

A semantic framework for epidemiological research

Álvaro Sánchez¹  and Regina Motz¹ 

Facultad de Ingeniería - Universidad de la Republica, Av. Julio Herrera y Reissig 565,
11300, Montevideo, Uruguay
`alv_sanchez@hotmail.com`, `rmotz@fing.edu.uy`

Abstract. This paper presents the software architecture of HCES (Historia Clínica Electrónica Semántica – Investigador Virtual), a semantic platform for epidemiological research. Although standards such as FHIR R4 and SNOMED CT provide well-defined clinical data structures and coded terminologies, the exploitation of their semantic potential for epidemiological analysis is notably enhanced through explicit enrichment based on Semantic Web technologies and the formal representation of the research process itself. HCES addresses this gap by integrating within a single platform three layers usually addressed separately, the standardized representation of clinical data through FHIR R4 and SNOMED CT, the semantic enrichment of each clinical resource into RDF linked to ontological concepts, and the formalization of the research process through the OBI ontology (Ontology for Biomedical Investigations). The architectural description follows a three-view model, with emphasis on four original contributions of the project, the JsonLD Traductor component that transforms each FHIR resource into a semantically enriched RDF subgraph at ingestion time, the Concept and View Manager (CVM) component that operates between the RDF modeling and the visual representation of knowledge within a research environment, a dual storage architecture that combines document-oriented persistence with RDF persistence, and a methodological traceability scheme that models each step of the research process as formal knowledge through OBI. The paper additionally presents lines of work related to interoperability, automated inference, and methodological analysis.

Keywords: software architecture, semantic web, FHIR, SNOMED CT, epidemiological research

Un software semántico para la investigación epidemiológica

Abstract. Este artículo presenta la arquitectura de software del sistema HCES (Historia Clínica Electrónica Semántica – Investigador Vir-

tual), una plataforma semántica para investigación epidemiológica. Si bien estándares como FHIR R4 y SNOMED CT proveen estructuras de datos clínicos y terminologías codificadas con semántica bien definida, la explotación del potencial semántico para análisis epidemiológico se favorece notoriamente mediante un enriquecimiento explícito basado en tecnologías de la web semántica y la representación formal del propio proceso de investigación. HCES aborda esta brecha integrando en una misma plataforma tres capas habitualmente tratadas por separado, la representación estandarizada de datos clínicos mediante FHIR R4 y SNOMED CT, el enriquecimiento semántico de cada recurso clínico en RDF vinculado a conceptos ontológicos y la formalización del proceso de investigación mediante la ontología OBI (Ontology for Biomedical Investigations). La descripción arquitectónica se desarrolla siguiendo un modelo de tres vistas, con énfasis en cuatro contribuciones originales del proyecto, el componente JsonLD Traductor que transforma cada recurso FHIR en un subgrafo RDF semánticamente enriquecido durante su ingesta, el componente Concept and View Manager (CVM) que opera entre el modelado RDF y la representación visual del conocimiento en un entorno de investigación, una arquitectura de almacenamiento dual que combina persistencia documental con persistencia en RDF y un esquema de trazabilidad metodológica que modela cada paso del proceso de investigación como conocimiento formal mediante OBI. Además, el artículo presenta líneas de trabajo vinculadas a la interoperabilidad, inferencia automática y análisis metodológico.

Keywords: arquitectura de software, web semántica, FHIR, SNOMED CT, investigación epidemiológica

1 Introducción

La adopción de estándares internacionales de interoperabilidad clínica como FHIR (Fast Healthcare Interoperability Resources) (Health Level Seven International, 2019), el estándar de HL7 para el intercambio de datos clínicos estructurados, y terminologías de referencia como SNOMED CT (SNOMED International, 2024), que organiza los conceptos clínicos en una jerarquía formal, ha producido en los últimos años un gran volumen de datos clínicos estructurados y codificados con un conjunto de vocabularios controlados. Sin embargo, disponer de datos codificados no equivale a disponer de conocimiento explotable para investigación epidemiológica. Los entornos tradicionales de investigación operan sobre conjuntos de datos exportados desde historias clínicas electrónicas en planillas que preservan los códigos clínicos como valores textuales. Los modelos de datos adoptan estructuras variadas donde se pierden las relaciones semánticas presentes en la fuente original y los procesos analíticos permanecen implícitos en el código de cada investigación. En consecuencia, el conocimiento previo que condujo a una investigación, qué conceptos clínicos se seleccionaron, qué métodos

se aplicaron, las transformaciones realizadas y qué relaciones se descubrieron, no queda representado con la formalidad con la que cuentan los datos sobre los que se investigó. Utilizar el potencial semántico de los datos clínicos estandarizados para investigación epidemiológica y formalizar el propio proceso de investigación que los explora, constituyen los objetivos que aborda este trabajo.

Este artículo presenta HCES (Historia Clínica Electrónica Semántica, Investigador Virtual), una plataforma semántica para investigación epidemiológica. La solución se plantea desde la integración en una misma arquitectura de tres capas habitualmente separadas, la representación estandarizada de datos clínicos mediante FHIR R4 y SNOMED CT, el enriquecimiento semántico apoyado en tecnologías de la web semántica (RDF/OWL, SPARQL, Apache Jena) y la formalización del proceso de investigación mediante OBI (Ontology for Biomedical Investigations) (Bandrowski et al., 2016), una ontología que modela los elementos de una investigación científica como el diseño, procesos realizados, datos utilizados o las conclusiones obtenidas, como entidades formales. La adopción de tecnologías abiertas, estándares W3C y software libre permite que la plataforma opere sin dependencia de licencias comerciales ni infraestructura propietaria.

Este artículo se enfoca en la descripción arquitectónica de HCES, haciendo énfasis en cuatro contribuciones originales que el proyecto incorpora respecto del estado actual, el componente JsonLD Traductor responsable del enriquecimiento semántico de los recursos FHIR, el componente Concept and View Manager (CVM) responsable de la transformación de representaciones semánticas en RDF a representaciones visuales en una plataforma tipo Wiki, una arquitectura de almacenamiento dual que combina persistencia documental sobre MongoDB con persistencia RDF sobre Apache Jena y un esquema de trazabilidad metodológica que modela el proceso de investigación mediante la ontología OBI.

La sección 2 presenta una visión general del sistema HCES, planteando la brecha semántica que motiva la propuesta y situándola respecto de propuestas como FHIR-Ontop-OMOP (Xiao et al., 2022), NERO (Pesquita et al., 2014) y OMOP CDM (Observational Health Data Sciences and Informatics, 2024) que abordan la integración semántica de datos clínicos desde perspectivas complementarias. La sección 3 desarrolla la arquitectura mediante vistas del modelo adoptado. La sección 4 presenta un caso de uso, la sección 5 las limitaciones detectadas y la sección 6 cuenta con conclusiones y líneas de trabajos a futuro.

2 HCES: Una visión general

HCES es una plataforma semántica para investigación epidemiológica organizada en torno a tres capas funcionales que operan de manera coordinada y persiguen objetivos complementarios. La capa de datos clínicos estandarizados se ocupa de la ingesta, validación y persistencia de recursos FHIR R4, ya sea mediante carga masiva o ingreso a través de formularios generados dinámicamente. A partir de la definición semántica de recursos FHIR R4, el sistema genera automáticamente las interfaces REST, el esquema de validación y el formulario de ingreso, sin requerir desarrollo específico para cualquier recurso FHIR. En segunda instan-

cia se tiene la capa de enriquecimiento semántico, que transforma los conceptos clínicos recibidos en formato JSON a su representación en RDF, vinculando de forma explícita cada concepto FHIR o código clínico con el concepto correspondiente de la terminología SNOMED CT y la referencia semántica de FHIR. Por último se tiene la capa de investigación, que proporciona un entorno colaborativo donde los investigadores construyen progresivamente el dominio de conocimiento de cada estudio, ejecutan métodos analíticos y registran sus hallazgos en una representación que vincula resultados, datos y metodología.

El aporte de la capa de enriquecimiento semántico merece una justificación específica, dado que FHIR R4 y SNOMED CT ya proveen estructuras de datos y terminologías con semántica definida. Sin embargo, un recurso FHIR en formato JSON contiene referencias a códigos SNOMED CT como cadenas de texto planas dentro de objetos Coding, sin vínculo formal con la ontología subyacente. Un recurso Condition con código de hipertensión arterial en formato FHIR JSON no conoce que ese concepto es subclase de trastorno del sistema circulatorio en la jerarquía SNOMED CT, ni puede generar inferencias con otros recursos. El enriquecimiento que realiza el componente JsonLD Traductor transforma ese concepto en un recurso de un subgrafo RDF donde el código clínico se conecta explícitamente con el concepto SNOMED CT como instancia owl:NamedIndividual y el recurso Condition queda clasificado bajo la jerarquía FHIR como instancia formal. Además, las propiedades del recurso se mapean al espacio de nombres RDF de FHIR. Esta transformación habilita consultas SPARQL federadas en distintos grafos, que exploran simultáneamente la estructura FHIR, la jerarquía SNOMED CT y la ontología OBI, capacidad que queda limitada cuando los datos permanecen en formato JSON documental.

Esta organización se reconoce también en sistemas con objetivos similares, como por ejemplo la virtualización de datos clínicos en FHIR-Ontop-OMOP, la federación de ontologías especializadas en NERO y la adopción de modelos de datos comunes en OMOP CDM. Lo que distingue a HCES es la integración de las tres capas en una arquitectura única. Mientras FHIR-Ontop-OMOP virtualiza datos existentes en OMOP como grafos RDF FHIR sin modificar la fuente, HCES enriquece semánticamente cada recurso durante la ingesta y lo persiste en ambos formatos. NERO provee una red de ontologías para anotación de recursos epidemiológicos, HCES opera directamente sobre datos clínicos estandarizados y formaliza el proceso de investigación mediante OBI. OMOP CDM estandariza datos para el análisis a gran escala, pero sobre un modelo relacional sin representación ontológica nativa, ni capacidad de consultas semánticas sobre los datos transformados. HCES, al integrar las capas mencionadas, permite que el enriquecimiento semántico, la exploración del dominio clínico y la trazabilidad del proceso de investigación operen sobre una base semántica, preservando el contexto semántico a lo largo de todo el ciclo de vida de la investigación.

Estos componentes operan de forma coordinada sobre dos plataformas centrales del sistema, Semantic HCE FHIR y Semantic TiddlyWiki(Ruston, 2024). En la Figura 1 se muestra el modelo conceptual, donde se observa el flujo de información desde las fuentes de datos clínicos hasta el espacio de investigación.

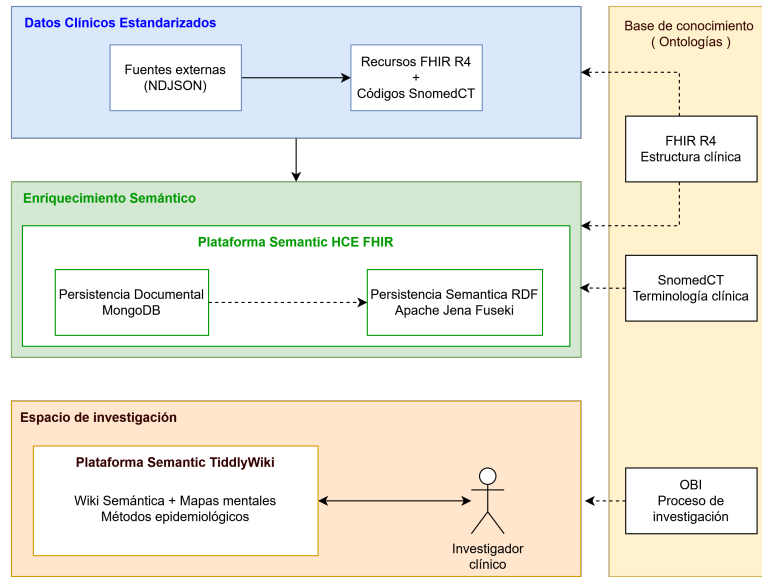


Fig.1. Modelo conceptual de HCES. Tres capas funcionales (Datos clínicos estandarizados, enriquecimiento semántico y espacio de investigación), soportadas sobre dos plataformas centrales (Semantic HCE FHIR y Semantic TiddlyWiki), conectadas por la base semántica (FHIR R4, SNOMED CT, OBI).

3 Arquitectura

La arquitectura del sistema HCES se describe mediante un modelo de vistas, marco metodológico ampliamente utilizado para documentar sistemas de software mediante un conjunto reducido de perspectivas complementarias. El presente artículo se concentra en la vista lógica que organiza los componentes y el conocimiento semántico, la vista de procesos que describe los flujos de ejecución en tiempo real y la vista de desarrollo que expone la organización del software. La vista física, que describe la topología de despliegue sobre la infraestructura, y una vista de escenarios quedan fuera del alcance de este artículo.

El marco metodológico del proyecto se apoya en un conjunto de decisiones transversales que, junto con los criterios de diseño detallados en la vista de desarrollo, constituyen los fundamentos que guiaron la arquitectura. La primera decisión transversal es el carácter híbrido del sistema, que combina un modelo orientado a servicios para operaciones transaccionales sobre recursos FHIR y un modelo orientado a eventos para la sincronización del conocimiento semántico entre el entorno de investigación y la base semántica. Esta combinación permite que las operaciones de ingesta de datos clínicos y las operaciones de exploración interactiva del dominio de investigación ocurran de manera desacoplada, donde cada operación se base en el patrón de comunicación más adecuado a sus requi-

sitos de comunicación y consistencia entre el conocimiento persistido en la base semántica y el disponible para el investigador.

La segunda decisión transversal es la arquitectura de almacenamiento dual, que persiste cada recurso clínico simultáneamente en formato documental sobre una base de datos MongoDB y en formato semántico RDF sobre Apache Jena Fuseki. Esta decisión arquitectónica atraviesa las tres vistas presentadas, aparece en la vista lógica como separación entre los componentes de gestión documental y de gestión semántica, en la vista de procesos como protocolo transaccional que coordina ambas persistencias y en la vista de desarrollo como capa de persistencia con dos lógicas heterogéneas. Este diseño se justifica en tres requisitos, trazabilidad de auditoría sobre el registro original sin transformaciones, capacidad de reconstrucción del grafo semántico ante cualquier inconsistencia durante el proceso de transformación semántica y la utilización de cada tecnología de almacenamiento en las operaciones donde cada una resulta más apropiada.

La tercera decisión transversal es la separación del conocimiento semántico en grafos RDF independientes alojados en una única instancia de Apache Jena, en lugar de una megaontología unificada en una misma instancia o distribuirlo en distintos SPARQL endpoints, uno por cada ontología. Esta decisión preserva la modularidad ontológica adoptada de NERO y evita la complejidad operativa de la federación distribuida.

3.1 Vista lógica

La vista lógica organiza los componentes del sistema HCES en función de sus responsabilidades sobre el conocimiento clínico y el proceso de investigación. La Figura 2 sintetiza esta organización, mostrando los componentes principales, sus dependencias y el flujo de información entre las plataformas centrales y la base semántica alojada en Apache Jena Fuseki donde se distinguen dos categorías de conocimiento. El conocimiento ontológico, constituido por FHIR R4, SNOMED CT y OBI, se carga en la inicialización y permanece estable, mientras que el generado por la ingesta de datos e investigaciones en curso evoluciona continuamente.

Organización del conocimiento semántico: El conocimiento semántico se estructura en cinco grafos RDF independientes alojados en una única instancia de Apache Jena Fuseki. La separación responde a tres criterios de diseño, el aislamiento en el ciclo de vida de cada operación, la eficiencia en la ejecución de consultas SPARQL al acotar el espacio de dominio de cada grafo y la preservación de la coherencia semántica de cada ontología, evitando que los axiomas de un modelo generen inconsistencias en las inferencias sobre otro.

El grafo de ontología FHIR contiene la especificación semántica del estándar FHIR R4, obtenida de la publicación oficial desde la web de HL7 en formato Turtle. En HCES, esta ontología cumple dos funciones, operar como esquema de validación semántica durante el enriquecimiento de los recursos clínicos y como fuente de metadatos para la generación automatizada de interfaces RESTful FHIR y formularios Web, generados durante la inicialización del sistema.

Cuando el sistema incorpora un nuevo perfil FHIR, la misma definición semántica de recurso genera simultáneamente y de forma automatizada el endpoint REST, el esquema de validación JSON Schema, el JSON Schema Form para el formulario de ingreso, y las restricciones que guían al JsonLD Traductor durante el enriquecimiento.

El grafo de terminología SNOMED CT alberga la jerarquía de conceptos clínicos codificados, generado mediante el `snomed-owl-toolkit`, una herramienta proporcionada por SNOMED International para la conversión de una publicación de tipo snapshot de la terminología a formato OWL. En la implementación actual se cargó la edición internacional completa de SNOMED CT, sin recorte por subconjunto. Esta representación OWL preserva las jerarquías, las relaciones entre los conceptos y los axiomas de equivalencia que habilitan razonamiento. Los conceptos SNOMED CT, referenciados desde los recursos FHIR, se modelan como instancias `owl:NamedIndividual` en lugar de mantenerlos exclusivamente como clases. Este metamodelado, presentado como parte de la propuesta del proyecto, habilita la capacidad de tratar a cada concepto como ítem de datos OBI analizable, coherente con el interés de operar sobre instancias clínicas concretas, además de poder razonar sobre cada jerarquía abstracta de cada ontología.

El grafo de datos clínicos contiene las instancias ingestadas de recursos FHIR, enriquecidas semánticamente. Cada recurso ingresado al sistema, tras su persistencia en la base de datos documental MongoDB, se transforma en un subgrafo RDF mediante el componente JsonLD Traductor, que preserva la estructura FHIR, enriqueciendo cada recurso y creando enlaces explícitos a los conceptos SNOMED CT correspondientes. Este grafo constituye la base empírica de datos del sistema sobre la que se ejecutan las consultas de análisis epidemiológico como análisis de prevalencia, agrupamiento jerárquico y análisis de co-ocurrencias.

El grafo de la ontología OBI, junto con sus ontologías fundacionales BFO (Basic Formal Ontology) e IAO (Information Artifact Ontology), provee el vocabulario formal mediante el cual cada investigación queda representada como un proceso formal, interpretable y reproducible. Sus clases `obi:investigation`, `obi:study_design`, `obi:data_item` y sub clases como `obi:measurement_datum`, `obi:data_transformation`, así como `obi:conclusion_based_on_data`, entre otras, constituyen la base de la trazabilidad metodológica que distingue a HCES de los enfoques centrados exclusivamente en datos.

Por último, el grafo de investigaciones modela cada investigación generada en la plataforma, como un grafo RDF conforme a la ontología OBI. Una investigación se representa como una instancia `obi:investigation`, vinculando a sus participantes que son modelados como instancias `foaf:Person` con soporte opcional de identificadores ORCID. Los ítems de datos seleccionados, dentro del alcance del proyecto, se representan como instancias `obi:measurement_datum`, `obi:diagnosis` o `obi:data_set` según el criterio de cada investigador en relación al rol de cada concepto. Los procesos analíticos planificados se representan como instancias `obi:planned_process` y los resultados generados como instancias `obi:conclusion_based_on_data`. De esta manera, la separación de este grafo respecto al grafo de datos clínicos, permite recuperar, por ejemplo, dentro del con-

texto de un hallazgo, qué investigaciones analizaron un concepto determinado, con qué metodología o en qué período se realizó la investigación, sin generar impacto de los datos clínicos en las consultas sobre el proceso de investigación.

Módulos funcionales: La arquitectura lógica agrupa estos componentes en cinco módulos funcionales, cada uno asociado a una fase del ciclo de vida del conocimiento. El módulo de inicialización, mediante los componentes Semantic Definitions FHIR y LoadResource, deriva de las definiciones semánticas FHIR los artefactos de interfaces y de servicios web, validación e ingesta de recursos. El módulo de interacción y presentación reúne el entorno de investigación de tipo Wiki con Semantic TiddlyWiki, con la visualización de relaciones mediante mapas mentales con TiddlyMaps, incorporando los formularios clínicos generados dinámicamente durante la inicialización. El módulo de gestión de conceptos y vistas articula el entorno mediante el Concept and View Manager que opera sobre catálogos de plantillas SPARQL y de visualización parametrizables. Esto permite incorporar nuevos conceptos o representaciones a partir de su definición semántica sin modificar los componentes existentes. El módulo de almacenamiento y consultas se compone de una persistencia dual, con el Knowledge Store sobre Apache Jena Fuseki y el Document Management sobre MongoDB, acoplada por el JsonLD Traductor, que enriquece semánticamente cada escritura documental y la propaga a la base semántica Apache Jena. Por último, el módulo analítico y relacional coordina la comunicación bidireccional entre la capa de presentación y la base semántica a través de los *listeners* Node Change, View Change y loadResearch, cuya interacción se detalla en la vista de procesos.

3.2 Vista de procesos

La vista de procesos describe cómo los componentes presentados en la vista lógica coordinan su operación en tiempo de ejecución. Esta vista se concentra en los flujos de control que atraviesan el sistema durante las tres situaciones representativas del ciclo de vida del conocimiento, la incorporación de nuevos recursos clínicos al sistema, la exploración iterativa del dominio por parte del investigador y la ejecución de métodos de análisis epidemiológico sobre el conocimiento acumulado.

Enriquecimiento semántico de recursos FHIR: El primer flujo transforma cada recurso FHIR que llega a las APIs REST FHIR a su representación RDF. El flujo se inicia cuando un recurso FHIR arriba a las APIs REST, ya sea por ingreso manual o por carga masiva desde archivos NDJSON. La API primero realiza una validación del recurso contra el JSON Schema derivado de las definiciones FHIR durante la inicialización del sistema y rechaza las peticiones malformadas. Una vez validada la petición, avanza a la primera fase de persistencia. El componente Logical Model FHIR construye el documento y lo persiste en MongoDB, operación que retorna el identificador único del recurso. A continuación,

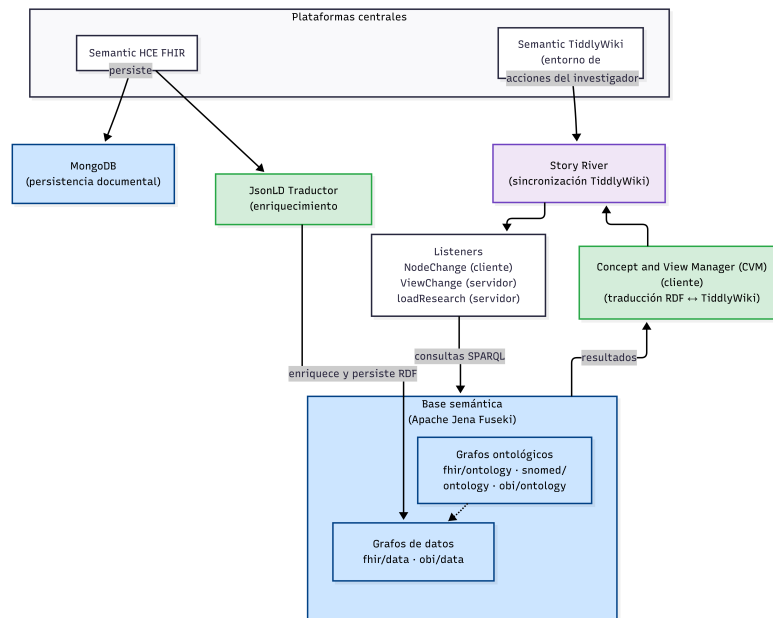


Fig. 2. Vista lógica de HCES. Componentes principales y flujos de información entre las dos plataformas centrales (Semantic HCE FHIR y Semantic TiddlyWiki) y la base semántica Apache Jena Fuseki. Los componentes JsonLD Traductor y CVM constituyen contribuciones originales del proyecto.

el documento persistido se transfiere al componente JsonLD Traductor, que ejecuta tres transformaciones sucesivas, primero inyecta el contexto JSON-LD que mapea las propiedades del recurso al espacio de nombres FHIR, luego genera las triplas que vinculan el recurso con los conceptos SNOMED CT correspondientes a partir de las propiedades de tipo Coding y luego clasifica el recurso bajo la jerarquía de FHIR correspondiente. El subgrafo resultante se persiste en el grafo fhir/data en la instancia de Apache Jena Fuseki.

Exploración del dominio de investigación: El segundo flujo permite al investigador construir progresivamente el dominio de conocimiento de cada investigación mediante la expansión iterativa del conocimiento en el espacio de trabajo de Semantic TiddlyWiki. Este flujo está orientado a eventos y su mecanismo de coordinación cliente-servidor merece una descripción específica por su carácter distintivo. Cuando el investigador referencia un concepto para su investigación o realiza un doble clic sobre un nodo de un mapa mental en TiddlyMaps, la extensión desarrollada abre el *tiddler* correspondiente al concepto en el *story river* de TiddlyWiki. *Story river* es la lista ordenada de *tiddlers* visibles que se constituye en el modelo de presentación principal del entorno. Si el concepto no existe aún como *tiddler* en la wiki, el sistema lo crea marcado con el parámetro

know=true, una bandera que indica que su contenido debe recuperarse desde la base semántica. Esta creación es un cambio en el estado persistente de la wiki, mecanismo de sincronización estándar de TiddlyWiki que propaga automáticamente de forma bidireccional entre el servidor y cada cliente a través de la sincronización del propio modelo de datos compartido entre ambos entornos. De esta manera, el story river actúa como bus de eventos, en el que las acciones del investigador se ejecutan como transiciones de estado observables por los componentes del lado del servidor. El *listener*¹ View Change, desplegado en el servidor TiddlyWiki, detecta la aparición del *tiddler* nuevo con know=true, identifica el concepto al que refiere y selecciona la plantilla SPARQL correspondiente al grafo al que pertenece el recurso, luego parametriza con la URI del concepto dicha consulta. Esta consulta se ejecuta en la base semántica fhir si se trata de una clase FHIR o snomedct si se trata de un concepto clínico, representada en la Figura 2. Los resultados recuperados llegan al cliente y el *listener* Node Change detecta estas actualizaciones y entrega este nuevo conocimiento al Concept and View Manager, que los traduce en *tiddlers* representativos utilizando plantillas para cada ontología. Además, incorpora cada recurso como nodos adicionales del mapa mental que corresponda.

Ejecución de análisis epidemiológico: El tercer flujo permite aplicar métodos formales de investigación como consultas SPARQL genéricas y parametrizables, ejecutadas sobre el conocimiento acumulado y persiste sus resultados como entidades OBI en el grafo de investigaciones. El flujo se dispara cuando el investigador solicita desde el espacio de trabajo la ejecución de uno de los métodos de análisis soportados sobre conceptos vinculados a la investigación en curso. El módulo analítico y relacional recupera la plantilla SPARQL correspondiente al método solicitado y la parametriza con las instancias obi asociadas al análisis. La consulta se ejecuta sobre la instancia de Apache Jena accediendo al grafo fhir/data. Una vez obtenidos los resultados, el sistema registra la trazabilidad metodológica mediante la persistencia en el grafo obi/data de tres elementos vinculados, una instancia de obi:data_transformation que representa el método aplicado, los ítems de datos utilizados como entrada mediante obi:has_specified_input, y una instancia de obi:conclusion_based_on_data con el resultado del análisis vinculada mediante obi:has_specified_output. Los resultados se entregan al investigador y se renderizan en las estructuras visuales apropiadas al método seleccionado.

3.3 Vista de desarrollo

La vista de desarrollo describe la organización del software del sistema HCES desde la perspectiva de su construcción y mantenimiento. La arquitectura cuenta con dos plataformas independientes que pueden evolucionar por separado pero que cooperan en tiempo de ejecución a través de la base semántica común. La

¹ Componente de software que observa eventos en el sistema y ejecuta acciones en respuesta.

plataforma Semantic HCE FHIR responsable de la ingesta y el enriquecimiento semántico de los recursos clínicos y la plataforma Semantic TiddlyWiki que provee el entorno de investigación y la visualización del conocimiento.

Plataforma Semantic HCE FHIR: La plataforma Semantic HCE FHIR se implementa como una aplicación Node.js organizada en cuatro capas con responsabilidades bien diferenciadas. La capa de inicialización genera los esquemas derivados de las definiciones FHIR. Sobre estos esquemas, la librería *express-restify-mongoose* construye dinámicamente los endpoints REST, los modelos de persistencia documental sobre Mongoose para MongoDB y la incorporación del cliente HTTP para Apache Jena, resolviendo en un mismo artefacto la capa de servicios y la capa de persistencia dual. El *JsonLD Traductor* se incorpora como función de extensión que se ejecuta luego de cada persistencia en la base documental, delegando el resultado RDF a Apache Jena Fuseki vía SPARQL Update sobre HTTP.

Plataforma Semantic TiddlyWiki: La plataforma Semantic TiddlyWiki extiende el sistema TiddlyWiki estándar con capacidades semánticas mediante plugins que registran los *listeners* en el sistema de eventos para capturar las modificaciones relevantes del story river, definen las macros que permiten ejecutar consultas SPARQL desde el lenguaje wiki, proveen los widgets que renderizan los resultados e implementa el conector con Apache Jena. El Concept and View Manager se organiza en un componente servidor que procesa los resultados SPARQL y genera las estructuras de datos que describen los *tiddlers* a crear, un componente cliente en el navegador, que interpreta esas estructuras y crea los *tiddlers*. La integración del plugin TiddlyMaps provee las capacidades de mapas mentales que permiten al investigador navegar en las relaciones semánticas.

Criterios de diseño: Cuatro criterios de diseño se reconocen transversalmente en la construcción del sistema. El criterio de generación sobre configuración se aplica en los artefactos derivados de las definiciones semánticas FHIR que se generan automáticamente durante la inicialización del sistema en lugar de mantenerse como código fijo, garantizando la coherencia entre las capas semántica, lógica y de presentación. El criterio de separación semántica organiza el conocimiento RDF en grafos independientes según su procedencia y su ciclo de vida. El criterio de conocimiento progresivo orienta la sincronización entre Semantic TiddlyWiki y Apache Jena, en lugar de cargar todo el conocimiento disponible al iniciar una investigación, el sistema carga solo el concepto seleccionado por el investigador y permite la expansión bajo demanda. El criterio de extensibilidad mediante plantillas permite que el sistema soporte nuevos métodos de análisis epidemiológico o nuevos tipos de concepto de nuevas ontologías, sin modificar los componentes principales.

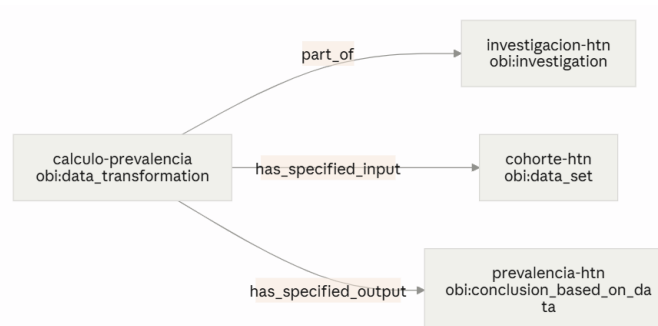


Fig. 3. Traza metodológica en el grafo de investigaciones, que enlaza el método, los datos y la conclusión dentro de la investigación que les da contexto.

4 Caso de uso

La arquitectura descrita se implementó en un prototipo funcional que cubre la ingesta y validación de recursos FHIR R4, su enriquecimiento semántico mediante el JsonLD Traductor, la persistencia dual sobre MongoDB y Apache Jena Fuseki, y la ejecución de métodos de análisis sobre el conocimiento acumulado. Para ilustrar la operación conjunta de estos componentes se desarrolla un caso de uso mínimo de análisis de prevalencia, con datos de carácter ilustrativo. No constituye una validación empírica, sino una demostración del flujo metodológico con el alcance actual.

A partir de una cohorte de recursos `Condition`, el investigador solicita el método de prevalencia sobre el concepto de hipertensión, código SNOMED CT 38341003. El módulo analítico recupera la plantilla SPARQL del método, la parametriza con la URI del concepto clínico en análisis y la ejecuta sobre el grafo de datos clínicos, como se muestra a continuación:

```

PREFIX fhir: <http://hl7.org/fhir/>
PREFIX sct: <http://snomed.info/id/>

SELECT (COUNT(DISTINCT ?cond) AS ?prevalencia)
FROM <http://hces/graph/fhir/data>
WHERE {
  ?cond a fhir:Condition ;
        fhir:code [ fhir:coding ( [ a sct:38341003 ] ) ] .
}
  
```

Durante la ingesta, el JsonLD Traductor agrega sobre el nodo `Coding` una afirmación de tipo (`rdf:type`) hacia los conceptos SNOMED CT, haciendo explícita una referencia que de otro modo permanece implícita en los literales `system` y `code`. Esta afirmación permite contabilizar los recursos del concepto y, extendiendo el recorrido a través de la propiedad `rdfs:subClassOf*` sobre el grafo de términos, agrupar jerárquicamente sus subconceptos.

Luego de obtener el resultado, el sistema lo incorpora al grafo de investigaciones. El método de análisis queda registrado como un proceso de la investigación con su procedencia, como se muestra en la Figura 3. Esta representación permite asociar el método, el cálculo, los datos que lo originaron y la conclusión resultante, recuperables, dentro de la investigación que les da sentido, aislando las consultas sobre el proceso de investigación de los datos clínicos.

5 Limitaciones

La arquitectura presentada, acompañada de un prototipo funcional, cuenta con un alcance que tiene límites que conviene explicitar. En primer lugar, no se dispone aún de una evaluación empírica sistemática, el caso de la Sección 4 ilustra el flujo metodológico, pero no constituye un estudio de validación con datos clínicos reales, métricas de desempeño ni comparación cuantitativa con otros enfoques de grafos de conocimiento clínico.

En segundo lugar, el enriquecimiento semántico que produce el JsonLD Traductor se genera de forma programática y su corrección se verificó mediante revisión experta. La incorporación de validación de conformidad sobre la salida RDF constituye una línea de trabajo a futuro necesaria.

En tercer lugar, el desempeño y la escalabilidad de Apache Jena Fuseki frente a volúmenes clínicos de magnitud real no han sido caracterizados, y representan un aspecto a evaluar antes de un despliegue en producción. Finalmente, como se señala en la Sección 6, la arquitectura persiste el conocimiento ontológico completo pero todavía no explota las capacidades de razonamiento OWL, lo que acota el análisis a las relaciones explícitamente representadas dentro del potencial previsto en la propuesta.

6 Conclusiones

Este artículo presenta la arquitectura de una plataforma semántica para investigación epidemiológica que aborda la brecha entre la disponibilidad de datos clínicos estandarizados y su explotación efectiva mediante tecnologías de la web semántica. Las cuatro contribuciones presentadas, el componente JsonLD Traductor que enriquece semánticamente cada recurso FHIR transformándolo en un subgrafo RDF consultable, el Concept and View Manager que traduce representaciones ontológicas en vistas navegables para el investigador, la arquitectura de almacenamiento dual que preserva el registro original mientras habilita la explotación semántica y la trazabilidad metodológica mediante OBI que formaliza el proceso de investigación como conocimiento recuperable, conforman una arquitectura integrada que preserva y explota el potencial semántico de los datos clínicos a lo largo de todo el ciclo de investigación epidemiológica.

Sobre la base de la arquitectura presentada se identifican dos líneas de trabajos a futuro de impacto significativo. La primera es la interoperabilidad bidireccional con repositorios basados en OMOP CDM mediante FHIR como estándar puente, lo que permitiría consumir y producir datos conformes al modelo de

datos observacionales más extendido a nivel global. La segunda es la incorporación de mecanismos de inferencia automática sobre la base semántica, dado que la arquitectura actual persiste el conocimiento ontológico completo pero no explota las capacidades de razonamiento OWL de Apache Jena para generar nuevo conocimiento derivado, como por ejemplo la clasificación automática de pacientes en cohortes a partir de axiomas SNOMED CT o la detección de relaciones implícitas entre diagnósticos que no fueron explícitamente vinculados por el investigador. Adicionalmente, la estructura de grafos interconectados de HCES abre oportunidades para la aplicación de técnicas de aprendizaje automático como predicción de enlaces sobre el conocimiento acumulado, la representación formal de procesos asistenciales mediante ontologías de procesos y la incorporación de fuentes de datos provenientes de dispositivos de monitoreo remoto.

References

- Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M. H., Bug, B., Chibucos, M. C., Clancy, K., Courtot, M., Derom, D., Dumontier, M., Fan, L., Fostel, J., Fragoso, G., Gibson, F., Gonzalez-Beltran, A., Haendel, M. A., He, Y., Heiskanen, M., Hernandez-Boussard, T., . . . Zheng, J. (2016). The ontology for biomedical investigations. *PLoS ONE*, *11*(4), e0154556. <https://doi.org/10.1371/journal.pone.0154556>
- Health Level Seven International. (2019). HL7 FHIR release 4 (R4) [Consultado el 15 de abril de 2026]. <https://www.hl7.org/fhir/R4/>
- Observational Health Data Sciences and Informatics. (2024). OMOP common data model [Consultado el 15 de abril de 2026]. <https://ohdsi.github.io/CommonDataModel/>
- Pesquita, C., Faria, D., Falcão, A. O., Lord, P., & Couto, F. M. (2014). The epidemiology ontology: An ontology for the semantic annotation of epidemiological resources. *Journal of Biomedical Semantics*, *5*, 4. <https://doi.org/10.1186/2041-1480-5-4>
- Ruston, J. (2024). TiddlyWiki — a non-linear personal web notebook [Consultado el 15 de abril de 2026]. <https://tiddlywiki.com/>
- SNOMED International. (2024). SNOMED CT [Consultado el 15 de abril de 2026]. <https://www.snomed.org/>
- Xiao, G., Pfaff, E., Prud'hommeaux, E., Booth, D., Sharma, D. K., Huo, N., Yu, Y., Zong, N., Ruddy, K. J., Chute, C. G., & Jiang, G. (2022). FHIR-Ontop-OMOP: Building clinical knowledge graphs in FHIR RDF with the OMOP common data model. *Journal of Biomedical Informatics*, *134*, 104201. <https://doi.org/10.1016/j.jbi.2022.104201>