UNIVERSIDAD DE MÁLAGA

E.T.S. INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE MÁLAGA

Trabajo Almacenes de Datos

Estándares para la Integración de Datos Clínicos en un Data Warehouse

Realizado por
De Pablo Diego y Soriano Juan

Profesor encargado:
Luque Baena Rafael Marcos

Departamento
Lenguajes y Ciencias de la Computación

MÁLAGA, noviembre de 2024

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

ESTUDIANTES DE INGENIERÍA BIOINFORMÁTICA

Estándares para la Integración de Datos Clínicos en un Data Warehouse

Almacenes de Datos

Realizado por
De Pablo Diego y Soriano Juan

Profesor encargado:
Navas Luque Baena Rafael Marcos

Departamento
Lenguajes y Ciencias de la Computación

UNIVERSIDAD DE MÁLAGA
MÁLAGA, NOVIEMBRE DE 2024

Contents

1 Introducción a la Integración de Datos Clínicos en un Data Warehouse

La integración de datos clínicos en un warehouse es fundamental para la centralización, normalización y accesibilidad de datos clínicos provenientes de diversas fuentes en un único repositorio [8][1].

La complejidad de integrar datos clínicos radica en la diversidad de los sistemas y formatos utilizados por las diferentes organizaciones de salud, como hospitales, laboratorios, y clínicas especializadas[22]. Con el fin de proponer una solución para este problema, se desarrollaron frameworks de interoperabilidad, como HL7, FHIR, SNOMED CT y LOINC, que permiten estructurar la información para su correcta integración en un data warehouse[17]. Este proceso asegura que los datos almacenados se mantengan consistentes y precisos[8].

El objetivo de este trabajo es analizar cómo se estandariza la información para la integración de datos clínicos en un data warehouse, herramientas utilizadas, y el impacto de esta tecnología en la gestión y uso de la información de salud.

2 Fundamentos y Conceptos Clave de los Data Warehouses en el Ámbito Clínico

Un **data warehouse** (almacén de datos) es un sistema diseñado para la recolección, almacenamiento y análisis de grandes volúmenes de datos provenientes de diversas fuentes. Esta plataforma reúne diversas tecnologías y componentes aprovechando al máximo los datos. Permite almacenar una gran cantidad de datos, así como también su tratamiento y análisis. El objetivo es **transformar los datos brutos en informaciones útiles**, y volverlos disponibles y accesibles para los usuarios [21].

La función de un data warehouse es actuar como un repositorio centralizado donde se recopilan datos de múltiples fuentes, como bases de datos transaccionales, sistemas ERP o archivos externos. Estos datos pueden ser estructurados (tablas relacionales), semiestructurados (como archivos XML o JSON) o no estructurados (texto libre, imágenes)[21].

Una vez los datos llegan al data warehouse, pasan por un proceso de **ETL** (Extracción, Transformación y Carga), donde son limpiados, organizados y transformados para su análisis. Los usuarios pueden acceder a estos datos mediante herramientas de **Business Intelligence** (BI), consultas SQL o visualizaciones interactivas.

Al centralizar la información, las organizaciones obtienen una visión completa y coherente de sus datos, lo que facilita la toma de decisiones basadas en hechos. Además, el data warehouse habilita técnicas de **data mining**, que permiten descubrir patrones ocultos y tendencias para mejorar estrategias comerciales o de operación[15].

En el ámbito clínico, los **sistemas de información sanitaria** (SIS) recopilan volúmenes crecientes de datos provenientes de la atención médica rutinaria. Esta fuente de **datos del mundo real** (RWD, por sus siglas en inglés) ofrece un gran potencial para mejorar la calidad de la atención médica. Por un lado, estos datos aportan beneficios directos al paciente —usos primarios— al ser fundamentales para el desarrollo de la medicina personalizada. Al mismo tiempo, brindan beneficios

indirectos —usos secundarios— al acelerar y mejorar la generación de conocimiento sobre patologías, condiciones de uso de productos y tecnologías sanitarias, así como la evaluación de su seguridad, eficacia y utilidad en la práctica diaria. Estos datos también son útiles para medir el impacto organizacional de las tecnologías de salud. El **manejo eficiente de grandes volúmenes de datos** a través de un data warehouse es una de las ventajas más importantes que permite maximizar el uso de esta información para mejorar los resultados en salud.

2.1 Fases del Flujo de Datos en un Clinical Data Warehouse (CDW)

El **Clinical Data Warehouse (CDW)** es una infraestructura que permite consolidar datos provenientes de uno o varios **Sistemas de Información Médica** (HIS, por sus siglas en inglés) en formatos homogéneos, independientemente del marco organizativo o del origen de los datos. Esta estructura es esencial para facilitar la reutilización de la información en diversos contextos como la gestión, la investigación y la atención médica. [6]

La **Figura 1** ilustra las cuatro fases clave del flujo de datos que conforman un CDW, desde la recopilación de las fuentes originales hasta los usos finales, destacando el proceso de transformación e integración de los datos.

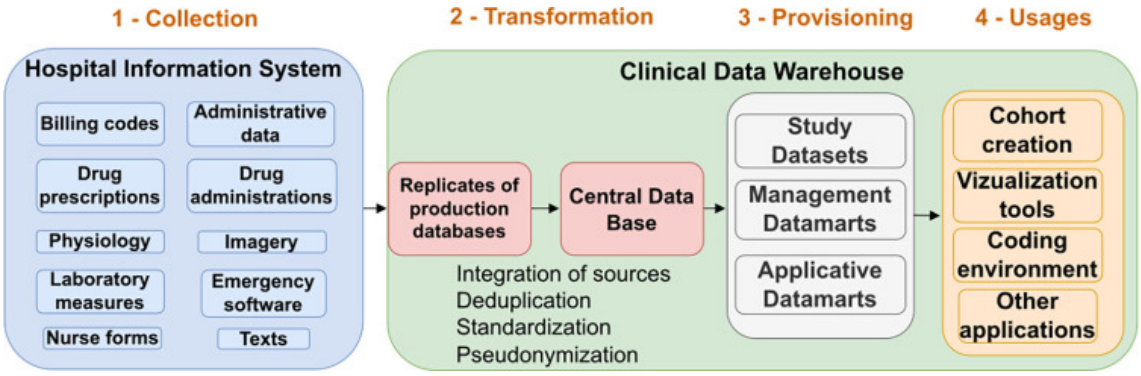


Figure 1: Cuatro pasos del flujo de datos del Sistema de Información Hospitalaria: (1) recopilación, (2) transformaciones y (3) aprovisionamiento. CDW, almacén de datos clínicos.[6]

1. **Recopilación (Collection)** La primera fase consiste en extraer los datos de las distintas fuentes que componen el HIS. Estas fuentes pueden incluir datos administrativos, prescripciones de medicamentos, mediciones fisiológicas, resultados de laboratorio, formularios de enfermería, entre otros. El objetivo principal es capturar la mayor variedad de datos posibles para que el CDW sirva como un repositorio central que refleje de manera precisa la información clínica almacenada en los sistemas hospitalarios. Cada una de estas fuentes puede estar almacenada en diferentes formatos, lo que subraya la necesidad de normalización en las fases posteriores.

2. **Transformación (Transformation)** Tras la recopilación, los datos pasan por un proceso de transformación donde se integran y armonizan para asegurar su coherencia. En esta fase se distinguen varias etapas clave:

- *Integración de fuentes*: Los datos de diferentes bases son unificados en una base central.
- *Desduplicación de identificadores*: Se eliminan las duplicaciones y se normalizan los identificadores que se puedan repetir en diversas fuentes.
- *Estandarización*: Un modelo de datos único, independiente de los sistemas de origen, asegura que los datos se almacenen y estructuren bajo un esquema común. Este proceso es crucial para asegurar la interoperabilidad y puede incluir la adopción de nomenclaturas estándar como SNOMED CT o LOINC.
- *Pseudonimización*: Con el fin de proteger la privacidad del paciente, los datos sensibles son anonimizados mediante técnicas de pseudonimización, eliminando o codificando cualquier información directamente identificativa.

3. **Provisión (Provisioning)** Una vez que los datos han sido transformados, se almacenan en una base central y se dividen en subconjuntos específicos o *datamarts*, que permiten su reutilización tanto en usos primarios como secundarios. Estos *datamarts* pueden estar orientados a diferentes objetivos, como estudios específicos, gestión administrativa o aplicaciones clínicas. El uso de *datamarts* especializados optimiza el acceso a los datos y facilita la creación de *cohortes*, la utilización de *herramientas de visualización*, así como el acceso a un *entorno de codificación* y otras aplicaciones.

4. **Usos (Usages)** En esta última fase, los usuarios del CDW pueden acceder a los datos ya procesados y almacenados en los *datamarts* a través de aplicaciones y herramientas especializadas. Los principales usos incluyen:

   - *Creación de cohortes*: Selección de subpoblaciones específicas basadas en criterios clínicos para su análisis en estudios de investigación.
   - *Herramientas de visualización*: Plataformas que permiten explorar los datos mediante gráficos, tablas y otras representaciones visuales para facilitar el análisis.
   - *Entornos de codificación*: Espacios diseñados para realizar análisis avanzados y procesar los datos de manera eficiente mediante técnicas de *machine learning* o análisis estadístico.
   - *Otras aplicaciones*: Los datos también pueden ser reutilizados en diversas aplicaciones, desde la mejora de procesos clínicos hasta la evaluación de nuevas tecnologías o tratamientos.

3 Estándares de Interoperabilidad en Datos Clínicos

El aspecto más importante de los estándares de integración de datos clínicos es la interoperabilidad para la integración efectiva en un data warehouse, ya que permite que sistemas de salud, que a menudo operan de forma independiente, puedan compartir información coherente y confiable. [17] Permite a los profesionales de la salud y a los investigadores disponer de un acceso centralizado a la información, independientemente de las plataformas o sistemas en los que se generaron originalmente los datos.[17] Detallaremos los estándares de integración más utilizados:

3.1 Estándares de mensajería y estructura de datos:

Hay varias formas de estructurar la información:

3.1.1 HL7 (Health Level 7)

HL7 es uno de los estándares de interoperabilidad más utilizados en el ámbito clínico. Desarrollado por la organización Health Level Seven International. Permite el intercambio de datos entre aplicaciones de salud mediante mensajes estandarizados que contienen información crítica, como datos de pacientes, resultados de laboratorio y diagnósticos médicos. Es ampliamente adoptado en sistemas de salud debido a su capacidad para manejar grandes volúmenes de datos de forma eficaz.

Ejemplo de Uso de HL7: Un hospital necesita compartir los resultados de laboratorio de un paciente con su proveedor de atención primaria. Utilizando HL7, los resultados de laboratorio se codifican en un mensaje HL7 que contiene los campos de información relevantes, como la identificación del paciente, el tipo de prueba realizada y los resultados numéricos de la prueba.

Un mensaje HL7 típico, con el estándar HL7 V2, podría tener un formato similar a este[18]:

```
MSH|^~\&|LABSYSTEM|HOSPITAL|CLINIC|12345|202410261230||ORU^R01|123|P|2.3
PID|1||123456^^^HOSPITAL^MR||Doe^John||19650215|M||2106-3^White^HL70005
OBR|1|12345|67890|CBC^Complete Blood Count^HL70001|202410261200|||||||
OBX|1|NM|59462-2^Hemoglobin^LN||13.5|g/dL|13.5-17.5|N|F
OBX|2|NM|718-7^White Blood Cell Count^LN||6.8|x10^3/uL|4.5-11.0|N|F
```

En este mensaje:

- MSH es el segmento de encabezado del mensaje.
- PID contiene la información del paciente.
- OBR y OBX detallan el tipo de prueba y los resultados obtenidos.

Este mensaje HL7 se envía al sistema del proveedor de atención primaria, quien recibe los datos y los integra en el expediente médico del paciente de forma automatizada, permitiendo así una consulta rápida y eficiente en cualquier momento.

3.1.2 FHIR (Fast Healthcare Interoperability Resources)

FHIR es un estándar relativamente nuevo y es considerado la evolución de HL7. Diseñado por la misma organización que HL7, FHIR permite el intercambio de datos rápido y seguro utilizando formatos ás utilizados como JSON y XML. La modularidad de FHIR, a través de recursos (resources) individuales que representan entidades clínicas, como pacientes, observaciones, y procedimientos, facilita la integración de distintos tipos de datos clínicos., Permite que sistemas diversos accedan a datos precisos y estructurados en tiempo real, lo que es ideal para aplicaciones web y móviles. En esta página (https://www.hl7.org/fhir/resourcelist.html) se puede encontrar que campos son necesarios dependiendo de nuestras necesidades.

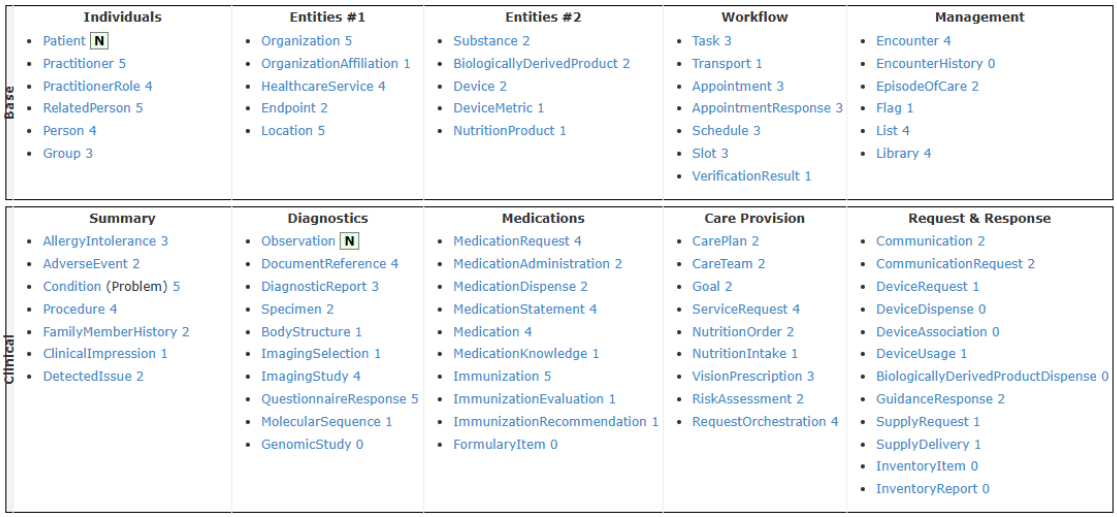


Figure 2: Recursos de FHIR

Ejemplo de Uso de FHIR: Imaginemos una aplicación móvil que permite a los pacientes visualizar su historial médico. Utilizando FHIR, la aplicación puede solicitar datos específicos al sistema de EHR (Electronic Health Record) del hospital

en tiempo real.

Si quisiéramos transferir un paciente de un hospital a otro, sería necesario que el json, que contiene la información del paciente tuviera algunos de los campos especificados en la imagen:

Name	Flags	Card.	Type	Description & Constraints
Patient	N		DomainResource	Information about an individual or animal receiving health care services Elements defined in Ancestors: id, meta, implicitRules, language, text, contained, extension, modifierExtension
identifier	Σ	0..*	Identifier	An identifier for this patient
active	?! Σ	0..1	boolean	Whether this patient's record is in active use
name	Σ	0..*	HumanName	A name associated with the patient
telecom	Σ	0..*	ContactPoint	A contact detail for the individual
gender	Σ	0..1	code	male | female | other | unknown Binding: AdministrativeGender (Required)
birthDate	Σ	0..1	date	The date of birth for the individual
deceased[x]	?! Σ	0..1		Indicates if the individual is deceased or not
deceasedBoolean			boolean	
deceasedDateTime			dateTime	
address	Σ	0..*	Address	An address for the individual
maritalStatus		0..1	CodeableConcept	Marital (civil) status of a patient Binding: Marital Status Codes (Extensible)
multipleBirth[x]		0..1		Whether patient is part of a multiple birth
multipleBirthBoolean			boolean	
multipleBirthInteger			integer	

Figure 3: Campos de Paciente

Un recurso FHIR en JSON para obtener la información de un paciente podría verse así:

```
{
 "resourceType": "Patient",
 "id": "123",
 "identifier": [
 {
  "use": "usual",
  "system": "http://hospital.org/patient",
  "value": "12345"
 }
 ],
 "name": [
 {
  "use": "official",
  "family": "Doe",
  "given": [
  "John"
  ]
 }
 ],
 "gender": "male",
 "birthDate": "1965-02-15"
}
```

Hay otros recursos como Observation (resultados de laboratorio) o MedicationRequest (ver recetas).

3.2 Estándares de codificación de términos:

Estos estándares funcionan para estandarizar los datos de diagnóstico de enfermedades y síntomas, entre otros, mediante su codificación.

3.2.1 SNOMED CT

SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) se utiliza para codificar términos clínicos de manera uniforme. Permiten describir con precisión diagnósticos, procedimientos, síntomas y otros datos clínicos en un formato estandarizado. Al usar SNOMED CT, se asegura que los términos médicos sean interpretables por diferentes sistemas y usuarios en diversas regiones y entornos [20].

Ejemplo de Uso de SNOMED CT: Un médico registra el diagnóstico de un paciente en su expediente electrónico, indicando que tiene diabetes tipo 2. En lugar de escribir "diabetes tipo 2" de forma libre, el sistema de EHR convierte este diagnóstico en un código SNOMED CT específico, como 44054006, que representa "diabetes mellitus tipo 2"[3].

Al registrar el diagnóstico con el código SNOMED CT:

- Los sistemas pueden realizar búsquedas de manera uniforme; por ejemplo, todos los pacientes con 44054006 pueden ser agrupados para analizar estadísticas de prevalencia.
- Los profesionales médicos en distintas instituciones entienden el diagnóstico sin importar el idioma o variaciones regionales, ya que el código es único y universal.
- Permite realizar análisis de datos clínicos sobre enfermedades, tratamientos y evolución de pacientes, basados en términos clínicos estandarizados y consistentes.

3.2.2 LOINC

LOINC es un estándar utilizado principalmente para codificar resultados de laboratorio y observaciones clínicas. Cada prueba de laboratorio y observación tiene un código único en LOINC que permite compararlos de forma consistente entre distintas instituciones de salud. Esto es especialmente útil cuando los datos se centralizan en un data warehouse y se necesita comparar resultados de laboratorio obtenidos en diferentes lugares o con diferentes equipos [17].

Ejemplo de Uso de LOINC: Un laboratorio realiza un análisis de glucosa en sangre y obtiene un valor de 105 mg/dL. El resultado se codifica con el código LOINC 2345-7, que representa una "Glucosa en suero o plasma"[16].

El uso de LOINC permite que:

- Los resultados de glucosa en sangre se almacenan en un data warehouse con el mismo código LOINC, sin importar el hospital o laboratorio que haya realizado la prueba.

- Al realizar consultas y comparaciones en el data warehouse, todos los valores de glucosa se identifican por el mismo código 2345-7, permitiendo así obtener una visión coherente de los niveles de glucosa en diferentes pacientes o en el mismo paciente a lo largo del tiempo.

- Por ejemplo, al analizar el historial de un paciente con el código 2345-7, es posible ver la evolución de su glucosa en sangre y evaluar si ha habido mejoría o empeoramiento en sus niveles. Esto puede ayudar a ajustar tratamientos y realizar diagnósticos más precisos.

4 Arquitectura del Data Warehouse Clínico

La arquitectura de un **data warehouse clínico** es fundamental para el análisis de datos de salud. Esta arquitectura se compone de varias capas que interactúan entre sí para garantizar que los datos sean accesibles, confiables y utilizables para la toma de decisiones clínicas. A continuación, se explican las capas y los componentes esenciales de esta arquitectura

4.1 Capas de la Arquitectura del Data Warehouse

La **Figura 4** ilustra una simplificación de la arquitectura típica de un data warehouse clínico se puede dividir en las siguientes capas:

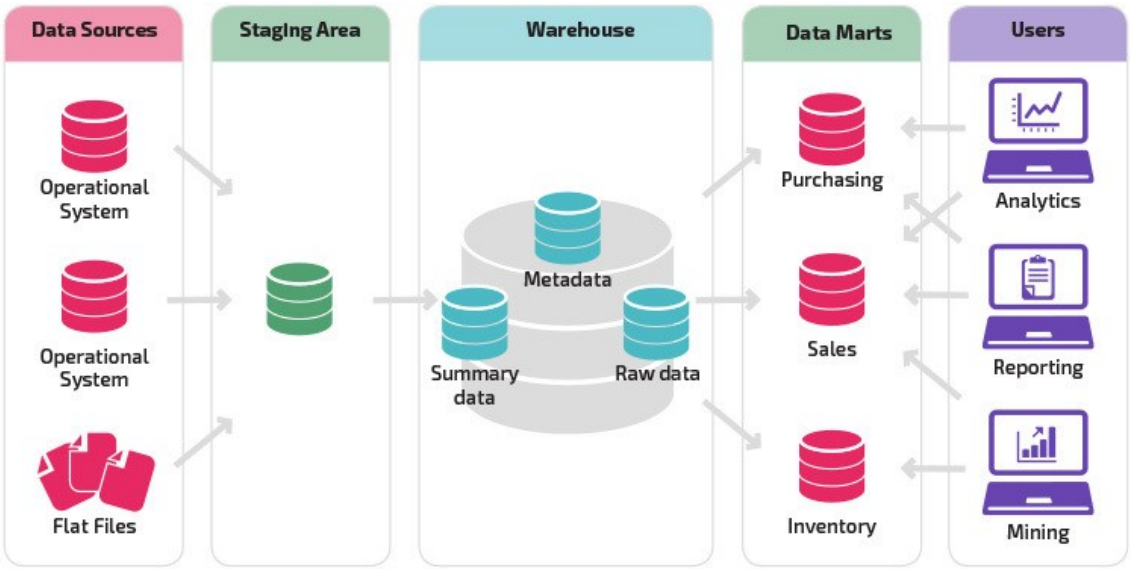


Figure 4: 5 puntos de la arquitectura del Data Warehouse: (1) Fuentes de Datos, (2) Carga (ETL), (3) Warehouse o almacenamiento, (4) Procesamiento y (5) Presentación.

- **Capa de Fuentes de Datos:** Esta es la base del data warehouse y está compuesta por diversas fuentes de datos clínicas, como lo pueden ser:

  - *Registros Electrónicos de Salud (EHR):* Documentos digitales que contienen información sobre la salud de los pacientes.

  - *Sistemas de Información de Salud (HIS):* Plataformas que integran y gestionan datos clínicos y administrativos.
  - *Sistemas de Laboratorio:* Sistemas que gestionan datos de pruebas y resultados de laboratorio.
  - *Dispositivos de Monitoreo:* Equipos que recogen datos de salud en tiempo real, como monitores de signos vitales.

- **Capa de Extracción, Transformación y Carga (ETL):** Esta capa es crucial para preparar los datos antes de ser almacenados en el data warehouse. Las actividades incluyen:
  - *Extracción:* Obtención de datos desde las diferentes fuentes mencionadas.
  - *Transformación:* Normalización, limpieza y enriquecimiento de datos para asegurar la calidad y consistencia. Esto puede incluir la conversión de formatos, eliminación de duplicados y la integración de datos de distintas fuentes.
  - *Carga:* Inserción de los datos transformados en el repositorio del data warehouse.

- **Capa de Almacenamiento:** En esta capa se encuentran los datos organizados y estructurados, donde se utilizan diferentes modelos de datos:
  - *Modelo Estrella:* Un esquema que organiza los datos en una tabla de hechos central conectada a varias tablas de dimensiones. Este modelo facilita consultas rápidas y análisis.
  - *Modelo Copo de Nieve:* Una variante del modelo estrella, donde las tablas de dimensiones están normalizadas para reducir la redundancia de datos.
  - *Data Mart:* Subconjuntos de un data warehouse que están diseñados para un área específica, como el manejo de enfermedades crónicas o análisis de laboratorio.

- **Capa de Procesamiento:** Esta capa se encarga de procesar y analizar los datos almacenados. Se puede incluir:
  - *Minería de Datos:* Técnicas que permiten descubrir patrones, correlaciones y tendencias en grandes volúmenes de datos clínicos.
  - *Análisis Predictivo:* Modelos estadísticos y algoritmos que utilizan datos históricos para predecir eventos futuros, como la probabilidad de enfermedades.
  - *Informes y Dashboards:* Herramientas que permiten a los usuarios visualizar datos y métricas a través de gráficos interactivos y resúmenes visuales.

- **Capa de Presentación:** Esta capa proporciona acceso a los datos a los usuarios finales. Los usuarios pueden interactuar con los datos a través de:
  - *Interfaces de Usuario:* Aplicaciones web o móviles que permiten a los profesionales de la salud consultar y analizar datos de manera intuitiva.

– *Herramientas de BI (Business Intelligence):* Software que ayuda en la toma de decisiones a través de la creación de informes, análisis de tendencias y generación de métricas de rendimiento.

Además de la arquitectura, también se deben considerar otros componentes clave para asegurar la calidad y eficiencia de un data warehouse clínico. Esto incluye la gobernanza de datos, que establece políticas para garantizar la calidad y privacidad de la información; la gestión de calidad de datos, que utiliza herramientas para asegurar la precisión y completitud de los datos mediante limpieza y monitoreo; la seguridad de datos, que protege la confidencialidad e integridad de la información a través de autenticación y cifrado; la interoperabilidad, que permite el intercambio efectivo de datos entre sistemas mediante estándares como HL7 y FHIR; y el mantenimiento y soporte, que asegura la continuidad operativa mediante actualizaciones y soporte técnico.

5 Proceso ETL para la Integración de Datos Clínicos

Explica las etapas de Extracción, Transformación y Carga (ETL) para normalizar y consolidar datos clínicos de diversas fuentes. La integración de datos clínicos en un data warehouse requiere de un proceso ETL (Extracción, Transformación y Carga) robusto que pueda manejar la heterogeneidad de las fuentes de datos y normalizar la información para su análisis y uso posterior.

5.1 Extracción

La etapa de extracción implica la recolección de datos desde múltiples fuentes heterogéneas, como sistemas de gestión hospitalaria (HIS), sistemas de registros médicos electrónicos (EHR), sistemas de laboratorio, y bases de datos de farmacia. Estas fuentes suelen tener estructuras de datos diferentes y utilizar estándares específicos, como HL7 o FHIR, y muchas veces incluso formatos propietarios[13]. Extracción en Datos Clínicos:

- **Acceso a fuentes variadas:** Conectar con diferentes sistemas, incluyendo sistemas de registros médicos electrónicos, dispositivos médicos y aplicaciones clínicas.
- **Acceso seguro:** Asegurar que la extracción de datos cumpla con las normativas de privacidad de datos de salud (como HIPAA en EE. UU. [7] o GDPR en Europa [4]).
- **Manejo de formatos de datos:** Leer e interpretar formatos como XML, JSON, y mensajes HL7 V2, o FHIR, así como datos de bases de datos SQL y archivos CSV.

Ejemplo de Extracción en Datos Clínicos: Un sistema de EHR contiene información de consultas médicas de pacientes en formato HL7. Para extraer esta información, el proceso ETL se conecta con el sistema de EHR mediante una interfaz segura y descarga los datos en formato HL7 V2 para luego proceder con la transformación.

5.2 Transformación

La transformación es la etapa más complicada del proceso ETL. Implica la normalización y estandarización de los datos extraídos. Los datos clínicos suelen estar codificados en diferentes sistemas (ej., SNOMED CT, LOINC, ICD-10) [14].El objetivo principal será convertir estos datos en un formato común que permita su análisis conjunto. Esta etapa incluye varios subprocesos clave:

- **Estandarización de formatos:** Convertir los datos a un formato común, como JSON o CSV, y unificar estructuras y tipos de datos.[2].

- **Normalización y limpieza de datos:** Corregir valores inconsistentes, eliminar duplicados y manejar valores nulos o faltantes. [5]

- **Codificación:** Mapear términos médicos a un estándar de codificación unificado, como SNOMED CT para diagnósticos y LOINC para resultados de laboratorio. [12]

- **Integración de datos:** Combinar datos de distintas fuentes en una estructura unificada, lo que puede implicar el mapeo de distintos campos de datos al mismo campo en el data warehouse.

Ejemplo de Transformación en Datos Clínicos: Imaginemos que se tienen registros de diferentes laboratorios en formatos y códigos distintos. Algunos laboratorios usan términos de texto libre para describir las pruebas de laboratorio, mientras que otros utilizan LOINC. En la fase de transformación, el proceso ETL asigna códigos LOINC a las pruebas descritas en texto libre y convierte todos los registros a un mismo formato de archivo (ej., JSON o CSV) para que puedan ser integrados en el data warehouse.

Además, durante esta etapa, los datos de un paciente con identificadores distintos en cada sistema fuente se unen mediante un identificador único (como el número de historia clínica o ID de paciente en el hospital), consolidando así el historial del paciente.

5.3 Carga

La etapa de carga implica mover los datos transformados y normalizados al data warehouse clínico, donde estarán disponibles para el análisis y consulta. En esta fase, se deben considerar aspectos como la frecuencia de actualización de los datos, el diseño del modelo de datos en el data warehouse, y el control de calidad.

Tipos de Carga:

- **Carga completa:** La primera vez que se cargan los datos al data warehouse, se hace de manera completa, es decir, se carga todo el histórico de datos clínicos.

- **Carga incremental:** Para actualizaciones periódicas, solo se cargan los datos nuevos o los que hayan sido modificados desde la última carga, optimizando el uso de recursos y el tiempo de carga.

Ejemplo de Carga en Datos Clínicos: Una vez que los registros de laboratorio han sido transformados y estandarizados, se insertan en el data warehouse en tablas relacionadas con los datos clínicos del paciente, como en una tabla de "Pruebas de Laboratorio". El sistema puede configurarse para realizar cargas incrementales diarias, de forma que cualquier nuevo resultado de laboratorio se agregue automáticamente al data warehouse, permitiendo a los analistas y médicos consultar la información más reciente.

6 Beneficios de la Integración de Estándares Datos Clínicos

La integración de datos clínicos a través de un *Data Warehouse* clínico ofrece múltiples beneficios para el manejo de la información sanitaria que han sido mencionadas anteriormente. Al centralizar los datos de diversas fuentes y garantizar su interoperabilidad mediante estándares como HL7 o FHIR, se facilita el acceso a información precisa y en tiempo real. Esto permite a los médicos y gestores de salud tomar decisiones más informadas y basadas en hechos, lo que mejora directamente la calidad de la atención. El uso de datos integrados también acelera la capacidad de realizar análisis avanzados, permitiendo a los sistemas de salud anticiparse a complicaciones, identificar tendencias en tiempo real y planificar de manera más eficiente los recursos sanitarios.

Además, los *Data Warehouses clínicos* ayudan a la investigación médica y el cumplimiento normativo. Al consolidar grandes volúmenes de datos en un formato estandarizado, los investigadores pueden realizar estudios más robustos, como análisis epidemiológicos o ensayos clínicos, que no solo apoyan la investigación científica sino que también permiten descubrir patrones que podrían mejorar las prácticas médicas. Asimismo, la capacidad de cumplir con las normativas y generar informes precisos para entidades regulatorias o aseguradoras se ve facilitada, promoviendo una mayor transparencia y control de calidad en la prestación de servicios médicos.

7 Relación con el Curso de Almacenes de Datos

Durante el curso se ha proporcionado una base sólida sobre el manejo, diseño y uso de *data warehouses*. Muchos de los conceptos aplicados en este trabajo, como la estructura multidimensional y los procesos ETL, están fundamentados en los conocimientos adquiridos en clase. Hemos aprendido a diferenciar entre los sistemas de bases de datos transaccionales (OLTP) y los almacenes de datos (OLAP), entendiendo cómo el cambio de paradigma, que se enfoca en agregar datos en lugar de actualizarlos o eliminarlos, permite crear una visión histórica completa. Esta visión es esencial en sistemas clínicos, donde el mantenimiento del historial íntegro de los datos de los pacientes es clave para mejorar el análisis, apoyar en la toma de decisiones, y avanzar en el cuidado de la salud.

La experiencia de trabajar con *data warehouses* durante el curso nos ha mostrado su relevancia directa en el ámbito de la salud. Aplicar técnicas como la extracción, transformación y carga (ETL) sobre datos clínicos reales permite a los ingenieros en salud centralizar y estructurar información médica de manera eficiente. Esto no solo

facilita el análisis predictivo y la identificación de patrones en grandes volúmenes de datos clínicos, sino que también permite diseñar soluciones más efectivas para optimizar recursos en hospitales y mejorar el manejo de pacientes. Lo que en clase se aborda desde una perspectiva técnica, en el contexto clínico se traduce en herramientas que impactan directamente la calidad de la atención médica y la investigación científica.

8 Ejemplos-Demo de Uso de Integración de Datos Clínicos con Estándares

8.1 Ejemplo con estándar FHIR:

En este caso lo que intentaremos es postear la información de un paciente con el estándar de datos FHIR en formato JSON en un servidor de FHIR.

8.1.1 Extracción y Transformación

Dado que partimos de un json y el servidor de FHIR trabaja con formato json, no hace falta transformarlo.

Primero necesitamos crear un archivo recurso FHIR y un json que cumpla con los estándares de mensaje de FHIR, por lo que para ello, primero nos descargaremos un archivo que nos ayudará a crear el json de un paciente.

Iremos al siguiente enlance nos iremos a https://hl7.org/fhir/ al apartado de documentación y le daremos a Downloads - Schemas, Code, Tools (https://hl7.org/fhir/downloads.html) y nos descargaremos un zip (https://hl7.org/fhir/fhir.schema.json.zip), que contiene un fhir.schema.json con toda la información del estándar de FHIR sobre cómo crear un json.

Creamos un setting.json, que se relacione con el json descargado:

```
{
    "json.schemas": [
        {
            "fileMatch": [
                "*.fhir.json"
            ],
            "url": "./fhir.schema.json"
        }
    ]
}
```

Figure 5: settings.json

Ahora cuando creemos un json nos aparecerán recomendaciones de los campos del estándar FHIR:

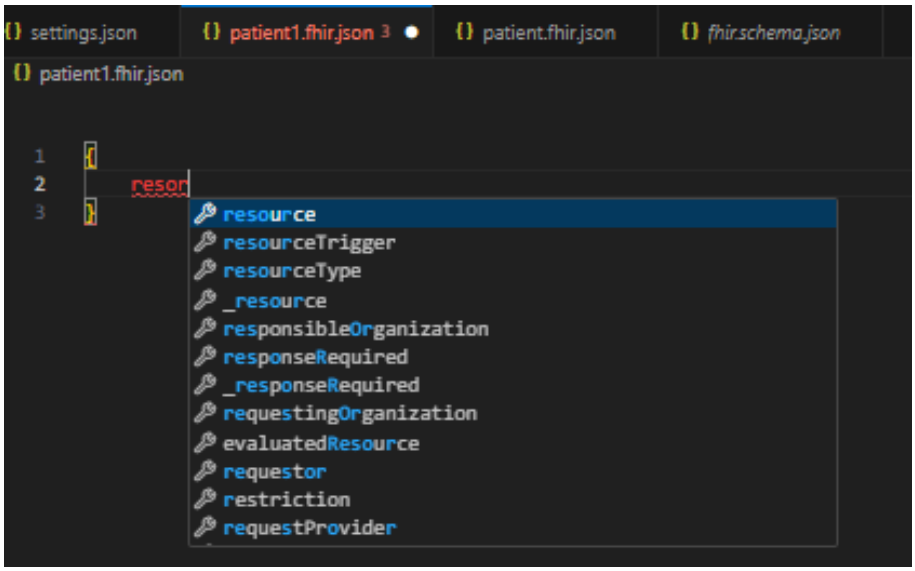


Figure 6: VS recomienda estructuras del estándar de FHIR

En este caso escogemos el resourceType y ponemos el valor “Patient”, ya que queremos insertar un paciente en el servidor.

Después de crear el json entero siguiendo el estándar para un json de FHIR:

```
{
 "resourceType": "Patient",
 "name": [
 {
  "use": "official",
  "given": ["Juan"],
  "family": "Soriano"
 }
 ],
 "gender": "male",
 "birthDate": "2003-09-08",
 "telecom": [
 {
  "value": "987654310",
  "use": "mobile",
  "system": "phone"
 },
 {
  "system": "email",
  "value": "tornadoalert@gmail.com"
 }
 ],
 "address":[
 {
  "line":[
  "213, Diamond Residency"
```

```
    ],
    "city": "Málaga",
    "state": "Andalucia",
    "postalCode": "567104"
  }
  ]
}
```

[11]

8.1.2 Carga

Se dispondrá de un servidor dado por FHIR para la prueba:

Figure 7: Servidor de FHIR [9]

Será a esta url [10] a la que le lanzaremos un post para meter nuestro cliente en el servidor.

Ponemos el json como cuerpo del mensaje. Si hacemos una petición get a https://hapi.fhir.org/baseR4/Patient/45106909, nos devolverá nuestro paciente añadido:

Figure 8: Resultado de la petición get

Como tal no hemos integrado los datos a un datawarehouse aplicando ETL, ya que no tenemos almacenes y no hacía falta aplicar ninguna codificación de los datos, pero probamos la efectividad del estándar FHIR para integrar datos clínicos en un repositorio único o una base de datos.

8.2 Ejemplo con estándar HL7:

Para este ejemplo utilizaremos Mirth Connect, que es una plataforma de integración de datos clínicos ampliamente utilizada en el ámbito de la salud para facilitar la interoperabilidad entre sistemas. Permite que distintos sistemas de información en

salud (como EHR, HIS, LIS, y Data Warehouses clínicos) intercambien y transformen datos de manera segura y eficiente[19]. Está diseñado para manejar datos en diversos formatos y estándares, como HL7, FHIR, X12, DICOM, y más, Mirth Connect permite transformar y adaptar los datos a los requisitos específicos de cada sistema, asegurando que se interpreten y utilicen de manera coherente[19].

Mirth connect proporciona la posibilidad de crear canales para poder mandar mensajes de datos clínicos y cargarlos en un datawarehouse.

En este caso, suponiendo que tenemos, ya extraída la información de un paciente siguiendo el estándar de mensajería de HL7, transformarlo y adaptar esa información, para integrarlo a una base de datos de mySQL, que simularía nuestro datawarehouse.

Para el ejemplo crearemos una base de datos llamada patient, con una tabla paciente:

Haremos un query:

```
CREATE DATABASE patient;
USE patient;
CREATE TABLE pacientes (
id INT AUTO_INCREMENT PRIMARY KEY,
nombre VARCHAR(255) NOT NULL,
fecha_admision DATETIME NOT NULL,
diagnostico VARCHAR(255)
);
```

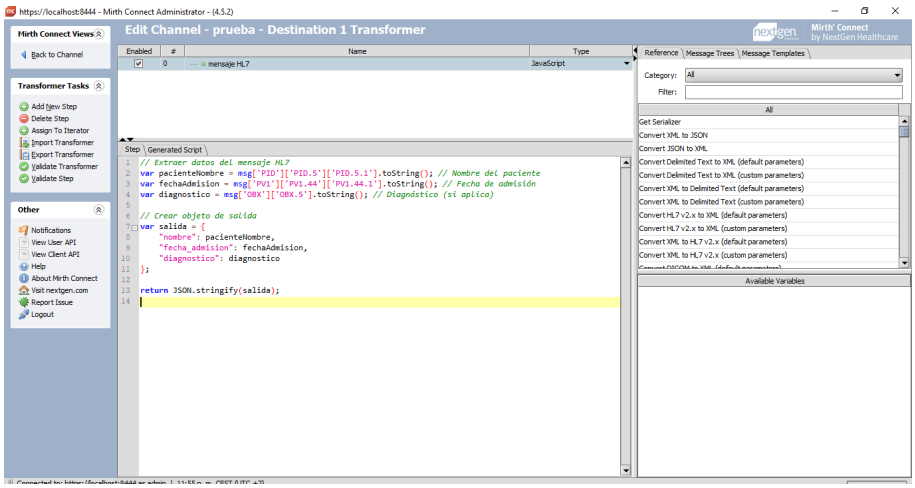


Figure 9: Script para feedback

8.2.1 Fase de extracción y transformación

Este es el msg template utilizado con el estándar de HL7:

```
MSH|^~\&|SendingApplication|SendingFacility|ReceivingApplication|ReceivingFacilit
EVN|A01|202310261200
PID|1||123456^^^Hospital^MR||Doe^John^||19800101|M|||
```

```
PV1|1|I|WardA^Room101^Hospital|U|1|1234^Smith^Jane^^Dr.| | | | |202310261200|
OBX|1|CE|DIAG^Diagnóstico||Hipertensión Arterial|||F
```

Con este msg template, estamos avisando al canal que por él, mandaremos mensajes HL7 con esa estructura.

8.2.2 Transformer HL7:

El transformador nos ayudará a captar la información que nos interesa del mensaje HL7, y adaptarla a nuestra base de datos mySQL. Esta transformación se llevará a cabo mediante el siguiente script de java:

```
[language=JAVA]
  // Extraer datos del mensaje HL7
  var pacienteNombre = msg['PID']['PID.5']['PID.5.2'].toString();
  var diagnostico = msg['OBX']['OBX.5']['OBX.5.1'].toString();

  // Log para depuración
  logger.info('Nombre: ' + pacienteNombre);
  logger.info('Diagnóstico: ' + diagnostico);

  // Almacenar solo datos necesarios en un objeto temporal
  var datosParaGuardar = {
   'pacienteNombre': pacienteNombre,
   'diagnostico': diagnostico
  };

  // Log del objeto temporal
  logger.info("Datos para guardar: " + JSON.stringify(datosParaGuardar));

  // Almacenar los datos extraídos en msg
  msg['pacienteNombre'] = pacienteNombre;
  msg['diagnostico'] = diagnostico;

  \subsubsection{Crear variables:}
  Con esta creación de variables decimos al canal, que cada vez que mandemos un men
  firstname
  msg['PID']['PID.5']['PID.5.2'].toString()
  diagnostico
  msg['OBX']['OBX.5']['OBX.5.1'].toString()
```

8.2.3 Carga

Configuracion de destination:

- 'com.mysql.cj.jdbc.Driver': Clase del controlador JDBC para MySQL.

- ‘jdbc:mysql://127.0.0.1:3306/patient‘: URL de conexión a la base de datos MySQL.

En el script de destination, ponemos lo que se va a insertar.

```
INSERT INTO pacientes (nombre, fecha_admision, diagnostico)
VALUES (${nombre}, ${fecha_admision}, '${diagnostico}');
```

```
GRANT ALL PRIVILEGES ON *.* TO 'root'@'172.17.0.1' IDENTIFIED BY 'Venecia1234'
WITH GRANT OPTION;
FLUSH PRIVILEGES;
ALTER USER 'root'@'%' IDENTIFIED WITH mysql_native_password BY 'NuevaContraseña';
```

Ahora si enviamos a nuestro canal:

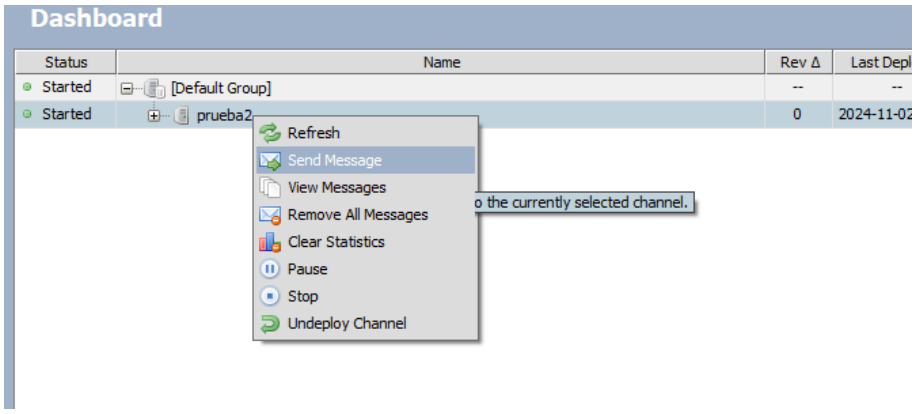


Figure 10: Mandar mensaje al canal

El siguiente mensaje:

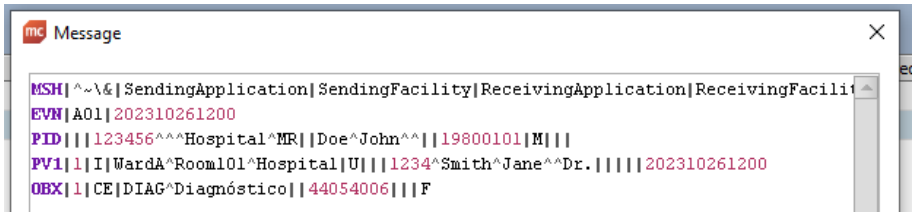


Figure 11: Confirmación de la inserción

Se confirma que en efecto se han captado los campos que nos interesaban:

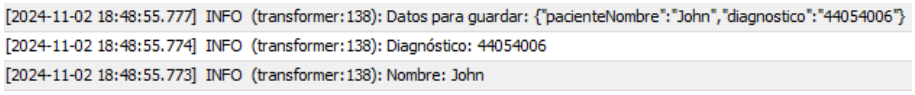
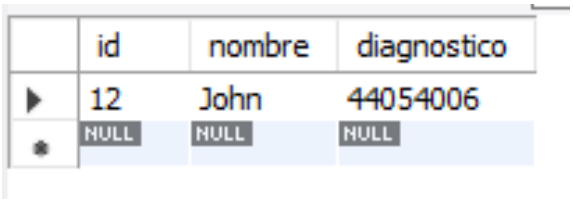


Figure 12: Dato insertado en MySQL

Y como vemos queda insertado en:



	id	nombre	diagnostico
▶	12	John	44054006
*	NULL	NULL	NULL

Figure 13: Mensaje HL7

Y como vemos queda insertado el paciente John con Diabetes tipo 2 (utilizando el estándar de codificación de diagnóstico de SNOMED CT)

9 Conclusiones y Perspectivas Futuras en la Integración de Datos Clínicos y referencias

La integración de datos clínicos en un *Data Warehouse* es una solución esencial para la centralización y reutilización de grandes volúmenes de información en el ámbito de la salud. A través del uso de estándares de interoperabilidad como *HL7*, *FHIR*, *SNOMED CT* y *LOINC*, se garantiza que los datos clínicos, provenientes de diversas fuentes, sean accesibles, consistentes y útiles para la toma de decisiones. Estos estándares permiten la comunicación entre distintos sistemas y aseguran que la información se mantenga precisa y estandarizada.

El diseño de un *Clinical Data Warehouse* (CDW) facilita el análisis de datos históricos, lo que no solo mejora la toma de decisiones clínicas, sino que también optimiza la planificación de recursos hospitalarios y promueve la investigación médica. Las distintas capas del CDW, desde la recopilación y transformación de los datos hasta su análisis y presentación, permiten un manejo eficiente de la información, lo que resulta en una atención médica de mayor calidad y en avances en la investigación basada en datos reales.

De cara al futuro, se espera que la integración de datos clínicos continúe evolucionando, facilitando un acceso aún más eficiente a información crítica para mejorar la atención médica y avanzar en la medicina personalizada. Tecnologías emergentes como la inteligencia artificial y el análisis predictivo, combinadas con el acceso centralizado a datos estructurados, prometen transformar la gestión sanitaria y la investigación en salud, permitiendo una atención más personalizada, basada en datos y optimizada para resultados clínicos.

References

[1] Astera. *Integración de datos en el sector salud.* [Último acceso: 2 de noviembre de 2024]. 2024. URL: https://www.astera.com/es/type/blog/healthcare-data-integration/.

[2] Astera Software. *¿Qué es la Estandarización de Datos? Definición, Beneficios y Mejores Prácticas.* https://www.astera.com/es/type/blog/data-standardization/. Accessed: 2024-11-03. 2023.

[3] BioPortal. *Concepto SNOMED CT: 44054006*. [Último acceso: 2 de noviembre de 2024]. 2024. URL: https://bioportal.bioontology.org/ontologies/SNOMEDCT?p=classes&conceptid=44054006.

[4] Comisión Europea. *Protección de datos en la UE*. [Último acceso: 2 de noviembre de 2024]. 2024. URL: https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_en.

[5] Datos Maestros. *Técnicas de Limpieza de Datos: Cómo limpiar datos de manera eficiente*. https://datosmaestros.com/tecnicas-de-limpieza-de-datos/. Accessed: 2024-11-03. 2023.

[6] Mathieu Doutreligne et al. "Good practices for clinical data warehouse implementation: A case study in France". In: *PLOS Digital Health* 2.7 (2023). Erratum in: PLOS Digit Health. 2023 Sep 29;2(9):e0000369, e0000298. DOI: 10.1371/journal.pdig.0000298.

[7] FBI. *HIPAA: Ley de Portabilidad y Responsabilidad de Seguros de Salud*. [Último acceso: 2 de noviembre de 2024]. 2024. URL: https://www.fbi.gov/file-repository/hipaa_spanish1.pdf.

[8] Founderz. *¿Qué es un Data Warehouse?* [Último acceso: 2 de noviembre de 2024]. 2024. URL: https://founderz.com/es/blog/data-warehouse-que-es/.

[9] HAPI FHIR. *HAPI FHIR R4*. https://hapi.fhir.org/baseR4. Accessed: 2024-11-02. 2024.

[10] HAPI FHIR. *Patient Resource - HAPI FHIR R4*. https://hapi.fhir.org/baseR4/Patient. Accessed: 2024-11-02. 2024.

[11] HL7. *Patient*. https://www.hl7.org/fhir/patient.html. Accessed: 2024-11-02. 2024.

[12] ITDO. *Terminología estándar en salud: SNOMED CT como vocabulario clínico estructurado*. https://www.itdo.com/blog/terminologia-estandar-en-salud-snomed-ct-como-vocabulario-clinico-estructurado/. Accessed: 2024-11-03. 2024.

[13] KeepCoding. *¿Cómo funciona la fase de extracción?* [Último acceso: 2 de noviembre de 2024]. 2024. URL: https://keepcoding.io/blog/funciona-fase-de-extraccion/.

[14] KeepCoding. *¿Cómo funciona la fase de transformación en ETL?* https://keepcoding.io/blog/funciona-fase-transformacion/. Accessed: 2024-11-03. 2023.

[15] Ralph Kimball and Margy Ross. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. 3rd. Wiley, 2013.

[16] LOINC. *LOINC Code 2345-7*. [Último acceso: 2 de noviembre de 2024]. 2024. URL: https://loinc.org/2345-7/.

[17] Meditecs. *Estándares de interoperabilidad en sanidad*. [Último acceso: 2 de noviembre de 2024]. 2024. URL: https://www.meditecs.com/es/kb/estandares-interoperabilidad-sanidad/.

[18] Meditecs. *Estándares HL7*. [Último acceso: 2 de noviembre de 2024]. 2024. URL: https://www.meditecs.com/es/kb/estandares-hl7/.

[19] Meditecs. *Tutorial de Mirth Connect*. [Último acceso: 2 de noviembre de 2024]. 2024. URL: https://www.meditecs.com/es/kb/tutorial-mirth-connect/.

[20] Ministerio de Sanidad de España. *Preguntas frecuentes sobre SNOMED CT*. [Último acceso: 2 de noviembre de 2024]. 2024. URL: https://www.sanidad.gob.es/areas/saludDigital/interoperabilidadSemantica/factoriaRecursos/snomedCT/preguntas.htm.

[21] S. G. Portal and M. de J. M. Jaramillo. "Diseño de un data warehouse para medir el desarrollo disciplinar en instituciones académicas". In: *Investigación Bibliotecológica* 31.72 (2017), pp. 16–45. URL: https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0187-358X2017000200161.

[22] Revista Médica. *Manejo de la información clínica*. [Último acceso: 2 de noviembre de 2024]. 2024. URL: https://revistamedica.com/manejo-informacion-clinica/.

UNIVERSIDAD DE MÁLAGA | uma.es

E.T.S de Ingeniería Informática
Bulevar Louis Pasteur, 35
Campus de Teatinos
29071 Málaga

E.T.S. DE INGENIERÍA INFORMÁTICA