UNIVERSIDAD DE MÁLAGA

Proyecto Final

Estándares de Datos Abiertos e Integración de Datos

Autores

Achraf Ousti El Moussati
Sebastián Rozenblum
Anabel Yu Flores Moral
Gabriela Milenova Yordanova
Karen Michell Herrera Sierra

Grado en Ingeniería de la Salud
E.T.S.I. Ingeniería Informática
Universidad de Málaga

Curso 2025/2026

Abstract

HOLA. PRUEBA.

Keywords: (e.g.Mobile computing, security, complexity, algorithms, image analysis, machine learning, information retrieval)

Contents

Abstract i

Contenidos ii

1 Introducción 1

1.1 Marco teórico 1

1.2 Objetivos del proyecto 2

1.3 Relevancia en bioinformática 3

1.4 Estructura del documento 3

2 Literature Review 5

3 Methodology 6

3.0.1 Colección: `patients` 6

3.0.2 Colección: `samples` 7

3.0.3 Colección: `variants` 7

4 Results 9

Chapter 1

Introducción

1.1 Marco teórico

En el campo de la bioinformática, la gestión eficiente y el análisis de grandes volúmenes de datos genómicos y clínicos son fundamentales para avanzar en la comprensión y el tratamiento de enfermedades como el cáncer. Los investigadores y clínicos suelen trabajar con información que proviene de múltiples fuentes: historiales clínicos de pacientes, análisis de muestras biológicas, secuenciación genómica, anotaciones moleculares y bases de datos de conocimiento oncológico. Esta diversidad de datos, junto con su naturaleza altamente interconectada y jerárquica, plantea importantes limitaciones para los sistemas de bases de datos relacionales tradicionales. Este proyecto se centra en el diseño y la implementación de una base de datos NoSQL, utilizando MongoDB, para albergar y organizar los datos del estudio *Acral Melanoma (TGEN, Genome Res 2017)*. Este conjunto de datos, disponible a través del cBioPortal for Cancer Genomics, ofrece una rica fuente de información genómica y clínica de pacientes con melanoma acral, un subtipo raro y agresivo de melanoma (cáncer de piel).

Dicho repositorio ofrece un conjunto de datos rico y representativo que incluye información clínica detallada de pacientes, características de muestras tumorales, y miles de variantes genómicas identificadas mediante secuenciación del exoma completo. Este tipo de datos biomédicos presenta características únicas que requieren soluciones tecnológicas específicas:

- **Estructura altamente anidada**: la información clínica de un paciente contiene múltiples niveles de detalle (datos demográficos, historial de tratamientos, información de seguimiento), cada uno con su propia estructura interna.
- **Relaciones complejas entre entidades**: un paciente puede tener múltiples muestras, cada muestra puede contener miles de variantes, y cada variante afecta a genes específicos con implicaciones clínicas conocidas.
- **Necesidad de enriquecimiento continuo**: los datos genómicos requieren integración con bases de datos externas (como OncoKB para significancia clínica de mutaciones, o UniProt para información proteica) que evolucionan constantemente.

- **Consultas multidimensionales**: los investigadores necesitan realizar análisis que cruzan información de pacientes, muestras y variantes de forma simultánea.

1.2 Objetivos del proyecto

Este proyecto aborda la problemática descrita mediante el diseño e implementación de un sistema integrado de gestión de datos clínicos y genómicos, aplicando tres tecnologías complementarias de estándares de datos abiertos e integración:

1. **Bases de datos NoSQL (MongoDB)** para el almacenamiento flexible y escalable de datos biomédicos jerárquicos.
2. **Tecnologías de transformación semántica (XML/XSLT)** para la generación automática de reportes clínicos visuales.
3. **Web Semántica y ontologías (OWL/SPARQL)** para la representación formal del conocimiento y consultas avanzadas.

El objetivo principal es demostrar cómo estas tecnologías, habitualmente estudiadas de forma aislada, pueden integrarse en un flujo de trabajo completo que va desde la captura de datos crudos hasta la consulta semántica del conocimiento biomédico. Específicamente, el proyecto implementa:

- Un **modelo de datos NoSQL** con tres colecciones principales (`patients`, `samples`, `variants`) más dos colecciones de enriquecimiento externo (`oncokb_genes`, `uniprot`), cada una con estructuras anidadas de hasta tres niveles que capturan la complejidad inherente a los datos clínico-genómicos.
- Un **pipeline ETL automatizado** que limpia, reestructura y enriquece los datos originales del estudio de melanoma acral, integrándolos con información actualizada de APIs públicas de relevancia oncológica.
- Un **sistema de generación de reportes** que consulta la base de datos MongoDB, transforma los resultados a XML y aplica plantillas XSLT para producir dashboards HTML interactivos que faciliten la visualización de relaciones complejas entre pacientes, muestras y variantes.
- Una **ontología OWL formal** diseñada en Protégé que modela todo el dominio del conocimiento clínico-genómico representado en la base de datos, incluyendo clases, propiedades de objeto, propiedades de datos y restricciones.
- **Capacidades de razonamiento automático** mediante reasoners como HermiT, que permiten inferir nuevo conocimiento (por ejemplo, clasificar automáticamente pacientes según su estado clínico o identificar muestras metastásicas).
- **Consultas SPARQL avanzadas** que explotan la semántica de la ontología para realizar búsquedas que serían difíciles o imposibles en bases de datos tradicionales

- **Generación automática de grafos RDF** a partir de los datos almacenados en MongoDB, creando un puente entre el almacenamiento NoSQL y la representación semántica.

1.3 Relevancia en bioinformática

La aproximación presentada en este trabajo es particularmente relevante para la bioinformática moderna por varias razones:

Gestión de heterogeneidad: Las bases de datos NoSQL permiten almacenar datos biomédicos sin forzarlos a esquemas rígidos predefinidos. Esto es crucial en investigación oncológica, donde nuevos biomarcadores, tratamientos o metodologías de secuenciación pueden requerir modificaciones frecuentes del modelo de datos.

Trazabilidad y reproducibilidad: La transformación de datos mediante tecnologías estándar como XML/XSLT garantiza que los reportes clínicos sean generados de forma determinista y auditab, esencial para la validación de resultados de investigación y para cumplir con regulaciones de datos clínicos.

Interoperabilidad semántica: El uso de ontologías OWL permite que diferentes sistemas y bases de datos biomédicas puedan compartir e integrar conocimiento de forma inequívoca. Un paciente clasificado como `PacienteFallecido` en nuestra ontología puede ser automáticamente reconocido y procesado por otros sistemas que utilicen estándares ontológicos compatibles.

Consultas basadas en conocimiento: SPARQL permite formular preguntas complejas que van más allá de la simple recuperación de datos. Por ejemplo, ”encontrar todos los pacientes con variantes oncogénicas en genes asociados a resistencia terapéutica que además presentaron recurrencia de la enfermedad” es una consulta que explota tanto los datos como el conocimiento representado en la ontología.

Escalabilidad hacia medicina personalizada: La arquitectura propuesta sienta las bases para sistemas más complejos de apoyo a la decisión clínica, donde la integración de datos genómicos, clínicos y de bases de conocimiento externas es fundamental para identificar estrategias terapéuticas personalizadas.

1.4 Estructura del documento

Este documento está organizado de la siguiente manera:

- El Capítulo 1 introduce el contexto general del trabajo, exponiendo la motivación, los objetivos, la estructura del documento y las tecnologías empleadas.
- El Capítulo 2 describe la metodología empleada, detallando el diseño de la base de datos MongoDB, el pipeline ETL y de enriquecimiento, el sistema de generación de reportes mediante XML/XSLT, el modelado ontológico en Protégé, y la implementación de los scripts de generación de grafos RDF y ejecución de consultas SPARQL.

- El Capítulo 3 presenta los resultados obtenidos, incluyendo estadísticas descriptivas de los datos almacenados, ejemplos de reportes HTML generados, la ontología resultante con sus inferencias, y los resultados de las consultas SPARQL sobre el grafo RDF.
- El Capítulo 4 discute las conclusiones del trabajo, reflexiona sobre las limitaciones encontradas, y propone líneas de trabajo futuro para extender este sistema hacia aplicaciones clínicas reales y su integración con otras fuentes de datos biomédicos.
- Finalmente, el Capítulo 5 expone las conclusiones del trabajo, incluyendo los aprendizajes alcanzados, las limitaciones enfrentadas y posibles direcciones a seguir en trabajos futuros dentro de esta línea.

Chapter 2

Literature Review

Chapter 3

Methodology

Para estructurar la información de manera coherente y facilitar consultas complejas, se ha optado por un diseño que consta de **tres colecciones interconectadas**: `patients`, `samples` y `variants`. Esta estructura nos permitirá no solo capturar la información de cada paciente de forma individual, sino también trazar las relaciones entre los pacientes, las muestras biológicas obtenidas de ellos y las variantes genómicas identificadas en dichas muestras.

A continuación, se detalla la estructura propuesta para cada una de las colecciones:

3.0.1 Colección: `patients`

Esta colección almacena la información demográfica y clínica de cada paciente incluido en el estudio.

- **Nivel 1:** Información básica del paciente.
  - `patient_id`: Identificador único del paciente.
  - `sex`: Sexo del paciente.
  - `race_category`: Categoría racial del paciente.
  - `age_at_diagnosis`: Edad del paciente en el momento del diagnóstico.
- **Nivel 2:** Historial clínico y de tratamiento.
  - `clinical_history`: Objeto con información sobre el historial médico del paciente.
    - `initial_diagnosis_date`: Fecha del diagnóstico inicial.
    - `primary_tumor_site`: Localización del tumor primario.
  - `treatments`: Array de objetos que detalla los tratamientos recibidos.
    - `treatment_type`: Tipo de tratamiento (e.g., "Ipilimumab", "Interferon").
    - `start_date`: Fecha de inicio del tratamiento.
    - `end_date`: Fecha de finalización del tratamiento.
- **Nivel 3:** Seguimiento y estado de la enfermedad.

  - `follow_up`: Objeto con información de seguimiento.
    * `disease_free_months`: Meses libre de enfermedad.
    * `disease_free_status`: Estado de la enfermedad (e.g., "0:DiseaseFree", "1:Recurred/Progressed").

3.0.2 Colección: `samples`

Contendrá información detallada sobre cada muestra biológica extraída de los pacientes.

- **Nivel 1:** Identificación y tipo de muestra.
  - `sample_id`: Identificador único de la muestra.
  - `patient_id`: Identificador del paciente al que pertenece la muestra (referencia a la colección `patients`).
  - `sample_type`: Tipo de muestra (e.g., "Primary", "Metastasis").
- **Nivel 2:** Detalles de la recolección y procesamiento.
  - `collection_info`: Objeto con detalles de la recolección.
    * `collection_date`: Fecha de recolección de la muestra.
    * `collection_method`: Método de recolección.
  - `processing_info`: Objeto con información del procesamiento.
    * `processing_date`: Fecha de procesamiento.
    * `sequencing_type`: Tipo de secuenciación realizada (e.g., "Whole Exome Sequencing").
- **Nivel 3:** Datos de análisis molecular.
  - `molecular_data`: Objeto que alberga datos moleculares.
    * `mutation_count`: Número de mutaciones identificadas.
    * `copy_number_alterations`: Array de objetos con información sobre alteraciones en el número de copias.

3.0.3 Colección: `variants`

Esta colección albergará la información específica de cada variante genómica identificada en las muestras.

- **Nivel 1:** Identificación de la variante.
  - `variant_id`: Identificador único de la variante.
  - `sample_id`: Identificador de la muestra en la que se encontró la variante (referencia a la colección `samples`).

  - `gene_symbol`: Símbolo del gen afectado.

- **Nivel 2:** Características de la variante.
  - `variant_details`: Objeto con las características de la variante.
    - `chromosome`: Cromosoma donde se localiza la variante.
    - `start_position`: Posición de inicio de la variante.
    - `end_position`: Posición de finalización de la variante.
    - `reference_allele`: Alelo de referencia.
    - `alternate_allele`: Alelo alternativo.

- **Nivel 3:** Anotación funcional y predicciones.
  - `functional_annotation`: Objeto con la anotación funcional.
    - `variant_classification`: Clasificación de la variante (e.g., "Missense_Mutation", "Nonsense_Mutation").
    - `protein_change`: Cambio en la proteína resultante.
    - `sift_prediction`: Predicción del impacto de la variante por SIFT.
    - `polyphen_prediction`: Predicción del impacto de la variante por PolyPhen.

Chapter 4

Results