios finales.

Los estándares de interoperabilidad en salud definen mecanismos comunes para organizar y acceder a los datos y compartirlos de forma adecuada y segura en todo el espectro de uso, en todos los entornos aplicables y con las partes interesadas pertinentes, incluida la persona. Así, aplicar un estándar de interoperabilidad de información implica la realización de interfaces, donde se mapea la información desde la estructura de datos local de un determinado sistema hacia una estructura estándar (y viceversa). Así, todas las aplicaciones que mapeen sus estructuras locales a la estándar podrán comunicarse de manera exitosa entre sí y con el resto, limitando la cantidad de interfaces necesarias.

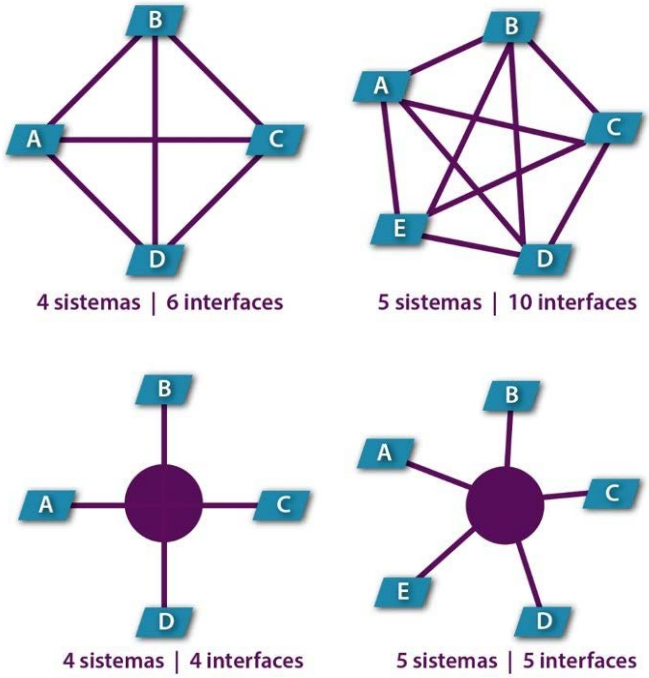

Ilustración 2 Esquemas de interoperabilidad con y sin modelos comunes

Existen múltiples iniciativas de modelos conceptuales, algunas de las cuales se describen en detalle en la sección 5 de este documento.

Asimismo, no debemos perder de vista la necesidad de cumplimiento del principio de interoperabilidad de acuerdo a lo que los principios FAIR nos indican, a saber:

I1. Los metadatos y los datos deben utilizan un lenguaje formal, accesible, compartido y ampliamente aplicable para la representación del conocimiento.

I2. Los metadatos y los datos utilizan vocabularios que siguen los principios FAIR

I3. Los metadatos y los datos incluyen referencias cualificadas a otros metadatos y otros datos.

La iniciativa más importante que se está desarrollando en este sentido es la Guía de Implementación FHIR de HL7 para el cumplimiento de los principios FAIR, fundamental para tener en cuenta en IMPaCT Data.

2.6. Calidad de los datos

La evaluación de la calidad de los datos de la es fundamental de cara a construir procesos de obtención de datos útiles en investigación y otros fines secundarios. Para ello, en primer lugar, se hace necesario definir qué se entiende por 'calidad del dato' y por qué indicadores está compuesto este término. Así, estudios previos han realizado una revisión de la literatura en la que se discute la metodología de evaluación de la calidad de los datos de la HCE, identificando siete dimensiones a evaluar:

- **Unicidad**. Esta dimensión mide la existencia de registros replicados que representen una misma entidad. Responde a la pregunta "¿Existen datos repetidos?"
- **Completitud**. Esta dimensión mide el grado de registro de los datos en la HCE. Responde a la pregunta "¿Faltan datos?"
- **Consistencia**. Esta dimensión mide si un elemento registrado en la HCE cumple los requisitos establecidos sobre el concepto al que hace referencia. Responde a la pregunta "¿Cumplen los datos las reglas establecidas?"
- **Corrección**. Esta dimensión mide si un elemento registrado en la HCE es cierto. Responde a la pregunta "¿Existen datos aparentemente anómalos?"
- **Concurrencia**. Esta dimensión mide si un elemento es registrado en la HCE en un periodo de tiempo válido para su utilización. Responde a la pregunta "¿En qué intervalo temporal dispongo del dato?".
- **Estabilidad temporal**. Esta dimensión mide la variabilidad esperada del dato a lo largo del tiempo. Responde a la pregunta "¿Existe variabilidad de los datos a lo largo del tiempo de concurrencia?"
- **Concordancia**. Esta dimensión mide si un elemento registrado en la HCE es coherente con otros elementos y fuentes de datos. Responde a la pregunta "¿Tienen variabilidad los datos dependiendo de su origen?"

Así mismo, se identificaron el conjunto de métodos para evaluar las dimensiones de calidad de datos anteriormente descritas:

- **Comparación con gold standard**. Un conjunto de datos extraídos de otras fuentes se utiliza como referencia para ser comparados con los de la HCE. Es útil para medir la unicidad, completitud y corrección de los datos.
- **Comparación de elementos de datos de la HCE**. Se comparan dos o más elementos dentro de una HCE para ver si reportan la misma información o información compatible. Es útil para medir la completitud, corrección y concordancia de los datos.
- **Comparación de distribuciones estadísticas**: se hace uso de estadísticas resumidas de datos agregados de HCE que se comparan con las distribuciones esperadas para los conceptos clínicos de interés. Es útil para medir la completitud, consistencia, estabilidad temporal, concordancia de los datos y el valor predictivo de los mismos.
- **Verificación de restricciones**: los datos de la HCE se evalúan para verificar si cumplen una serie de restricciones establecidas. Es útil para medir la corrección y consistencia de los datos.
- **Revisión de metadatos de registro**: se examinan los metadatos sobre las prácticas reales de entrada de datos (por ejemplo, fechas, horas, ediciones). Es útil para medir la concurrencia de los datos.

3. Normas de metadatación

Para un efectivo cumplimiento de una parte importante del objetivo técnico 3 del plan estratégico de IMPaCT Data, sobre el desarrollo e implementación de mecanismos de FAIRificación, se requiere una propuesta de normas de metadatación orientada a dicho cumplimiento.

En este sentido, debemos tener en cuenta que dichas normas den respuesta a los propios principios FAIR respecto a lo que dichos metadatos deben ofrecer.

El diseño y realización de proyectos de investigación con datos en entornos colaborativos y distribuidos, en los que se utilizan datos procedentes de distintas fuentes no homogéneas, suele verse frecuentemente limitado por una falta de información adecuada y completa que describa los conjuntos de datos con los que se va a trabajar.

En el proceso de ejecución de un proyecto de investigación con datos, hay dos fases en las que el investigador requiere de una manera especial una adecuada descripción de la estructura y contenido de los conjuntos de datos con los que va a trabajar. La primera fase es la fase de búsqueda de información. Es la fase de diseño de un proyecto en la que el investigador necesita conocer qué fuentes de datos, de las disponibles en la red, contienen datos que encajen con la definición de caso planteada, o pueden ser elementos a analizar para responder a las preguntas científicas planteadas.

La segunda fase en la que es imprescindible una correcta y completa descripción de los datos a utilizar, es la descripción detallada de los conjuntos de datos que van a ser intercambiados, agregados, o analizados de forma distribuida, pero respondiendo a una misma estructura de información. Para esta segunda fase, la utilización de modelos comunes de datos (CDM, véase sección 5) facilita considerablemente este trabajo. Pero los modelos comunes de datos no siempre tienen capacidad de especificar el detalle semántico necesario, limitándose a una definición sintáctica y estructural de los modelos de datos. El detalle semántico debe hacerse, con mucha frecuencia, apoyándose en terminologías y ontologías de dominio (véase sección 4). En este contexto, disponer de normas claras para la elaboración de los metadatos que describen los conjuntos de datos, integrando la información sintáctica (modelos de datos) y semántica (ontologías), y proporcionando interfaces de acceso que permitan un acceso automatizado a los metadatos, resulta imprescindible para lograr la interoperabilidad de los datos.

Las herramientas / softwares implementadas para la anotación de metadatos deben contar con una serie de requisitos, que a continuación se detallan, con el fin de afrontar los retos que se presentan en el dominio clínico: (i) Código abierto y accesible, el software debe estar disponible bajo una licencia de código abierto (*open source*) o al menos ser de libre acceso y generalista, que permita su adaptación al caso de uso concreto; (ii) Compatibilidad con formatos de datos comunes, capacidad de procesamiento de diferentes formatos de intercambio de datos existentes, especificados y estandarizados (ej.; HL7 FHIR y CDISC ODM); (iii) Cumplimiento de los principios FAIR para la integración de servicios

terminológicos, los vocabularios y terminologías utilizadas para la anotación también deben ser FAIR (como se indica en el principio de interoperabilidad FAIR): (iv) Sugerencia de anotaciones y búsqueda de terminología, se persigue una funcionalidad simple, siendo preferible rechazar las anotaciones creadas de forma autónoma que buscar manualmente el concepto adecuado; (v) Soporte de integración de ontologías/terminologías personalizadas o configuración de conexión a un servicio de terminología/ontología (ej.; BioPortal).

A continuación, se detallan algunas de las herramientas actualmente disponibles para llevar a cabo la tarea de anotación de metadatos semánticos, las cuales son candidatas a utilizarse en el marco de este proyecto [16]:

- **ODMedit [17]**: aplicación web de acceso gratuito para crear modelos de datos con anotaciones semánticas uniformes. Utiliza un repositorio público de metadatos, para alcanzar la uniformidad en la codificación, desde el cual se sugieren anotaciones que pueden ser seleccionados por los expertos (enfoque semi automatizado); así como las terminologías UMLS (*Unified Modeling Language System*) y SNOMED CT, ofreciendo al usuario los enlaces a los sitios web del metatesauro UMLS o del NCI (*National Cancer Institute*) en caso de no disponer del código apropiado.

- **RightField [18]**: aplicación de escritorio de libre acceso que permite insertar anotaciones desde ontologías definidas en plantillas de hojas de cálculo de Microsoft Excel. Permite a los usuarios introducir sus datos de forma coherente sin conocer en detalle las ontologías utilizadas, al restringir las celdas a rangos específicos de clases o instancias de vocabularios estándar. Las anotaciones pueden realizarse utilizando ontologías de BioPortal del NCBO (https://bioportal.bioontoloby.org) o utilizando un archivo local en formatos OWL, OBO, RDFS y RDF.

- **eleMAP [19]**: herramienta basada en servicios web desarrollada dentro de la Red de Registros Médicos Electrónicos y Genómicos (eMERGE), que permite el mapeo semiautomático de elementos de datos a vocabularios biomédicos estandarizados (como NCI-T y SNOMED CT) y registros de metadatos (como caDSR). Dispone de una interfaz RESTful que consulta el caDSRS (Cancer Data Standards Registry and Repository) y reutiliza los mapeos. Igualmente, puede utilizarse el servicio REST de BioPortal para identificar conceptos, mapearlos a elementos de datos y enviar los mapeos al repositorio de metadatos caDSR.

- **CEDAR [20]**: suite de soluciones de libre acceso basadas en Web y API REST desarrolladas por el El Center for Expanded Data Annotation and Retrieval (CEDAR), que permite a los usuarios construir plantillas de metadatos, o rellenar plantillas para generar y compartir metadatos de alta calidad, teniendo en cuenta los principios FAIR. Los metadatos están disponibles en formatos JSON, JSON-LD o RDF para facilitar la integración en aplicaciones científicas y la reutilización de los mismos. La construcción de las plantillas de metadatos se apoya en las ontologías de BioPortal para la especificación semántica de los datos, mediante un servicio de búsqueda interactiva vinculado a BioPortal. Este servicio permite a los diseñadores de plantillas encontrar conjuntos de términos ontológicos para anotar las plantillas y los campos. CEDAR se está usando actualmente para el desarrollo de los talleres Metadata for Machines

(M4M) que desarrolla GOFAIR en los últimos años (https://www.gofairfoundation.org/m4m/). Los talleres M4M son eventos ágiles, tipo hackathon, que reúnen a expertos de dominio (que pueden y quieren representar a una comunidad de dominio) con expertos en metadatos FAIR (administradores de datos) que guían un debate que conduce a los requisitos de metadatos que satisfacen las necesidades de datos FAIR de esa comunidad de dominio. Se ha usado para construir el FAIR COVID Health Portal de Dinamarca (https://www.zonmw.nl/en/research-and-results/fair-data-and-data-management/open-science-in-covid-19-research/).

- **SAP [21]**: Pipeline de anotación semántica basado en el sistema CEDAR, desarrollado para automatizar la anotación semántica de metadatos procedentes de repositorios de datos públicos. Transforma los metadatos de los repositorios en el formato CEDAR JSON-LD antes de añadir anotaciones a los metadatos con formato CEDAR. Emplea la herramienta de reconocimiento de entidades de Apache UIMA, ConceptMapper [22] para mapear los metadatos a términos de ontologías de BioPortal. Actualmente la información sobre esta herramienta es muy limitada.

- **D2Refine [23]**: plataforma web construida sobre la herramienta de código abierto OpenRefine (anteriormente Google Refine) [24] para la estandarización y armonización de los diccionarios de datos de estudios de investigación clínica, OpenRefine, emplea el contenido de bases de datos, como la Base de Datos de Genotipos y Fenotipos (dbGaP), la Base de Conocimiento de Fenotipos (PheKB) y el Atlas del Genoma del Cáncer (TCGA) para procesar y transformar grandes conjuntos de datos en una interfaz similar a una hoja de cálculo. D2Refine, en su caso, amplía el mecanismo de exportación de OpenRefine (CSV, TSV, HTML, Excel, ODF, XML, RDF) permitiendo serializar modelos en representaciones estándar como el Archetype Definition Language (ADL) de OpenEHR, el Archetype Modeling Language (AML) de OMG, las Shape Expressions (ShEx) de W3C y los perfiles FHIR de HL7. Ofrece enfoque semiautomático (sugiriendo anotaciones basadas en la implementación de los Servicios Terminológicos Comunes 2 - CTS2) y manual (opción de realizar la búsqueda para obtener anotaciones alternativas).

- **Prototipo de anotación semántica de Wiktorin [25]**: herramienta colaborativa aún en desarrollo para la anotación semántica de datos médicos con los estándares terminológicos SNOMED CT, LOINC y ATC. Las anotaciones se apoyan en mapeos ya existentes y similares, y en la realización de mapeos cruzados del metatesauro UMLS. No proporcionan interfaces para la integración de la terminología y al encontrarse aún en desarrollo no ha hecho público el código fuente ni la información sobre los formatos de importación y exportación

A pesar de que ninguna de las herramientas anteriores cumple por completo con los requisitos previamente indicados, SAP y eleMAP son las que mejores sugerencias de anotaciones aportan, sin embargo, aún no está disponible su código fuente. Las herramientas de anotación SAP, RightField y CEDAR muestran una amplia cobertura al proporcionar interfaces con portales ontológicos (BioPortal u OLS), pero no disponen de interfaces para la integración de terminología adicional, siendo RightField y D2Refine las únicas que permiten cargar archivos

locales, ficheros excel en el caso de RightField al soportar actualmente solo este formato. eleMAP y ODMedit, adaptadas a casos de uso específicos, tampoco parecen tener interfaces para la integración de terminología adicional. Respecto a la generación de sugerencias de anotaciones: ODMedit, eleMAP y Prototype Wiktorin lo realizan de forma directa, y SAP mediante el uso de una herramienta de reconocimiento de conceptos.

La importancia de anotar correctamente los conjuntos de datos con el fin de lograr la interoperabilidad, normalización y harmonización de la información queda reflejada en numerosas acciones europeas, a continuación, se detallan algunas de ellas:

- **DCAT-AP2 (DCAT Application Profile for Data Portals en Europa)**[26]: estándar utilizado a nivel europeo, basado en el Data Catalogue Vocabulary desarrollado por W3C, para describir los conjuntos de datos abiertos en Europa y facilitar así la homogeneización y la búsqueda cruzada mediante el uso de metadatos. Entre sus características más relevantes destacan la descripción de servicios de datos (APIS o Servicios Web), la especificación de roles de agente; versionado, procedencia, granularidad y gobernanza de los conjuntos de datos; así como uso de vocabularios controlados y validación y control de calidad de grafos RDF.
- **MIABIS** [27]: desde la Infraestructura de Investigación de Biobancos y Recursos Biomoleculares de Suecia (BBMRI.se) con el objetivo de facilitar la reutilización de los recursos biológicos y los datos asociados, se desarrolló en el 2012 el estándar sobre la Información Mínima sobre Intercambio de Datos de Biobancos (MIABIS). Su gran aceptación por parte de la comunidad científica, desencadenó en una segunda versión, constituida por 22 atributos que describen los biobancos, las colecciones de muestras y los estudios, y considerándose como la información mínima necesaria para iniciar colaboraciones entre biobancos y permitir el intercambio de muestras y datos biológicos.
- **MICA** [28] : herramienta utilizada para crear portales web de datos para estudios epidemiológicos a gran escala o consorcios de estudios múltiples. Mica2 es el sucesor de Mica. Ayuda a los estudios a proporcionar una visibilidad de los datos y una presencia en la web sin un esfuerzo significativo, proporcionando una descripción estructurada de consorcios, estudios, diccionarios de datos anotados y con capacidad de búsqueda, y gestión de solicitudes de acceso a datos. Proporciona anotaciones a nivel de conjunto de datos, como por ejemplo, cómo, cuándo, dónde, por quién y en qué condiciones se han recogido los datos, información que posteriormente se publica en un portal web. Aunque Mica implementa identificadores únicos, éstos no son persistentes, y tampoco están explícitamente definidos en los metadatos. Esto plantea problemas de accesibilidad y reutilización de los datos..MICA ha sido utilizado recientemente en el proyecto BEAT-COVID en Leiden University Medical Center de Holanda [29] .

Existen además multitud de normas de metadatación específicas de dominio, que puede ser necesario tener en cuenta en proyectos específicos. Así, en el ámbito de la imagen médica es imprescindible trabajar con el estándar de comunicación y gestión de imagen digital médica **DICOM**. DICOM no sólo especifica la forma de almacenar y transferir imágenes médicas, sino también cómo metadatar las imágenes y los informes clínicos asociadas. DICOM es objeto de tratamiento detallado en el entregable E4.4 de IMPaCT-Data. En el ámbito de datos geoespaciales, la directiva europea **INSPIRE** define el contenido y formato de los metadatos para información geográfica, que puede ser necesaria en aquellos estudios que impliquen georreferenciar a los pacientes, por ejemplo, en estudios epidemiológicos o en los que la componente geográfica pueda suponer un condicionante de salud. Igualmente, la norma **CESSDA**, definida para metadatar información en el ámbito de la investigación en ciencias sociales, puede ser necesaria para la realización de estudios en los que se combina la información clínica con, por ejemplo, factores socioeconómicos como condicionantes de salud. El detalle de estas y otras normas de nicho queda fuera del alcance del presente documento.

4. Estándares terminológicos y de clasificación

4.1. SNOMED-CT

Actualmente, la terminología clínica de referencia más completa y más utilizada a nivel mundial es **SNOMED-CT**[53] . Está basada en conceptos, sus descripciones y relaciones que describen la vinculación jerárquica y los atributos de dichos conceptos. Representa el conocimiento a través de una ontología médica, lo cual permite una gran riqueza semántica y facilidad de crecimiento evolutivo. De esta forma, con SNOMED-CT se puede desarrollar contenido médico completo y de alta calidad.

El modelo lógico de SNOMED-CT define el modo en el que se relaciona y representa cada tipo de componente de la terminología (conceptos, descripciones y relaciones) y sus derivados. [53]

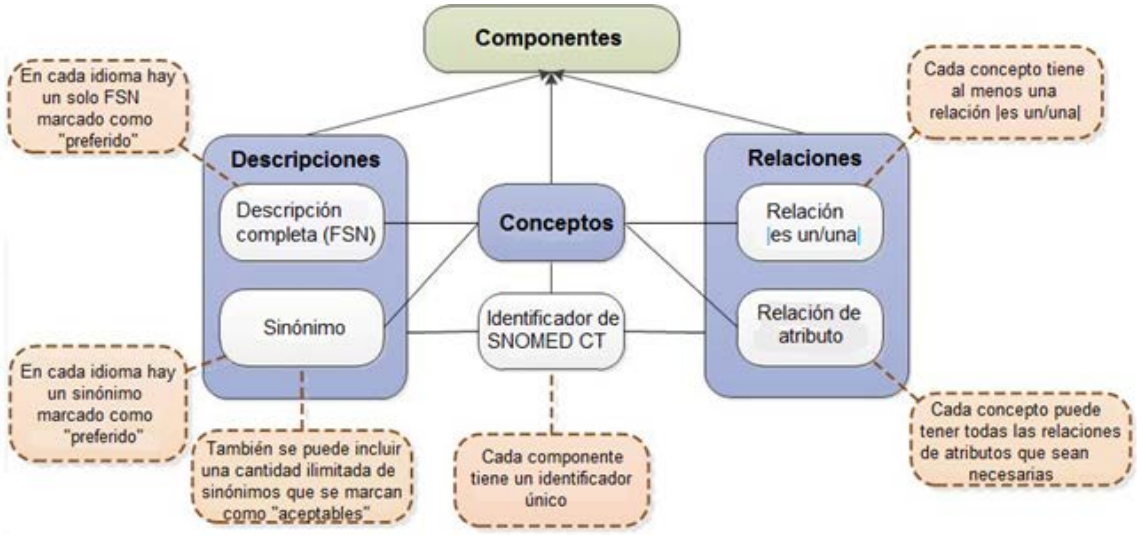

Ilustración 3 Modelo básico de la ontología de SNOMED-CT

Fuente: extraído de "Guía de introducción a SNOMED-CT", 2018.

A su vez, el modelo conceptual de SNOMED-CT es el que especifica cómo se definen los conceptos de la terminología mediante una combinación de lógica formal y reglas editoriales. Los atributos que describen a los conceptos deben tener un dominio (la jerarquía a la que pertenecen) y un rango (los conceptos de SNOMED-CT permitidos como valores del atributo). De las 19 jerarquías que tiene la terminología (conceptos ubicados en el nivel superior), los atributos permiten representar el significado de conceptos de nueve de ellas.

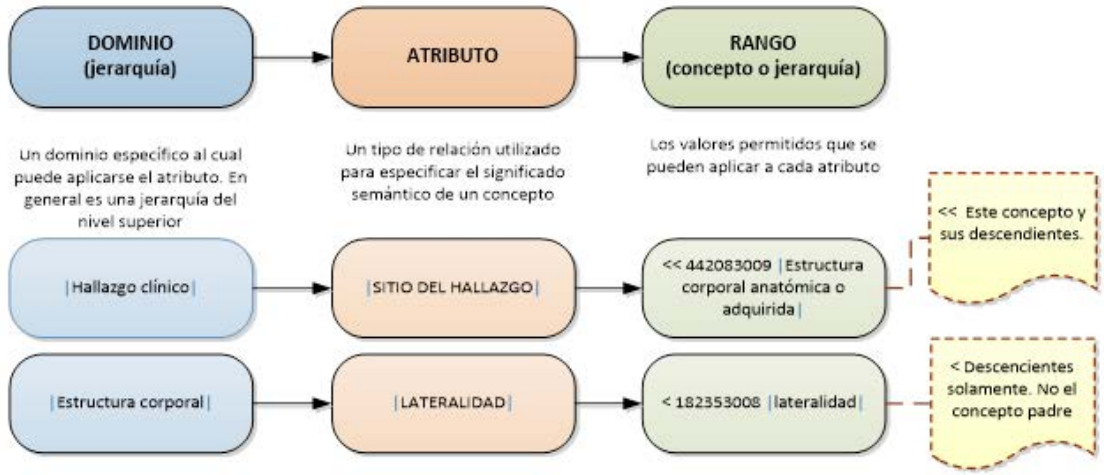

Ilustración 4 Ejemplos de dominio y rango especificados para los atributos |sitio del hallazgo| y |lateralidad|

Fuente: extraído de "Guía de introducción a SNOMED CT", 2018 (2).

Por último, cabe mencionar que, a partir de la postcoordinación, SNOMED-CT permite representar frases clínicas, aunque el concepto exacto no esté definido en la terminología, mediante la construcción de expresiones que contienen dos o más identificadores de concepto.

SNOMED es la base de interoperabilidad en múltiples proyectos e iniciativas. Es una de las terminologías de referencia en modelos comunes como OMOP, y es también la base de la implementación de proyectos de interoperabilidad a nivel nacional como la Historia Clínica Digital del Sistema Nacional de Salud (HCDSNS) y la receta electrónica interoperable del Sistema Nacional de Salud. Así mismo, España es país Licenciatario de SNOMED-CT, siendo el Centro Nacional de Referencia de SNOMED-CT del Ministerio de Sanidad (CNR) el encargado de gesitonar, a nivel nacional, esta terminología y dispone de dos ediciones nacionales: "Extensión para España" y "Extensión para España de Medicamentos".

4.2. LOINC

En cuanto a las mediciones, observaciones y documentos de salud, la terminología más ampliamente utilizada a nivel mundial es **LOINC** [54]. Este estándar permite mapear códigos locales de las instituciones a estándares universales que puedan fácilmente ser intercambiados entre distintos sistemas. Se utiliza principalmente para la identificación unívoca de pruebas diagnósticas (particularmente de laboratorio), aunque también puede usarse para signos vitales y documentos clínicos, entre otros. Cada código LOINC distingue seis dimensiones para una determinada observación: componente o analito, propiedad, intervalo de tiempo, espécimen, escala y, opcionalmente, método de obtención de la observación.

4.3. Clasificación Internacional de Enfermedades (CIE)

La clasificación más ampliamente utilizada a nivel nacional e internacional es la **Clasificación Internacional de Enfermedades (CIE)** [55], de la Organización Mundial de la Salud (OMS). Con más de 100 años de evolución, la versión más ampliamente distribuida e implementada es la CIE-10. Se trata de una clasificación monoaxial, donde la jerarquía de términos está basada en una raíz común y cada término pertenece únicamente a una clase. Este sistema fue diseñado para informar diagnósticos médicos, aunque las versiones modificadas (CIE-9-CM, CIE-10-CM) también incluyen procedimientos.

Clasificación de Enfermedades y Lesiones	
01. (A00-B99)	CIERTAS ENFERMEDADES INFECCIOSAS Y PARASITARIAS
02. (C00-D48)	TUMORES [NEOPLASIAS]
03. (D50-D89)	ENFERMEDADES DE LA SANGRE Y DE LOS ÓRGANOS HEMATOPOYÉTICOS, Y CIERTOS TRASTORNOS QUE AFECTAN EL MECANISMO DE LA INMUNIDAD
04. (E00-E90)	ENFERMEDADES ENDOCRINAS, NUTRICIONALES Y METABÓLICAS
05. (F00-F99)	TRASTORNOS MENTALES Y DEL COMPORTAMIENTO
06. (G00-G99)	ENFERMEDADES DEL SISTEMA NERVIOSO
07. (H00-H59)	ENFERMEDADES DEL OJO Y SUS ANEXOS
08. (H60-H95)	ENFERMEDADES DEL OÍDO Y DE LA APÓFISIS MASTOIDES
09. (I00-I99)	ENFERMEDADES DEL SISTEMA CIRCULATORIO
10. (J00-J99)	ENFERMEDADES DEL SISTEMA RESPIRATORIO
11. (K00-K93)	ENFERMEDADES DEL SISTEMA DIGESTIVO
12. (L00-L99)	ENFERMEDADES DE LA PIEL Y DEL TEJIDO SUBCUTÁNEO
13. (M00-M99)	ENFERMEDADES DEL SISTEMA OSTEOMUSCULAR Y DEL TEJIDO CONJUNTIVO
14. (N00-N99)	ENFERMEDADES DEL SISTEMA GENITOURINARIO
15. (O00-O99)	EMBARAZO, PARTO Y PUERPERIO
16. (P00-P96)	CIERTAS AFECCIONES ORIGINADAS EN EL PERÍODO PERINATAL
17. (Q00-Q99)	MALFORMACIONES CONGÉNITAS, DEFORMIDADES Y ANOMALÍAS CROMOSÓMICAS
18. (R00-R99)	SÍNTOMAS, SIGNOS Y HALLAZGOS ANORMALES CLÍNICOS Y DE LABORATORIO, NO CLASIFICADOS EN OTRA PARTE
19. (S00-T98)	TRAUMATISMOS, ENVENENAMIENTOS Y ALGUNAS OTRAS CONSECUENCIAS DE CAUSAS EXTERNAS
20. (V01-Y98)	CAUSAS EXTERNAS DE MORBILIDAD Y DE MORTALIDAD
21. (Z00-Z99)	FACTORES QUE INFLUYEN EN EL ESTADO DE SALUD Y CONTACTO CON LOS SERVICIOS DE SALUD
22. (U00-U99)	CÓDIGOS PARA PROPÓSITOS ESPECIALES

Ilustración 5 Ejemplo de códigos CIE-10

Fuente: extraído de "eCIE10 - Edición electrónica de la CIE-10", 2010.

La versión más reciente es la CIE-11, aunque aún no se ha desplegado significativamente. Además de agregar nuevos capítulos y modificar algunos antiguos, tiene dos principales cambios con respecto a la CIE-10. Por un lado, ya no es una clasificación estrictamente monoaxial, sino que un término puede pertenecer a más de una clase. Además, los códigos principales pueden ser postcoordinados (agregado de códigos de extensión), de manera tal de poder agregar granularidad manteniendo la estandarización.

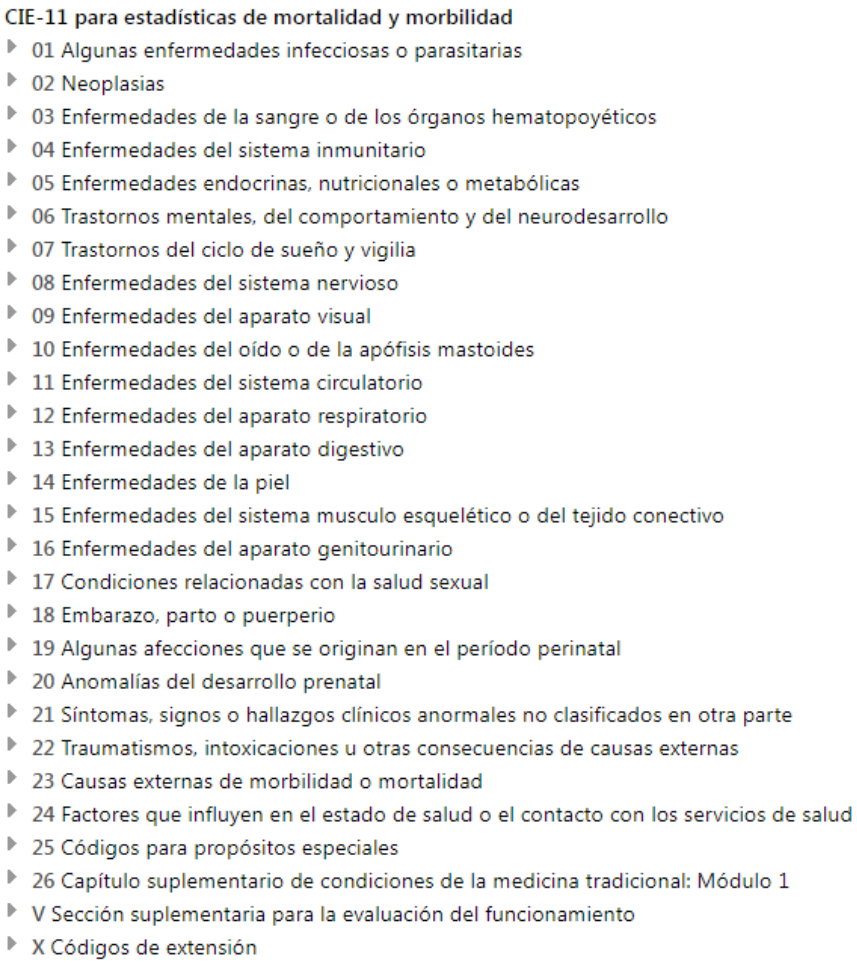

Ilustración 6 Ejemplo de códigos CIE-11

Fuente: extraído de "CIE-11 para estadísticas de mortalidad y morbilidad", 2021.

4.4. Clasificador Internacional de Atención Primaria (CIAP)

El **Clasificador Internacional de Atención Primaria (CIAP)** [56 se utiliza para capturar y organizar la información clínica en dicho ámbito, y es reconocido como un estándar por la familia de clasificaciones internacionales de la OMS (WHO-FIC). Al estar mapeado a la CIE, se puede intercambiar información entre las dos clasificaciones, a la vez que usarse complementariamente, si bien este mapeo no tiene una correspondencia exacta, debido a que el enfoque aportado por CIAP, con frecuencia sintomático, sistémico o sindrómico, es muy distinto al utilizado en atención especializada, mucho más detallado semánticamente, y de orientación epistemológica.

El CIAP posee una estructura biaxial y consta de 17 capítulos, incluyendo algunos con problemas sociales, muy relevantes en la atención primaria. Su uso permite representar toda la consulta del paciente, desde el motivo de consulta hasta el procedimiento o tratamiento.

4.5. Clasificación de medicamentos

Al igual que en otras áreas del conocimiento médico como las que se han comentado en los apartados anteriores, también se aplican a los medicamentos y tóxicos. Entre las diversas clasificaciones internacionales, destaca por su relevancia, el Anatomical Therapeutic Chemical (ATC) Classification System que se utiliza para clasificar los principios activos e ingredientes de los medicamentos. Se trata de una clasificación elaborada y mantenida por la **Organización Mundial de la Salud** (OMS). Esta clasificación permite conocer de una forma estandarizada qué medicamentos se utilizan por ejemplo en un sistema de información como la Historia Clínica Electrónica.

El ATC está divide los medicamentos en diferentes grupos en función del órgano o sistema sobre el que actúa de forma terapéutica y según las características químicas de los productos. Se describen los productos de forma individualizada pero también si existe su presentación en el mercado, combinados con otros medicamentos.

La clasificación de los medicamentos se realiza en diferentes grupos y hasta en cinco niveles y llevan asociados unos códigos determinados:

- Un primer nivel que incluye 14 grupos principales.
- Un segundo nivel que hace referencia a los subgrupos farmacológicos a los que pertenece.
- El tercer y cuarto nivel hacen referencia a los subgrupos terapéuticos y farmacológicos en los que se pueden incluir.
- El quinto nivel es la sustancia química.

Por ejemplo, en el caso del grupo A de medicamentos para el tracto alimentario y metabolismo podríamos encontrarnos un subgrupo de medicamentos utilizados en la diabetes mellitus y concretamente del grupo terapéutico de la Biguanidas y concretamente una de ellas como es la Metformina:

A	Tracto alimentario y metabolismo (primer nivel)
A10	Medicamentos utilizados en la diabetes (segundo nivel)
A10B	Medicamentos para reducir los niveles de glucos (excluyendo insulinas) (tercer nivel)
A10BA	Biguanidas (cuarto nivel)
A10BA02	Metformina (quinto nivel)

Debemos tener en cuenta que un mismo componente o sustancia medicamentosa puede clasificarse en diversos lugares según las diferentes aplicaciones terapéuticas y vías de administración. Así por ejemplo la Prednisolona (un corticoide) puede presentar las siguientes codificaciones:

A07EA01	Agentes antiinflamatorios intestinales (enemas)
C05AA04	Antihemorroidales de uso tópico (supositorios)
D07AA03	Preparados dermatológicos (cremas, lociones)
H02AB06	Corticoides de uso sistémico (tabletas, inyectables)
R01AD02	Descongestionantes nasales (spray/gotas nasales)

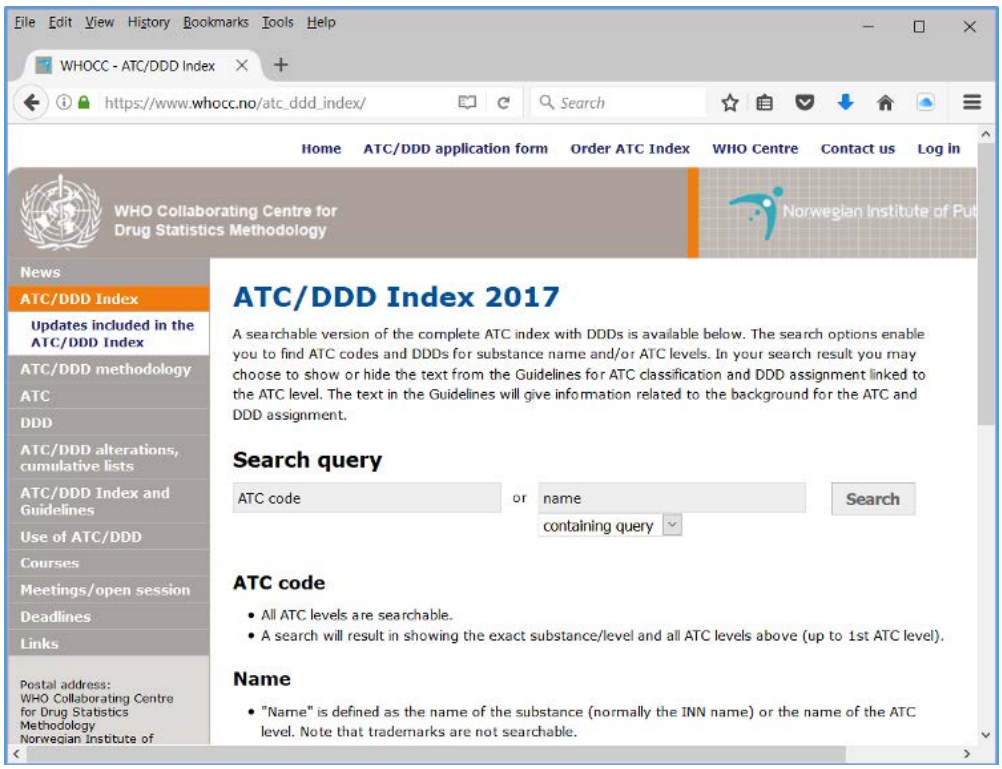

Ilustración 7 Página web del ATC/DDD de la Organización Mundial de la Salud

4.6. Clasificación de exploraciones radiológicas

El Catálogo de Exploraciones de la Sociedad Española de Radiología Médica (SERAM) (https://seram.es/catalogo-seram/), actualmente en su edición de 2016, fue adoptado por el Consejo Interterritorial como el Catálogo del Sistema Nacional de Salud (SNS) y desde 2010

forma parte del conjunto mínimo de datos del informe de resultados de las pruebas de imagen del SNS. Sirvió de base para el proyecto europeo ERDDS (European Radiologic Digital Data System), 2004-2006 [30][30][30].

Este catálogo, actualmente, utiliza códigos de 8 dígitos. Los dos primeros se refieren a la técnica, adaptado a la clasificación del SIAE (sistema de información de atención especializada) con la excepción del intervencionismo general que correspondería a otros procedimientos. Los dos siguientes se refieren al órgano o sistema correspondiente con una estructura semejante a los códigos ACR, con la excepción del vascular-intervencionista, que sigue una estructura propia. Los dos últimos codifican la prueba y el procedimiento. Por ejemplo, el código 01030101 corresponde a “ABDOMEN SIMPLE AP” en la seccción “RADIOLOGÍA SIMPLE”.

El catálogo incluye 8 columnas:

- Código SERAM de 8 dígitos.
- Prueba: Grupos de pruebas, con agrupaciones de técnicas (p.ej. “Radiología Simple” o “Mamografía e Interv. de mama”) y subgrupos ordenados tanto por criterios topográficos (tórax, abdomen, cráneo y cara), como por tecnología o técnica (“Mamografía con tomosíntesis” o “biopsia con aguja gruesa mama”.
- Procedimiento: Se utiliza para detallar aún más procedimientos, modalidades, o localizaciones anatomómicas (p. Ej. “BAV mama por esterotaxia”).
- Tiempos de ocupación de sala (TOS)
- Tiempos médicos (TM)
- Unidad de actividad radiológica (URA): Tiene en cuenta el consumo de recursos humanos y el tiempo de ocupación de sala.
- Unidad relativa de valor (URV): Su objetivo es establecer un coste económico imputable a cada prueba concreta de la cartera de procedimientos de un servicio.

4.7. HPO – Ontología de Fenotipos Humanos

En el contexto de IMPaCT-Data, uno de cuyos objetivos es establecer los criterios y mecanismos que permitan juntar datos clínicos e imágenes médicas con información genómica, la codificación de determinada información clínica sobre la base de una ontología de fenotipos probablemente sea de gran ayuda. En este sentido, la ontología más utilizada en el ámbito de la genética médica es HPO [31].

La Ontología de Fenotipos Humano (Human Phenotype Ontology - HPO) proporciona un vocabulario normalizado de las anormalidades fenotípicas encontradas en las patologías humanas. Cada término en HPO describe un fenotipo anormal, como por ejemplo “defecto septal atrial”. HPO está siendo desarrollada a partir de la literatura médica, de Orphanet, DECIPHER y OMIM, que son normas y bibliotecas centradas en la herencia genética humana

y la caracterización de enfermedades raras de origen genético. HPO contiene actualmente más de 13.000 términos y alrededor de 156.000 anotaciones de enfermedades hereditarias. HPO está mantenida y gestionada por un consorcio financiado por el National Institute of Health denominado Monarch Initiative, y es un componente central de varios proyectos desarrollados en el seno de la Global Alliance for Genomics and Health (GA4GH).

4.8. UMLS

Unified Medical Language System (UMLS) es un servicio gestionado y mantenido por la National Library of Medicine americana, que integra y distribuye estándares terminológicos, de clasificación y codificación para promover la creación de sistemas y servicios de información biomédica interoperables. UMLS relaciona y permite mapear términos entre multitud de terminologías y ontologías, entre las que están muchas de las mencionadas anteriormente (SNOMED-CT, LOINC, RxNORM, ATC, HPO, CIE...), por lo que con frecuencia se utiliza como herramienta para la traducción y normalización terminológica. Sin embargo, dado que UMLS asigna un identificador propio a cada concepto mapeado a cualquiera de las restantes ontologías y clasificaciones, puede ser considerada como una clasificación más, y en algunos proyectos se ha utilizado como terminología “per se”.

4.9. Otras terminologías “de nicho”

Además de las ontologías y terminologías explicadas en las secciones anteriores, existen otras muchas terminologías “de nicho”, que se usan en dominios o usos específicos de información sanitaria, con mayor o menos nivel de madurez. Aquí estarían terminologías como ORPHA , para la catalogación de enfermedades raras, junto con la ontología ORDO, o la codificación de OMIM [58] para genética humana (tanto genotipos como fenotipos), o terminologías como MeSH/DeCS [59], utilizada en etiquetado de literatura biomédica. BioPortal[52] recoge más de 900 ontologías distintas, y entrar en el detalle de todas ellas excede el objeto de este documento.

5. Modelos comunes de datos

5.1. Observational Medical Outcomes Partnership (OMOP)

Un ejemplo de modelo común de datos de aplicación en investigación observacional es el Observational Medical Outcomes Partnership (OMOP). OMOP se inició como asociación público-privada, presidida por la Administración de Drogas y Alimentos (FDA) de los EE. UU. y financiada por un consorcio de compañías farmacéuticas que colaboraron con investigadores académicos y socios de datos de salud para establecer un programa de investigación. que buscaba avanzar en la ciencia de la vigilancia activa de la seguridad de los productos médicos utilizando datos observacionales de atención médica.

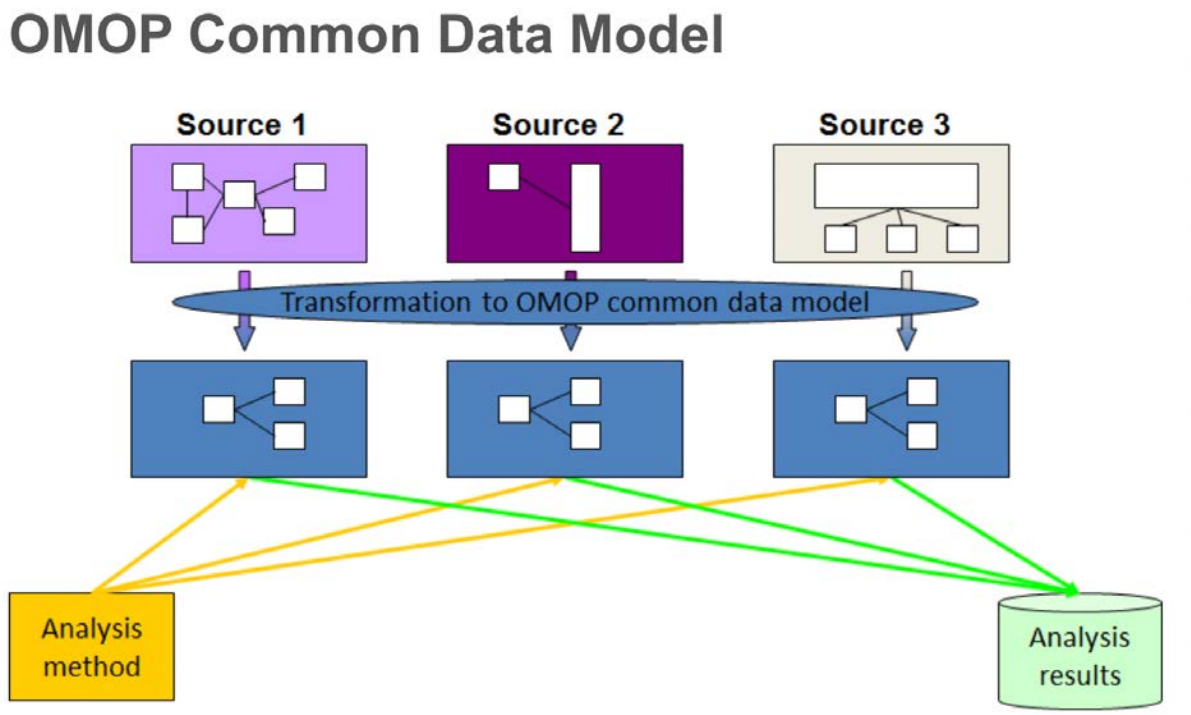

Ilustración 8 Proceso de estandarización de datos en un modelo comú

Propuesto por el OMOP Common Data Model

Actualmente, el modelo común de datos OMOP [34] es responsabilidad del programa OHDSI (Observationsl Health Data Sciences and Informatics), Este programa, en el que participan múltiples entidades, tanto públicas como privadas, es un programa de colaboración interdisciplinar que busca extraer valor de los datos de salud mediante su análisis a gran escala. Para ello, OHDSI ha establecido una red internacional de investigadores y bases de datos observacionales, coordinada desde la Universidad de Columbia.

Para conseguir sus objetivos, OHDSI no solo se encarga de mantener y evolucionar el modelo común de datos OMOP, con la colaboración de todos sus socios y colaboradores, sino que desarrolla gran cantidad de herramientas en código abierto, para facilitar el trabajo de recolección, persistencia, búsqueda y mapeo terminológico. Entre estas herramientas detacan HADES (Health Analytics Data-to-Evience Suite), que está constituido por un conjunto de librerías de R que permiten la realización de análisis bioestadísticos avanzados big data, y ATLAS, una herramienta web que facilita el diseño, ejecución de análisis de forma estandarizada de los datos clínicos de todas las bases de datos clínicas de la comunidad internacional OHDSI que se encuentran en formato OMOP. Esta colaboración permite el acceso a millones de datos de pacientes para el desarrollo de investigación en red.

En Europa, además de las múltiples instituciones que participan directamente en la iniciativa OHDSI, se ha creado el consorcio EHDEN (European Health Data Evidence Network), financiado por un programa IMI, y cuyo objetivo fundamental es promover la adopción de OMOP como modelo común de datos de salud para proyectos de análisis, y la promoción de proyectos de investigación en red, poniendo en contacto a sus *data partners*. EHDEN cuenta en la actualidad con más de 250 *data partners* en toda Europa, y más de 20 solo en España, incluyendo hospitales, centros de investigación e incluso servicios autonómicos de salud completos.

El modelo de datos comunes de OMOP permite el análisis sistemático de bases de datos observacionales de gran diversidad. El objetivo de este enfoque es transformar los datos contenidos dentro de diversas bases de datos como la historia clínica electrónica, en un formato común (modelo de datos), así como una representación común (terminologías, vocabularios, esquemas de codificación), que permitan además realizar análisis sistemáticos utilizando una biblioteca de rutinas analíticas estándar basadas en un formato común.

Entre los objetivos generales de la utilización del formato OMOP de datos clínicos que forma parte de la iniciativa internacional OHDSI destacan:

- la reproductibilidad: para la mejora de la salud es esencial disponer de la evidencia que permita ser reproducible y precisa para evitar errores.
- promover la creación de una comunidad internacional, tanto profesionales como pacientes o investigadores.
- favorecer la colaboración de la comunidad internacional para dar respuesta a los diferentes problemas que plantean el mundo real.
- facilitar la utilización de herramientas en abierto y gratuitas para la favorecer la investigación y la generación de evidencia.
- proteger los derechos de los individuos y de las organizaciones en el uso de los datos.

El modelo de datos de OMOP se basa en una ontología relativamente sencilla (ver Ilustración 9), y sus herramientas facilitan el trabajo de mapeo entre los datos contenidos en historias

clínicas electrónicas y registros clínicos, y las entidades del CDM OMOP. Este mapeo, que a su vez está basado en terminologías y ontologías estandarizadas, tales como SNOMED, LOINC o CIE, facilitan la interoperabilidad semántica entre distintos repositorios, al tener una estructura de entidades y atributos muy definida y apoyada en una ontología. El modelo conceptual, a su vez, se soporta sobre un esquema relacional definido (soportado en distintos gestores de base de datos como MySQL, Oracle o Postgres), lo que facilita a su vez el desarrollo de código analítico que pueda ejecutarse sin problemas en distintos nodos, dado que todos ellos comparte en mismo esquema relacional.

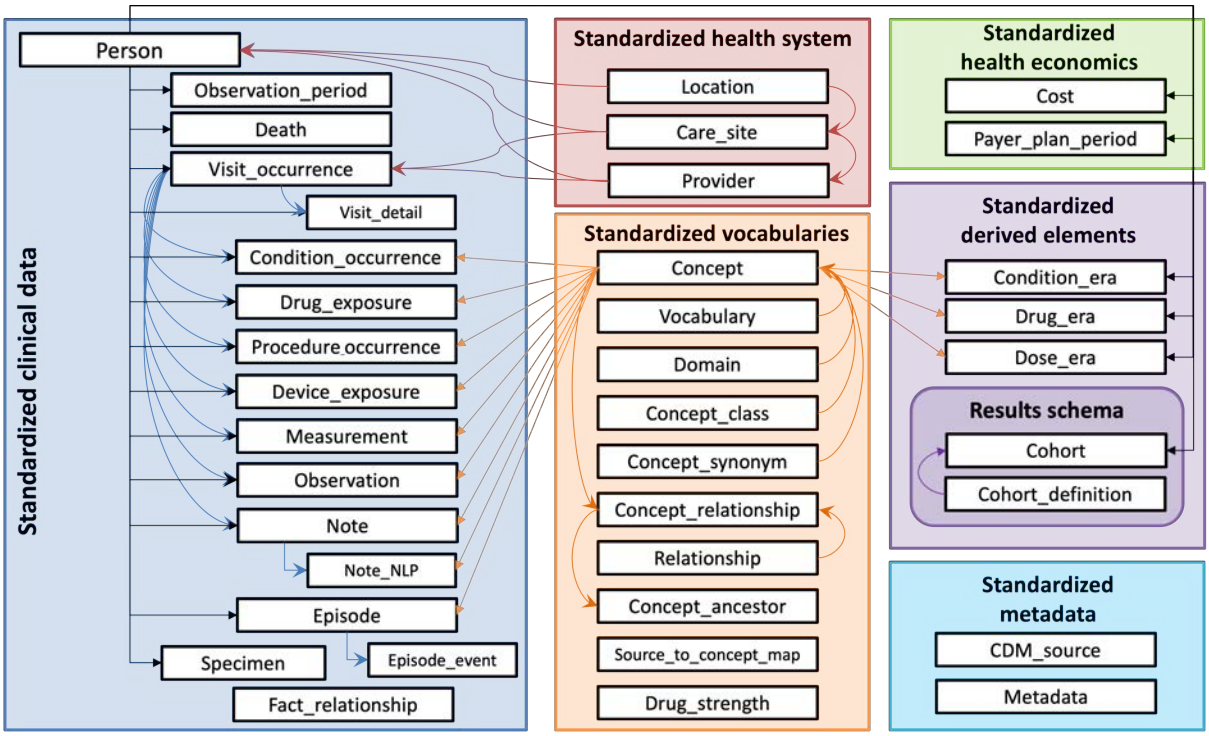

Ilustración 9 Modelo de datos OMOP CDM

El problema que puede existir con OMOP es que, al estar basado en una ontología relativamente simple, y que su origen ha estado muy vinculado al desarrollo de proyectos relacionados con la industria farmacéutica y los estudios post-autorización, con cierta frecuencia no es capaz de cubrir el 100% de los conceptos que se necesitan compartir para un proyecto de investigación concreto. Esta carencia se puede suplir, no obstante, mediante guías de interoperabilidad que especifique la forma en la que se han mapeado los conceptos no estándar, y a través de los mecanismos que ofrece OHDSI para extender el modelo OMOP CDM de una forma ordenada.

Lo cierto, en cualquier caso, es que OMOP se está convirtiendo en el estándar *de facto*. como modelo de datos de referencia para proyectos que requieren compartir datos de salud o ejecución federada de análisis sobre los mismos.

5.2. *Integrating Biology and the Bedside (i2b2)*

Integrating Biology and the Bedside (i2b2) es una plataforma modular para el almacenamiento y análisis de datos de salud procedentes de múltiples orígenes, desarrollada por la Facultad de Medicina de Harvard con financiación de los National Institutes of Health de los Estados Unidos (NIH) [32][32][32]. El componente principal de esta herramienta es el repositorio de datos clínicos [33], cuya base de datos se basa en un diseño de esquema en estrella, formado por la tabla central Observation_Fact, que almacena todas las observaciones de salud (entendiendo observación como cualquier evento en la salud del paciente), y un conjunto de tablas adicionales, enlazadas a esta, que le aportan información adicional. Estas tablas se denominan *dimensiones*, e incluyen:

- Patient_Dimension, para la información del paciente sobre el que se realiza la observación;
- Visit_Dimension, para la información de la visita en el que se realiza la observación;
- Provider_Dimension, para la información del proveedor de atención sanitaria que realiza la observación;
- Concept_Dimension, para la información de la entidad observable sobre la que se reporta el dato; y
- Modifier_Dimension, para información adicional de la entidad observable sobre la que se reporta el dato.

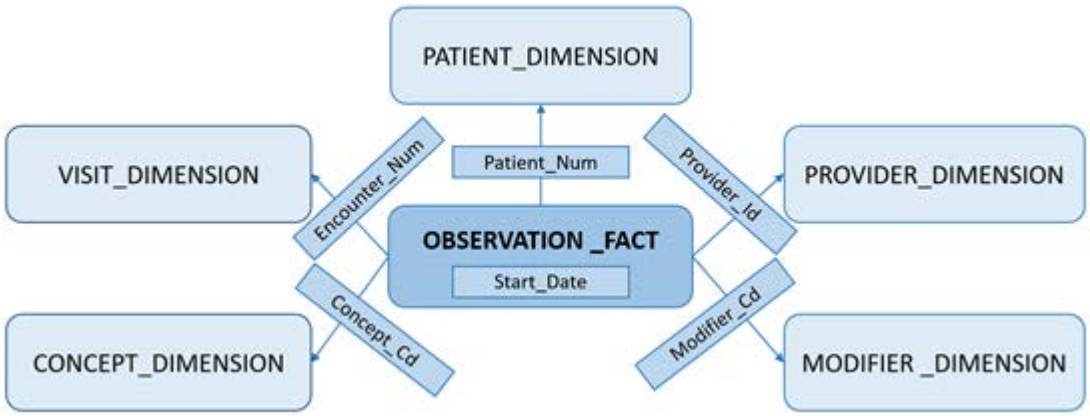

Ilustración 10 Modelo de datos básico del repositorio i2b2

Fuente: Hospital Universitario 12 de Octubre

Así, i2b2 normaliza el modelo de datos, pero no establece un marco común de conceptos del dominio clínico para el contenido del mismo, como sí hacen otras propuestas de repositorios normalizados como OMOP CDM [34]. Esto hace a i2b2 flexible en la incorporación de

estándares terminológicos y de clasificación de acuerdo con las necesidades específicas de cada caso de uso y los orígenes de datos. Por contra, esto implica que cada proyecto debe acordar previamente este marco conceptual entre las organizaciones que carguen sus datos al repositorio. La estructura jerárquica de dichos conceptos se define en la ontología de i2b2, también denominada *metadatos*, la cual se implementa en una base de datos distinta a la del repositorio y puede estar repartida en una o varias tablas comunes.

Los repositorios i2b2 se utilizan frecuentemente en el ámbito sanitario junto a estándares de interoperabilidad para normalizar la carga de datos sobre el mismo. Así, diferentes estudios han implementado mecanismos de cargas de conceptos y datos y hacia estos a través de diferentes mecanismos. Por un lado, el estudio llevado a cabo por Haarbrandt B et al ha logrado la creación automatizada de ontologías i2b2 a partir de arquetipos y plantillas y la integración de instancias ***OpenEHR*** de 903 pacientes de una unidad de cuidados intensivos pediátricos [36] . Del mismo modo, el estudio de Solbrig HR et al. propone la combinación de los datos de la HCE partiendo del estándar FHIR a un repositorio de datos integrados de i2b2, de forma que dichos datos puedan ser representados y consultados por un investigador clínico de manera apropiada [37]. A sí mismo, se han desarrollado diferentes estudios para la conversión del modelo i2b2 a otros modelos de repositorios de uso secundario. Así, en el programa de investigación ***All Of Us***, se pretende construir una cohorte de alrededor de un millón de pacientes en un repositorio OMOP CDM a partir de un repositorio de datos conforme a i2b2 [38]. Por último, diversos proyectos de datos de salud emplean esta especificación como pieza nuclear para centralizar y combinar información procedente de diferentes nodos. Entre otros, el proyecto EHR4CR diseñó su plataforma de datos en torno a un repositorio i2b2. Esta permite realizar consultas distribuidas para ayudar a evaluar la viabilidad de los ensayos clínicos y a reclutar pacientes, así como proveer a los investigadores de datos e indicadores útiles para el desempeño de su actividad [39].

6. Estándares de interoperabilidad

6.1. Reference Information Model (RIM) de HL7 v.3

Se trata de un modelo de datos complejo, cuyo alcance son todos los dominios en el ámbito de la atención de la salud. Define mensajes, documentos, reglas y plantillas que permiten representar información clínica, administrativa, financiera, de salud pública, etc.

El RIM de HL7 v.3 está compuesto de dos elementos conceptuales principales: el acto (acciones que se realizan y documentan durante el proceso de cuidado de la salud) y la entidad (personas y objetos intervinientes en dicho proceso). A estos se suman la participación, el rol, las relaciones entre actos, los vínculos entre roles y la información necesaria para el intercambio de mensajes y para expresar el formato en los documentos clínicos.

6.2. Fast Healthcare Interoperability Resources (FHIR) de HL7

Tras la baja adopción de HL7 v.3, se buscó crear un estándar abierto, fácil de implantar, semánticamente sólido (que pueda ser mapeado al RIM v.3, openEHR, etc.), y creado con tecnologías modernas. El alcance de FHIR también son todos los aspectos de la atención sanitaria, y soporta la interoperabilidad mediante cuatro arquitecturas: mensajería, documentos, servicios y REST.

Su unidad básica de interoperabilidad son los recursos que representan conceptos del mundo sanitario. Estos pueden utilizarse en su forma más sencilla o agruparse en forma de mensajes, documentos o servicios.

Cabe destacar el nuevo proyecto VULCAN de HL7 para la identificación y registro de fenotipos orientado a acelerar la investigación clínica y traslacional basada en FHIR

6.3. UNE-EN ISO 13606

Este estándar basa su arquitectura en el modelo dual, en el cual el modelo de información de referencia estable constituye un primer nivel de modelado, mientras que las definiciones formales de contenido clínico (arquetipos y plantillas) conforman el segundo nivel. Los arquetipos constituyen una expresión de la semántica del dominio vinculable a terminologías de referencia, independiente de la tecnología, y autocontenida en una sola fuente.

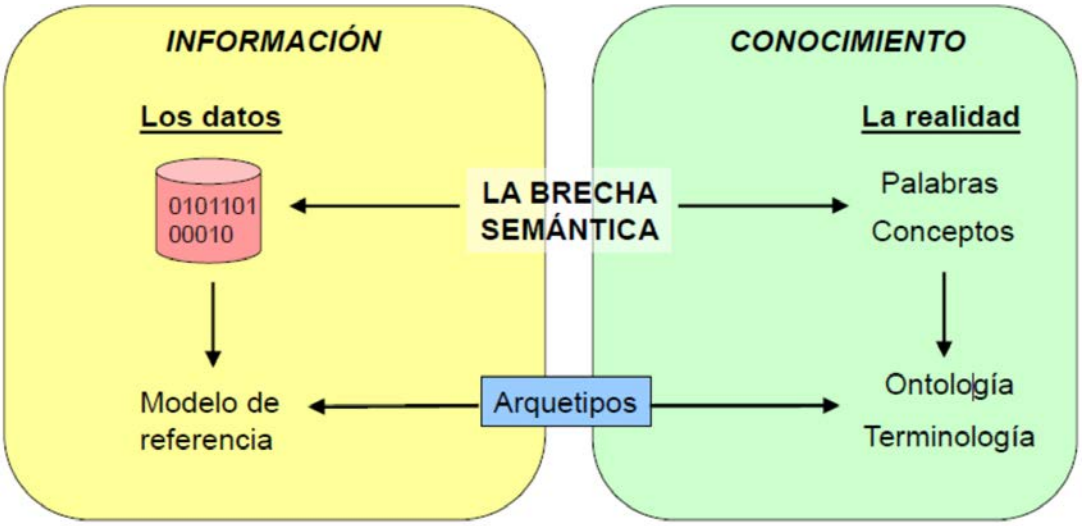

Ilustración 11 El modelo dual

UNE-EN ISO 13606 se divide en cinco partes: modelo de referencia, especificación de arquetipos, arquetipos de referencia y listas de términos, seguridad y modelos de intercambio. Este estándar proporciona el mapeo a HL7 CDA y openEHR para facilitar la interoperabilidad.

6.4. openEHR

OpenEHR es una especificación de código abierto que no solo prescribe el contenido de los artefactos que se intercambian, sino que también prescribe cómo tiene que construirse el software. Es decir, es un modelo para la construcción de un sistema de historia clínica electrónica. Se incluye este estándar de todas formas en el apartado ya que, al igual que UNE-EN ISO 13606, se basa en un enfoque de modelado dual para separar los conceptos clínicos del modelo de información. Todos los tipos de datos y estructuras de datos en openEHR se adhieren a las especificaciones estándares dadas por UNE-EN ISO 13606 para estructuras particulares.

7. Marco de implementación de los principios FAIR basado en estándares

Tal como se dice en el Plan Estratégico de IMPaCT-Data, la propuesta de normas internacionales de información de HCE que se aporta en este documento debe asegurar un cumplimiento efectivo de los principios FAIR tal como se requiere en el 3º objetivo técnico de IMPaCT-Data y tal como se indica en la LET1 de Integridad Científica que pretende establecer los procedimientos "para garantizar los principios FAIR en la organización y gestión de la información en IMPaCT."

Todas las aportaciones que se incluyen en este entregable deben ir orientadas a dicho cumplimiento.

En este sentido, el marco de implementación de los principios FAIR para la información de HCE de IMPaCT Data se establece en 2 ejes:

Eje 1 sobre una adaptación del **Modelo de Madurez de datos FAIR** de la Research Data Alliance RDA seleccionando de los 41 indicadores del modelo aquellos de obligado cumplimiento para disponer del nivel mínimo de madurez, así como un conjunto de indicadores de interoperabilidad ya que en el modelo de madurez no hay obligatorio ningún indicador relacionado con la interoperabilidad. Se identificarán aquellos indicadores cuya evidencia de ser implementados completamente (nivel 4 del modelo) se demuestra con el uso efectivo de los estándares que aquí se proponen [40] .

Eje 2 sobre la generación de un **Perfil de Implementación FAIR (FIP) para los datos de HCE de IMPaCT Data** [41], siguiendo la herramienta Data Steward Wizard (https://ds-wizard.org/), cumplimentaremos el cuestionario FIP de GOFAIR y tendremos como resultado un FIP que podrá ser exportado, incluso machine-readable. El propio FIP será un FAIR Digital Object.

La cumplimentación de DSW además supondrá la incorporación del IMPaCT Data HCE FIP a la matriz de convergencia FAIR [42] que se desarrolla de manera paralela al modelo de perfiles de implementación FAIR de GOFAIR.

La generación del IMPaCT Data HCE FIP salvará los obstáculos y aportará una armonización respecto a lo que supone una implementación de un CDE (Common Data Elements) específico para la información de HCE en IMPaCT Data, teniendo en cuenta un subconjunto de estándares de los propuestos en este entregable, y lo que este CDE supondría de limitante para una efectiva implementación de los principios FAIR [43].

8. Anonimización y otros aspectos ELSI

Los datos relativos a la salud de personas son datos especialmente protegidos por la legislación vigente, tanto española como europea, ya que afectan de forma directa al derecho fundamental a la privacidad de las personas físicas. El uso primario de estos datos, es decir, el uso de la información relativa a la salud de los pacientes para la prestación de la asistencia sanitaria y la preservación de la salud, tienen base de legitimación en la propia normativa de protección de datos, y en la específica sanitaria (Ley General de Sanidad [44] y Ley de Autonomía del Paciente [45]), y el consentimiento del paciente resulta implícito al mero hecho de la prestación asistencial.

Pero en el caso de un uso secundario de dicha información, la legitimación para su uso no es tan directa y evidente, aunque los beneficios derivados del análisis masivo de datos de salud para la investigación y la mejora de los tratamientos de los pacientes sean indiscutibles.

La definición de usos de la historia clínica se encuentra en los apartados 1,2,3 y del artículo 16 de la ley de autonomía del paciente:

1. La historia clínica es un instrumento destinado fundamentalmente a garantizar una asistencia adecuada al paciente.
2. Cada centro establecerá los métodos que posibiliten en todo momento el acceso a la historia clínica de cada paciente por los profesionales que le asisten
3. El acceso a la historia clínica con fines judiciales, epidemiológicos, de salud pública, de investigación o de docencia, se rige por lo dispuesto en la legislación vigente en materia de protección de datos personales, y en la Ley General de Sanidad [44], y demás normas de aplicación en cada caso. (entre ellas la ley de investigación clínica)
4. El personal de administración y gestión de los centros sanitarios sólo puede acceder a los datos de la historia clínica relacionados con sus propias funciones.

Según los precedentes existentes el apartado 1 y el 4 representan un uso primario. El apartado 3 (judiciales, epidemiológicos, salud pública, investigación y docencia) representa un uso secundario.

Esta nueva infraestructura podría representar un cambio de paradigma en cuanto a la asistencia sanitaria y el soporte a la toma de decisiones en tiempo real basadas en datos, de ser acompañado de la correspondiente normativa y regulación, podríamos ofrecer estos servicios de uso primario no sólo a gerentes sino también a clínicos.

En el plano de la protección de datos, la principal referencia es el Reglamento Europeo (UE) 2016/679 de Protección de Datos (RGPD), donde quedan reguladas de forma general las garantías y excepciones aplicables al tratamiento con fines de investigación científica, las circunstancias en las que es posible el tratamiento de datos personales de salud y también las excepciones a la normativa. Este tratamiento ha sido específicamente desarrollado en la normativa nacional de privacidad, la Ley Orgánica 3/2018, de 5 de diciembre, de Protección

de Datos Personales y garantía de los derechos digitales (LOPDGDD) que, por un lado regula las condiciones específicas en las que es posible un uso de datos personales de salud en la investigación biomédica y, por otro lado, establece la obligatoriedad para las entidades del sector público de implementar las medidas de índole técnica y organizativa de seguridad conformes con el Real Decreto 3/2010, de 8 de enero, por el que se regula el Esquema Nacional de Seguridad (ENS) en el ámbito de la Administración Electrónica.

Desde esta misma óptica de la seguridad de la información, se contempla la integración de las medidas del ENS con las requeridas para el cumplimiento del Real Decreto 43/2021, de 26 de enero, por el que se desarrolla el Real Decreto-ley 12/2018, de 7 de septiembre, de seguridad de las redes y sistemas de información, y que define la transposición de la Directiva NIS de ámbito europeo al ordenamiento jurídico español.

No obstante, toda la normativa legal detallada aquí hace referencia expresa al tratamiento de datos personales de salud, y por ende, de datos identificados o identificables. Sin embargo, la redacción del Reglamento General de Protección de Datos abre ciertas puertas al uso secundario de los datos de salud para investigación sanitaria, siempre que se cumplan una serie de condiciones. De acuerdo con el artículo 9 del RGPD, el tratamiento de datos especialmente protegidos, como es el caso de los datos de salud, no es posible sin una base de legitimación suficiente, que de forma general se basa en el consentimiento informado, libre y explícito por parte del paciente. Sin embargo, el apartado 9.2 de dicho artículo establece ciertas excepciones a este principio, entre las que cabe destacar:

i) el tratamiento es necesario por razones de interés público en el ámbito de la salud pública, como la protección frente a amenazas transfronterizas graves para la salud, o para garantizar elevados niveles de calidad y de seguridad de la asistencia sanitaria y de los medicamentos o productos sanitarios, sobre la base del Derecho de la Unión o de los Estados miembros que establezca medidas adecuadas y específicas para proteger los derechos y libertades del interesado, en particular el secreto profesional

j) el tratamiento es necesario con fines de archivo en interés público, fines de investigación científica o histórica o fines estadísticos, de conformidad con el artículo 89, apartado 1, sobre la base del Derecho de la Unión o de los Estados miembros, que debe ser proporcional al objetivo perseguido, respetar en lo esencial el derecho a la protección de datos y establecer medidas adecuadas y específicas para proteger los intereses y derechos fundamentales del interesado.

Es decir, la obtención del consentimiento podrá quedar exceptuada en aquellos supuestos en los que el tratamiento de datos se utilice con la finalidad de investigación biomédica, no sea posible la identificación del sujeto porque se han anonimizado sus datos, siempre previa autorización del Comité Ético de investigación, así como en aquellos casos en los que se traten los datos de salud en una investigación biomédica ulterior compatible con la inicial para la que se obtuvo el consentimiento. Se podrán tratar datos de salud, sin el consentimiento del

paciente, con fines de interés público en el ámbito de la salud pública y con fines de investigación científica, respectivamente; y de acuerdo al art. 89 RGPD, cabe realizar el tratamiento de datos de salud con fines de archivo en interés público, fines de investigación científica o histórica o fines estadísticos, sin la necesidad de obtener el consentimiento, siempre que se ofrezcan las garantías adecuadas para los derechos y libertades de los interesados. Entre estas garantías destacan todas aquellas medidas técnicas y organizativas necesarias para cumplir con el principio de minimización de los datos personales (seudonimización, por ejemplo) que permita alcanzar tales fines a través de un tratamiento ulterior sin la identificación del interesado.

En el mismo sentido, la disposición adicional decimoséptima de la Ley Orgánica de Protección de Datos y Garantía de los Derechos Digitales establece que:

> *Se considera lícito el uso de datos personales seudonimizados con fines de investigación en salud y, en particular, biomédica.*
>
> *El uso de datos personales seudonimizados con fines de investigación en salud pública y biomédica requerirá:*
>
> *1) Una separación técnica y funcional entre el equipo investigador y quienes realicen la seudonimización y conserven la información que posibilite la reidentificación.*
>
> *2) Que los datos seudonimizados únicamente sean accesibles al equipo de investigación cuando:*
>
> *a) Exista un compromiso expreso de confidencialidad y de no realizar ninguna actividad de reidentificación.*
>
> *b) Se adopten medidas de seguridad específicas para evitar la reidentificación y el acceso de terceros no autorizados.*
>
> *Podrá procederse a la reidentificación de los datos en su origen, cuando con motivo de una investigación que utilice datos seudonimizados, se aprecie la existencia de un peligro real y concreto para la seguridad o salud de una persona o grupo de personas, o una amenaza grave para sus derechos o sea necesaria para garantizar una adecuada asistencia sanitaria.*

8.1. Seudonimización, k-anonimidad, y privacidad diferencial

A continuación se listan algunos conceptos que permitirán acometer las tareas de anonimización con mayor claridad.

Dato anonimizado o irreversiblemente disociado: Dato que no puede asociarse a una persona identificada o identificable por haberse destruido el nexo con toda información que identifique al sujeto, o porque dicha asociación exige un esfuerzo no razonable, entendiendo portal el empleo de una cantidad de tiempo, gastos y trabajo desproporcionados.

Dato seudonimizado: Dato sin información explícita que pueda conducir a la identificación de la persona identificada o identificable pero que puede asociarse a dicha persona por existir un nexo con la información que identifique al sujeto.

Conviene resaltar que los datos a los que tendrán acceso los usuarios de este repositorio serán datos anonimizados, se argumenta que desde el punto de vista del usuario que accede al repositorio siempre existirá una separación técnica y funcional respecto al equipo que ha realizado la anonimización, siendo dicho equipo el único que conserva la información que posibilita la reidentificación; de esta forma la capacidad de reidentificar a un paciente concreto por parte del usuario del repositorio se minimiza. Esta probabilidad se calculará, en todo caso, de una forma objetiva usando herramientas de **k-anonimización** que nos indicarán de manera numérica las probabilidades de reidentificación, pudiendo hacer estimaciones que nos permitan establecer umbrales de seguridad en cuanto a mínima posibilidad de reidentificación.

Según la AEPD [47] la **k-anonimidad** es una propiedad de los datos anonimizados que permite cuantificar hasta qué punto se preserva la anonimidad de los sujetos presentes en un conjunto de datos en el que se han eliminado los identificadores. Dicho de otro modo, es una medida del riesgo de que agentes externos puedan obtener información de carácter personal a partir de datos anonimizados.

Se dice que un individuo es k-anónimo dentro del conjunto de datos en el que se encuentra incluido si, y sólo si, para cualquier combinación de los atributos cuasi-identificadores asociados, existen al menos otros k − 1 individuos que comparten con él los mismos valores para esos mismos atributos. Hay que tener en cuenta que la k-anonimidad no se centra en los atributos sensibles de los registros, sino en los atributos cuasi-identificadores que pueden permitir la vinculación.

Estas técnicas, junto con otros métodos relacionados, como la introducción de ruido estadístico, la agregación o la minimización de datos, permiten garantizar la **privacidad diferencia**l [46]. Esta disciplina, relativamente reciente, trata de estimar de forma objetiva, mediante formulación matemática, el nivel de privacidad de cada registro en un conjunto de datos. Esta privacidad se ve especialmente comprometida a medida que vamos construyendo conjuntos de datos más extensos, pues la probabilidad de reidentificar a un paciente anonimizado crece a medida que se incrementa la información que tenemos de él en el conjunto de datos.

La minimización de datos, es decir, extraer y utilizar solo los datos necesarios para cada estudio en concreto, el uso de datos agregados siempre que no sea imprescindible el uso de datos individuales, la seudonimización de los registros, la generalización multinivel –por ejemplo, utilizando edad o tramo quinquenal en vez de fecha de nacimiento-, la utilización de intervalos relativos de tiempo en lugar de tiempos absolutos, o la introducción de cierta cantidad de ruido estadístico en algunas variables, permiten incrementar la privacidad diferencial de los registros, reducir la probabilidad de reidentificación de los pacientes, y lo

que es más importante, evitar que alguien acceda a información personal referida a terceros a la que no debería tener acceso.

En el proceso de la seudonimización de los registros clínicos, y habida cuenta de que los sistemas de información clínicos en general se construyen en la práctica mediante la agregación lógica de múltiples sistemas, tanto centrales como departamentales, es muy frecuente que la información clínica de cada paciente se encuentre distribuida entre múltiples aplicativos y bases de datos. A la hora de integrar toda esa información en un único repositorio para uso secundario es importante tener en cuenta algunos aspectos que tienen implicaciones tanto en la privacidad de los datos como en su calidad de la información:

- Realizar un correcto enlazado de los datos de cada paciente entre distintas fuentes, así como los datos de un paciente en distintos cortes temporales. Para ello es imprescindible que el proceso de seudonimización no dependa de cada fuente de información, sino que sea un proceso único central a nivel de la organización que integra toda la información, sea ésta un hospital, un consorcio de hospitales, un servicio de salud autonómico o nacional, o cualquier otro proveedor de datos.
- Utilizar mecanismos y reglas de control de calidad que eviten la redundancia de información, por ejemplo recogiendo repetidamente un mismo diagnóstico recogido en más de un sistema de información, bien mediante la identificación de episodios duplicados, o por coincidencia temporal.
- Propagar al repositorio seudonimizado aquellos eventos administrativos que modifiquen la imputación de información clínica a un paciente, como son los procesos de fusión de pacientes o de episodios clínicos.

8.2. De-identificación / anonimización en informes clínicos

Según la Regulación Europea (EU) 2016/67, los informes clínicos y registros sanitarios necesitan estar, al menos, seudonimizados como paso previo a su reutilización para usos distintos a la asistencia sanitaria individual. Estos textos médicos contienen información relevante para la investigación, pero también detalles personales de pacientes y profesionales sanitarios que pueden poner en peligro su privacidad. La mayoría de trabajos sobre de-identificación de textos médicos están centrados en el inglés pero, dada la existencia de repositorios nacionales en castellano u otros idiomas estatales, resulta de interés hacer uso de estos datos para adaptar las soluciones ya existentes a los idiomas comunitarios.

La de-identificación de textos tradicionalmente se ha abordado mediante la aplicación de técnicas de reconocimiento de patrones y expresiones regulares, que dependen del idioma. Además, el contexto de una palabra y los errores tipográficos provocan que estos métodos sean propensos a errores cuando se trabaja con textos médicos y textos no corregidos, como por ejemplo los registros sanitarios. Una alternativa al reconocimiento de patrones es la aplicación de redes neuronales artificiales diseñadas para Procesamiento de Lenguaje Natural (NLP, Natural Language Processing). Este enfoque no tiene un bajo rendimiento

cuando trata con errores tipógraficos y puede incorporar el contexto de una palabra como característica, diferenciando entre “Parkinson” cuando se usa como apellido y cuando se usa como nombre de enfermedad. Aunque el NLP también depende del idioma, se puede adaptar más fácilmente que un conjunto de reglas de reconocimiento de patrones a otro idioma. La principal desventaja de esta tecnología es que requiere de un corpus relativamente grande de texto anotado para poder entrenar las redes neuronales.

Debido a la naturaleza de los proyectos internacionales, la recomendación es utilizar un enfoque de NLP:

- Es una metodología robusta para informes médicos en texto libre, sin encabezados de metadatos e incluyendo errores tipográficos.
- El modelo de NLP se puede adaptar para detectar nuevas categorías o entidades de interés, tanto para borrarlas como para preservarlas si son de interés clínico o científico.
- Aunque requiere un corpus de entrenamiento anotado, una cantidad de entre 500 y 1000 informes médicos es suficiente para generar un modelo robusto de NLP.
- No es obligatorio disponer de una infraestructura informática elevada ni para el entrenamiento de modelos ni para el despliegue de de-identificación.

Para garantizar la fácil aplicación de la tecnología recomendada, se proponen métodos de de-identificación basados en el reconocimiento de entidades con nombre (NER) de una manera que no depende de los recursos de procesamiento de lenguaje natural ya existentes para cualquier idioma. El reconocimiento de entidades nombradas es una subtarea de NLP de extracción de información que se centra en la ubicación y clasificación de entidades nombradas incluidas en texto no estructurado en categorías predefinidas, como las categorías de información de salud protegida (PHI) definidas por HIPAA Ilustración 12. Metodología de DisMed para la anonimización de informes radiológicos). La metodología propuesta, denominada DiSMed [51] se ha testado en informes radiológicos españoles, y es fácilmente extensible al menos a otras lenguas romances y germánicas derivadas con un relativo pequeño conjunto de datos de entrenamiento anotado. Por tanto, sería aplicable a la mayoría de los informes médicos europeos.

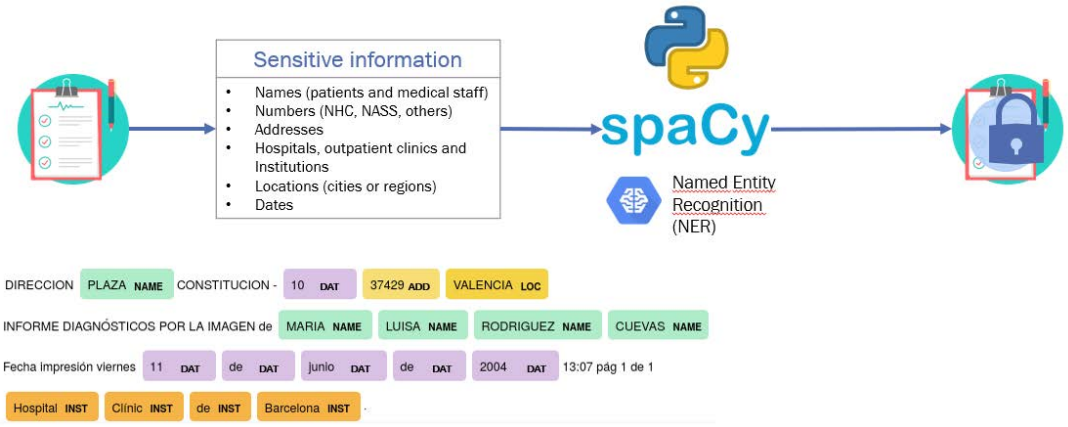

Ilustración 12. Metodología de DisMed para la anonimización de informes radiológicos

Independientemente de la tecnología aplicada para detectar palabras que contengan información sensible, tradicionalmente se han propuesto dos metodologías principales para de-identificar: (1) la supresión de palabras y (2) la sustitución de palabras por una etiqueta. Estas dos estrategias no tienen en cuenta que es imposible archivar un rendimiento perfecto, siendo las mejores estadísticas alrededor de un 3% de palabras de identificación que quedan en el texto después del etiquetado y borrado de entidades nombradas, lo que podría no ser suficiente para de-identificar por completo el texto. Las palabras restantes serían fácilmente reconocibles si se aplica una metodología de supresión de palabras o sustitución de palabras por una etiqueta. Para evitar la extracción de estas palabras, proponemos una metodología basada en la aleatorización de entidades nombradas, que se ha aplicado a la metodología DiSMed.

La metodología de aleatorización de entidades nombradas depende del idioma, pero se puede aplicar a cualquier idioma siempre que estén disponibles bases de datos públicas para las diferentes entidades nombradas detectadas. Si no están disponibles públicamente, es recomendable generarlos para aplicar la aleatorización de entidades nombradas a los informes médicos. Aunque generar estas bases de datos requiere mucho tiempo y esfuerzo, esta metodología garantiza la de-identificación y minimiza la fuga de datos de pacientes.

Para el caso concreto de informes radiológicos se presentará en el punto 2.5.3 del entregable E4.4.

9. Implementaciones de referencia

9.1. Casos de éxito en España

Arquitectura normalizada de datos clínicos para uso secundario en investigación: COVID-19

Este proyecto, que actualmente se encuentra en su fase final, tiene como objetivo consolidar en una base de datos OMOP información clínica y demográfica de los pacientes hospitalizados por COVID-19 en el Hospital Clínic de Barcelona (HCB) y el Hospital 12 de Octubre (H12O) de Madrid.

El primer paso fue la definición, junto a *stakeholders* clínicos, de las variables a incluir en el estudio. Para lograr escalabilidad en futuros proyectos y evitar la pérdida de granularidad inherente al uso de modelos comunes de datos como OMOP, se definió como paso previo la utilización de arquetipos basados en la norma UNE-EN ISO 13606 para el modelado de las variables. Posteriormente, se definieron de manera conjunta entre miembros de ambas instituciones los arquetipos que se utilizarían.

Ambas instituciones incluyeron datos provenientes de repositorios locales y de terceros. De esta forma, se pudieron utilizar tanto datos estructurados como no estructurados (por ejemplo, a partir de la extracción por procesamiento de lenguaje natural de entidades reconocidas en informes y cursos clínicos).

A partir de los datos de cada institución, se crearon extractos UNE-EN ISO 13606 con los datos de los pacientes, para luego transformarlos al CDM de OMOP y consolidarlos en una base de datos en el H12O.

Carga desde HCE en CRFs de investigación y Plataformas RWD en COVID-19

El Hospital Universitario 12 de Octubre (H12O) preparó, desde el inicio de la pandemia, su HCE para que los datos registrados fueran útiles en la asistencia de los pacientes COVID-19, así como en otros usos, denominadas secundarios, que incluyen actividades como la investigación sanitaria. Este proceso ha sido publicado en la Medical Informatics Europe Conference (MIE2021) [48]. Bajo este paradigma de reutilización de los datos asistenciales, se formalizó una metodología para la obtención de datos útiles en fines secundarios a partir de la HCE, esto es, sin que sea necesario realizar registro manual en sistemas específicos. Esta metodología fue publicada en marzo de 2021 en la revista “Journal of Biomedical Informatics”, de primer cuartil en el área de la informática médica [49]. Esta metodología ha permitido que el Hospital Universitario 12 de Octubre participe en diversos proyectos y

consorcios internacionales de datos de salud, lo cual no habría sido posible de requerir un registro manual y específico en cada uno de ellos.

Como parte de la aplicación de esta metodología, se realizó la definición de modelos de información estándares, basados en normas internacionales adoptadas por el Ministerio de Sanidad, que permiten que los datos sean registrados, intercambiados y combinados sin pérdida de significado. Estos modelos han sido registrados como propiedad intelectual bajo licencia Creative Commons:

- Modelo dual de observaciones de interés en COVID-19: https://www.safecreative.org/work/2102196969609-h12o-modelo-dual-de-observaciones-covid19.
- Modelo de arquetipos de observaciones de interés en COVID-19: https://www.safecreative.org/work/2102196969593-h12o-covid-19-observations-archetypes.

Así mismo, estos modelos de información desarrollados han sido adoptados por el **COVID-19 Data Portal** como conceptos de referencia para hacer posible el uso de la HCE en investigación [50].

El proyecto STOP-CORONAVIRUS

Desarrollado en el H12O, y financiado por el ISCIII, ha sido el proyecto pionero en la aplicación de la metodología de reutilización de la HCE, realizándose cargas de datos directamente desde la HCE al cuaderno de recogida de datos del proyecto, implementado en REDCap. Esto ha permitido optimizar el trabajo de las personas que trabajan en la gestión de los datos, pasando a ser registrados una única vez en la HCE para ser utilizados en múltiples propósitos. Así mismo, se ha realizado un análisis de calidad de los datos generados, midiendo la completitud y validez de los mismos, sobre el que se está escribiendo una publicación compartida entre el grupo de Ciencia de datos y el Servicio de Medicina Interna.

Del mismo modo, esta reutilización de la HCE ha permitido que el Hospital Universitario 12 de Octubre participe en diversos proyectos y consorcios internacionales de datos de salud, lo cual no habría sido posible de requerir un registro manual y específico en cada uno de ellos:

- **Consorcio ISARIC**. Consorcio internacional de enfermedades respiratorias agudas que al inicio de la pandemia lanzó un Case Report Form (CRF) para casos COVID-19 y un sistema REDCap para su cumplimentación por los diferentes nodos participantes. El Hospital 12 de Octubre propuso el envío de datos directamente desde la HCE, siendo el único que no realiza registro manual en el proceso.
- **Consorcio 4CE**. Consorcio internacional liderado por la Universidad de Harvard para la compartición de datos agregados de pacientes COVID-19 desde los diferentes nodos participantes, y la ejecución de análisis sobre el conjunto. El Hospital 12 de Octubre contribuye en el envío de datos, y ha participado en diferentes publicaciones en el marco de este proyecto.

- **Consorcio EHDEN**. Consorcio europeo (IMI2) para la implantación de una red federada de repositorios normalizados bajo el modelo OMOP. Cada participante implementa su propio repositorio de datos para ser utilizado en la ejecución de estudios distribuidos.El Hospital 12 de Octubre participa, junto a Atención Primaria de Madrid, en la implementación de un repositorio regional para investigación en COVID-19.
- **Plataforma TriNetX**. Plataforma de datos internacional, compuesta por una red federada de repositorios normalizados conformes al modelo i2b2, que permite la construcción de cohortes de pacientes y la ejecución de análisis sobre los datos almacenados sobre la misma. El Hospital 12 de Octubre ha sido la referencia europea en el modelado y carga de datos COVID-19, estos son, las diferentes pruebas diagnósticas y las alertas clínicas implementadas durante la pandemia.

Covid Data Save Lives

HM Hospitales ha puesto a libre disposición de la comunidad médica y científica internacional un dataset anonimizado con toda la información clínica disponible sobre los pacientes tratados en nuestros centros hospitalarios por el virus SARS-CoV-2. Este dataset clínico recoge las distintas interacciones en el proceso de tratamiento del COVID-19, incluyendo información pormenorizada sobre diagnósticos, tratamientos, ingresos, pasos por UCI, pruebas diagnósticas por imagen, resultados de laboratorio, alta o deceso, entre otros muchos registros.

OPEN DATA COVID

Sanitas pone a disposición de la comunidad científica y académica datos relativos a los pacientes COVID-19 que han sido ingresados en los centros médicos de la compañía. Esto incluye tanto datos demográficos como información clínica relevante que facilitará la investigación sobre el mayor desafío sanitario de los últimos 100 años. La información de los pacientes puesta a disposición de la sociedad por parte de Sanitas es segura y está anonimizada, para evitar así la identificación de ningún individuo. Este proyecto forma parte de Sanitas Data4Good, iniciativa de Open Data de Sanitas que nace con el fin de contribuir a la sociedad a través de los datos, especialmente en el ámbito de la salud y el bienestar.

BIMCV-COVID-19

La Society for Imaging Informatics in Medicine (SIIM), junto a la Fundació per al Foment de la Investigació Sanitària i Biomèdica de la Comunitat Valenciana (Fisabio) y la Sociedad Radiológica de Norteamérica (RSNA) organizaron un desafío para la detección y localización de neumonía por COVID-19 mediante el uso de la Inteligencia Artificial.Esta competición es

una continuación del proyecto que recibió financiación de la Conselleria de Innovación en la Llamada al Sistema de Innovación e Investigación.

El desafío científico se llevo a cabo a través de la plataforma Kaggle, y utilizo conjuntos de imágenes de radiografías convencionales de la Medical Imaging Data Resource Center (MIDRC) - RSNA International COVID-19 Open Radiology Database (RICORD) y del BIMCV-COVID-19 Dataset (Banco de Imagen Médica de la Comunidad Valenciana). El dataset fue entregado en formato MIDS (Medical Imaging Data Structure).

El desafío de detección y localización de COVID-19 se celebró en mayo del 2021 y los ganadores se presentaron en septiembre del año pasado en la Conferencia 2021 sobre inteligencia artificial en imágenes médicas (CMIMI).

9.2. Casos de éxito a nivel internacional

A nivel nacional e internacional, podemos encontrar algunas iniciativas de open data sanitario similares a la que aquí planteamos que están obteniendo una gran repercusión y reconocimiento, se listan en orden cronológico, según su fecha de aparición. A nivel Internacional, la base de datos clínica más usada (sobre todo para entrenar modelos de inteligencia artificial orientados a abordar preguntas clínicas) es MIMIC, MIMIC ha sido desarrollada por el Laboratory for Computational Physiology (LCP). El Laboratorio de Fisiología Computacional (LCP) del MIT, bajo la dirección del profesor Roger Mark, investiga sobre la mejora de la atención sanitaria a través de enfoques nuevos y refinados para la interpretación de datos. Algunos de los investigadores del grupo tienen formación médica; otros tienen formación en informática, ingeniería eléctrica, física o matemáticas; y otros tienen formación que abarca varias de estas disciplinas. La investigación actual del laboratorio incorpora la fisiología, la informática, la ingeniería y las matemáticas aplicadas. Utilizando enfoques modernos de modelado, procesamiento de señales, reconocimiento de patrones y aprendizaje automático, los investigadores del laboratorio desarrollan y perfeccionan métodos para analizar datos -por ejemplo, de pacientes en unidades de cuidados intensivos- y para generar modelos predictivos que ayuden a la atención de los pacientes. La investigación actual del laboratorio comprende dos grandes proyectos, el proyecto MIMIC, parte de PhysioNet.

Estados Unidos

Casos de éxito con mayor trayectoria del LCP (aunque se han desarrollado dentro de EEUU, lo han hecho conforme al RGPD)

- MIMIC (Medical Information Mart in Intensive Care) es el resultado de la colaboración entre el Beth Israel Deaconess Medical Center y el Instituto Tecnológico de Massachusetts (MIT). Contiene datos de más de 50.000 pacientes entre los años 2001 y 2012 y es accesible previa firma de un acuerdo de uso.

- eICU, la base de datos de investigación colaborativa eICU está poblada con datos de una combinación de muchas unidades de cuidados críticos de todo el territorio

continental de Estados Unidos. Los datos de la base de datos colaborativa abarcan a los pacientes que ingresaron en las unidades de cuidados críticos en 2014 y 2015.

Europa

Ha habido otros repositorios abiertos colaborativos clínicos que tomando como referencia la labor del LCP han surgido otras a nivel europeo:

- AMDS; El Consorcio de Hospitales Universitarios de Amsterdam (Amsterdam UMC) es el primero en Europa en hacer disponibles los datos de pacientes que han estado hospitalizados en la unidad de cuidados intensivos para la investigación y mejorar los servicios de salud. Al hacerlo,los pacientes individuales no son razonablemente identificables. Casi mil millones de puntos de datos estarán disponibles en total, la mayoría de ellos proviene del equipo de monitoreo. Gracias a esta gran cantidad de datos (Big Data), los médicos e investigadores de todo el mundo podrán desarrollar algoritmos con técnicas de inteligencia artificial como aprendizaje automático (Machine Learning). Estas técnicas garantizarán que futuros pacientes en la unidad de cuidados intensivos reciban un tratamiento adecuado y aún más rápido que hasta ahora. Y esto es absolutamente necesario puesto que actualmente el 30% de los pacientes admitidos en la unidad de cuidados intensivos mueren aun recibiendo cuidados médicos óptimos. Eso significa cientos de muertos en Europa cada día.
- The Dutch ICU Data Warehouse Es un proyecto promovido por el Consorcio de Hospitales Universitarios de Amsterdam (Amsterdam UMC) y en el que las unidades de cuidados intensivos de los Países Bajos han comenzado a colaborar mediante la compartición de grandes cantidades de datos recogidos de forma rutinaria. Actualmente, 38 UCIs están participando en el proyecto, estimando que por cada paciente hay disponible más de 30.000 data points por día. El proyecto fue iniciado por la Asociación Europea para la Medicina de Cuidados Intensivos (European Society of Intensive Care Medicine - ESICM) con el establecimiento de su sección de Data Science y cuenta con el apoyo de la asociación holandesa de cuidados intensivos (NVIC).
- HiRID es un conjunto de datos abierto que contiene información de 33.000 pacientes admitidos en el Departamento de Medicina Intensiva del Hospital Universitario de Berna (Suiza). Este proyecto fue promovido de forma colaborativa con el Swiss Federal Institute of Technology (ETH) Zürich.
- El FAIR COVID Health Portal de Dinamarca, es una experiencia real de ciencia abierta y de cumplimiento efectivo de los principios FAIR en Europa. En colaboración con GOFAIR Foundation, ha aplicado la herramienta CEDAR y los talleres M4M para la metadatación de la colección de datos del portal.
- HONEUR. La iniciativa en red federada Haematological Outcomes Network in Europe (HONEUR), constituye una colaboración de diversos países de Europa para la investigación basada en Real-World Data de patologías como el Mieloma Múltiple en los que las intituciones participantes deben disponer de la HCE en un formato común

de datos como OMOP. En España participan el Hospital 12 de Octubre y el Hospital del Mar y su Instituto de Investigaciones Médicas (IMIM).

- EHDEN COVID-19 Rapid Colaboration. Proyecto promovido por el consorcio EHDEN (European Health Data Evidence Network) en el que participan más de 20 entidades en España, se han implementado e integrado en formato OMOP Common Data Model (CDM) historias clínicas y otros respositorios de datos como el Registro de Tumores, haciendo posible la colaboración internacional en el desarrollo de investigación en red sobre COVID-19, mediante la utilización de herramientas analíticas y de visualización desarrolladas en el consorcio internacional Observational Health Data Sciences and Informatics (OHDSI).

10. Conclusiones

Se ha expuesto a lo largo de las sucesivas secciones del presente documento una extensa relación de normas, estándares y modelos de datos que están siendo utilizados tanto en el conjunto de sistemas de información asistenciales (uso primario) como en el contexto de uso secundario y mixto (investigación, ensayos clínicos y registros de patologías específicas), por parte de los distintos servicios públicos de salud de España, así como por instituciones públicas e institutos de investigación biomédica, tanto a nivel nacional como internacional. Este catálogo permite tener un mapa más o menos completo de las distintas soluciones existentes en la actualidad como punto de partida para la construcción de un espacio de datos de salud para investigación en España, por lo menos en lo que respecta a dato clínico.

Es cierto que en el ámbito del uso primario (asistencial), la disparidad de sistemas, normas y modelo, e incluso la carencia de éstos, está muy generalizada. Pero cada vez existe más conciencia de que la explotación masiva y sistemática de la información clínica recogida en estos sistemas es una fuente de información inestimable para la investigación biomédica y el desarrollo de políticas sanitarias, y por tanto de un enorme impacto final en la salud de los pacientes y en la sostenibilidad del sistema sanitario. Por esta razón, cada vez más los sistemas de información sanitario se van construyendo desde el diseño sobre la base de modelos, estándares y ontologías que faciliten, no solo la actividad asistencial, sino también la reutiliación de esa información.

Los distintos modelos y estándares recogidos en este documento cubren el espacio de soluciones necesarias para la interoperabilidad de datos sanitarios en uso secundario, pero no de una forma perfecta ni probablemente completa. Hay soluciones que se solapan total o parcialmente con otras en este espacio de soluciones. Hay estándares de nicho, que responden a temáticas o problemáticas muy concretas, mientras que otros tienen una vocación mucho más universal, lo cual quizá los haga menos precisos a la hora de abordar algunas cuestiones concretas. La tipología de los casos de uso que se propongan en el futuro, también será condicionante a la hora de seleccionar unas y otras soluciones. No es lo mismo abordar proyectos de aprendizaje federado, en los que los algoritmos se ejecutan de forma distribuida en nodos que, necesariamente, tendrán que tener modelos de datos totalmente

homogéneos, que proyectos de agregación centralizada, en la que el modelo de datos para el análisis tiene una sola ubicación, pero en la que es preciso establecer mecanismos para la transferencia de los datos, y sobre todo para lograr un perfecto mapeo conceptual de este modelo central con los distintos modelos existentes en cada uno de los proveedores de datos.

En un trabajo posterior, será necesario analizar más a fondo algunos de estos modelos y estándares, estudiarlos en el contexto de los distintos casos de uso, evaluar su potencialidad, completitud, cobertura de implantación y facilidad de uso, con objeto de poder proponer una o varias pilas de soluciones que puedan ser adoptadas por los distintos nodos que participan en el proyecto IMPaCT-Data. Deberá tenerse en cuenta además que IMPaCT-Data no contempla el uso aislado de datos clínicos, sino su integración con imagen médica para caracterizar fenotipos y datos genómicos, por lo que las pilas de soluciones propuestas deberán ser compatibles con las soluciones propuestas en las respectivas tareas, y poder construir así una solución global consistente y eficaz.

Referencias

[1] Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. Published 2016 Mar 15. doi:10.1038/sdata.2016.18

[2] FORCE 11. The FAIR data principles. https://www.force11.org/group/fairgroup/fairprinciples

[3] S. Hodson, S. Jones, S. Collins, F. Genova, N. Harrower, L. Looksonen, D. Mietchen, R. Petrauskaité, P. Wittenburg. (2018) Turning FAIR into reality. Final Report and Action Plan from the European Commission Expert Group on FAIR Data. Directorate-General for Research and Innovation, European Commission. Brussels.

[4] Strategic Research and Innovation Agenda (SRIA) of the European Open Science Cloud (EOSC). https://www.eosc.eu/sites/default/files/EOSC-SRIA-V1.0_15Feb2021.pdf

[5] European Commission. Cost of not having FAIR research data - Cost-Benefit analysis for FAIR research data. 2018. https://op.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1

[6] Página web Proyecto FAIR4Health https://www.fair4health.eu/

[7] Espacio Europeo de Datos Sanitarios https://ec.europa.eu/health/ehealth/dataspace_es

[8] RDA FAIR Data Maturity Model WG https://www.rd-alliance.org/groups/fair-data-maturity-model-wg

[9] RDA https://www.rd-alliance.org/groups/raising-fairness-health-data-and-health-research-performing-organisations-hrpos-wg

[10] RDA Raising in health data and health research performing organisations (HRPOs) WG FAIRness https://www.rd-alliance.org/groups/raising-fairness-health-data-and-health-research-performing-organisations-hrpos-wg

[11] Vesteghem C, Brøndum RF, Sønderkær M, Sommer M, Schmitz A, Bødker JS, Dybkær K, El-Galaly TC, Bøgsted M. Implementing the FAIR Data Principles in precision oncology: review of supporting initiatives. Brief Bioinform. 2020 May 21;21(3):936-945. doi: 10.1093/bib/bbz044. PMID: 31263868; PMCID: PMC7299292. https://pubmed.ncbi.nlm.nih.gov/31263868/

[12] Datamed: Buscador de datos biomédicos https://datamed.org/

[13] GO FAIR FAIRification Process. https://www.go-fair.org/fair-principles/fairification-process/

[14] Sinaci, A. A. et al. (2020). From Raw Data to FAIR Data: The FAIRification Workflow for Health Research. Methods of Information in Medicine, 59(S 01), e21-e32.

[15][15] (Schultes E., Magagna B., Hettne K.M., Pergl R., Suchánek M., Kuhn T. (2020) Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence. In: Grossmann G., Ram S. (eds) Advances in Conceptual Modeling. ER 2020. Lecture Notes in Computer Science, vol 12584. Springer, Cham. https://doi.org/10.1007/978-3-030-65847-2_13).

[16] Sasse, J., Darms, J., & Fluck, J. (2022). Semantic Metadata Annotation Services in the Biomedical Domain—A Literature Review. Applied Sciences, 12(2), 796.

[17] https://odm.uni-muenster.de/

[18] https://rightfield.org.uk/

[19] Pathak, J., Wang, J., Kashyap, S., Rongling, L., Masys, D. R., & Chute, C. G. (2010). eleMAP: an online tool for harmonizing data elements using standardized metadata registries and biomedical vocabularies. Am Med Inform Assoc, 1214.

[20] https://more.metadatacenter.org/tools-training/cedar-metadata-tools

[21] https://more.metadatacenter.org/tools-training/outreach/sap-cedar-based-pipeline-semantic-annotation-biomedical-metadata

[22] https://uima.apache.org/downloads/sandbox/ConceptMapperAnnotatorUserGuide/ConceptMapperAnnotatorUserGuide.html

[23] GitHub. D2Refine URL: https://github.com/caCDE-QA/D2Refine [accessed 2018-04-30] [WebCite Cache ID 6z4YIIQV9]

[24] OpenRefine. OpenRefine URL: http://openrefine.org/index.html [accessed 2018-07-11] [WebCite Cache ID 6z5n0mnNq]

[25] Wiktorin, T. Semantische Annotation im Gesundheitswesen—Prototypische Entwicklung und Evaluation eines Kollaborativen Werkzeugs zur Semantischen Annotation Medizinischer Daten. Master's Thesis, Hochschule Bonn-Rhein-Sieg, University of Applied Sciences, University Hospital Bonn (UKB), Bonn, Germany, 2021.

[26] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe

[27] Norlin L, Fransson MN, Eriksson M, Merino-Martinez R, Anderberg M, Kurtovic S, Litton JE. A Minimum Data Set for Sharing Biobank Samples, Information, and Data: MIABIS. Biopreserv Biobank. 2012 Aug;10(4):343-8. doi: 10.1089/bio.2012.0003. PMID: 24849882.

[28] https://www.obiba.org/pages/products/mica/

[29] https://doi.org/10.1101/2021.08.13.21262023

[30] https://www.sciencedirect.com/science/article/abs/pii/S0221036306876766

[31] HPO: https://hpo.jax.org/app/

[32] I2B2: https://www.i2b2.org/

[33] González L, Pérez-Rey D, Alonso E, Hernández G, Serrano P, Pedrera M, Gómez A, De Schepper K, Crepain T, Claerhout B. Building an I2B2-Based Population Repository for Clinical Research. Stud Health Technol Inform. 2020 Jun 16;270:78-82. doi: 10.3233/SHTI200126. PMID: 32570350.

[34] https://www.ohdsi.org/data-standardization/the-common-data-model/

[35] https://community.i2b2.org/wiki/display/getstarted/3.7+Metadata+Tables

[36] Birger Haarbrandt, Erik Tute and Michael Marschollek. Automated population of an i2b2 clinical data warehouse from an openEHR-based data repository. Journal of Biomedical Informatics. https://doi.org/10.1016/j.jbi.2016.08.007

[37] Solbrig HR, Hong N, Murphy SN, Jiang G. Automated Population of an i2b2 Clinical Data Warehouse using FHIR. AMIA ... Annual Symposium proceedings. AMIA Symposium. 2018 ;2018:979-988. PMID: 30815141; PMCID: PMC6371332.

[38] https://www.researchallofus.org/data-tools/methods/

[39] https://www.imi.europa.eu/projects-results/project-factsheets/ehr4cr

[40] FAIR Data Maturity Model Working Group. (2020). FAIR Data Maturity Model. Specification and Guidelines (1.0). https://doi.org/10.15497/rda00050

[41] Schultes E., Magagna B., Hettne K.M., Pergl R., Suchánek M., Kuhn T. (2020) Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence. In: Grossmann G., Ram S. (eds) Advances in Conceptual Modeling. ER 2020. Lecture Notes in Computer Science, vol 12584. Springer, Cham. https://doi.org/10.1007/978-3-030-65847-2_13

[42] Hana Pergl Sustkova, Kristina Maria Hettne, Peter Wittenburg, Annika Jacobsen, Tobias Kuhn, Robert Pergl, Jan Slifka, Peter McQuilton, Barbara Magagna, Susanna-Assunta Sansone, Markus Stocker, Melanie Imming, Larry Lannom, Mark Musen, Erik Schultes. The FAIR Convergence Matrix: Optimizing the reuse of existing FAIR-related resources. Data Intelligence 2(2020), 158–170. doi: 10.1162/dint_a_00038

[43] R.D. Kush, D. Warzel, M.A. Kush, A. Sherman, E.A. Navarro, R. Fitzmartin, F. Pétavy, J. Galvez, L.B. Becnel, F.L. Zhou, N. Harmon, B. Jauregui, T. Jackson, L. Hudson, FAIR data sharing: The roles of common data elements and harmonization, Journal of Biomedical Informatics, Volume 107, 2020, 103421, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2020.103421

[44] Ley 14/1986 General de Sanidad. https://www.boe.es/buscar/act.php?id=BOE-A-1986-10499

[45] Ley 41/2002 de Autonomía del paciente. https://www.boe.es/buscar/act.php?id=BOE-A-2002-22188

[46] Cynthia Dwork and Aaron Roth, The Algorithmic Foundations of Differential Privacy. Foundations and TrendsR in Theoretical Computer Science Vol. 9, Nos. 3–4 (2014) 211–407

[47] https://www.aepd.es/sites/default/files/2019-09/nota-tecnica-kanonimidad.pdf

[48] https://ebooks.iospress.nl/doi/10.3233/SHTI210114

[49] https://www.sciencedirect.com/science/article/pii/S1532046421000265?via%3Dihub.

[50] https://www.covid19dataportal.es/health-variables/

[51] Pérez-Díez I., Pérez-Moraga R., López-Cerdán A., Caparrós Redondo M., Salinas-Serrano J.M., de la Iglesia-Vayá M. (2020). De-identifying Spanish medical texts – Named Entity Recognition applied to radiology reports. medRxiv, 2020.04.09.20058958.

[52] BioPortal: https://bioportal.bioontology.org/

[53] SNOMED: https://www.snomed.org/

[54] LOINC: https://loinc.org/

[55] CIE: https://www.who.int/standards/classifications/classification-of-diseases

[56] CIAP: https://icpc.global/

[57] Orphanet: https://www.orpha.net/

[58] OMIM, Online Mendelian Inheritance in Man: https://www.omim.org/

[59] Descriptores de Ciencia de la Salud: https://decs.bvsalud.org/E/homepagee.htm

Acrónimos y Abreviaturas

AEPD	Agencia Española de Protección de Datos
API	Application Programming Interface
ATC	Anatomic, Therapeutic, Chemical (clasificación de principios activos farmacéuticos)
CDISC	Clinical Data Interchange Standards Consortium
CDM	Common Data Model
CESSDA	Consortium of European Social Science Data Archives
CIAP	Codificación Internacional de Atención Primaria (ICPC)
CIE	Codificación Internacional de Enfermedades (ICD)
CMBD	Conjunto Mínimo Básico de Datos
DCAT-AP	Data Application profile for data portals in Europe
DICOM	Digital Image Communication
ELSI	Ethical, Legal and Social Issues
EOSC	European Open Science Cloud
FAIR	Findable, Accesible, Interoperable, Reusable
FHIR	Fast Health Interoperability Resources
FIP	FAIR Implementation Profile
HCE	Historia Clínica Electrónica
HIPAA	Health Insurance Portability and Accountability Act
HL7	Health Level 7
IMPaCT	Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología
IMPaCT-Data	Programa de ciencia de datos de IMPaCT
INSPIRE	Infrastructure for Spatial Information in Europe
JSON	Javascript Object Notation
JSON-LD	Javascript Object Notation - Linked Data
LOINC	Logical Observation Identifiers Names and Codes
LOPDGDD	Ley Orgánica de Protección de Datos y Garantía de Derechos Digitales
NLP	Natural Language Processing
OMOP	Observational Medical Outcomes Partnership
openEHR	Open Electronic Health Record
RDF	Resource Description Framework
REST	Representational State Transfer
RGPD	Reglamento General de Protección de Datos
RIM	Reference Information Model
SNOMED-CT	Systematic Nomenclature on Medicina – Clinical Terms
UMLS	Unified Medical Language System