E5.2-3 Demostrador de Integración de Datos Biomédicos. Arquitectura, despliegue y evaluación v 1.1

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Instituto de Salud Carlos III

IMPaCT

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

Programa	IMPacT: Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología			
Nombre Proyecto	IMPaCT-Data: Programa de Ciencia de Datos de IMPaCT			
Expediente	IMP/00019			
Duración	Enero 2021 – Diciembre 2023			
Página web	impact-data.bsc.es			
Paquete Trabajo	WP5. Integración de Datos			
Tarea	Tarea 5.1 Integración de datos genómicos y radiómicos con datos médicos estructurados para su uso secundario			
Entregable	E5.2-3. Demostrador de Integración de Datos Biomédicos Biomédicos. Arquitectura, despliegue y evaluación (fusión de entregables E5.2 y E5.3)			
Versión	1.1			
Fecha Entrega	23/12/2024		***Fecha Aprobación***	31/12/2024
Responsable	FPS			
Nivel Diseminación	X	PU	Público	
		CO-IMP	Confidencial, sólo participantes de los pilares de IMPaCT, incluyendo la comisión de evaluación de IMPaCT.	
		CO-DATA	Confidencial, sólo participantes de IMPaCT-Data, incluyendo la comisión de evaluación de IMPaCT.	

Autores		
Organización	**Nombre**	**Rol**
Acrónimo	**Nombre y Apellidos**	**Coordinación / Autor / Revisor**
FPS	Joaquin Dopazo	Coordinador/Autor
FPS	Javier Perez Florido	Coordinador/Autor
CNIO	Fátima Al-Sharour	Revisor
FJD	Pablo Mínguez	Revisor
FPS	Juan Carlos Garrido	Autor
FPS	Gema Roldán	Autor

Historial de versiones			
Nro.	**Fecha**	**Descripción**	**Autor**
v 0.1	29/11/2024	Borrador del índice	Joaquín Dopazo (FPS) Javier Pérez Florido (FPS)
v 0.1	29/11/2024	Índice revisado y aprobado	Fátima Al-Sharour (CNIO) Pablo Mínguez (IIS-FJD)
v 0.1	29/11/2024 - 12/12/2024	Elaboración de contenidos	Javier Pérez Florido (FPS) Juan Carlos Garrido (FPS) Gema Roldán (FPS)
v 0.2	12/12/2024	Revisión interna	Joaquín Dopazo (FPS)
v 0.2	13/12/2024	Envío a revisión	-
v 0.2	13/12/2024	Revisión y visto bueno	Pablo Mínguez (IIS-FJD)
v 0.2	18/12/2024	Revisión y visto bueno	Fátima Al-Sharour (CNIO)
v 0.3	23/12/2024	Adaptación con sugerencias de revisores	Javier Pérez Florido (FPS)
v 1.0	23/12/2024	Revisión y versión final	Javier Pérez Florido (FPS) Joaquín Dopazo (FPS)

Contenido

Contenido 3
Tablas 5
Figuras 5
Resumen Ejecutivo 7
Introducción 8
Audiencia 8
Ámbito 8
Relación con otros Entregables 8
Estructura Entregable 8
1 Motivación y esquema general de integración 10
2 Descubrimiento de datos biomédicos 11
2.1 Sistemas *Beacon* 11
2.2. Descubrimiento de datos genómicos 12
2.3. Descubrimiento de datos clínicos 13
2.4. Descubrimiento de datos de imagen 14
3. Integrador de datos biomédicos 15
3.1. Descripción general y objetivos 15
3.2. Arquitectura del integrador 16
3.2.1. Componentes de servicio 16
3.2.2. Flujo de consulta de datos biomédicos 18
3.3. Diseño del integrador 19
3.3.1. Componentes externos 19
3.3.1.1. Bases de conocimiento 19
Bases de datos genómicos 20
Bases de datos clínicos 23
Bases de datos de imagen 24
3.3.1.2. Componentes *Beacon* 26
Beacon genómico 26
Beacon clínico 28
3.3.1.4. *Beacon* de imagen 29
3.3.2. Componentes internos 31
3.3.2.1. Aplicación de navegador 31

3.3.2.2. Servicio de *query* integrada 35
3.4. Despliegue y evaluación del integrador 36
3.4.1. Despliegue de las bases de conocimiento 36
3.4.2. Despliegue de los servicios del integrador 37
3.4.3. Análisis y evaluación de servicios *Beacon* 38
3.4.3.1. *Beacon* genómico 38
3.4.3.2. *Beacon* clínico 38
3.4.3.3. *Beacon* de imagen 39
3.4.4. Análisis y evaluación del servicio integrador 40
3.5. Casos de uso del integrador de datos biomédicos 41
3.5.1. Ejemplo 1. Descubrimiento de pacientes haciendo uso de las tres fuentes de datos 41
3.5.2. Ejemplo 2. Descubrimiento de pacientes haciendo uso de las tres fuentes de datos: búsqueda de pacientes con neumonía asociada a ventilación mecánica y mutación específica en genoma viral 46
4. Discusión y trabajo futuro 48
5. Conclusiones 49
Referencias 50
Acrónimos, abreviaturas y glosario de términos 51

Tablas

Tabla 1. Colecciones y atributos de la base de datos genómicos. 22

Tabla 2. Colecciones y atributos de la base de datos de imagen. 25

Tabla 3. *Endpoints* del servicio Beacon genómico. 28

Tabla 4. *Endpoints* del servicio Beacon clínico. 29

Tabla 5. *Endpoints* del servicio Beacon de imagen. 31

Tabla 6. *Endpoints* de la query integrada. 35

Figuras

Figura 1. Esquema general del proceso de descubrimiento, acceso y análisis de datos en un entorno seguro de investigación que permite estudios federados con repositorios locales. ... 10

Figura 2. Diagrama general de descubrimiento de datos clínicos en Beacon-OMOP 13

Figura 3. Descubrimiento de datos clínicos con el sistema B4OMOP mediante una consulta específica. 14

Figura 4. Diagrama general del integrador de datos biomédico basado en sistemas de descubrimiento tipo Beacon. 15

Figura 5. Arquitectura de servicios a alto nivel del integrador de datos biomédico desarrollado. 17

Figura 6. Diagrama de flujo de consulta al integrador de datos biomédicos. 18

Figura 7. Diagrama entidad-relación de los modelos Beacon v2. 21

Figura 8. Esquema de datos OMOP-CDM. 24

Figura 9. Opción de implementación escogida de un sistema de descubrimiento Beacon v2 para datos genómicos 26

Figura 10. Diagrama del componente de Beacon genómico. 27

Figura 11. Diagrama de componente Beacon clínico. 28

Figura 12. Estructura JSON de ejemplo para consultas al Beacon clínico. 29

Figura 13. Arquitectura del componente Beacon de imagen. 30

Figura 14. Estructura JSON de ejemplo para consultas al Beacon de imagen. 30

Figura 15. Vista de consulta del integrador de datos biomédico desarrollado. 31

Figura 16. Ejemplo de vista de resultados genómicos. 33

Figura 17. Ejemplo de vista de resultados clínicos. 33

Figura 18. Ejemplo de vista de resultados de imagen. 33

Figura 19. Pestaña ABOUT del integrador de datos biomédicos desarrollado. 34

Figura 20. Arquitectura de despliegue del demostrador de integración de datos biomédico desarrollado. 36

Figura 21. Filtros empleados en la búsqueda. 42

Figura 22. Resultados devueltos por el integrador en la categoría Genómica. 42

Figura 23. Resultados devueltos por el integrador en la categoría clínica. 42

Figura 24. Resultados devueltos por el integrador en la categoría de imagen. 43

Figura 25. Filtros empleados en una nueva búsqueda, restringiendo las variantes genómicas a la región 6954-6968 del genoma de referencia de SARS-CoV-2. 43

Figura 26. Resultados devueltos por el integrador en la categoría Genómica en una región más restringida. 44

Figura 27. Resultados devueltos por el integrador en las categorías clínica y de imagen teniendo en cuenta una región genómica más restringida. 44

Figura 28. Filtros empleados en una nueva búsqueda, restringiendo las variantes genómicas a un cambio concreto 6954 T > C. 45

Figura 29. Resultados genómicos, clínicos y de imagen restringiendo la búsqueda genómica a una variante en concreto. 45

Figura 30. Búsqueda a través del integrador de datos biomédico de los individuos que contienen una variante específica del linaje B.1.1.7 de SARS-CoV-2. 46

Figura 31. Resultados genómicos, clínicos y de imagen restringiendo la búsqueda genómica a una variante en concreto de interés. 47

Figura 32. Búsqueda a través del integrador de datos biomédico de los individuos varones que contienen una variante específica del linaje B.1.1.7 de SARS-CoV-2 y con neumonía asociada a ventilación mecánica. 47

Figura 33. Resultados genómicos, clínicos y de imagen restringiendo la búsqueda genómica a una variante en concreto de interés en pacientes varones con neumonía asociada a ventilación mecánica. 48

Resumen Ejecutivo

El **Demostrador de Integración de Datos Biomédicos** desarrollado en el marco del proyecto IMPaCT-Data tiene como objetivo facilitar el descubrimiento y la integración de datos genómicos, clínicos y de imagen médica para su uso secundario en investigaciones biomédicas. Este entregable representa un avance significativo hacia la interoperabilidad de datos biomédicos en España, mediante el desarrollo de una arquitectura basada en sistemas tipo *Beacon* sin comprometer la privacidad del conjunto de datos.

El integrador combina herramientas de descubrimiento basadas en datos heterogéneos, permitiendo consultas unificadas a través de tres servicios *Beacon* (genómico, clínico y de imagen). Como prueba de concepto, se utilizaron datos genómicos y clínicos reales del circuito de secuenciación de SARS-CoV-2 en Andalucía, complementados con datos de imagen sintéticos, demostrando la capacidad del sistema para responder a preguntas heterogéneas de investigación mediante filtros integrados.

Por tanto, este entregable describe los requisitos técnicos y casuísticas para la implementación y despliegue de un integrador de datos biomédicos basado en sistemas *Beacon.*

Introducción

Audiencia

Este entregable está dirigido a los participantes del proyecto IMPaCT-Data como referencia a procedimientos y tecnologías para la implementación de sistemas que integren información clínica, genómica y de imagen médica para su uso en investigación clínica sin comprometer la privacidad de los datos. Se proporciona la información básica del despliegue y puesta a punto de un descubridor basado en sistemas *Beacon* para la integración de datos biomédicos (genómicos, clínicos y de imagen). El documento será de utilidad para cualquier grupo de investigación o institución interesada en implementar o ampliar la implementación de un sistema que integre dicha información.

Ámbito

Este entregable recoge el trabajo sobre implementación realizado sobre estándares de los paquetes de trabajo 3 y 4 del proyecto IMPaCT-Data y que puede ser utilizado como una referencia general para la realización de estudios que requieran de soluciones tipo *Beacon* para la integración de dos o más tipos de datos, tanto en los casos de uso propuestos por el paquete 6, como en las infraestructuras provistas por el paquete 2 y en otros proyectos de las convocatorias de medicina personalizada.

Relación con otros Entregables

Al tratar sobre integración de datos, este entregable guarda relación con otros entregables. Entre ellos, los más relacionados serían: E3.4. Análisis genómico en entornos sanitarios; E4.4. Normas Internacionales de Anotación de Información de Imagen Médica; E4.1. Normas Internacionales de Anotación de Información de HCE; E4.2. Comparación de técnicas de gestión de Información de HCE; E4.5. Comparación de Técnicas de Gestión de Información de Imagen Médica; E4.3-6. Pruebas de Concepto de Extracción de Información de HCE e Imagen; E5.1. Técnicas de Integración de Datos Biomédicos; E5.4. Requisitos técnicos para la puesta en marcha de sistemas Beacon; E5.5. Especificaciones revisadas para la inclusión de marcadores biomédicos de imagen médica; E6.4. Aspectos de seguridad en el manejo de datos sensibles.

Estructura Entregable

El entregable tiene la siguiente estructura:

1. Motivación y esquema general de integración
2. Descubrimiento de datos biomédicos: sistemas *Beacon*

3. Integrador de datos biomédicos: descripción general y objetivos, arquitectura y diseño del integrador, despliegue y evaluación del mismo. Ejemplos de uso.
4. Discusión, trabajo futuro y conclusiones

1 Motivación y esquema general de integración

El principal objetivo de **IMPaCT-Data** es crear la primera iteración, a modo de prueba de concepto, de una infraestructura para el uso secundario de los datos de los sistemas sanitarios españoles, que incluyen historias clínicas electrónicas, datos de imagen médica y datos depositados en repositorios genómicos. Uno de los objetivos técnicos de IMPaCT-Data es el desarrollo de prototipos de integración de los resultados del análisis genómico y de imagen con la información normalizada extraída de las Historias Clínicas Electrónicas (HCE). Estos prototipos, ponen a disposición de los investigadores/as herramientas que le permitan descubrir, mediante consultas integradas, si los repositorios de datos disponen de los datos necesarios para realizar un determinado estudio. Una vez detectados los datos de interés, el investigador deberá solicitar el acceso de acuerdo con lo que marca la regulación y las normativas de acceso a datos del repositorio y si éste es concedido, los datos se podrán usar bajo las condiciones autorizadas.

La Figura 1, muestra el esquema general de integración de datos biomédicos. Tal y como se indicó en el entregable E5.1 “Técnicas de integración de datos biomédicos”, partimos de diferentes tipos de datos (clínicos, genómicos y de imagen) distribuidos en diversas entidades dentro del sistema de salud.

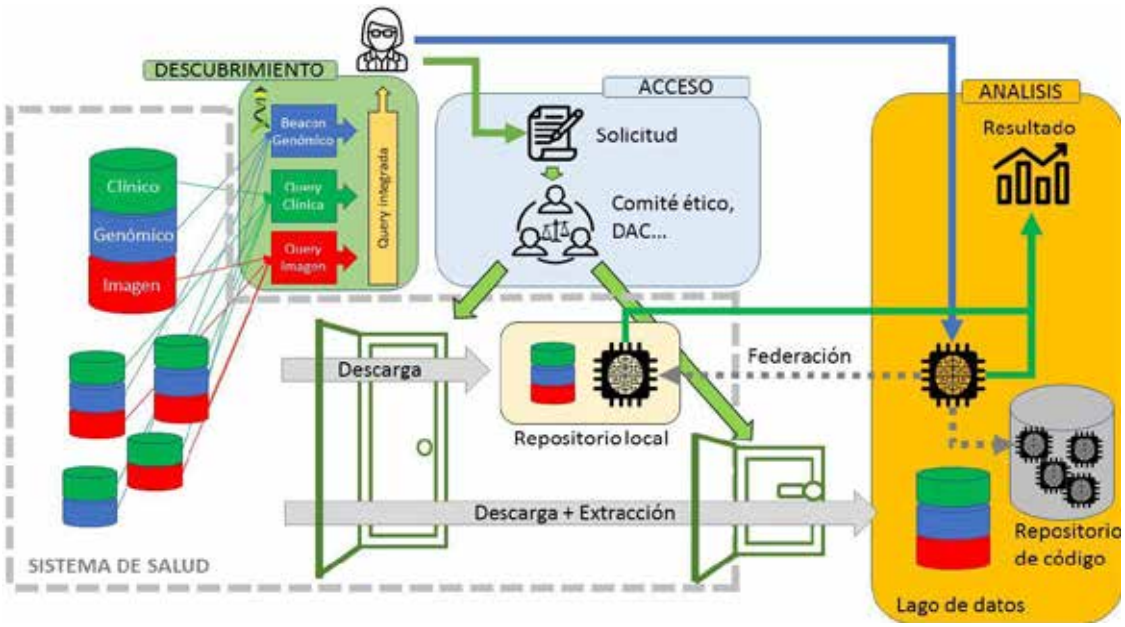

Figura 1. Esquema general del proceso de descubrimiento, acceso y análisis de datos en un entorno seguro de investigación que permite estudios federados con repositorios locales

Podemos distinguir 3 etapas en el proceso de integración:

1. **Descubrimiento,** donde mediante consultas a datos genómicos, clínicos y de imagen mediante sistemas *Beacon,* el investigador podrá descubrir si los repositorios de datos de las entidades disponen de los datos necesarios para un estudio determinado. Este sistema de integración mediante descubrimiento, permitirá consultas de tipo AND/OR, de forma que permite realizar consultas integradas teniendo en cuenta estos tipos de datos. Las consultas sólo obtienen respuestas con datos anonimizados.
2. **Acceso**. Como los datos están sujetos a Regulación General de Protección de Datos (RGPD) [1], una vez detectados los datos de interés, el investigador deberá solicitar el acceso de acuerdo con lo que marca la regulación y las normativas de acceso a datos del repositorio.
3. **Análisis de los datos y visualización**. Una vez se ha obtenido una autorización de uso de los datos de acuerdo al punto anterior, el investigador puede usar herramientas de análisis y visualización para llevar a cabo el estudio autorizado.

En este entregable nos centramos en el desarrollo de un integrador de datos biomédicos a nivel de ***Descubrimiento*** basado en sistemas tipo ***Beacon***.

2 Descubrimiento de datos biomédicos

2.1 Sistemas *Beacon*

De acuerdo a lo especificado en el entregable "E5.4 Requisitos Técnicos para puesta en marcha de sistemas ***Beacon***" y a modo de recordatorio, un sistema *Beacon* es un protocolo estandarizado diseñado para facilitar el descubrimiento de datos genómicos y biomédicos a través de múltiples conjuntos de datos en todo el mundo. Está desarrollado bajo la iniciativa GA4GH (Global Alliance for Genomics and Health) y de ELIXIR. Permite a investigadores y clínicos determinar la presencia o ausencia de variantes genéticas específicas en bases de datos participantes, promoviendo el acceso a información relevante sin comprometer la privacidad de los pacientes ni la propiedad de los datos.

Las características principales de un sistema ***Beacon*** son:

- **Descubrimiento de datos**: Permite a los usuarios consultar conjuntos de datos para consultar la existencia de variantes genéticas específicas, mejorando la capacidad de localizar información genómica relevante.
- **Preservación de la privacidad**: Proporciona un método seguro para compartir datos que respeta los consentimientos de los pacientes y cumple con los requisitos legales, garantizando la protección de información sensible.
- **Estandarización**: Ofrece un marco API consistente que facilita la interoperabilidad entre diferentes recursos de datos genómicos, promoviendo el intercambio fluido de información.

En 2022, GA4GH lanzó la versión 2 (v2) del protocolo ***Beacon*** con importantes mejoras:

- **Consultas ampliadas**: soporta consultas más complejas que incluyen metadatos biológicos y técnicos, como fenotipos, códigos de enfermedades, sexo o edad.
- **Respuestas detalladas**: proporciona información más completa sobre los conjuntos de datos, incluyendo anotaciones de variantes y detalles de cohortes, cuando están disponibles.
- **Acceso facilitado a datos**: Incluye mecanismos para guiar a los usuarios sobre cómo acceder a los datos, especificando puntos de contacto o condiciones de uso de los datos.

Estas mejoras han transformado el sistema *Beacon* en una herramienta más robusta para el descubrimiento y acceso a datos genómicos, alineándose con las necesidades en evolución de las comunidades de investigación y clínicas.

Aunque *Beacon* ha sido principalmente concebido como una herramienta de descubrimiento para conjuntos de datos genómicos, se han realizado numerosos avances en su adaptación a otro tipo de datos biomédicos, como datos clínicos y datos de imagen. En las siguientes secciones nos centraremos en los sistemas tipo *Beacon* de descubrimiento para estos tipos de datos.

2.2. Descubrimiento de datos genómicos

Uno de los principales desafíos en la investigación genética, tanto en humanos como en otros organismos, es la falta de acceso a los datos, no por su escasez, sino por las diversas limitaciones en su compartición dentro de la comunidad científica. Los datos genómicos, al ser identificables, requieren protección, y la ausencia de infraestructuras de seguridad y buenas prácticas para su manejo desincentiva su compartición. Esto limita el conocimiento de los avances logrados por otros investigadores.

Las soluciones tipo ***Beacon***, desarrolladas bajo las iniciativas de la GA4GH y ELIXIR, abordan esta problemática al permitir la búsqueda de variantes genómicas y su información asociada sin comprometer la privacidad de los datos. La API de Beacon facilita que instituciones de investigación u hospitales compartan sus datos genómicos de manera segura, aumentando el acceso a un mayor volumen de información y beneficiando a toda la comunidad científica, a los pacientes y la sociedad. En [2] podemos encontrar la documentación de la implementación de referencia ***Beacon v2*** (*Beacon v2 Reference Implementation*, B2RI) para el descubrimiento de datos genómicos.

2.3. Descubrimiento de datos clínicos

El descubrimiento de datos clínicos se centra en el desarrollo de métodos y herramientas que actúan sobre las HCEs, en tanto que son la fuente principal de datos sobre la que trabajar. En este sentido, tal y como se indicó en el entregable E5.1 "*Técnicas de integración de datos biomédicos*", es deseable que los datos estructurados recogidos de los sistemas de información primarios (HCE y otros sistemas asistenciales) estén estructurados en un modelo común de datos como OMOP-CDM. El modelo de datos común (CDM) de *Observational Medical Outcomes Partnership* (OMOP) es una estructura de base de datos estandarizada para organizar y analizar datos de atención médica de diferentes fuentes. Esta base de datos de uso común está diseñada para facilitar la investigación observacional a gran escala y tiene como objetivo permitir a los investigadores generar evidencia confiable para la toma de decisiones en materia de atención médica. OMOP CDM organiza los datos en tablas estandarizadas con formatos predefinidos, lo que facilita la armonización de datos de fuentes dispares y la realización de análisis en diversos conjuntos de datos de atención médica.

En este sentido, para el descubrimiento de datos clínicos, se ha desarrollado dentro de IMPaCT-Data un componente local denominado **Beacon-OMOP** (B4OMOP) [3] (Figura 2). B4OMOP es un desarrollo derivado del software **B2RI** [2] para el descubrimiento de datos genómicos y fenotípicos, que permite la integración de un *Beacon* en cualquier base de datos de OMOP. El desarrollo de esta herramienta ha sido liderado por BSC-CNS (Barcelona Supercomputing Center-Centro Nacional de Supercomputación) con la orientación y soporte de CRG (Centre de Regulació Genómica).

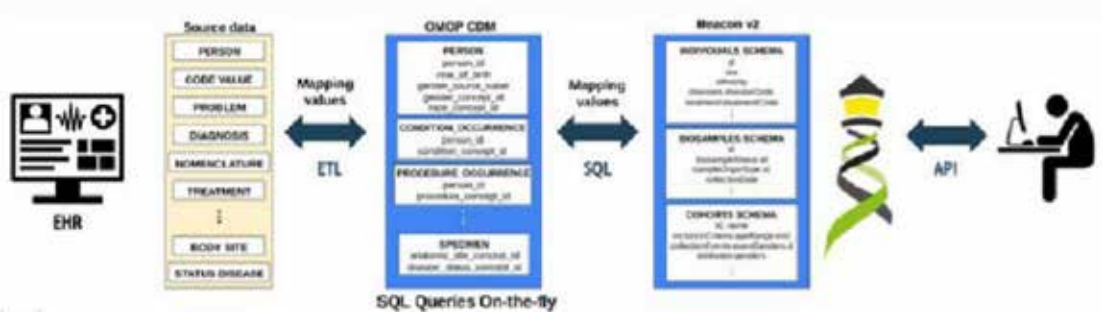

Figura 2. Diagrama general de descubrimiento de datos clínicos en *Beacon*-OMOP

El procedimiento de implementación funciona de la misma manera que el B2RI, donde iniciar el contenedor o realizar una instalación manual configura instantáneamente un sistema *Beacon*, donde la aplicación tiene una infraestructura de *backend* que realiza consultas a una base de datos relacional basada en OMOP CDM.

El mapeo entre los modelos OMOP CDM y *Beacon* garantiza la alineación de elementos como diagnósticos, tratamientos, resultados de laboratorio, exposiciones y datos de muestras biológicas. De esta manera, las instituciones que dispongan de un CMD-OMOP pueden

implementar un sistema *Beacon* sobre el conjunto de los datos contenidos en la HCE de manera que éstos sean descubribles (Figura 3).

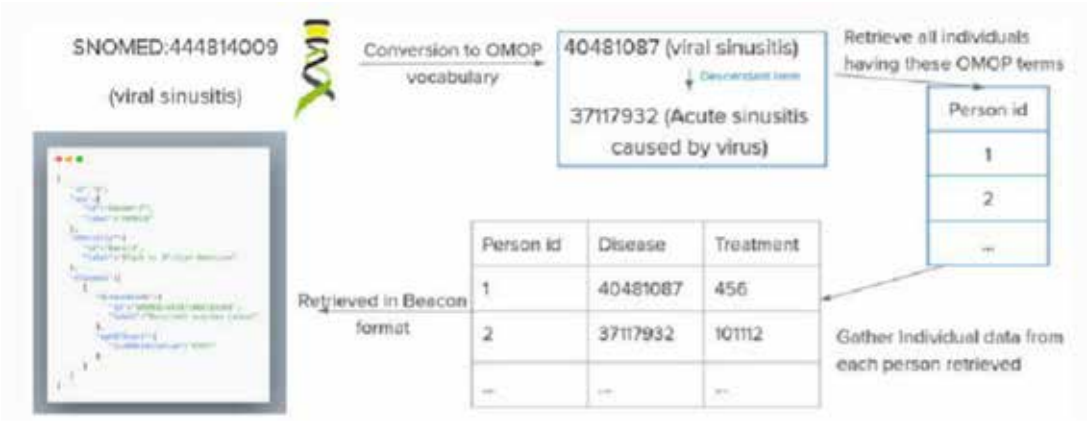

Figura 3. Descubrimiento de datos clínicos con el sistema B4OMOP mediante una consulta específica

2.4. Descubrimiento de datos de imagen

Al contrario que los datos genómicos, la imagen médica es un tipo de dato cuyo uso está bien consolidado dentro del sistema nacional de salud (SNS). Tal y como se describe en el entregable E5.1 "*Técnicas de Integración de datos biomédicos*", los hospitales y centros sanitarios utilizan los sistemas PACS *(Picture Archiving and Communication System*) para gestionar la información de las imágenes médicas (almacenamiento, gestión, visualización y distribución de las mismas y su información relacionada). En entornos de investigación (fuera del entorno asistencial), los repositorios suelen ser los principales sistemas de gestión de la imagen médica.

El descubrimiento de datos de imagen médica (búsquedas sobre metadatos radiómicos) requiere de un proceso de extracción de información relevante a partir de la imagen médica, un almacenamiento de dicha información en una base de datos y establecer un sistema de consulta/descubrimiento sobre ese conjunto de datos. Como en el caso de datos clínicos, es deseable que los datos extraídos a partir de la imagen médica estén contenidos en un modelo común de datos como OMOP-CDM, de forma que los datos estén organizados en tablas estandarizadas con formatos predefinidos. Sin embargo, el modelo de datos OMOP-CDM actual no dispone de una estructura que permita almacenar de forma estructurada la información relevante a los metadatos de las imágenes médicas y la radiómica que se puede extraer de ellas En este sentido, la OHDSI (*The Observational Health Data Sciences and Informatics*) ha presentado recientemente una propuesta que busca ampliar su alcance hacia la integración de datos de imágenes médicas con una extensión del OMOP-CDM, tal y como se indicó en el entregable E4.3-6. "*Pruebas de Concepto de extracción de información de HCE e imagen médica*". De esta manera, se enriquecerá de forma significativa la capacidad de análisis y comprensión de datos en el ámbito de la salud. Así, de forma similar al *Beacon* clínico de la sección anterior, un sistema tipo *Beacon* podría contemplar una infraestructura

de *backend* que realizara consultas a una base de datos basada en OMOP-CMD con la extensión del éste para contemplar la integración de datos de imágenes médicas.

A falta de la materialización de esta propuesta por parte de la OHDSI, durante el desarrollo del proyecto IMPaCT-Data se ha realizado el desarrollo de un primer prototipo denominado ***Beacon 4 Images*** [4] de descubrimiento de datos de imagen mediante un sistema *Beacon* que realiza consultas sobre una base de datos que contiene datos extraídos de imágenes médicas. Este prototipo nos permite, de esta forma, desarrollar el demostrador de integración de datos biomédicos completo que integre todas las fuentes de datos: genómicos, clínicos y de imagen.

3. Integrador de datos biomédicos

3.1. Descripción general y objetivos

El integrador de datos biomédicos es una herramienta creada con el objetivo de facilitar el descubrimiento y la interoperabilidad, de manera unificada, de tres tipos de datos (genómicos, clínicos y de imagen) cuyas estructuras de datos, modos y servicios de consulta se estructuran y operan de manera diferente.

Así, a partir de diferentes desarrollos de sistemas de descubrimiento tipo *Beacon* sobre dichos tipos de datos, se ha implementado una interfaz de navegador y un servicio de API integrada, que permite a los investigadores y profesionales de la salud consultar y descubrir información relevante en conjuntos de datos genómicos, clínicos y de imagen (Figura 4).

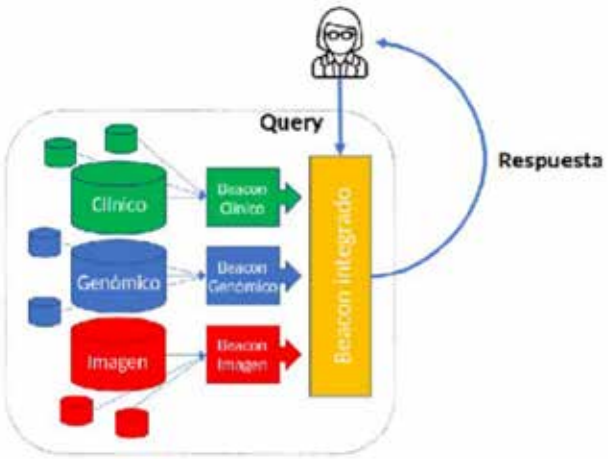

Figura 4. Diagrama general del integrador de datos biomédico basado en sistemas de descubrimiento tipo *Beacon*

El principal objetivo del demostrador es, por tanto, integrar estos datos heterogéneos bajo un único modelo de consulta, y evidenciar que, gracias a esta integración, es posible disponer de herramientas de descubrimiento más potentes, que permita una investigación más ágil y eficiente para conocer si los repositorios de datos disponen de la información necesaria para, en su caso, realizar un determinado estudio.

Como prueba de concepto del integrador y aunque puede ser extensible a otras áreas de investigación biomédica, se ha hecho uso de un conjunto de datos genómicos, clínicos y de imagen (estos últimos sintéticos) relacionados con el circuito de secuenciación genómica de SARS-CoV-2 de Andalucía [5] y descrito también en el entregable "E5.1. Técnicas de integración de datos biomédicos". Los detalles del conjunto de datos utilizado se encuentran en la sección en la sección "3.4.1. Despliegue de las bases de conocimiento".

3.2. Arquitectura del integrador

En esta sección se definen los componentes necesarios para poder ofrecer el servicio integrador de consultas de datos biomédicos, la interdependencia entre los mismos, y la localización física de cada servicio.

3.2.1. Componentes de servicio

El integrador de datos biomédicos está compuesto por los componentes de servicio indicados a continuación:

- Aplicación de navegador
- *Query* integrada
- *Beacon* genómico
- *Beacon* clínico
- *Beacon* de imagen

Adicionalmente, el integrador requiere conexión con diferentes bases de conocimiento para tener acceso a datos biomédicos a través de los componentes *Beacon*. Así, las bases de conocimiento son:

- Base de datos genómicos
- Base de datos clínicos
- Base de datos de imagen

La figura 5 representa los componentes de servicio, sus conexiones y el nivel de accesibilidad de los datos en función de su localización física.

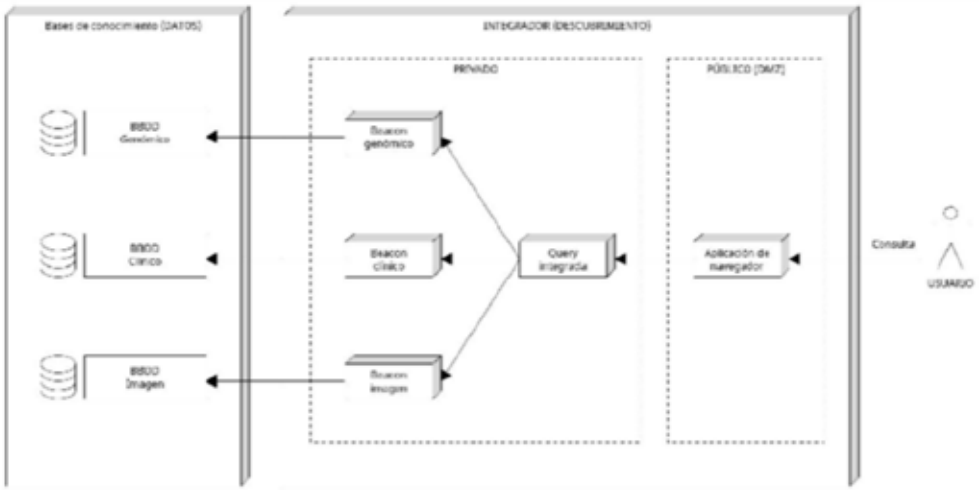

Figura 5. Arquitectura de servicios a alto nivel del integrador de datos biomédico desarrollado

Así, los componentes del integrador son:

- **Aplicación de navegador**: servicio que ofrece una interfaz de usuario accesible via web, permitiendo realizar consultas de descubrimiento de datos biomédicos, utilizando filtros genómicos, clínicos y de imagen, combinados o de forma independiente.
- ***Query* integrada**: Servicio interno de API. Recibe e interpreta la consulta introducida por el usuario en la aplicación de navegador, y realiza las consultas necesarias a los servicios de *Beacon* para recopilar y entregar los resultados de la consulta a la aplicación de navegador.
- ***Beacon* genómico**: Servicio interno de API. Atiende peticiones de la *query* integrada, realizando consultas a la base de datos genómicos.
- ***Beacon* clínico**: Servicio interno de API. Atiende peticiones de la *query* integrada, realizando consultas a la base de datos clínicos.
- ***Beacon* de imagen**: Servicio interno de API. Atiende peticiones de la *query* integrada, realizando consultas a la base de datos de imagen.
- **Base de datos genómicos**: Base de conocimiento de datos genómicos, estructurada como base de datos NoSQL (mongoDB).
- **Base de datos clínicos**: Base de conocimiento de datos clínicos, estructurada como base de datos relacional bajo estándar OMOP.
- **Base de datos de imagen**: Base de conocimiento de datos de imagen, estructurada como base de datos relacional bajo estándar OMOP.

3.2.2. Flujo de consulta de datos biomédicos

En la Figura 6 se representa el flujo de información, desde que el usuario realiza una consulta a través de la aplicación de navegador, hasta que recibe la respuesta del sistema con la información de descubrimiento.

Figura 6. Diagrama de flujo de consulta al integrador de datos biomédicos

Para realizar una consulta, el usuario accede a la aplicación de navegador e introduce o selecciona los parámetros deseados de filtro, tales como:

- Clínicos: Género del paciente, enfermedad, historial médico (condiciones de salud pasadas y actuales del paciente, tratamientos, cirugías y otros eventos médicos ocurridos a lo largo de su vida) y datos de analíticas (suero, plasma o sangre).
- Genómicos: mutaciones genéticas del genoma viral.
- De imagen: Diagnóstico y región anatómica.

Una vez aplicados los filtros de búsqueda, la aplicación de navegador traslada la consulta a la *query* integrada, la cual reconstruye y realiza consultas independientes, en base a los filtros, a cada *Beacon* correspondiente. Debido a la naturaleza de las consultas y los datos accesibles por cada servicio de *Beacon*, la *query* integrada realiza las consultas a los servicios *Beacon* de manera secuencial, haciendo uso de los resultados de las consultas a unos servicios *Beacon* para acotar la búsqueda en las siguientes consultas que sean dependientes de los resultados anteriores, hasta que todos los filtros son satisfechos.

Una vez que la *query* integrada ha obtenido el resultado final de la consulta, construye la respuesta para presentarla al usuario a través de la aplicación de navegador. Finalmente, una vez recibida la respuesta de la consulta, la aplicación de navegador muestra los resultados al usuario.

3.3. Diseño del integrador

La presente sección describe los componentes externos e internos del integrador, profundizando en detalles relevantes de diseño de cada servicio.

3.3.1. Componentes externos

En esta subsección se describe el diseño de los componentes externos al integrador, como son las bases de conocimiento (datos genómicos, clínicos y de imagen). Los componentes *Beacon* se consideran aquí como componentes externos, al haber partido su desarrollo de proyectos o especificaciones de equipos externos, aunque, a nivel de arquitectura, están considerados como componentes internos del integrador.

3.3.1.1. Bases de conocimiento

Las bases de conocimiento del integrador de datos biomédicos corresponden a:

- Bases de datos genómicos: contienen información sobre variantes genéticas.
- Bases de datos clínicos: almacenan información relacionada con datos clínicos y biométricos de pacientes, como diagnósticos, tratamientos, historiales médicos y resultados de pruebas de laboratorio.

- Bases de datos de imágenes: almacenan información relacionada con imágenes biomédicas, como radiografías.

Bases de datos genómicos

La base de datos genómicos utiliza la misma estructura que la implementación de referencia [6]. En la Figura 7 se muestra el diagrama entidad-relación de los diferentes modelos implementados en *Beacon* v2 [7]. Estos modelos describen el conjunto de conceptos incluidos en la versión de *Beacon* actual, como *Individuo* y *Biomuestra*, así como las relaciones entre ellos (ver detalles en el documento E5.4 "*Requisitos técnicos para puesta en marcha de sistemas Beacon*").

En la Tabla 1 se detallan brevemente los atributos de aquellas entidades de relevancia para el conjunto de datos genómicos de SARS-CoV-2. Para más información acerca de los atributos, se puede consultar la documentación oficial de *Beacon* [8].

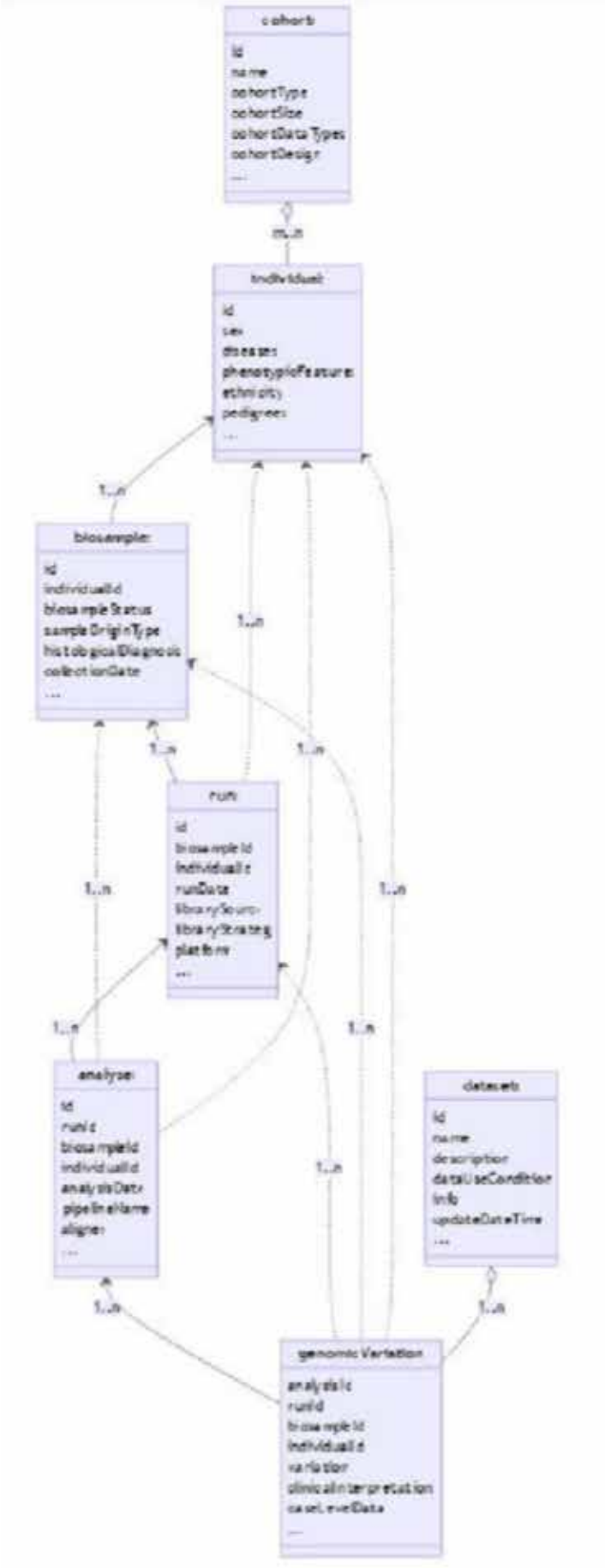

Figura 7. Diagrama entidad-relación de los modelos *Beacon v2*

Tabla 1. Colecciones y atributos de la base de datos genómicos.

Colección	Atributos	Tipo	Observaciones
analyses	N/A		
biosamples	_id	ObjectId	
	biosampleStatus: id label	Object String String	ontology BFO
	individualId	String	id OMOP
	sampleOriginType: id label	Object String String	ontology OBI
cohorts	N/A		
counts	N/A		
datasets	_id : ObjectId	ObjectId	ontology NCIT
	createDateTime	DateTime	
	dataUseConditions: duoDataUse: - id label version modifiers: - id label description id name updateDateTime version	Object Array (Object) String String String Array (Object) String String String String String DateTime String	ontology DUO ontology DUO
filtering_terms	_id	ObjectId	
	type	String	"ontology"
	id	String	ontology BFO
	label	String	
	scopes	Array (String)	["biosample"]
genomicVariations	_id	ObjectId	
	caseLevelData: - biosampleId	Array (Object) String	
	identifiers: genomicHGVSId	Object String	
	molecularAttributes: aminoacidChanges geneIds molecularEffects: - id label	 Array (String) Array (String) Array (Object) String String	ontology SO
	variantInternalId	String	
	variation: location:	Object Object	

	type: sequence_id interval: type start: type value end: type value alternateBases referenceBases variantType	String String Object String Object String Integer Object String Integer String String String	"SequenceLocation" "SequenceInterval" "Number" “Number”
individuals	_id	ObjectId	
	id	String	id OMOP
	sex: id label	Object String String	 ontology NCIT
runs	N/A		
similarities	N/A		
synonyms	N/A		
user	N/A		

La información de la Tabla 1 se ha estructurado en las siguientes columnas:

- Colecciones: nombre de las agrupaciones de datos.
- Atributos: campos que conforman cada colección.
- Tipo: Formato o naturaleza del dato, tipo de información que se puede almacenar (cadena de texto, numérico, fecha, etc)
- Observaciones: incluye detalles adicionales, si aplican, como el código de la ontología utilizada o valores por defecto.

Bases de datos clínicos

La base de datos clínicos utiliza una estructura de datos bajo el estándar OMOP CDM v5.4 [9], bajo un servicio PostgreSQL con la estructura de tablas ilustrada en la Figura 8.

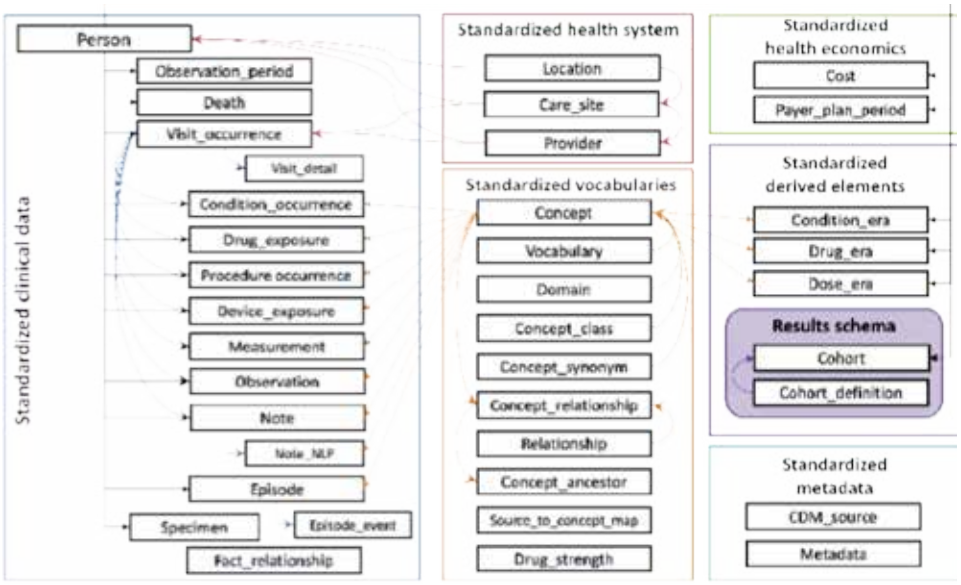

Figura 8. Esquema de datos OMOP-CDM (extraído del entregable E4.3-6)

OMOP, para los datos clínicos, estructura las tablas haciendo dos distinciones fundamentales, aquellas que incluyen datos clínicos, y aquellas que incluyen identificadores de vocabulario. Las tablas de datos clínicos requieren, para cada una de sus entradas, que toda referencia, bien sea de medida analítica, enfermedad, procedimiento, etc, esté definida en las tablas de vocabulario. Si se requiere la introducción de, por ejemplo, una nueva enfermedad, ha de referenciarse con un identificador e incluir una entrada para dicho identificador en las tablas de vocabulario.

OMOP mantiene los datos organizados, gracias a esta estructura, de manera coherente, relacionando los mismos entre sí mediante identificadores, y asociando cada dato al individuo. Asimismo, esta estructura ofrece la capacidad de generar, si es necesario para distintos fines, nuevas tablas de datos clínicos, siempre que el servicio que hace uso de los mismos lo necesite, y los datos estén asociados a identificadores presentes en las tablas de vocabulario.

Las tablas para la base de datos clínicos del proyecto se estructuran en dos esquemas, '*vocabularies*' y '*cdm*', incluyendo en el primero las tablas '*concept*' y '*concept_ancestor*', que forman parte de las tablas de vocabulario y en el segundo las restantes.

Bases de datos de imagen

La base de datos de imagen estructura las colecciones de datos bajo un esquema NoSQL, de la forma representada en la Tabla 2.

Tabla 2. Colecciones y atributos de la base de datos de imagen.

Colección	Atributos	Tipo	Observaciones
conditions	_id condition_occurrence_id person_id concept_id concept_name condition_start_date condition_start_datetime condition_end_date condition_end_datetime condition_type_concept_id condition_status_concept_id provider_id visit_occurrence_id condition_source_value condition_source_concept_id condition_status_source_value	ObjectId String String String String Date DateTime Date DateTime String String String String String String String	
datasets	_id : ObjectId	ObjectId	ontology NCIT
	createDateTime	DateTime	
	dataUseConditions: duoDataUse: - id label version description id ids: - individuals name updateDateTime version	Object Array (Object) String String String String String Object Array (String) String DateTime String	 ontology DUO id OMOP
measurements	N/A		
occurrences	_id imaging_occurrence_id person_id procedure_occurrence_id wadors_uri imaging_occurrence_date imaging_study_uid imaging_study_series modality anatomic_site_concept_id anatomic_site_concept_name	ObjectId String String String String String String String String String String	 id OMOP

La información se ha estructurado en la tabla en las siguientes columnas:

- Colecciones: nombre de las agrupaciones de datos.
- Atributos: campos que conforman cada colección.

- Tipo: Formato o naturaleza del dato que se puede almacenar (cadena, numérico, fecha, etc)
- Observaciones: incluye detalles adicionales, si aplican, como el código de la ontología utilizada o valores por defecto.

Las colecciones identificadas en la tabla son las descritas a continuación:

- *Conditions*: Incluye enfermedades diagnosticadas a través de pruebas de imagen.
- *Datasets*: Incluye la lista de individuos de los cuales se tienen registros.
- *Measurements*: Incluye mediciones tomadas en pruebas de imagen, como podrían ser las dimensiones de un daño óseo o de cualquier otro tipo.
- *Occurrences*: Incluye eventos de pruebas de imagen. Estos eventos registran un identificador del tipo de prueba realizada, fecha de realización y región anatómica, entre otros.

3.3.1.2. Componentes *Beacon*

Beacon genómico

El servicio *beacon* genómico toma como base la implementación de B2RI [2], compuesta por una base de datos NoSQL poblada de datos genómicos de humano, y una API REST que proporciona una forma estandarizada el envío y respuesta de las peticiones realizadas por el usuario. En el caso del integrador propuesto en este entregable, se ha realizado una adaptación específica a datos genómicos de **SARS-CoV-2** y en base al modelo de implementación 'C', propuesto en la documentación original [10] e ilustrado en la figura 9.

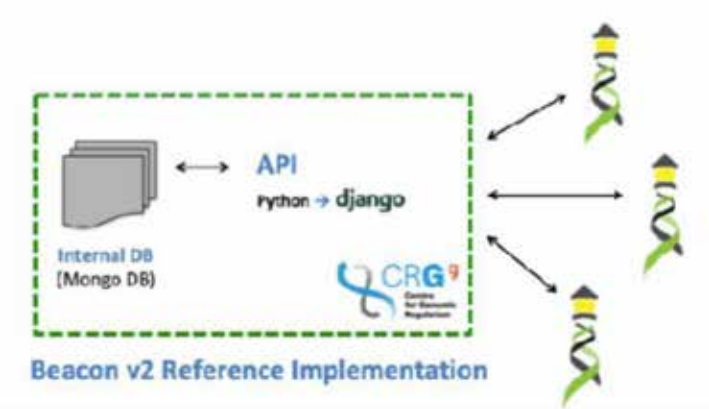

Figura 9. Opción de implementación escogida de un sistema de descubrimiento *Beacon v2* para datos genómicos

El modelo de implementación 'C' consta de una API donde los usuarios pueden realizar las peticiones de consultas y una base de datos interna NoSQL (MongoDB) donde se almacena la información. Además, incluye una herramienta para la importación de los datos donde a partir de archivos VCF de variantes genómicas y de un fichero de metadatos en formato .xls o .json, se introducen dichos metadatos en la base de datos.

En la Figura 10 podemos ver el diagrama del componente de *Beacon* genómico.

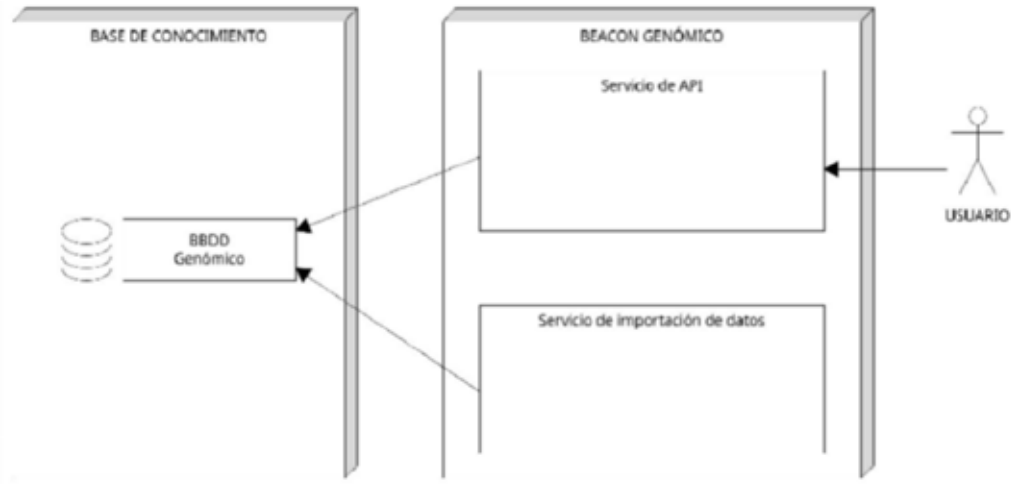

Figura 10. Diagrama del componente de *Beacon* genómico

Con objeto de adaptar la implementación de referencia a los objetivos del servicio para datos genómicos enfocados a SARS-CoV-2, han sido necesarios los siguientes pasos:

- Instalación y preparación del proyecto base [11].
- Adaptación del servicio de importación de datos (**beacon2-ri-tools**). Conviene recordar que la implementación original está orientada a datos genómicos de humano. Sin embargo, dada la naturaleza de los datos genómicos utilizados en el integrador (SARS-CoV-2), ha sido necesaria una adaptación de un sistema específico de variantes genéticas humanas a un sistema con variantes genéticas del virus SARS-CoV-2. En concreto:
    - Adición de términos de la ontología SO (*Sequence Ontology*) [12] no contemplados en la implementación original, como por ejemplo el término SO:0001826 etiquetado como *disruptive_inframe_deletion*
    - Recogida de datos del fichero de entrada VCF: adaptación de lectura de los campos de anotación en el conjunto de datos utilizado
    - Identificadores específicos de SARS-CoV-2: la secuencia de referencia del virus corresponde a MN908947.3
    - Cuestiones relacionadas con la conexión con base de datos, de forma que permita configurar el nombre de la base de datos

- Modificación del servicio de API, donde se han llevado a cabo diversas adaptaciones, como la ampliación de límites preestablecidos para las repuestas a las consultas (devolución de 1000 registros en lugar de 100) o la configuración de la base de datos.

El servicio de API habilita las llamadas identificadas en la especificación de referencia [13]. De todas las llamadas posibles, destacamos en la Tabla 3 aquellas que se han utilizado en el contexto del integrador de datos biomédico.

Tabla 3. *Endpoints* del servicio *Beacon* genómico

Endpoint	Descripción
'api/g_variants'	Variantes disponibles
'api/biosamples'	Biomuestras disponibles

Beacon clínico

El servicio *Beacon* clínico toma como base la implementación de referencia Beacon V2, y adapta la gestión de los datos en base al modelo de datos OMOP, con datos estructurados mediante un esquema relacional bajo un servicio PostgreSQL de acuerdo a la implementación descrita en [3].

El componente *Beacon* clínico incluye un servicio API para dar servicio a las consultas del usuario, y conecta con las bases de datos clínicos para consultar la información (figura 11).

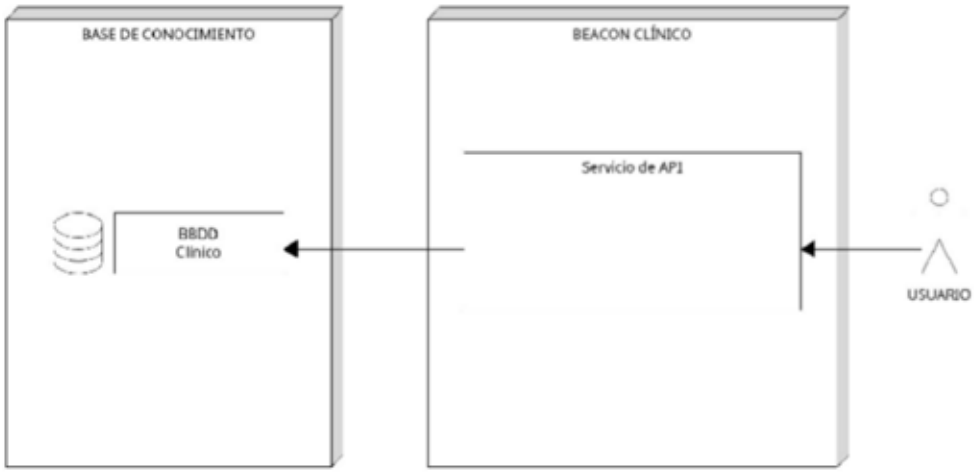

Figura 11. Diagrama de componente *Beacon clínico*

El servicio atiende a peticiones JSON con la estructura mostrada en el ejemplo de la Figura 12 para datos clínicos de individuos, cohortes y biomuestras. En dicho ejemplo, se realiza una consulta de datos clínicos para individuos, filtrando por aquellos individuos que sean mujeres (*Gender:F*) y que hayan sufrido una conmoción cerebral sin pérdida de conocimiento (*SNOMED:62106007*). Adicionalmente, se indica que se busca un máximo de 10 resultados

(*limit:10*), y que el objetivo de la consulta es obtener los registros, no solo el número de resultados encontrados (*requestedGranularity: record*).

Figura 12. Estructura JSON de ejemplo para consultas al *Beacon* clínico

De entre todas las llamadas originalmente desarrolladas en el componente *Beacon*, el servicio habilita un subconjunto de llamadas a la API, para dar servicio a consultas que busquen como resultado pacientes que cumplan con diferentes parámetros de filtrado, siendo éste el uso necesario para los objetivos del integrador (Tabla 4).

Tabla 4. *Endpoints* del servicio *Beacon clínico*

Endpoint	Descripción
'/api/filtering_terms'	Términos de filtrado disponibles
'/api/individuals'	Consulta sobre individuos con filtros

3.3.1.4. *Beacon* de imagen

El servicio *Beacon* de imagen toma como base la implementación de referencia *Beacon* V2, y adapta la gestión de los datos en base al modelo de datos OMOP, con datos estructurados mediante un esquema relacional bajo un servicio PostgreSQL [4].

El componente *Beacon* de imagen incluye un servicio API para dar servicio a las consultas del usuario y conecta con las bases de datos de imagen para consultar la información (Figura 13).

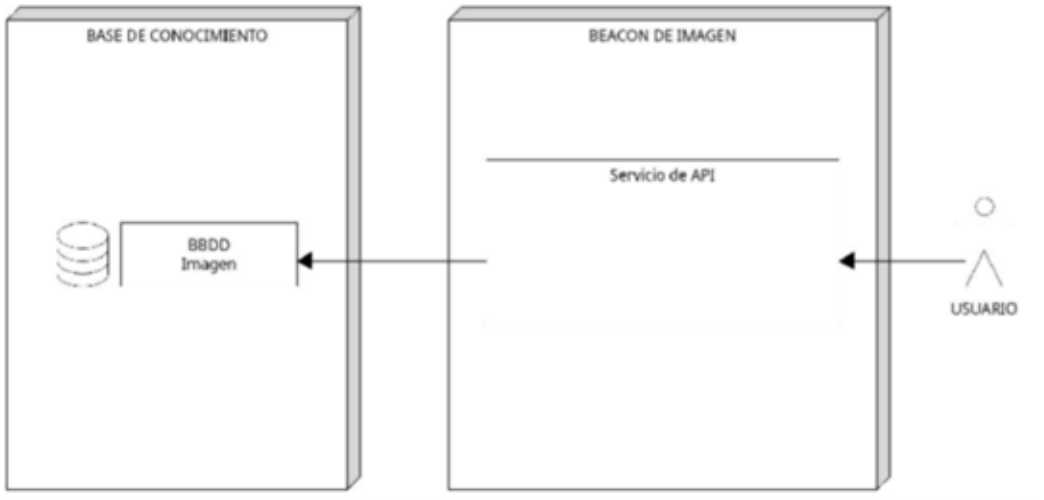

Figura 13. Arquitectura del componente *Beacon* de imagen

El servicio atiende a peticiones JSON con la estructura mostrada en el ejemplo de la Figura 14 para datos de imagen sobre individuos. En dicho ejemplo, se realiza una consulta de todos los datos de imagen disponibles para un individuo, filtrando por el individuo registrado con el id "*312437*". Adicionalmente, se indica que se busca un máximo de 100 resultados (*limit:100*), y que el objetivo de la consulta es obtener los registros, no solo el número de resultados encontrados (*requestedGranularity: record*).

```
"meta":
{
    "apiVersion": "2.0"
},
"query":
{
    "requestParameters":
    {
    },
    "filters": [
      {
        "id": "312437"
      }
    ],
    "includeResultsetResponses": "HIT",
    "pagination":
    {
      "skip": 0,
      "limit": 100
    },
    "testMode": false,
    "requestedGranularity": "record"
}
```

Figura 14. Estructura JSON de ejemplo para consultas al *Beacon* de imagen

De entre todas las llamadas originalmente desarrolladas en el componente *Beacon*, el servicio habilita un subconjunto de llamadas a la API (tabla 5), para dar servicio a consultas que busquen como resultado aquellos resultados de imagen que cumplan con diferentes filtros de pruebas de imagen y diagnósticos, siendo este el uso requerido para los objetivos del integrador.

Tabla 5. *Endpoints* del servicio *Beacon de imagen*

Endpoint	Descripción
'/api/filtering_terms'	Términos de filtrado disponibles
'/api/occurrences'	Pruebas de imagen registradas
'/api/conditions'	Diagnósticos registrados

3.3.2. Componentes internos

En esta sección se describe el diseño de los componentes internos del integrador, como son la aplicación de navegador y el servicio de *query* integrada (Figura 5). Como se ha expuesto anteriormente, no se recogen los componentes *Beacon* como componentes internos, al haber partido su desarrollo de proyectos o especificaciones de equipos externos, aunque estos componentes *Beacon*, a nivel de arquitectura, están considerados como componentes internos del integrador.

3.3.2.1. Aplicación de navegador

Para hacer uso del integrador de datos biomédicos de una forma más amigable, se ha creado una aplicación web en https://www.clinbioinfosspa.es/tools/virus-beacon/ donde se puede realizar las consultas a los distintos *Beacons* de forma visual y sencilla.

Así, la vista de navegador (Figura 15) está dividida en 2 pestañas: “SEARCH” y “ABOUT”.

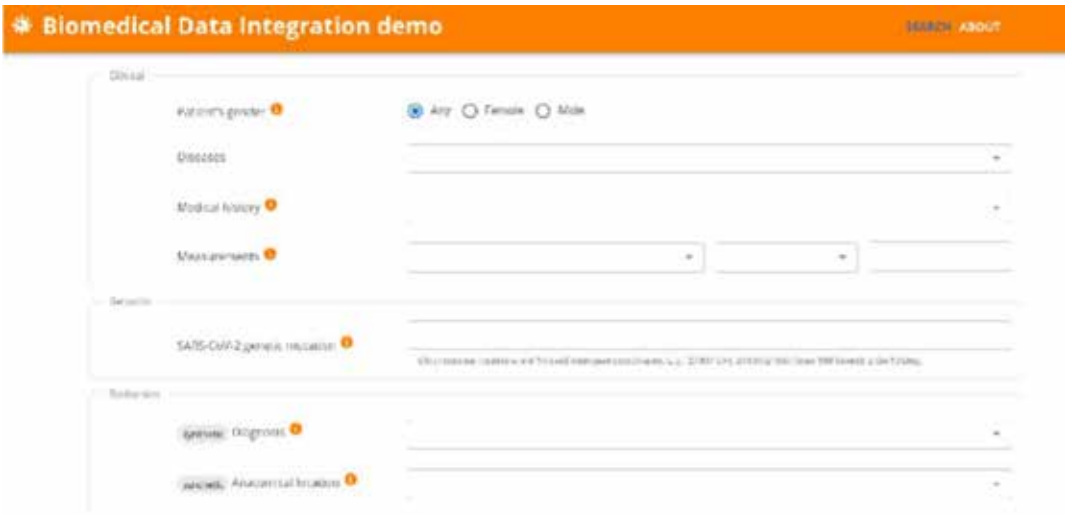

Figura 15. Vista de consulta del integrador de datos biomédico desarrollado

Pestaña *SEARCH*

Este formulario de búsqueda está organizado en tres secciones, permitiendo diferenciar sobre qué datos se van a realizar los filtros. Así, disponemos de filtros:

- Clínicos (*Clinical):*
  - *Patients's gender:* Identidad de género o sexo biológico del paciente según consta en su historia clínica.
  - *Diseases*: Enfermedades diagnosticadas.
  - *Medical history*: Condiciones de salud, tratamientos, cirugías y otros eventos médicos pasados y actuales del paciente que hayan ocurrido a lo largo de su vida.
  - *Measurements*: Datos cuantitativos recogidos de suero, plasma o sangre. Ejemplos:
    - ☐ pH de la sangre >= 7339
    - ☐ Plaquetas [#/volumen] en sangre = 165
- Genómicos (*Genomic*):
  - SARS-CoV-2 genetic mutation: mutación genética del genoma viral del paciente con respecto al genoma de referencia (MN908947.3). Ejemplos:
    - 21618 C>T
    - 44 C>T
    - 27259 A>C
    - 27382 GAT>CTC
    - p.Ser135Arg
    - p.Phe924Phe
    - p.Pro13Leu
    - p.Arg203Lys
    - Rangos específicos (limitado a 100 bases). Por ejemplo 1-100 o 200-300
- Imagen/Radiómicos (*Radiomics*):
  - *Diagnosis*: Enfermedad o condición a través de imágenes médicas.
  - *Anatomical location*: Posición o área específica dentro del cuerpo donde se toman las imágenes.

De manera similar, los resultados devueltos por el integrador también se han dividido en 3 secciones para adaptar la información de salida. Cada una de estas secciones desencadena una nueva consulta por parte del integrador, utilizando el mismo formulario de filtro. Esto es debido a que el orden de consulta a los respectivos servicios *beacon* puede variar, en función de la sección de resultados de interés para el usuario:

- GENOMIC: se muestran el listado de variantes que cumplen los filtros seleccionados, así como diversa información relacionada con la variante genómica: posición, referencia, alternativa, cambio de aminoácido, gen, efecto molecular y el número de individuos que poseen esa variante (Figura 16).

Figura 16. Ejemplo de vista de resultados genómicos

- CLINICAL: devuelve el número de individuos que cumplen los filtros seleccionados (Figura 17). A la hora de interpretar los datos, es preciso tener en cuenta que, para una búsqueda que devuelva resultados genómicos, el número de individuos devuelto en la vista de resultados clínicos no implica que cada uno de ellos incluya todas y cada una de las variantes encontradas en la vista de resultados genómicos, sino que cada individuo puede incluir una o más de estas variantes, siendo el conjunto de variantes encontradas la suma de variantes diferentes presentes en el total de individuos devuelto por la vista de resultados clínicos.

Figura 17. Ejemplo de vista de resultados clínicos

- RADIOMICS (imagen): devuelve el número de imágenes que cumplen los filtros seleccionados. Para este apartado actualmente no se dispone de datos reales, por lo que se ha generado de forma aleatoria la información y en la web se muestra la etiqueta "synthetic" (Figura 18). De manera similar a lo que ocurre con la vista de resultados clínicos, en la vista de resultados de imagen, el número de resultados es la suma de los resultados obtenidos del total de individuos devueltos por la vista de resultados clínicos.

Figura 18. Ejemplo de vista de resultados de imagen

Es importante señalar que se integran tres bases de datos distintas y algunas poseen un elevado volumen de datos, lo que implica ciertas limitaciones a la hora de mostrar información: para la sección de datos genómicos, sólo devuelve las 100 primeras variantes

o aquellas que transcurrido un determinado tiempo haya encontrado. Por lo tanto, para obtener números más precisos, es necesario indicar filtros más restrictivos. Del mismo modo, el número de individuos que poseen la variante puede ser aproximado en algunos casos, reflejando una estimación en lugar de una cifra exacta.

Pestaña *ABOUT*

En esta pestaña (Figura 19) podemos encontrar una descripción del integrador de datos biomédicos: componentes *Beacon* que utilizan, descripción del conjunto de datos, etc.

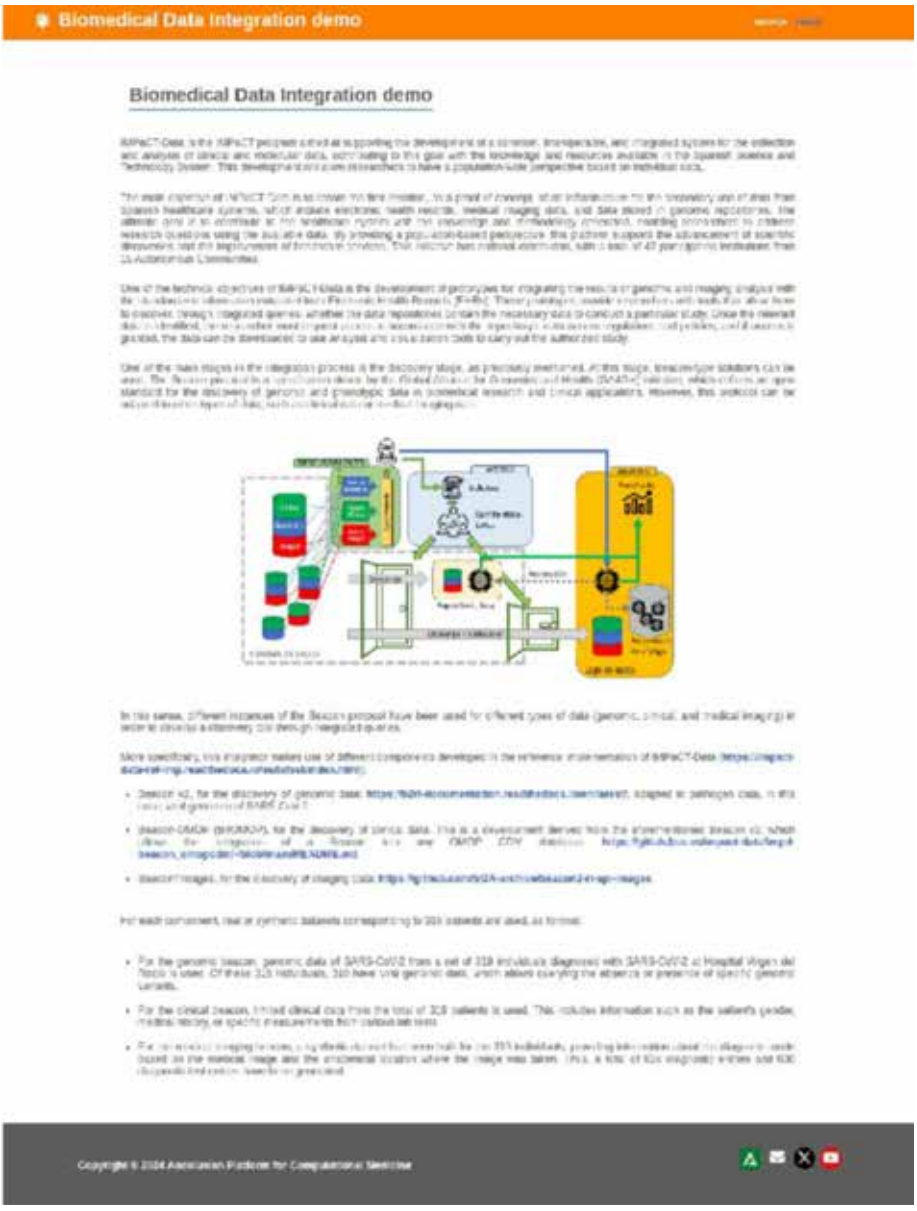

Figura 19. Pestaña *ABOUT* del integrador de datos biomédicos desarrollado

3.3.2.2. Servicio de *query* integrada

El servicio de *query* integrada interpreta la consulta con los filtros indicados por el usuario en la aplicación de navegador, y realiza las consultas correspondientes a cada servicio *Beacon*. Una vez obtenida la respuesta, este servicio construye y envía los datos que serán presentados de nuevo al usuario. Es, por tanto, el componente más importante en la cadena de la búsqueda de la información consultada por el usuario, abstrayendo la gestión de una consulta compleja y facilitando el acceso a la información.

En la Tabla 6 se detallan los servicios disponibles del integrador y el *Beacon* al que solicita información.

Tabla 6. *Endpoints* de la *query* integrada

Endpoint	Descripción	Beacon
/api/info	Información general integrador	N/A
/api/maps	Listado de los servicios disponibles	N/A
/api/images/filtering_terms	Listado de filtros de imagen	Imagen
/api/filtering_terms	Listado de filtros clínicos	Clínico
/api/biosamples	Lista de variantes que cumple todos los filtros solicitados	Genómico
/api/g_variants	Lista de variantes que cumple todos los filtros solicitados (filtros clínicos, genómicos o de imagen)	Clínico, Genómico y/o Imagen
/api/individuals	Número de individuos que cumple todos los filtros solicitados (filtros clínicos, genómicos o de imagen)	Clínico, Genómico y/o Imagen
/api/occurrences	Número de pruebas de tipo imagen que cumple todos los filtros solicitados (filtros clínicos, genómicos o de imagen)	Clínico, Genómico y/o Imagen

Así, para una consulta realizada por parte de un usuario, el proceso de búsqueda de resultados es el siguiente:

- Paso 1: Filtrado de individuos asociados al *Beacon* clínico. Si el usuario ha introducido valores en los filtros del *Beacon* clínico, se consultan los individuos existentes que cumplan dichos filtros
- Paso 2: Filtrado de individuos asociados al *Beacon* de imagen. Si el usuario ha introducido valores en los filtros del *Beacon* de imagen, se consultan los individuos existentes que cumplan dichos filtros.
- Paso 3: Generación de los individuos comunes del *Beacon* clínico y de imagen (intersección Paso 1 y Paso 2 anteriores)

- Paso 4: Generación del listado de *biomuestras* que pertenecen a los individuos obtenidos en el paso 3.
- Paso 5: Generación de variantes y de *biomuestras* asociadas al *Beacon* genómico
    - Paso 5.1: Generación de variantes: de acuerdo a los filtros introducidos por el usuario en el *Beacon* genómico, se solicita las variantes genómicas que cumplan dichos filtros
    - Paso 5.2: Generación de *biomuestras* asociadas a las variantes obtenidas en el paso 5.1
- Paso 6: Generación de resultados: dadas las *biomuestras* obtenidas en los pasos 4 y 5, se realiza la intersección de las mismas y se devuelve la información obtenida (variantes y número de individuos)

3.4. Despliegue y evaluación del integrador

Atendiendo al despliegue ideal de producción de los servicios implicados, se ha dispuesto, en equivalencia, para el desarrollo y las pruebas del integrador la arquitectura representada en la Figura 20.

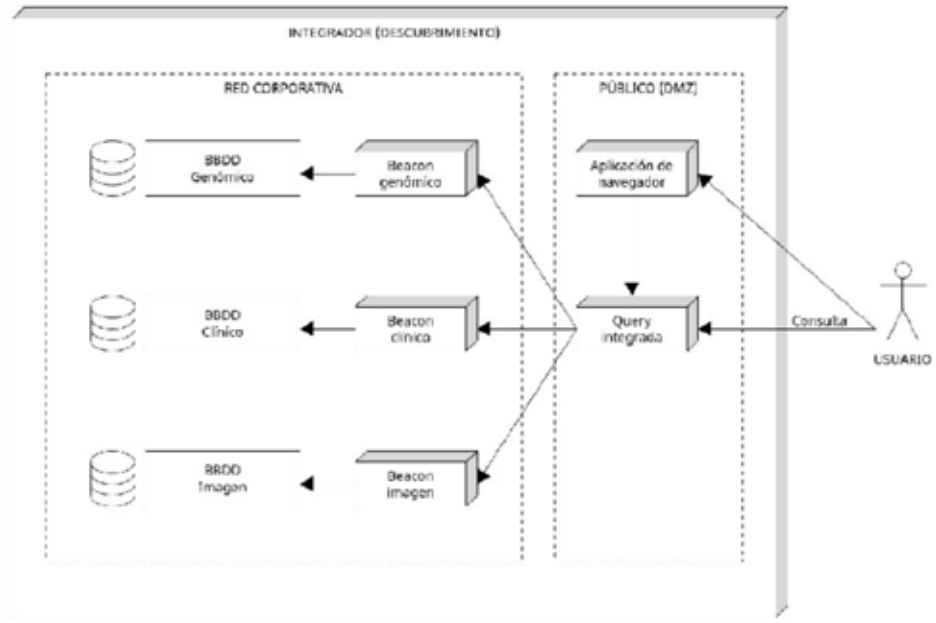

Figura 20. Arquitectura de despliegue del demostrador de integración de datos biomédico desarrollado

3.4.1. Despliegue de las bases de conocimiento

Durante el desarrollo y el despliegue de los servicios del integrador, han sido desplegadas las bases de datos genómicos, clínicos y de imagen. Más concretamente:

- Para el *Beacon* genómico, se ha poblado la base de datos con un subconjunto de 319 genomas de **SARS-CoV-2** secuenciados en el **Hospital Virgen del Rocío** en el contexto del **circuito de secuenciación genómica de Andalucía** [5].
- Para el *Beacon* clínico, se ha poblado la base de datos PostgreSQL, estructurada al estándar OMOP-CDM, con datos clínicos limitados de los 319 pacientes a los que se secuenció el virus SARS-CoV-2. Se dispone de información del género del paciente, de la historia clínica del mismo o de medidas específicas correspondientes a diferentes pruebas analíticas. Esta información clínica, ha sido proporcionada por el **Grupo de Innovación Tecnológica del Hospital Virgen del Rocío** y responsable del WP6, junto con el BSC.
- Para el *Beacon* de imagen médica, se ha construido un conjunto de datos sintéticos (tal y como se indica en el demostrador) para los 319 individuos, disponiendo información sobre el diagnóstico realizado sobre la imagen médica y el lugar anatómico donde se tomó esa imagen. Así, se han generado un total de 626 entradas de diagnósticos y 630 entradas de pruebas diagnósticas. Por otro lado, debido al estado de madurez del *Beacon* de imagen, éste se ha puesto en marcha sobre un servicio de MongoDB, simulando en estructura de datos a una base de datos relacional en base al estándar OMOP, que debería reflejar su despliegue ideal.

En cuanto al nivel de seguridad, todas las bases de conocimiento han sido desplegadas dentro de la red corporativa de la Plataforma de Medicina Computacional de la Fundación Progreso y Salud.

3.4.2. Despliegue de los servicios del integrador

Los componentes de servicio del integrador han sido desplegados, bien en red corporativa o bien en red de acceso externo (DMZ), en base al nivel de protección necesario de cada servicio, así como de la necesidad de acceso a los mismos desde el exterior de la red corporativa de la Plataforma de Medicina Computacional. De esta forma, tanto la aplicación de navegador como el servicio de *query* integrada se han dispuesto en DMZ, para ser accesibles desde fuera de la red corporativa, y los servicios de *Beacon* han sido desplegados dentro de la red corporativa.

Así, esta sería la relación de proyectos desplegados como servicios del integrador:
- Aplicación de navegador: https://www.clinbioinfosspa.es/tools/virus-beacon/
- Query integrada [14]
- Beacon genómico [15]
- Beacon clínico [3]
- Beacon de imagen [4]

3.4.3. Análisis y evaluación de servicios *Beacon*

Tras haber desplegado los servicios *Beacon*, en esta sección se destacan los puntos de interés identificados durante el análisis y las pruebas realizadas, que han servido para determinar las bondades de los sistemas en su estado actual de madurez.

3.4.3.1. *Beacon* genómico

El servicio *Beacon* genómico hace uso de una estructura de consultas *Beacon* v2, con una gestión de la búsqueda de información a través de una base de conocimiento estructurada en una base de datos NoSQL. El nivel de madurez del servicio ofrece la capacidad de realizar consultas identificando modificaciones genéticas sobre biomuestras de pacientes, devolviendo de esta forma todas aquellas biomuestras que cumplan con el criterio de filtrado.

Con objeto de establecer el nivel de detalle en las búsquedas que es posible obtener con la versión más reciente del servicio, se realizó una batería de pruebas, consistente en combinar filtros genómicos, tales como biomuestras con variante, aminoácido con individuo o intervalo de localización de las variantes con biomuestras.

Debido a la integración de resultados de este *Beacon* con el resto (clínico y de imagen), se detectó una demora significativa en la devolución de resultados, en función del rango de búsqueda introducido por el usuario en la consulta. Es por ello que, tras las pruebas, se ha tomado la decisión de limitar el rango máximo de búsqueda a 100 bases.

Tras el análisis, se concluye que el servicio de *Beacon* genómico cumple satisfactoriamente para una primera versión viable del integrador de datos en las búsquedas de variaciones genéticas sobre individuos.

3.4.3.2. *Beacon* clínico

El servicio *Beacon* clínico hace uso de una estructura de consultas *Beacon* v2, con una gestión de la búsqueda de información adaptada a bases de conocimiento bajo el estándar OMOP.
El nivel de madurez del servicio ofrece la capacidad de realizar consultas sobre pacientes, utilizando diferentes categorías de filtro como:

- Género
- Enfermedad
- Prueba analítica
- Procedimiento
- Tratamiento
- Edad de diagnóstico
- Edad de toma de medida analítica

Con objeto de establecer el nivel de detalle en las búsquedas que es posible obtener con la versión más reciente del servicio se realizó una batería de pruebas consistente en combinar filtros clínicos, tales como género con enfermedad, medida analítica con enfermedad o edad con enfermedad.

Las pruebas realizadas para filtros combinados mostraron coherencia en los datos devueltos. Aunque el integrador propuesto no requiere una alta complejidad en cuanto a combinación de filtros, se realizaron pruebas para consultas que podrían ser de interés a futuro. En este sentido, se detectaron algunas limitaciones para establecer filtros complejos, como consultas de más de un diagnóstico a pacientes con umbrales de edad diferentes. Por ejemplo, una búsqueda con un resultado acorde al esperado puede ser aquella que busca a todos aquellos pacientes que hayan sido diagnosticados de gripe con más de 75 años. Sin embargo, si a la búsqueda anterior se le añade el criterio adicional de que, además de gripe, estos pacientes hayan sido diagnosticados de SARS-CoV-2 con más de 71 años, esta búsqueda devolverá resultados inesperados, ya que añadirá aquellos pacientes que hayan sido diagnosticados de gripe con más de 71, así como aquellos que hayan sido diagnosticados de SARS-CoV-2 con más de 75. Esto se debe a que el término de filtro que gestiona la búsqueda mediante la edad de diagnóstico es un término independiente de la enfermedad.

De manera adicional, se advierte igualmente dificultad para consultar la lista de biomuestras de todos aquellos individuos a los que se les haya diagnosticado una enfermedad específica, o que cumplan con un filtro específico. Esta consulta de biomuestras es esencial para poder filtrar los resultados del *Beacon* genómico en base a los filtros clínicos, y la versión actual del servicio devuelve la lista de biomuestras para un paciente específico, por lo que se necesita obtener en primer lugar los pacientes que cumplen con los criterios clínicos, para posteriormente obtener la lista de identificadores de biomuestras de cada uno de ellos.

Tras el análisis, se concluye que el servicio de *Beacon* clínico cumple satisfactoriamente para una primera versión viable del integrador de datos en las búsquedas de pacientes, siempre que se aplique, a lo sumo, un filtro por categoría, así como un primer acercamiento a la búsqueda de biomuestras de pacientes que cumplan con diferentes filtros clínicos.

3.4.3.3. *Beacon* de imagen

El servicio *Beacon* de imagen hace uso de una estructura de consultas *Beacon v2*, con una gestión de la información a través de una base de conocimiento estructurada en una base de datos NoSQL (MongoDB), simulando una estructura relacional cercana al estándar OMOP.

El nivel de madurez del servicio ofrece la capacidad de realizar consultas sobre pacientes, utilizando diferentes categorías de filtro como:

- Diagnóstico
- Prueba analítica
- Región anatómica

Para el desarrollo del integrador se utilizaron el diagnóstico y la región anatómica como categorías de filtro, recibiendo así la lista de pruebas analíticas que cumplen con los filtros indicados tras la consulta al *Beacon*.

Con objeto de establecer el nivel de detalle en las búsquedas que es posible obtener con la versión más reciente del servicio se realizó una batería de pruebas consistentes en combinar filtros de imagen tales como diagnóstico con pruebas analíticas o diagnóstico con región anatómica

Se encontraron algunas limitaciones para establecer filtros mixtos, de manera que pueda realizarse una consulta directa indicando tanto el diagnóstico como la región anatómica. Para resolver la consulta, han de realizarse dos consultas independientes, una filtrando el diagnóstico y otra filtrando por la región anatómica para, posteriormente, escoger aquellos resultados que aparezcan en ambas consultas.

Tras el análisis, se concluye que el servicio *Beacon* de imagen cumple satisfactoriamente para una primera versión viable del integrador de datos en las búsquedas de pruebas analíticas en base a las categorías de filtro disponibles, aplicando un único filtro por categoría.

3.4.4. Análisis y evaluación del servicio integrador

El servicio integrador hace uso, internamente, de los tres servicios *Beacon* (genómico, clínico e imagen), fragmentando la consulta ingresada por el usuario en consultas independientes para cada servicio *Beacon*, y aplicándolas en el orden adecuado para obtener los resultados esperados (ver sección 3.3.2.2. *Servicio de query integrada*).

El nivel de madurez del servicio ofrece la capacidad de realizar consultas sobre datos clínicos de pacientes, biomuestras, y pruebas analíticas de imagen utilizando diferentes categorías de filtro como:

- Género del paciente
- Enfermedad
- Evento clínico (procedimiento, tratamiento, prueba analítica).
- Mutación genética
- Enfermedad
- Región anatómica de prueba analítica de imagen

Para el desarrollo, se utilizó un solo filtro por cada parámetro de filtrado disponible en cada sección del formulario de consulta de la aplicación de navegador, dado que esta limitación viene heredada de las limitaciones de los servicios *Beacon* de los que depende el integrador, que fueron señaladas en la sección 3.4.3. Para ilustrar esto, en la sección clínica del formulario, el integrador permite definir un filtro que incluya un evento clínico y una medida analítica. Sin embargo, no es posible indicar más de una medida analítica o más de un evento clínico, debido a las limitaciones advertidas durante el análisis del *Beacon* clínico, que

devuelve resultados no deseados cuando, por ejemplo, se indica un evento clínico y dos medidas analíticas, en cuyo caso devolvería como positivas aquellas coincidencias de individuos que hayan tenido eventos clínicos con cada una de las medidas analíticas por separado, en lugar de un único evento con las dos condiciones de medida indicadas al mismo tiempo.

Tras probar diferentes consultas, adaptar la estructura del formulario para cubrir un abanico útil de consultas y confirmar que los resultados devueltos corresponden con la respuesta esperada para las mismas, se confirmó que el integrador devuelve resultados de descubrimiento para la búsqueda correctos y alineados con los filtros indicados, tanto si se utilizan filtros de un único campo (clínico, genómico, de imagen), como si se utilizan combinaciones de los mismos.

Tras el análisis, se concluye que el servicio integrador cumple satisfactoriamente con un primer nivel de detalle en las búsquedas de datos biomédicos en base a las categorías de filtro disponibles, aplicando un único filtro por categoría.

3.5. Casos de uso del integrador de datos biomédicos

En esta sección, se muestran dos casos de uso o ejemplos diferentes del integrador de datos biomédico propuesto,disponible en https://www.clinbioinfosspa.es/tools/virus-beacon/ y haciendo uso de los conjuntos de datos descritos en la sección 3.4.1. *"Despliegue de las bases de conocimiento".*

3.5.1. Ejemplo 1. Descubrimiento de pacientes haciendo uso de las tres fuentes de datos

En este primer ejemplo, vamos a realizar en el conjunto de datos genómico, clínico y de imagen, una búsqueda de las variantes genómicas de virus SARS-CoV-2 secuenciado del conjunto de pacientes con una determinada enfermedad ("*Chronic obstructive pulmonary disease with acute lower respiratory infection*") y un diagnóstico de imagen ("*Acute respiratory distress*"). Así mismo, se mostrará el número de individuos existentes con las características clínicas y de imagen de acuerdo a estos filtros.

Pasos:

1. Seleccionar filtros, generando una consulta en los tres *Beacon* disponibles (figura 21):
    a. *Beacon* clínico: "Diseases": *"Chronic obstructive pulmonary disease with (acute) lower respiratory infection*"
    b. *Beacon* de imagen: "Diagnosis": "*Acute respiratory distress*"
    c. *Beacon* genómico: no se seleccionan filtros (aunque se realiza consulta)

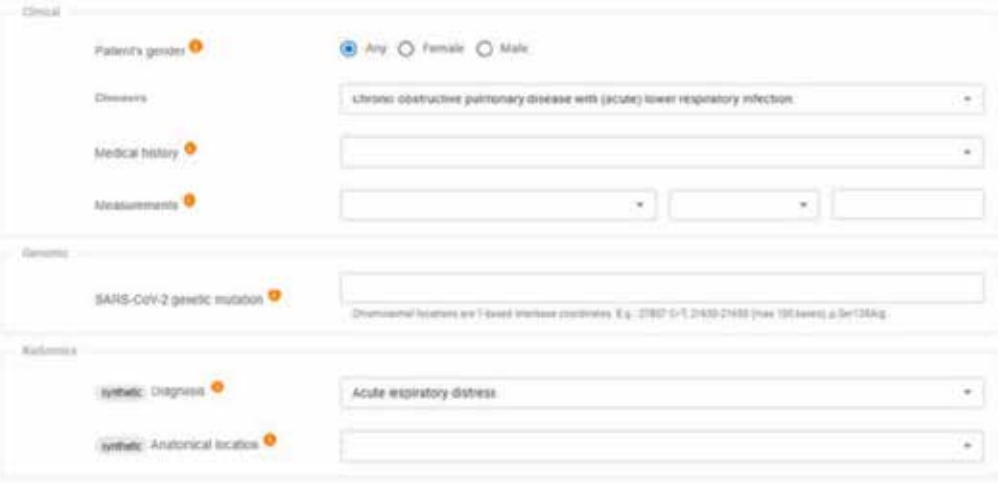

Figura 21. Filtros empleados en la búsqueda

2. Información devuelta por el integrador de acuerdo a los filtros seleccionados, divididos por categoría:
   a. “GENOMIC”: información sobre variantes encontradas (figura 22)

Figura 22. Resultados devueltos por el integrador en la categoría Genómica

   b. “CLINICAL”: número de individuos disponibles (figura 23)

Figura 23. Resultados devueltos por el integrador en la categoría clínica

   c. “RADIOMICS”: número de pruebas de imagen disponibles (figura 24)

Figura 24. Resultados devueltos por el integrador en la categoría de imagen

Así, para los filtros indicados y consultando las tres pestañas de resultados, observamos que existen 46 variantes. Las 46 variantes pertenecen a 5 individuos, de manera que en cada uno de estos individuos podrá aparecer una o más de estas variantes, siendo las 46 la suma de variantes diferentes encontradas para estos 5 individuos. Adicionalmente, hay 5 pruebas diagnósticas en total que cumplen el filtro, que pueden pertenecer a uno o más individuos.

Como se ha comentado anteriormente, en los resultados devueltos por el integrador de la categoría genómico, el número de individuos es parcial. Para obtener un número más exacto de individuos, es fundamental acotar los resultados mediante la adición de más filtros o siendo más restrictivos en éstos. Por ejemplo, acotar los resultados a una determinada región genómica del virus, por ejemplo, la comprendida en el rango 6954-6968 manteniendo los filtros clínicos y de imagen indicados anteriormente (Figura 25).

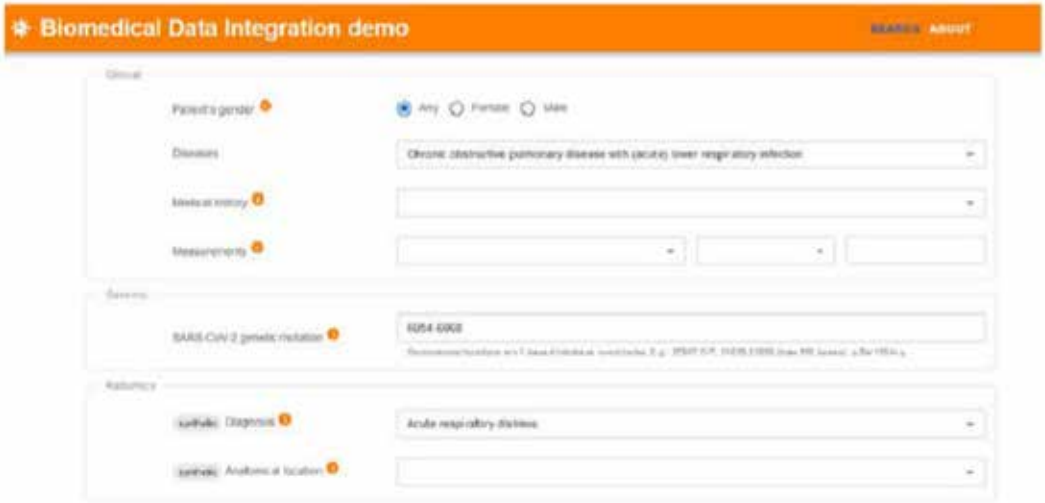

Figura 25. Filtros empleados en una nueva búsqueda, restringiendo las variantes genómicas a la región 6954-6968 del genoma de referencia de SARS-CoV-2

Con este filtro adicional, se han encontrado 2 variantes genómicas, donde podemos observar cómo de frecuente es en nuestro conjunto de datos (figura 26):

- El cambio T>C en la posición genómica 6954 está presente en 4 individuos
- El cambio C>T en la posición genómica 6968 está presente en 3 individuos

Figura 26. Resultados devueltos por el integrador en la categoría Genómica en una región más restringida

Por otro lado, respecto a los resultados devueltos por el integrador en las categorías clínica y de imagen, tenemos 4 individuos que cumplen con los filtros indicados y 4 pruebas de imagen disponibles (Figura 27).

Figura 27. Resultados devueltos por el integrador en las categorías clínica y de imagen teniendo en cuenta una región genómica más restringida

Si queremos ser más precisos, podemos filtrar por una variante en concreto (T>C en la posición 6954) y manteniendo los filtros clínicos y de imagen, tal y como vemos en el filtro de mutación de SARS-CoV-2 en la Figura 28.

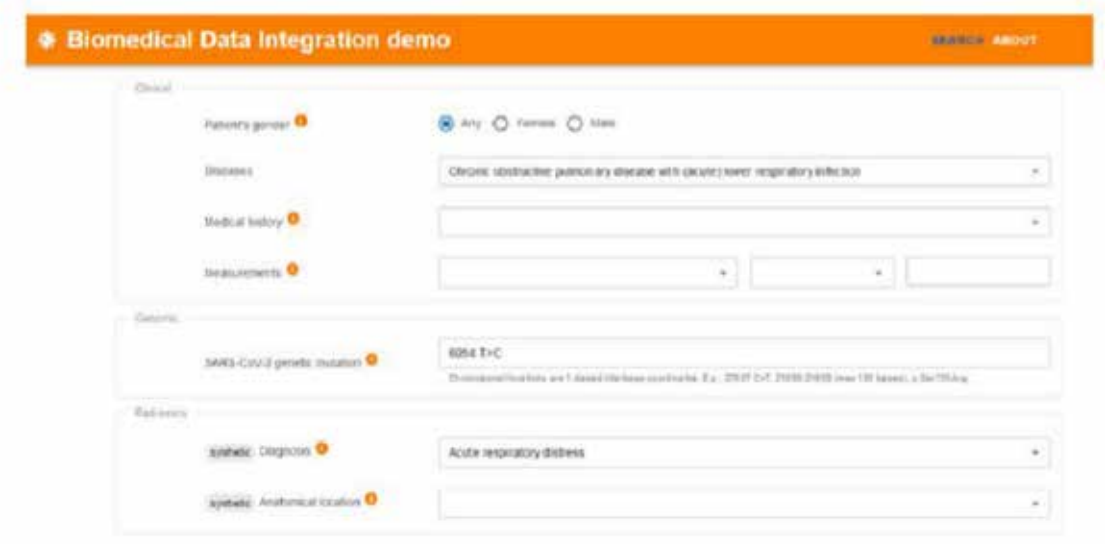

Figura 28. Filtros empleados en una nueva búsqueda, restringiendo las variantes genómicas a un cambio concreto 6954 T > C

De forma que podemos ver que existen 4 individuos en nuestro conjunto de datos con la variante genómica indicada en el filtro de mutación, así como 4 pruebas de imagen disponibles (Figura 29).

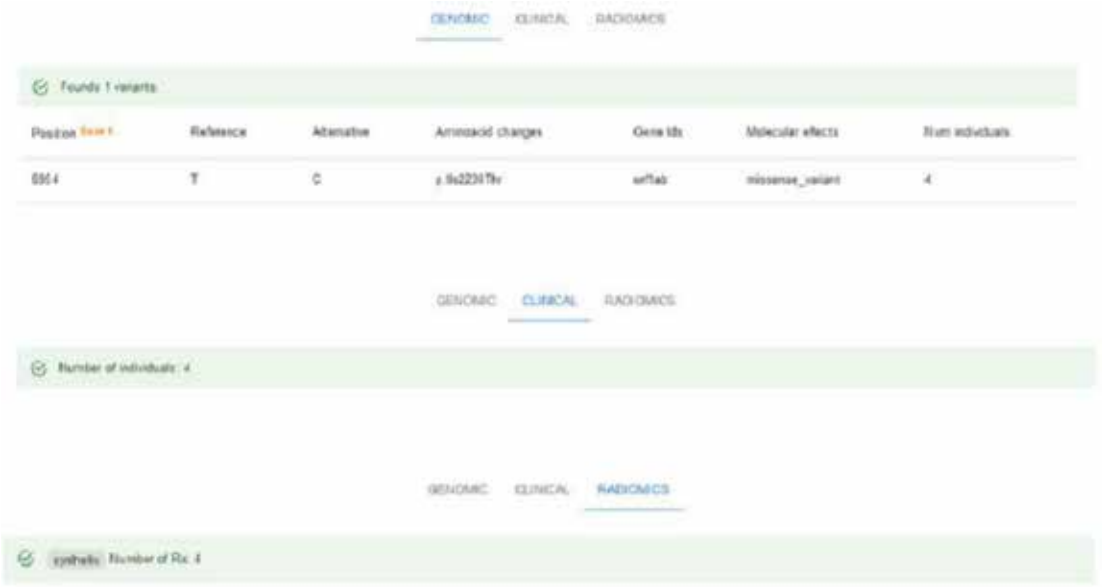

Figura 29. Resultados genómicos, clínicos y de imagen restringiendo la búsqueda genómica a una variante en concreto.

en marcha de un servicio completo y actualizado, sino que también permitiría una disminución de la complejidad del software, al tener disponibles estas relaciones. Es recomendable que la base de conocimiento de datos clínicos y de imagen se implementen siguiendo el estándar OMOP, así como disponer de un mecanismo automatizado de migración periódica de datos desde su origen.

El servicio *Beacon* genómico sufre de una limitación práctica en cuanto al tiempo de búsqueda de coincidencias en un rango amplio de bases. Aplicar una respuesta a esta limitación parece deseable a futuro, y probablemente esta respuesta deba venir de la mano de una optimización a nivel de arquitectura de servicio, que permita realizar consultas en paralelo sobre diferentes muestras, una entrega dinámica de resultados que muestre datos al usuario conforme se vayan encontrando resultados, o ambas.

En cuanto al servicio *Beacon* clínico, una nueva iteración de desarrollo que permita aplicar más de un filtro por categoría, así como un mayor nivel de profundidad a la hora de especificar categorías de filtro de segundo nivel, como la edad de diagnóstico o la edad de la toma de una medida analítica, y asociarlas a categorías de filtro de primer nivel, como las enfermedades o las medidas analíticas, aportaría un valor adicional de alto impacto.

Para el último de los servicios *Beacon*, el de imagen, sería muy interesante que continuase su desarrollo, adaptándose completamente a una base de datos bajo el estándar OMOP, de manera que pudieran incluirse las entradas de imagen de la misma forma que se incluyen las entradas clínicas, y pudiendo relacionar de esta manera de una forma más adecuada cada una de las entradas a los *person_id* registrados en el sistema. Asimismo, las capacidades de filtrado del servicio podrían mejorar notablemente tras una implementación que permita aplicar filtros de diferentes categorías en una misma consulta, como filtrar los resultados filtrando al mismo tiempo por diagnóstico y por región anatómica, incluso permitiendo más de un filtro por categoría.

Por parte del servicio integrador, una vez que hayan sido resueltas las limitaciones de los servicios *Beacon* y sea posible aplicar filtros más complejos, es potencialmente interesante añadir la capacidad de aplicar múltiples filtros por categoría, tanto a través del servicio de *query* integrada como desde la aplicación de navegador. Esto añadirá un nivel de profundidad mayor en las consultas, así como resultados más acotados y ajustados a las necesidades reales de análisis.

5. Conclusiones

El demostrador de Integración de Datos Biomédicos es una herramienta valiosa para el descubrimiento y la integración de datos clínicos, genómicos y de imagen médica. A través de una arquitectura basada en sistemas tipo *Beacon*, se ha logrado un primer nivel de interoperabilidad y funcionalidad que permite consultas integradas de manera segura.

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

3.5.2. Ejemplo 2. Descubrimiento de pacientes haciendo uso de las tres fuentes de datos: búsqueda de pacientes con neumonía asociada a ventilación mecánica y mutación específica en genoma viral

El integrador desarrollado nos permite realizar tareas de descubrimiento haciendo uso de las tres fuentes de datos con el objetivo posterior de responder a una pregunta clínica o de investigación específica. Por ejemplo, podemos identificar el número de individuos varones que tienen una variante genómica determinada de interés, por ejemplo la mutación p.Pro681His, y conocer cuántos individuos que contienen esta variante tienen registros clínicos y pruebas de imagen. La mutacion p.Pro681His está contenida en el linaje B.1.1.7 de SARS-CoV-2 (Alpha), aparecido a principios de 2021 y ha sido ampliamente estudiada debido a su impacto en la transmisión y la biología del virus (mayor transmisibilidad, mayor carga viral en víras respiratorias superiores y asociada a una mayor mortalidad en pacientes hospitalizados no vacunados)[16]. Así, el filtro en el integrador sería el siguiente (Figura 30)

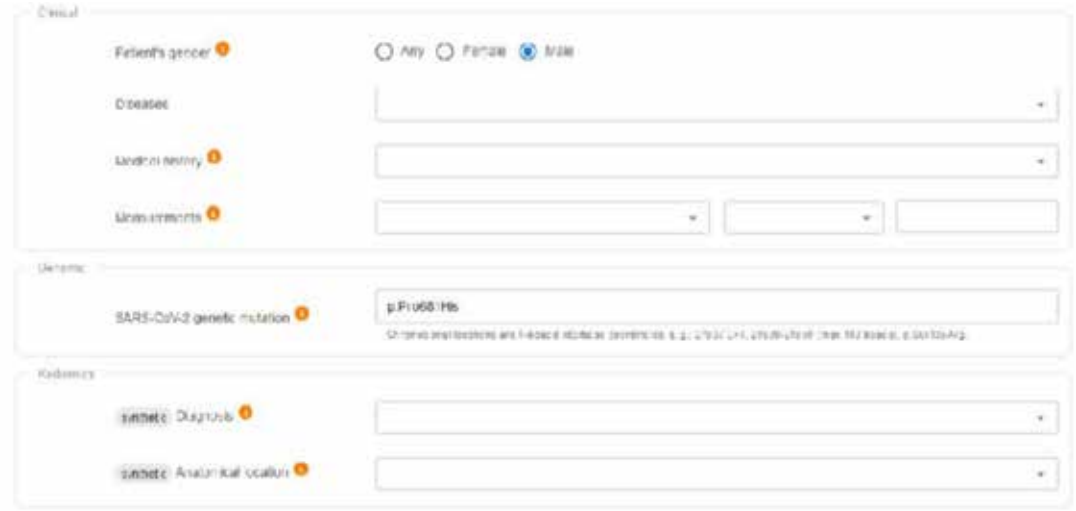

Figura 30. Búsqueda a través del integrador de datos biomédico de los individuos que contienen una variante específica del linaje B.1.1.7 de SARS-CoV-2

El integrador nos indica que existe un número aproximado de individuos (100) que contienen la variante p.Pro681His. Sin embargo, al ser una sola variante, el resultado devuelto por el *beacon* clínico, nos indica el número exacto de individuos que tienen la variante, es decir, 158. Los resultados relacionados con imagen nos indican que existen 114 pruebas de imagen disponibles (Figura 31).

Así, se destacan los siguientes puntos:

1. Integración funcional y aplicación práctica: Se ha implementado con éxito un integrador que permite descubrir datos biomédicos heterogéneos bajo un modelo común, evidenciando su utilidad en casos de uso próximos a la realidad.
2. Seguridad y privacidad: El sistema respeta los estándares de protección de datos, como el RGPD, y promueve la accesibilidad sin comprometer la privacidad de los individuos.
3. Limitaciones identificadas y perspectivas futuras. A pesar de los logros alcanzados, se han identificado aspectos de mejora clave que permitirán que el integrador evolucione hacia una solución más robusta, escalable y adaptada a las necesidades de investigación biomédica:
    1. Los *Beacons* genómico, clínico y de imagen han mostrado algunas limitaciones técnicas, como la dificultad para aplicar filtros complejos o manejar grandes volúmenes de datos. Es probable que el desarrollo actual de los componentes *Beacon* permitan salvar dichas restricciones a futuro.
    2. Es fundamental avanzar hacia una base de conocimiento más completa y actualizada, posibilitando la adopción completa del estándar OMOP-CMD en el *Beacon* de imagen.

Como valoración final, el integrador desarrollado constituye un punto de partida sólido para la integración y el descubrimiento de datos biomédicos, sentando las bases para un entorno de investigación más ágil y eficiente. Con las mejoras propuestas, se espera que el sistema evolucione hacia una solución aún más robusta y escalable, capaz de satisfacer las necesidades crecientes de la investigación biomédica y clínica.

Referencias

1 Reglamento general de protección de datos, https://rgpd.es/
2 Beacon v2 Reference Implementation (B2RI), https://b2ri-documentation.readthedocs.io/en/latest/what-is-beacon/
3 Repositorio Beacon clínico, https://gitlab.bsc.es/impact-data/impd-beacon_omopcdm
4 Repositorio Beacon de imagen, https://github.com/EGA-archive/beacon2-ri-api-images
5 Circuito de secuenciación genómica de Andalucía, https://www.clinbioinfosspa.es/COVID_circuit/
6 Beacon v2 general implementation docs, https://b2ri-documentation.readthedocs.io/en/latest/
7 Beacon v2 models, https://docs.genomebeacons.org/models/#introduction
8 Beacon v2 terms, https://docs.genomebeacons.org/schemas-md/beacon_terms/
9 OMOP CDM v5.4. https://ohdsi.github.io/CommonDataModel/cdm54.html
10 Beacon v2, implementation options. https://b2ri-documentation.readthedocs.io/en/latest/which-are-the-implementation-options/

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

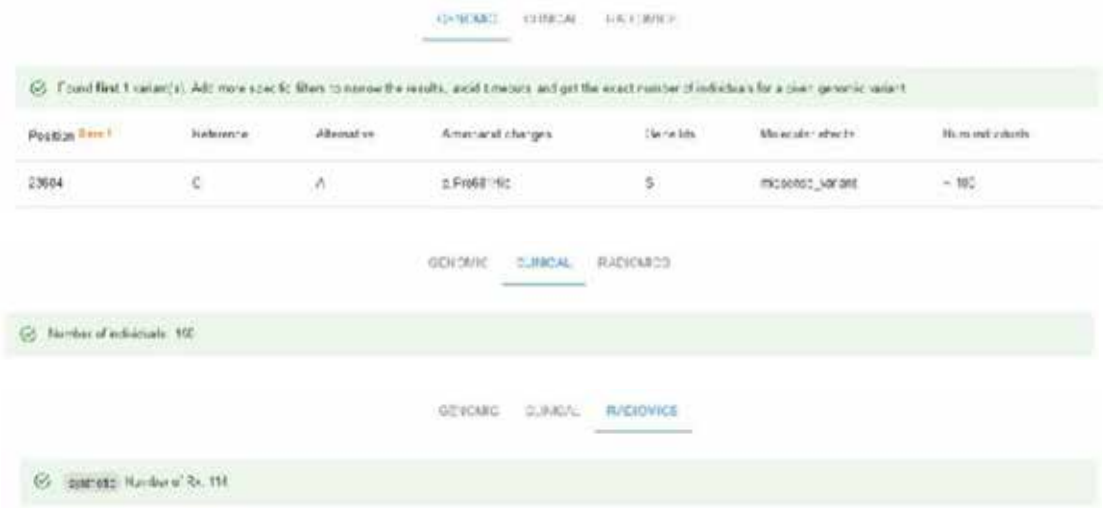

Figura 31. Resultados genómicos, clínicos y de imagen restringiendo la búsqueda genómica a una variante en concreto de interés.

Si restringimos la búsqueda a pacientes varones con neumonía asociada a ventilación mecánica (*Diseases: Ventilator associated pneumonia)* (Figura 32) el resultado se acota a 5 pacientes de los que hay disponibles 3 pruebas de imagen (Figura 33).

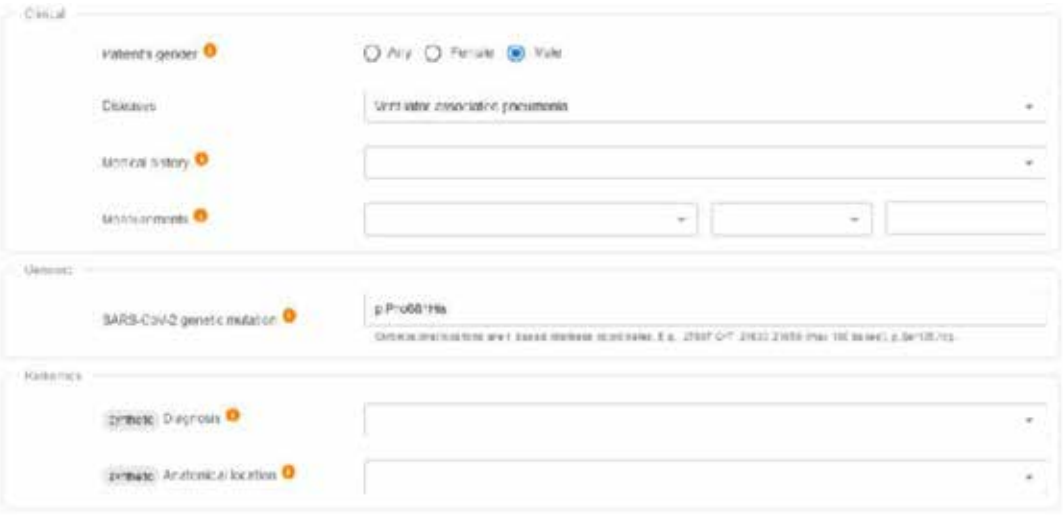

Figura 32. Búsqueda a través del integrador de datos biomédico de los individuos varones que contienen una variante específica del linaje B.1.1.7 de SARS-CoV-2 y con neumonía asociada a ventilación mecánica

11 Beacon v2, Installation and beacon2-ri-tools, https://b2ri-documentation.readthedocs.io/en/latest/download-and-installation/
12 Sequence Ontology Terms, http://www.sequenceontology.org/
13 Beacon v2 API endpoints, https://b2ri-documentation.readthedocs.io/en/latest/api/#api-endpoints]
14 Repositorio query integrada, https://github.com/babelomics/biomedical-integrator-impactData
15 Repositorio Beacon genómico, https://github.com/babelomics/beacon2-app-virus
16 Loucera C, Perez-Florido J, Casimiro-Soriguer CS, et al. Assessing the Impact of SARS-CoV-2 Lineages and Mutations on Patient Survival. Viruses. 2022;14(9):1893. Published 2022 Aug 27. doi:10.3390/v14091893

Acrónimos, abreviaturas y glosario de términos

GA4GH	Global Alliance for Genomics and Health
OMOP CDM	Observational Medical Outcomes Partnership Common Data Model
API	Application Programming Interface
Endpoint	Dirección de una llamada a la API que representa una función
REST (API)	Representational State Transfer
SQL	Structured Query Language
NoSQL	Not only SQL
BBDD	Base de datos
JSON	Javascript Object Notation
Individuo	Ser vivo u organismo. En este contexto, paciente.
Cohorte	Conjunto de entradas o agrupación de datos que comparten una misma propiedad, criterio. La cohorte define estas propiedades de agrupación.
Biomuestra	Cualquier material biológico recolectado de un organismo que se utiliza para análisis y estudio en investigaciones biomédicas o genómicas
DMZ	Demilitarized zone; zona de despliegue de servicios accesibles externamente, fuera de la red corporativa.
Ontología	Representación formal del conocimiento en la que los conceptos se describen por su significado y las relaciones que guardan entre ellos
SO	Sequence Ontology, proyecto colaborativo de ontología para la definición de características de secuencias utilizadas en la anotación de secuencias biológicas.
PACS	Picture Archiving and Communication System. Sistema de almacenamiento, recuperación y transmisión de imágenes médicas utilizado en hospitales y entornos clínicos
OHDSI	Observational Health Data Sciences and Informatics. Iniciativa global

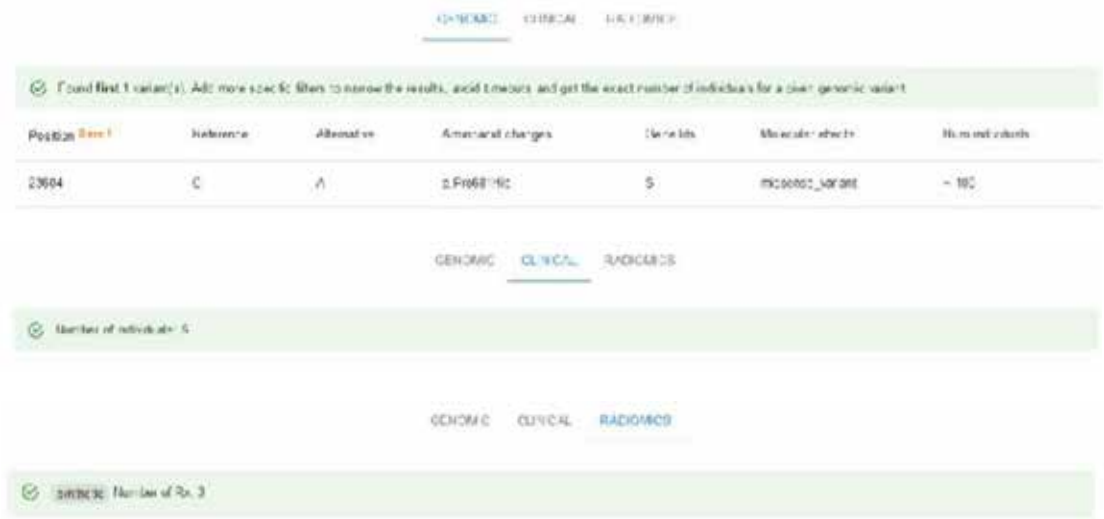

Figura 33. Resultados genómicos, clínicos y de imagen restringiendo la búsqueda genómica a una variante en concreto de interés en pacientes varones con neumonía asociada a ventilación mecánica.

Este último caso es un ejemplo claro de la utilidad de un integrador de datos como el desarrollado, donde se trata de identificar la existencia de casos donde mutaciones concretas del virus tienen una relevancia clínica importante en el curso de la COVID-19 y que permitan, en una segunda fase, solicitar el acceso a los datos completo para poder realizar un estudio de interés.

4. Discusión y trabajo futuro

El integrador de datos biomédico desarrollado, ha demostrado ser una herramienta útil en el contexto de descubrimiento científico, permitiendo a los investigadores identificar, mediante consultas integradas, si los repositorios de datos disponen de los datos necesarios para realizar un estudio determinado. Una vez identificados los datos de interés, el investigador deberá solicitar el acceso conforme a la regulación y las normativas de acceso a datos del repositorio. Si el acceso es concedido, los datos podrán descargarse para su análisis y visualización, facilitando así la realización del estudio autorizado.

Una vez que el integrador ha demostrado su potencial como herramienta de descubrimiento de datos biomédicos con los conjuntos de datos genómicos de SARS-CoV-2 y los conjuntos de datos clínicos y de imagen (en este último caso sintéticos) de los pacientes correspondientes, se hace patente la conveniencia de disponer, a futuro, de una base de conocimiento actualizada, con todos los datos disponibles del Sistema Nacional de Salud, y sin la necesidad de hacer uso de datos sintéticos, que disponga de datos relacionados entre los pacientes, sus datos genómicos, clínicos y de imagen. Esto no solo serviría para la puesta

	colaborativa que busca transformar los datos de salud observacionales en conocimiento útil y procesable mediante metodologías de análisis avanzadas y estándares comunes.
VCF	Variant Call Format. Formato de archivo ampliamente utilizado en bioinformática para almacenar información sobre variantes genómicas

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

52