GEOSTRATEGY

Guía de usuarios CDM de OMOP

Siglas, acrónimos y abreviaturas

CDM	Common data model (Modelo de datos común)
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
ETL	Extract, Transform and Load (Extracción, transformación y carga)
TI	Tecnología de la Información

Introducción

'Observational Health Data Sciences and Informatics (OHDSI)' es una red internacional de investigadores y bases de datos observacionales de salud, la cual tiene como objetivo crear y aplicar soluciones analíticas de datos de código abierto a una gran red de bases de datos clínicos para mejorar la salud y el bienestar humanos. Esta red de investigación surgió del 'Observational Medical Outcomes Partnership (OMOP)', que fue una asociación público-privada establecida en los EE. UU. para informar el uso apropiado de las bases de datos observacionales de atención médica para estudiar los efectos de los productos médicos [1].

OHDSI busca lograr el principio de inclusión mediante una colaboración abierta, por lo tanto, personas de todo el mundo participan en la red, brindando tiempo, datos o financiamiento. Actualmente, cuenta con más de 2000 colaboradores en 74 países y registros de salud de aproximadamente 810 millones de pacientes únicos de todo el mundo [1].

Para lograr la integración entre los datos de los múltiples participantes de la red, OHDSI utiliza el CDM desarrollado por OMOP. Este, permite el análisis sistemático de bases de datos observacionales dispares. El concepto detrás de este enfoque es transformar la información contenida en esas bases de datos en un formato común (modelo de datos), así como una representación común (terminologías, vocabularios, esquemas de codificación), y luego realizar análisis sistemáticos utilizando una biblioteca de rutinas analíticas estándar que se han escrito basadas en el formato común [2].

El CDM presenta las siguientes características [2]:

- Garantiza que los métodos de investigación puedan aplicarse sistemáticamente para producir resultados significativamente comparables y reproducibles.
- Está diseñado para apoyar una amplia gama de actividades de investigación observacional.
- Es un modelo relacional centrado en la persona, con dominios que incluyen datos demográficos, períodos de observación, exposición a drogas, ocurrencia de condiciones, procedimientos, visitas y observaciones clínicas
- Además de estandarizar la estructura de los datos, a través de los Vocabularios Estandarizados también estandariza la representación del contenido.
- Los códigos fuente se mantienen en el CDM para una trazabilidad completa.

En la Fig. 1 se muestra el esquema general de la versión actual (v5.4) del CDM de OMOP. Allí se observa las 39 tablas que tiene, la sección a la que pertenece cada tabla y algunas conexiones entre tablas. Es importante resaltar, que este es un modelo centrado en la persona, por ello todas las tablas van a tener como centro y se van a conectar a la table PERSON.

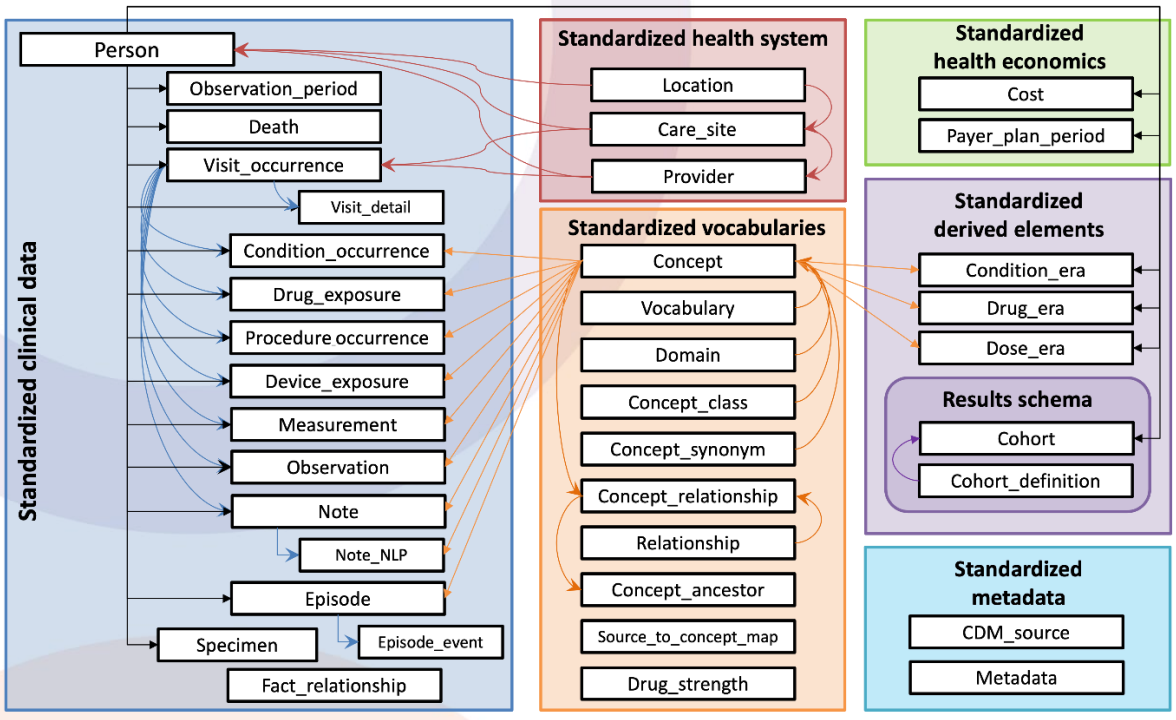

Fig. 1. Esquema de La versión actual (v5.4) de CDM de OMOP.
Nota: fuente https://ohdsi.github.io/CommonDataModel/index.html

Acceso a los datos del hospital

Para acceder a los datos del hospital, es necesario tener firmado un acuerdo de confidencialidad con el hospital, debido a que se está tratando con datos personales de los pacientes y se debe garantizar la seguridad de estos. Una vez se tenga esto listo, los pasos a seguir son:

1. Instalar dos softwares: 'FortiClient', el cual permite conectarse a la VPN del hospital y 'SQL Server 2014' para poder visualizar y manejar la base de datos. Además, se debe instalar en el celular la aplicación 'FortiToken Mobile'.
2. Con los softwares instalados, se ingresa primero a 'FortiClient' y se selecciona ACCESO REMOTO, allí se ingresa el usuario y contraseña que deben ser proporcionados por el hospital. Luego, se ingresa el token que es el número de 6 dígitos proporcionado por la aplicación 'FortiToken Mobile'. En la Fig. 2 se muestra cómo es el ingreso con el software y los campos a llenar.

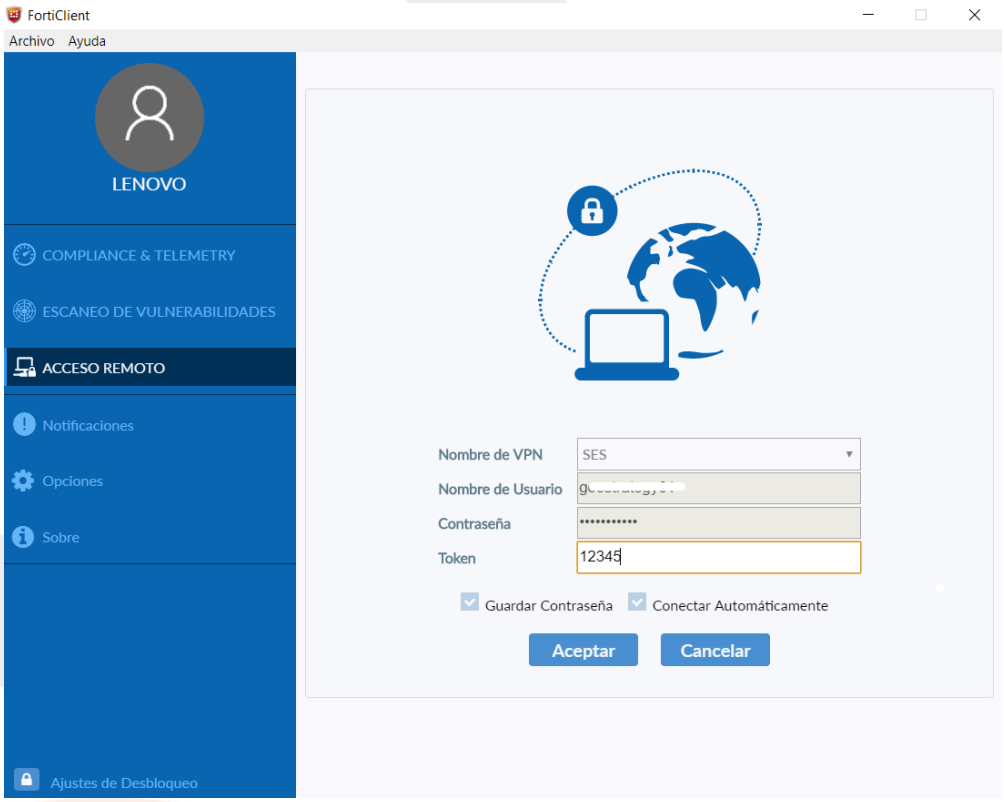

Fig. 2. Interfaz gráfica de FortiClient para ingresar a la VPN del hospital.

3. Cuando se esté conectado a la VPN del hospital, se ingresa al 'SQL Server 2014 Management Studio', en donde se introduce el nombre del servidor, usuario y contraseña también proporcionados por el hospital como se muestra en la Fig. 3.

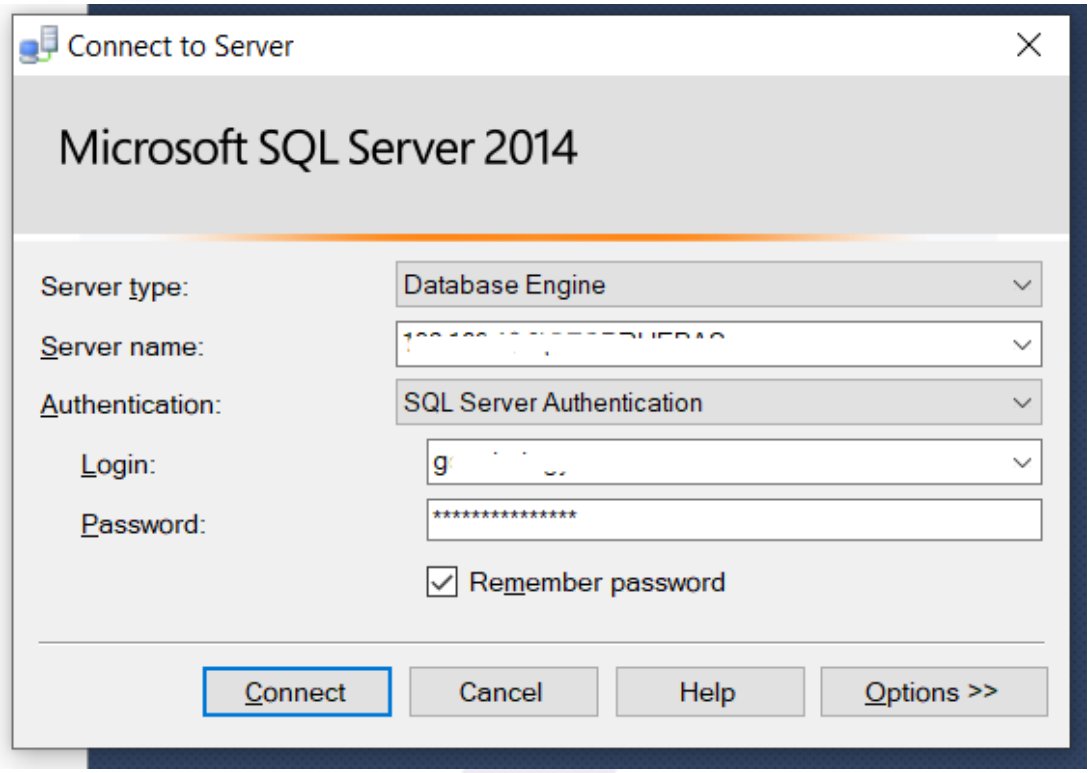

Fig. 3. conexión a la base de datos del hospital mediante SQL Server

4. Una vez conectado se pueden realizar consultas y ver la base de datos del hospital desde el sistema de gestión de base de datos relacional 'SQL Server'.

Diseño ETL

Teniendo en cuenta que el flujo general recomendado por el OHDSI para el ETL es el mostrado en la Fig. 4, se debe empezar por la lógica de la tabla PERSON, seguido de la tabla OBSERVATION_PERIOD y por último VISIT_OCURRENCE, antes de seguir con las demás tablas clínicas que se pueden realizar en el orden que se elija.

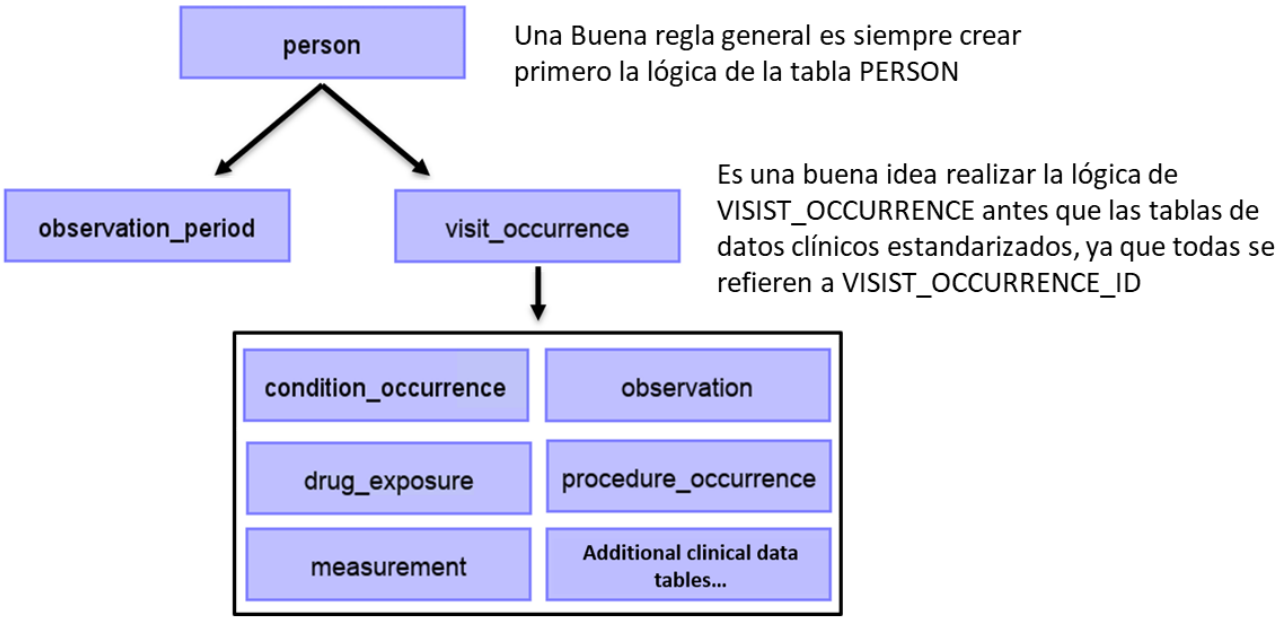

Fig. 4. Flujo general del ETL según el libro de OHDSI
Nota: Adaptado de https://ohdsi.github.io/TheBookOfOhdsi/ExtractTransformLoad.html

Una vez se haya elegido que tabla mapear, se debe tener una claridad sobre esta, es decir, que campos contiene y que significa cada uno. Para esto, se puede ver la descripción de cada tabla en la siguiente página: https://ohdsi.github.io/CommonDataModel/cdm54.html . Posteriormente y teniendo en cuenta la información que debe contener la tabla elegida en el CDM, se debe definir cuáles son las tablas de origen (de la base de datos del hospital) que poseen esta información, para lo cual se pueden hacer algunas preguntas al equipo TI del hospital.

Teniendo la claridad sobre cuáles son las tablas que se van a mapear y cuales tablas del hospital se necesitan, se deben seguir los siguientes pasos:

1. Escanear las tablas de origen para tener claridad de qué información hay en cada tabla, como los campos y tipos de datos que tiene. Para esto se puede usar el software 'White Rabbit' que es proporcionado por OHDSI para este propósito y el cual se puede descargar desde: https://github.com/OHDSI/WhiteRabbit/releases/tag/v0.10.7 .
2. Desde 'White Rabbit' se conecta a la base de datos con los mismos datos proporcionados por el hospital, que son el nombre del servidor, usuario y contraseña. En *Data Type* se selecciona 'SQL Server' y en 'Working Folder' se elige la carpeta en donde se quieran guardar los reportes, como se muestra en la Fig. 5.

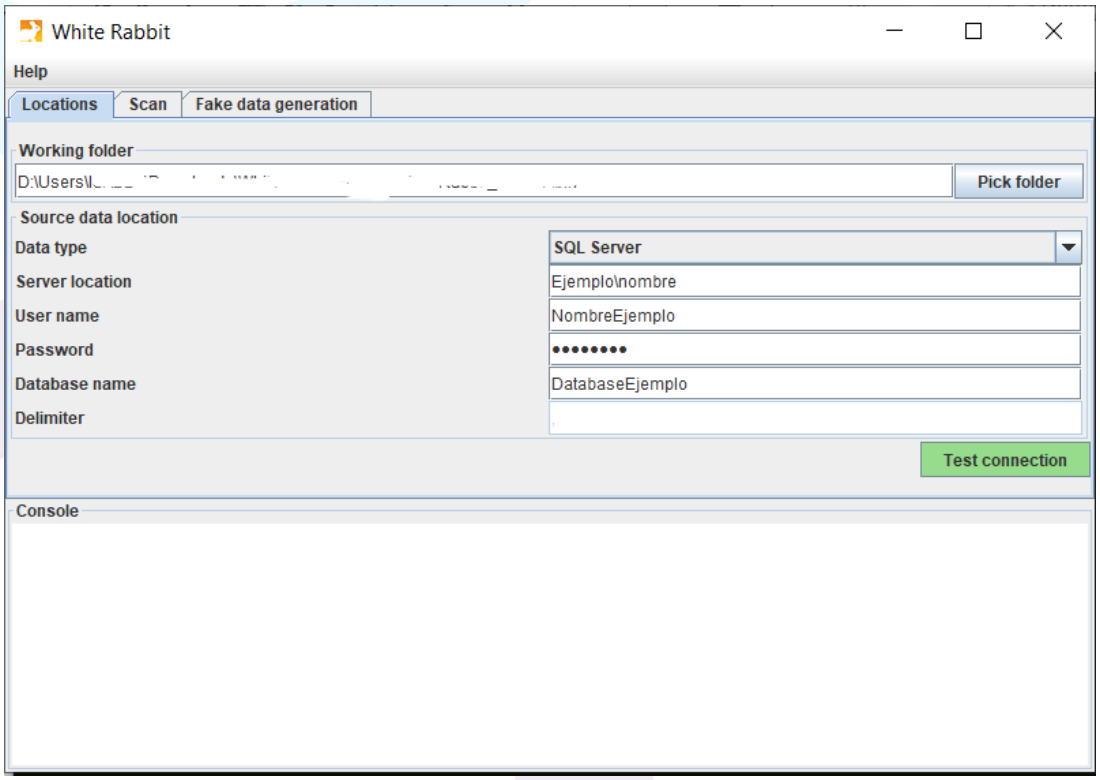

Fig. 5. Conexión a la base de datos del hospital desde 'White Rabbit'.

3. Una vez la conexión sea exitosa, desde la pestaña *Scan* se añaden las tablas seleccionadas anteriormente y que contiene la información de interés. Todos los valores que se muestran en la Fig. 6 se dejan por defecto. Luego, se selecciona el botón de *Scan Tables* y se espera a que se genere el reporte. Este reporte queda guardado como un archivo xlsx de nombre 'ScanReport' en la carpeta que se seleccionó en *Working folder.*

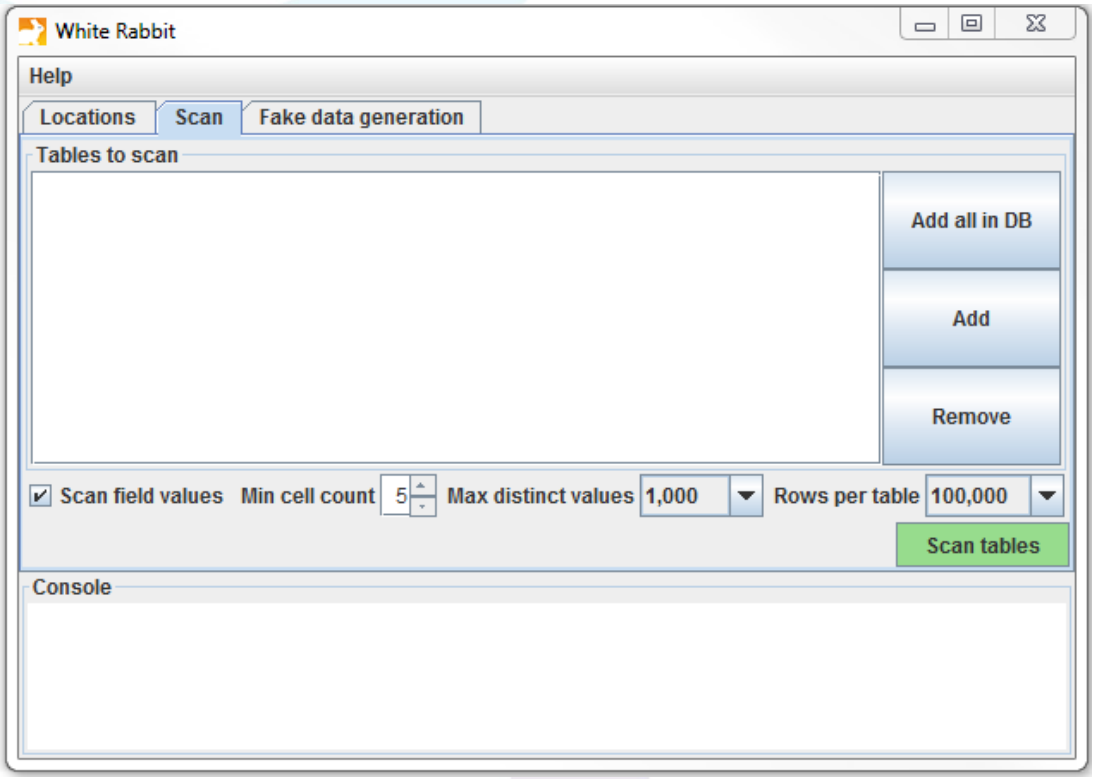

Fig. 6. Pestaña de escaneo de 'White Rabbit'.

Posteriormente, se debe realizar el diseño del ETL con la herramienta 'Rabbit in a hat' que vine en el mismo paquete que 'White Rabbit'.

a) Se importa el reporte generado por 'White Rabbit' en 'Rabbit in a hat'.
b) Se realizan las conexiones entre las tablas de origen y las tablas de CDM. Para esto se arrastra con el mouse desde el recuadro de la tabla de origen hasta el recuadro de la tabla del CDM, donde debe quedar una flecha como se ve en la Fig. 7, que muestra el ejemplo para las tres primeras tablas del CDM. También se puede ver que desde la misma herramienta, al seleccionar alguna tabla del CDM haciendo doble clic, se da información sobre los campos que tiene, el tipo de datos de cada campo y una breve descripción del campo. Si se selecciona alguna tabla de los datos de origen, también se proporciona la información de los campos y el tipo de variable, sin embargo, no proporciona alguna otra descripción de los campos. Por lo tanto, si se tienen dudas sobre los campos es importante preguntar al personal de TI del hospital.

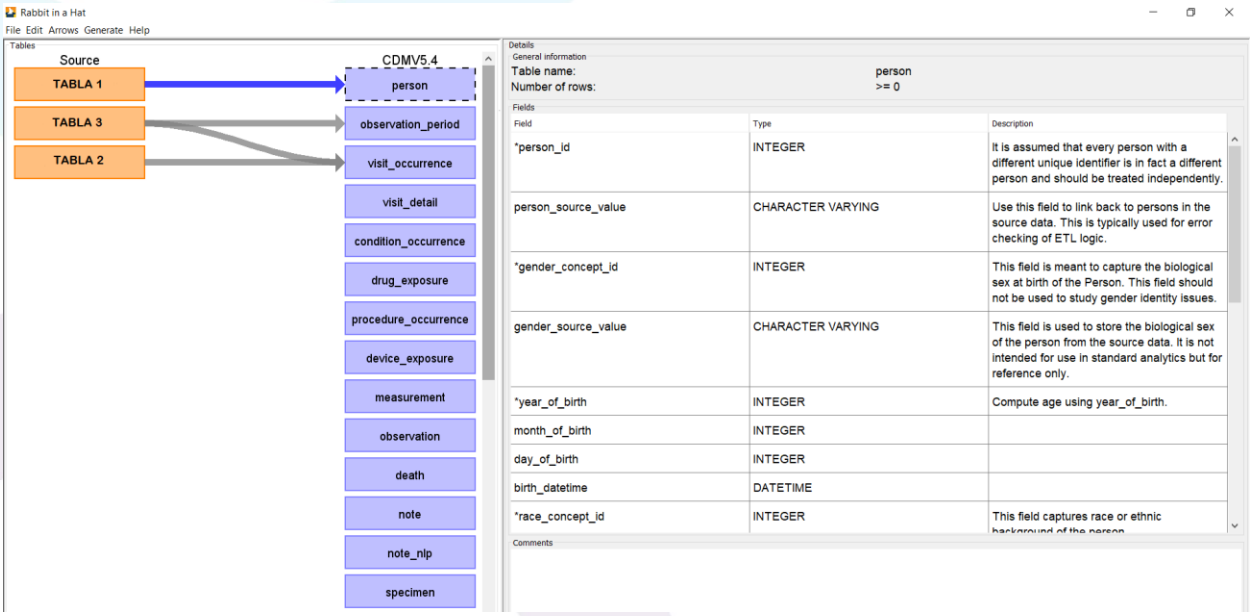

Fig. 7. Relacionamiento entre las tablas de origen y las tablas del CDM desde 'Rabbit in a hat'.

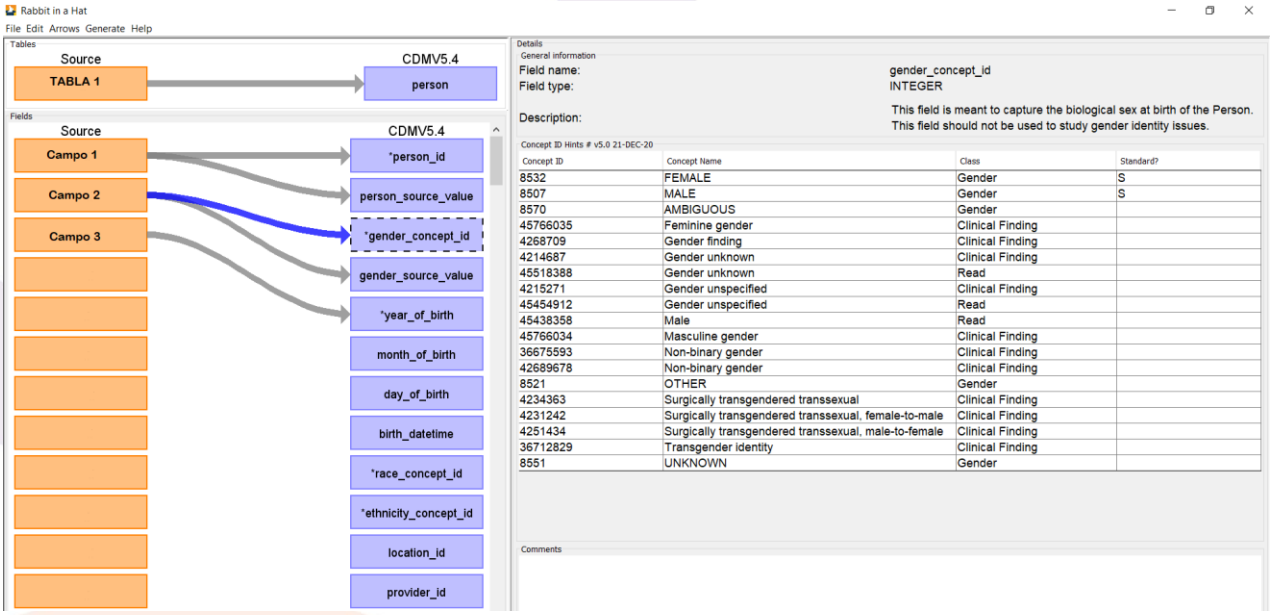

Fig. 8. Relacionamiento entre los campos de la tabla de origen con los campos de la tabla PERSON.

c) Posteriormente se realizan las conexiones entre los campos de las tablas, para esto es necesario dar doble clic sobre la línea que une las tablas en las que se quiere profundizar. En la Fig. 8 se muestra cómo sería el proceso para los campos de la tabla PERSON. El CDM de OMOP tiene como objetivo estandarizar toda la información con un vocabulario, es por esto que en algunos campos se deben colocar códigos específicos del vocabulario. En el ejemplo mostrado, al seleccionar el campo *gender_concept_id* en la interfaz aparece una lista a la derecha. Esta lista nos indica los códigos existentes, la idea es utilizar los que indican 'S' ya que estos son los términos y códigos estándar.

d) Una vez se tenga claro que código se debe utilizar, se debe dar doble clic en la flecha que une los dos campos para escribir la lógica que se quiere seguir. Por ejemplo, si se selecciona la flecha que indica el campo de género, que en el ejemplo mostrado en la Fig. 8 sería la que está de color azul, la lógica se podría escribir como se muestra a continuación en la Fig. 9. Aquí, en la fuente de origen se usan los números 1,2,3 para asignar los géneros masculino, femenino o indefinido. Entonces, la lógica escrita nos va a indicar que cada que tengamos un 1 en los datos de origen los vamos a cambiar por un 8507 en la tabla del CDM y así sucesivamente.

Fig. 9. Ejemplo de lógica para el campo *gender_concept_id.*

e) Luego de tener toda la lógica y las conexiones entre tablas, se genera el informe desde 'Rabbit in a hat'. Este informe es un archivo de word que indica cuales son las tablas que se relacionan y la lógica que se va a seguir. Es muy útil para escribir los scripts para insertar los datos y no tener que abrir otra vez la herramienta de 'Rabbit in a hat'.

Implementación ETL

Lo siguiente que se hace es la implementación del ETL, es decir, se escriben los scripts de SQL para insertar los datos. Para esto, se puede usar como guía el siguiente repositorio: https://github.com/OHDSI/ETL-Synthea , en donde se encuentra el código comentado utilizado para convertir los datos de Synthea en el CDM de OMOP.

Por último, una vez se hayan insertado los datos de origen en la tabla del CDM, se debe realizar una verificación. Para esto, se puede comprobar mediante una consulta en 'SQL Server' o escaneando las tablas nuevas con 'White Rabbit', la cantidad de registros en la tabla de origen y la cantidad de registros que se insertaron en la nueva tabla del CDM. Sin embargo, hay que tener cuidado, pues en muchos casos se excluyen registros que tengan valores nulos. También se puede verificar la transformación de la tabla siguiendo a una persona aleatoria y comprobando que los datos de esta persona estén correctos en la tabla del CDM, mediante consultas en 'SQL Server'.

Nota

Los instaladores de los softwares se pueden encontrar en la carpeta anexa a este documento.

Referencias

[1] G. Hripcsak *et al.*, "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers," *Stud. Health Technol. Inform.*, vol. 216, p. 574, 2015.

[2] C. Blacketer, "The Common Data Model," *The Book of OHDSI*. [Online]. Available: https://ohdsi.github.io/TheBookOfOhdsi/CommonDataModel.html. [Accessed: 31-Aug-2022].