Revista Latinoamericana de Difusión Científica
Volumen 8 – Número 14
Depósito Legal ZU2019000058 - ISSN 2711-0494

Revista Latinoamericana de Difusión Científica

Volumen 8 - Número 14
Enero – Junio 2026
Maracaibo – Venezuela

Los retos del almacenamiento de datos multidimensionales en Salud: Sobre la construcción de la infraestructura de TI para el biobanco de la Alcaldía Iztapalapa

DOI: https://doi.org/10.5281/zenodo.18444764

Jesús Hernández Guillén*
Ricardo Marcelín Jiménez**
Marco Antonio Núñez Gaona***
Francisco Javier Hernández Olvera****

RESUMEN

En la actualidad, la tecnología ha incrementado considerablemente la generación global de información digital. En el ámbito de la salud, esta información corresponde a expedientes médicos, los cuales son de carácter confidencial y deben resguardarse de manera segura, permitiendo su acceso únicamente al personal autorizado. Su volumen, complejidad y sensibilidad plantean desafíos técnicos, éticos y legales. De esta manera, el objetivo de este artículo es ofrecer una visión general de los principales retos que surgen en la organización y administración de grandes conjuntos de datos en el ámbito biomédico. Para ello, la metodología se fundamentó en el desarrollo de una solución de almacenamiento de información digital biomédica llamada Biobanco de la Alcaldía Iztapalapa. Esta plataforma se concibe como un modelo integral, confiable y sostenible para la gestión responsable de información poblacional destinada a la investigación y al desarrollo científico. Como resultado, se realizó la implementación de un modelo de gestión que prioriza la protección ética de la información, asegura su calidad científica y establece protocolos para un uso responsable. Este marco garantiza la privacidad de los pacientes y permite emplear los datos de forma segura para identificar riesgos sanitarios en la población.

PALABRAS CLAVE: Biobanco, Almacenamiento distribuido, DICOM, Privacidad de datos, Gobernanza de datos.

*Profesor. Universidad Autónoma Metropolitana, México. ORCID: https://orcid.org/0000-0002-4665-3010. E-mail: cbi2243802371@xanum.uam.mx

**Profesor. Universidad Autónoma Metropolitana, México. ORCID: https://orcid.org/0000-0002-5355-5830. E-mail: rmarcelin@izt.uam.mx

***Investigador. Instituto Nacional de Rehabilitación, México. ORCID: https://orcid.org/0000-0002-8450-6003. E-mail: mnunez@inr.gob.mx

****Profesor. Universidad Autónoma Metropolitana, México. ORCID: https://orcid.org/0000-0002-5890-8973. E-mail: jhernandez@alephdatasolutions.com

Recibido: 23/09/2025 Aceptado: 19/11/2025

The Challenges of Multidimensional Data Storage in Health: On the Construction of the IT Infrastructure for the Biobank of the Iztapalapa City Hall

ABSTRACT

Currently, technology has significantly increased the global generation of digital information. In the healthcare sector, this information corresponds to medical records, which are confidential in nature and must be securely safeguarded, allowing access only to authorized personnel. Their volume, complexity, and sensitivity pose technical, ethical, and legal challenges. Thus, the objective of this article is to provide an overview of the main challenges arising in the organization and management of large datasets in the biomedical field. To achieve this, the methodology was based on the development of a biomedical digital information storage solution called the Iztapalapa Mayor's Office Biobank. This platform is conceived as a comprehensive, reliable, and sustainable model for the responsible management of population-level information intended for research and scientific development. As a result, a management model was implemented that prioritizes the ethical protection of information, ensures its scientific quality, and establishes protocols for responsible use. This framework guarantees patient privacy and enables the secure use of data to identify health risks within the population.

KEYWORDS: Biobank, Distributed storage, DICOM, Data privacity, Data governance.

Introducción

En la última década se ha producido un incremento en la publicación de estudios científicos dedicados a comprender la salud de la población. Un ejemplo de este avance se encuentra en las investigaciones realizadas en el Reino Unido, donde se han abordado temas como las enfermedades crónicas, el envejecimiento, los riesgos asociados a la genética, así como la influencia del estilo de vida y del entorno en la salud de las personas. Este volumen de información ha sido posible gracias a una iniciativa de investigación impulsada por la colaboración entre el sector público y el privado, conocida como el Biobanco del Reino Unido (UK Biobank). Este proyecto reúne una extensa colección de datos biomédicos y funciona como una plataforma de investigación que permite a la comunidad científica analizar de manera integral cómo los genes, el ambiente y los hábitos cotidianos interactúan y condicionan la salud humana. De acuerdo con (Alkhatib y Gaede, 2024), este ofrece un enorme potencial para la investigación sobre la predicción temprana de enfermedades y la alineación de los fenotipos derivados de imágenes (IDP) con observaciones cognitivas, conductuales, genéticas y médicas.

Entre los hallazgos más importantes generados por este proyecto se pueden

destacar los estudios sobre los marcadores genéticos de la demencia o el COVID realizados por (Bahcall y O.G, 2018) y (Watts y Geoff, 2012). Por ejemplo, para hacerle frente a esta última, los investigadores pusieron en marcha diversos proyectos de investigación enfocados en el desarrollo de pruebas diagnósticas, vacunas y en la identificación de tratamientos eficaces. Para conseguir datos significativos, estos estudios requirieron el análisis de muestras biológicas provenientes de posibles pacientes de COVID-19, así como la correlación de estos resultados con datos clínicos y epidemiológicos.

En este contexto, los Biobancos representaron un puente entre la práctica médica y la investigación científica, siendo responsables de la recolección, procesamiento, conservación y distribución de las muestras, además de gestionar la información sanitaria vinculada. De acuerdo con (Juozapaité, 2023), entre sus principales aportes estan la disponibilidad de colecciones de muestras y datos, su experiencia en el manejo de especímenes, el cumplimiento de principios bioéticos y su capacidad para coordinar proyectos de investigación.

Segun (Lian y Geng, 2025), a la fecha, ha reunido más de 15 millones de muestras biológicas y una base de datos sobre 500 mil participantes voluntarios en edad adulta. El tamaño de estos registros supera los 30 petabytes. Por ejemplo (Bahcall y O.G, 2018), para ponerlo en perspectiva, esto implica que, si se usaran discos de 1 terabyte, se necesitarían más de 30 mil discos para almacenar este volumen de información, así como un sistema de indexación que permita correlacionar información de diversas fuentes. Además del UK Biobank, existen otras infraestructuras similares en otras partes del mundo.

Existen acepciones complementarias acerca del concepto de Biobanco. Por un lado, se entiende como la infraestructura necesaria para la preservación de muestras biológicas (como sangre, tejidos, ADN, etc.). De acuerdo con (Scapicchio y Gabelloni, 2021), también se entiende como las tecnologías de la información que hacen posible la gestión de la información derivada del análisis de las muestras biológicas o de otros estudios (tales como imagenología o señales biomédicas), así como la información ambiental y socioeconómica de la población que participa. La información que se resguarda en un Biobanco suele organizarse en distintas categorías fundamentales, entre las que se incluyen los datos clínicos de las personas donantes, la información demográfica, las características de las muestras biológicas y los datos administrativos asociados a su gestión.

Tabla 1: Biobancos en diferentes regiones del mundo (Bukreeva y Malsagova, 2024).

Nombre del Biobanco	Descripción	Tipo de biomaterial	Territorio cubierto
UK Post-cancer simple collection	El primer Biobanco virtual dedicado a la investigación del cáncer de próstata.	Tejido, muestras de sangre y DNA.	UK
DXConnect Virtual Biobank	Una plataforma que brinda acceso a muestras clínicas para investigadores. Permite consultar las colecciones registradas por cualquier institución en todo el mundo.	Diferentes tipos de muestras clínicas.	Mundial
APPRISE Virtual Bank	Biobanco virtual con información sobre muestras recolectadas de pacientes con una patología infecciosa.	Plasma, suero, células madre hematopoyéticas periféricas.	Australia
BioIVT	Una plataforma que ofrece acceso a un amplio repositorio de especímenes biológicos y a una base de datos clínica con fines de investigación.	Diferentes tipos de muestras biológicas.	Mundial
BioVitrina	Un repositorio web que ofrece acceso a información sobre especímenes biológicos bien anotados de diversas patologías.	Tejidos, orina, suero sanguíneo, plasma sanguíneo.	Rusia
BBMRI-ERIC	La infraestructura europea de Biobancos que ofrece acceso a bio-recursos y servicios para apoyar la investigación biomédica.	Diferentes tipos de muestras biológicas.	Unión Europea
iSpecimen Marketplace	Funciona como una plataforma centralizada para obtener bioespecímenes humanos de alta calidad a partir de una red global de proveedores.	Diferentes tipos de muestras biológicas.	Mundial

Uno de los principales retos que deben considerarse durante la etapa de desarrollo de estos sistemas es que gran parte de los datos clínicos se obtiene de manera manual a partir de historias médicas y cuestionarios. Este proceso requiere un esfuerzo considerable, tanto en términos técnicos como organizativos, especialmente al momento de diseñar una arquitectura de bases de datos adecuada para cada nuevo proyecto de investigación. Además, la ausencia de estándares comunes para la estructuración de

esta información dificulta la selección, integración y reutilización de los datos, lo que reduce las posibilidades de colaboración entre instituciones y limita el aprovechamiento pleno de los recursos disponibles en los Biobancos.

Por lo tanto, resulta fundamental destacar la importancia de estudiar a la población como conjunto, más que a los individuos de manera aislada. Este enfoque está directamente vinculado con los objetivos de este tipo de proyectos, cuyo propósito es desarrollar investigaciones sobre grupos de personas que comparten determinadas características, con el fin de identificar biomarcadores y reconocer patrones genéticos, clínicos, de exposición a riesgos y de vulnerabilidad, entre otros aspectos relevantes. Además, un elemento esencial en el diseño y funcionamiento de un Biobanco es la dimensión temporal. Se trata de proyectos concebidos para sostener investigaciones a largo plazo, que pueden extenderse durante varias décadas. Este horizonte permite, por ejemplo, estudiar los procesos de envejecimiento de un grupo poblacional e incluso analizar fenómenos que abarcan distintas generaciones de participantes.

Según (Scapicchio y Gabelloni, 2021) los Biobancos son fundamentales en la evolución de la investigación médica, ya que proporcionan los recursos necesarios para estudios de largo plazo y permiten la reutilización de muestras y datos para múltiples proyectos científicos. La creación y el funcionamiento de un Biobanco de datos implican diversos retos técnicos, estos son resultado de la complejidad de almacenar, gestionar y proteger tanto las muestras biológicas como la información que las acompaña. En este contexto, la calidad de los datos se convierte en un elemento central para el correcto aprovechamiento de un Biobanco.

La precisión, la integridad y la confiabilidad de la información son aspectos esenciales para preservar el valor científico de las muestras y de los datos asociados. Cuando la información es precisa, consistente y segura es posible sustentar resultados de investigación robustos y confiables. Sin embargo, pueden surgir dificultades cuando se presentan errores en la captura de datos, anotaciones inconsistentes o discrepancias entre distintas fuentes de información, lo que genera imprecisiones y posibles sesgos en los análisis. Para reducir estos riesgos, resulta indispensable implementar mecanismos de control y validación, estandarizar los procesos de registro de datos y llevar a cabo auditorías periódicas que aseguren la exactitud, coherencia y calidad de la información a lo largo del tiempo.

Un desafío relacionado con la gestión de los datos ocurre cuando están incompletos o faltantes, ya sea por fallas en la recolección, falta de seguimiento de los

participantes u otro tipo de errores. Para no comprometer la calidad, es necesario contar con protocolos que registren estas lagunas y reduzcan su impacto en los estudios. Además, un Biobanco debe integrar información de orígenes diversos, lo que hace necesario unificarlos para que sean consistentes y puedan trabajarse en conjunto. De acuerdo con (Alkhatib y Gaede, 2024) la solución pasa por adoptar formatos estandarizados, vocabularios comunes y técnicas de normalización que permitan integrar diferentes conjuntos de datos sin perder calidad en el proceso.

Además de la perspectiva técnica sobre las funciones que deben cumplir los Biobancos y del manejo de muestras biológicas y datos, de acuerdo con (Seebode y Ort, 2015) existe también un reconocimiento sobre la importancia de este tipo de herramienta de investigación y del cuidado que debe ponerse en la relación con los donantes. En este trabajo se busca presentar una visión panorámica sobre los retos técnicos asociados con la gestión de la información registrada en un Biobanco. Es importante subrayar que la construcción de una colección de expedientes como los que se resguardan en un Biobanco implica también el cumplimiento de regulaciones legales y procedimientos éticos que, aunque son fundamentales, escapan de los objetivos de este artículo.

1. Requerimientos de la solución de almacenamiento: Tipos de datos y su gestión

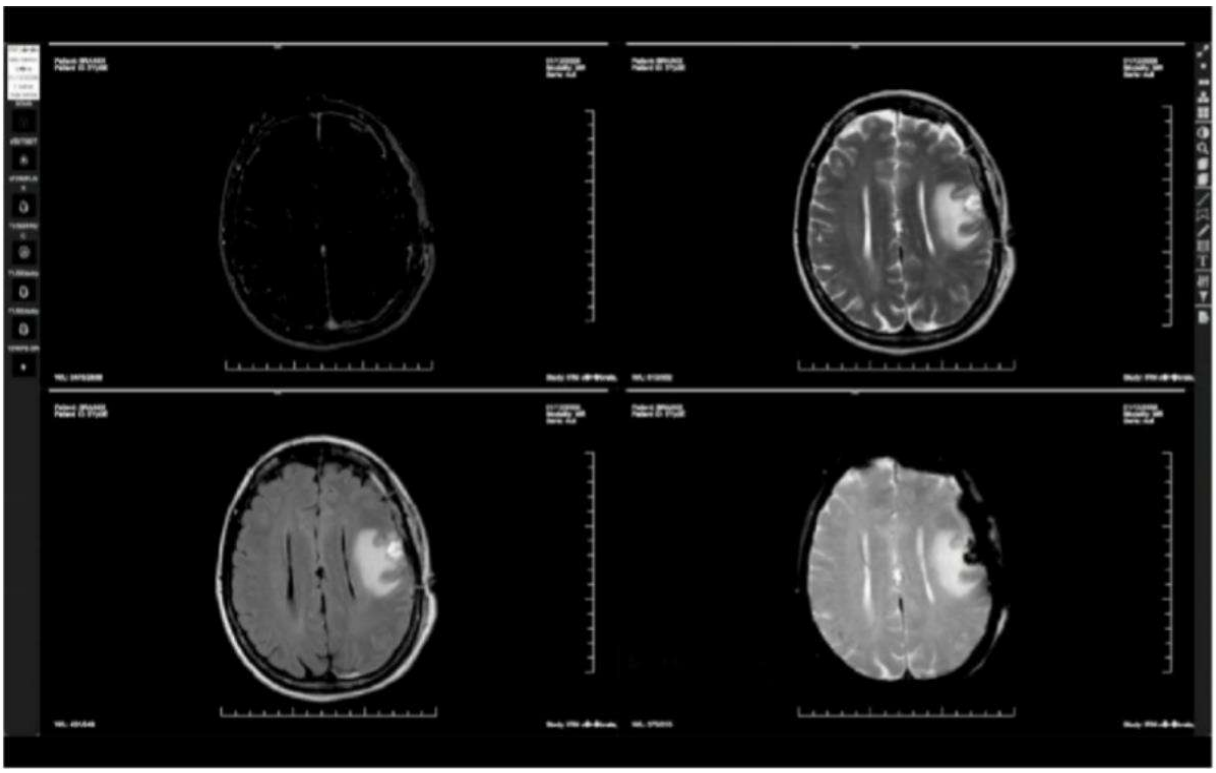
Según (Scapicchio y Gabelloni, 2021) y (Păscuțoiu y Ungureanu, 2025) todos los documentos o archivos digitales de imagenología obedecen a una norma internacional llamada el estándar Digital Imaging and Communications in Medicine (DICOM). Gracias al estándar DICOM, las imágenes y la información médica pueden ser compartidas, analizadas e interpretadas de manera universal. Un archivo DICOM no solo contiene la imagen, sino también un encabezado con metadatos estructurados en etiquetas específicas. Estas etiquetas catalogan cada dato (como el tipo de estudio o el informe), lo que asegura que, independientemente del equipo o laboratorio de origen, cualquier especialista con un sistema que cumpla la norma pueda acceder, intercambiar y analizar los estudios sin problemas de compatibilidad.

Esto implica que independientemente del laboratorio, o aparato desde el que se realice un estudio, existe la garantía de que éste pueda intercambiarse, consultarse e interpretarse por cualquier especialista y desde cualquier dispositivo que se apegue a la norma. Tal como lo señalan (Mileva y Caviglione, 2021) y (Bomewar y Baraskar, 2015), el estándar DICOM actúa como un protocolo que define la estructura para empaquetar

datos médicos. Este protocolo ofrece servicios de red para transferir e imprimir imágenes, gestionar flujos de trabajo, y garantizar la coherencia y calidad en la visualización, además de establecer los requisitos de compatibilidad para los equipos. La norma define los Objetos de Información (IOD), que son estructuras que detallan, mediante atributos, todas las características de una imagen.

Estos atributos cubren desde el tipo de imagen y los datos del paciente hasta los procedimientos, informes y la información técnica del equipo (fabricante, modelo, etc.). Es importante señalar que DICOM no especifica conexiones físicas, sino que utiliza el Protocolo de Capa Superior (ULP) del modelo ISO/OSI, lo que lo hace compatible con diferentes tipos de redes. Finalmente, las imágenes DICOM pueden visualizarse en diversas estaciones de trabajo, ya sea en escala de grises o color. Su correcta visualización está asegurada gracias a un encabezado que especifica la profundidad de bits y el tipo de compresión utilizado.

Figura 1: Imagen en formato DICOM hospedada en un sistema de almacenamiento observada a través de un visor Dicom (B. Priya y N. Deepan, 2023).



Por citar algunos ejemplos sobre la información almacenada, en promedio, una imagen digital de rayos X (RX) puede “pesar” entre 5 y 25 MB. Un estudio completo de tomografía computarizada (CT) puede producir entre 100 y 5000 imágenes, cada una de

entre 0.5 y 2 MB, lo que implica que el conjunto puede alcanzar incluso algunos cientos de MB. En la siguiente tabla, se muestra el tamaño de diferentes formatos de imágenes médicas.

Tabla 2: Tamaño en KB para distintos formatos de imágenes médicas (Dandu, 2008).

Formato de imagen	Matriz de imagen (en píxeles)	Rango dinámico (bits por pixel)	Tamaño del archivo (por imagen)
Imágenes de Resonancia Magnética (MRI)	256x256	16	131 KB
Tomografía Computarizada (CT Scan)	512x512	16	524 KB
Ultrasonido	512x512	8	262 KB
Color Doopler	768x576	8	442 KB
Radiografía digital	Hasta 3000x3000	Hasta 16	Hasta 18 MB
Mamografía digital	Hasta 3328x4096	14	27 MB
Radiografía computarizada	3520x4280	12	30 MB

Por otra parte, de acuerdo con una secuenciación de genoma completo (GWS) de un ser humano puede alcanzar un tamaño de 3 GB de datos. Según (Kamel y Belal, 2023), para la secuenciación del exoma (WES), el tamaño del archivo FASTQ de un genoma a profundidad media de 30x tiene un valor de entre 75 y 100Gb y el de un exoma a profundidad media de 100x varía entre de 8 y 14 GB. En contraste, con base en (J.H. y Oliver, 2011) un archivo de microarreglos de ADN pesa cerca de 30 MB. A diferencia de los estudios de imagenología, en genética no hay un estándar único, probablemente por la rapidez con la que han surgido las diferentes técnicas de secuenciación. Según (Hernaez, Weissman y Ochoa, 2019) se puede hablar de algunas prácticas comunes y estándares de facto, como el formato FASTQ para las lecturas crudas, el formato SAM/BAM para alineamiento, los archivos FASTA para secuencias de referencia y las referencias para anotaciones, o como la base de datos GENCODE tal como señalan (Mudge y Carbonell-Sala, 2025) y (Larson y Oberg, 2023). La gestión de grandes volúmenes de datos, se puede traducir en retos relacionados con su almacenamiento, organización y seguridad.

Algunas características adicionales del sistema son:

i. Confidencialidad y trazabilidad: este requisito implica la implementación de mecanismos técnicos que aseguren la protección de la información, garantizando que su acceso se realice de forma controlada y segura. Además, es indispensable anticipar y mitigar posibles contingencias, así como mantener un registro íntegro e inalterable de los usuarios que hayan accedido a los datos.

ii. Integración de diferentes fuentes (heterogéneas): se refiere a la necesidad de considerar que los estudios y registros pueden originarse en distintos laboratorios, corresponder a diferentes periodos temporales y seguir diversas normas de codificación. Sin embargo, en todos los casos, dicha información debe ser estandarizada para garantizar su correcta vinculación con los individuos que conforman la población de estudio.

iii. Integridad de la información: se refiere a la posibilidad de que los datos que codifican la información puedan verse alterados de manera no deseada, ya sea como consecuencia de una manipulación inadecuada o de fallos en los medios de almacenamiento. La preservación de la integridad implica que dichas alteraciones deben ser detectables y, en la medida de lo posible, reversibles.

iv. Escalabilidad del sistema: implica el crecimiento de la colección de expedientes bajo resguardo, lo que trae consigo el crecimiento del número de dispositivos de almacenamiento hasta alcanzar un número masivo que puede llegar a miles de discos u otra clase de dispositivos.

v. El horizonte de tiempo: tienen que ver con las previsiones que deben tomarse para garantizar las operaciones de un sistema sobre un período de tiempo que puede abarcar décadas ¿Qué puede ocurrir con un sistema de información en un lapso de tiempo como este? apagones y fallas, cambios tecnológicos, cambios en la administración, sólo por citar algunas posibilidades.

El sistema del Biobaco está diseñado para obtener un máximo aprovechamiento de las capacidades de almacenamiento disponibles. Usualmente, para garantizar la disponibilidad de información y hacerle frente a la caída en fallo de algun(os) nodo(s) de almacenamiento, muchas soluciones replican n-veces un documento, alojando cada copia en un servidor distinto. Cuando alguno cae en fallo, el documento puede recuperarse de algunas de las unidades que aún estén en línea y estar disponible para el usuario. Sin embargo, a pesar de parecer una técnica confiable, en realidad no es la mejor manera de aprovechar estas capacidades.

Si se considera el ejemplo de una GWS, la cual tiene un peso de 3 GB, si se replica esta información (por ejemplo, cuatro veces) el sistema almacenaría en realidad el mismo documento cinco veces. Esto implicaría el uso de 15 GB totales, lo que significa que para garantizar la replicación se emplearía un 400% de capacidad adicional. Si dos nodos de almacenamiento caen en fallo, el documento aún podría recuperarse de alguno de los tres restantes.

Por otra parte, la solución de almacenamiento propuesta contempla el uso de códigos de borrado, los cuales constituyen técnicas matemáticas destinadas a la protección de la información mediante esquemas de fragmentación y redundancia. Este enfoque permite la recuperación de los datos incluso ante fallos o pérdidas parciales. Dicho mecanismo se implementa mediante la incorporación de fragmentos de paridad o redundantes, lo que reduce el espacio de almacenamiento requerido en comparación con la replicación total de los datos, sin comprometer la tolerancia a fallos y facilitando la reconstrucción de la información en caso de daño o pérdida.

Mediante la técnica empleada en el Biobanco, usando estos códigos el mismo GWS se puede dividir en cinco partes, cada una de las cuales equivaldría en tamaño aproximadamente al 33% (≈ 1013 MB) del documento original. De esta manera, por los cinco fragmentos se almacenaría aproximadamente cerca de 5.068 GB lo cual representa un 66% adicional de espacio. Al igual que con la replicación, mediante códigos de borrado cada fragmento se almacenaría en una unidad diferente, permitiendo al usuario recuperar el documento original usando cualesquiera tres de los cinco fragmentos, por lo que también podría tolerarse la misma cantidad de fallos pero empleando cerca de una tercera parte del espacio consumido mediante replicacion.

2. Sobre la confidencialidad y la trazabilidad

El diseño, implementación y operación de la infraestructura de tecnologías de la información de un Biobanco deben centrarse en la protección de la información de los participantes. La salvaguarda de los datos personales implica un conjunto de procedimientos vinculados al tratamiento de la información de los voluntarios, que incluye, entre otros, datos de identificación, información sociodemográfica, resultados de pruebas de laboratorio, estudios de imagen y los datos derivados de investigaciones clínicas. Este último conjunto se denomina bioinformación del participante. Entre los procedimientos más relevantes asociados a su gestión se encuentran el consentimiento informado y la anonimización de los datos de identificación. El consentimiento informado

establece que cada participante debe recibir información clara sobre los objetivos de su participación y el uso que se dará a los estudios que empleen sus datos sociodemográficos y su bioinformación, para posteriormente otorgar su autorización expresa.

En el segundo caso, debe recibir la garantía de que ninguna persona pueda vincular los resultados de estos estudios con su identidad. De acuerdo con (Müller y Heimo, 2015), en algunas situaciones, esta última reserva puede plantearse en otros términos, por ejemplo, estableciendo que sólo algunas personas puedan vincular los resultados de estos estudios con su identidad. Considérese el siguiente escenario: un investigador que realiza estudios sobre una población, cuyos datos se encuentran registrados en un Biobanco, encuentra que las personas que reúnen una serie de características genéticas, son más proclives a sufrir un problema de salud que puede evitarse con un tratamiento oportuno.

El investigador trabaja exclusivamente con expedientes que no permiten establecer una vinculación directa con los participantes de origen, salvo mediante una clave o código que puede ser resguardado por una institución o persona distinta, responsable de conservar los expedientes originales y de asociar dicha clave con una identidad específica. En este contexto, el investigador puede proporcionar los resultados de sus hallazgos junto con las claves correspondientes a las instancias autorizadas. Corresponde a estos custodios o responsables de las claves decidir si resulta pertinente establecer contacto con alguna persona. Se considera que existe una anonimización de datos cuando el procedimiento que separa la identidad del expediente es irreversible; en caso contrario, se habla de pseudoanonimización de los datos.

Aun cuando la información contenida en un Biobanco se puede entender como un bien público resguardado para fines de investigación, también es claro que deben existir protocolos para solicitar el acceso a sus contenidos. Esto significa que un investigador podrá tener un acceso controlado siempre que justifique las razones por las que desea utilizar los expedientes que solicite y se comprometa con su manejo de acuerdo con las políticas del Biobanco. Lo anterior implica también que el acceso al Biobanco debe estar basado en diferentes soluciones técnicas de seguridad entre las cuales merece especial atención el cifrado de la información y la trazabilidad del acceso. De acuerdo con (Nadaf y Sumangala, 2023), el cifrado se refiere a que los documentos digitales no pueden almacenarse como archivos en claro (o texto plano), sino utilizando técnicas de criptografía de llave simétrica.

El mismo manejo de las llaves de cifrado debe considerar una gestión colegiada basada en protocolos de compartición de secretos (secret sharing) (Sankaranarayanan, 2024). Por su parte, la trazabilidad se refiere a la necesidad de que exista una bitácora donde se registre cada acceso y cada operación realizada, garantizando que este registro sea inviolable. Como lo plantean (Fadhil y Jawaher, 2024), para dar cumplimiento a este requerimiento parece pertinente pensar en una solución basada en la tecnología de cadena de bloques (blockchain) que ha servido para garantizar la permanencia o inviolabilidad de los contratos o acuerdos económicos basados en transacciones digitales.

3. La integración de fuentes de información

Por otra parte, es posible que la información sociodemográfica y la bioinformación de un participante provenga de diferentes fuentes. Esto puede ocurrir porque sus estudios clínicos, pruebas de laboratorio e imagenología, fueron practicados en diferentes momentos o en diferentes lugares. Según (Alkhatib y Gaede, 2024), además de garantizar que todos los datos correspondan a la misma persona, el otro reto en este punto tiene que ver con la integración o armonización de documentos que obedecen a diferentes normas o estándares. El enfoque para abordar este problema comienza por considerar el uso de bases de datos no relacionales que ofrecen esquemas flexibles para el manejo de grandes volúmenes de datos no estructurados.

De acuerdo con (Martínez y Zaldívar-Revé, 2022), en la búsqueda de soluciones, también se debe considerar lo que se ha hecho en el sector salud para la gestión de datos de pacientes en hospitales, por ejemplo, como los sistemas de gestión de información de laboratorio (LIMS) y los estándares de interoperabilidad. Como se aborda en (Gazzarata y Almeida, 2024), la referencia Health Level Seven (HL7), por ejemplo, es un conjunto de estándares que permiten el intercambio de información clínica en formato electrónico, denominado Modelo de Información de Referencia (RIM) en el dominio de la salud. Se especificó utilizando el lenguaje unificado de modelado (UML) y un metalenguaje extensible de marcado de etiquetas (XML), con el objetivo de definir el ciclo de vida de la mensajería dentro del flujo de trabajo en el ámbito hospitalario.

De acuerdo con (Ayaz y Pasha, 2021), como una evolución de los mecanismos de interoperabilidad propuestos por el mismo consorcio que ha desarrollado HL7, se cuenta con un marco de referencia impulsado por el Fast Healthcare Interoperability Resources (FHIR) que busca facilitar el intercambio de datos sobre salud entre diferentes

sistemas, mediante una representación en lenguajes como JSON, XML a través del enfoque arquitectónico REST, facilitando la implementación de aplicaciones médicas interoperables.

4. Sobre la integridad y escalabilidad del almacenamiento

De la mano con la codificación de la información se debe considerar que los documentos se registran en dispositivos de almacenamiento, como discos duros o cintas magnéticas, las cuales, de acuerdo con (Coughlin y Hoyt, 2024), con el tiempo pueden experimentar algún tipo de degradación que lleve a la pérdida de datos. En este contexto es importante planificar el almacenamiento pensando en la manera como se podrán detectar estas contingencias y hacerles frente. La clave está en guardar un exceso de información, o redundancia con la que se pueda, primero, detectar la corrupción de un documento y luego, recuperar los datos que se han dañado. La pregunta inevitable en este contexto será ¿cuánta información redundante debe considerarse?

En el caso de la replicación, este exceso de almacenamiento tiene consecuencias no sólo en el costo de la solución, sino que también tiene un impacto ambiental mayor, el cual se puede explicar por el exceso de dispositivos electrónicos que deben mantenerse en operación. ¿Será posible tener las mismas garantías de disponibilidad de la información con una cantidad menor de información redundante? la respuesta es: si, pero debe construirse. Según (Namvari-Tazehkand y Pashazadeh, 2021), otro factor que debe pesar sobre las decisiones de diseño de un Biobanco, es la escalabilidad. Esto es porque una colección puede crecer hasta alcanzar un volumen en el orden de miles de terabytes o petabytes.

Uno de los principales desafíos que enfrentan los centros de datos a gran escala es la gestión eficiente de datos distribuidos en numerosos nodos de almacenamiento. La administración de grandes volúmenes de información continúa siendo un reto crítico, ya que puede limitar tanto el rendimiento de las operaciones de entrada y salida, como la escalabilidad del sistema. Los sistemas de almacenamiento a gran escala suelen operar en entornos heterogéneos, donde los nodos presentan diferentes capacidades y características. En este contexto, un elemento clave es la estrategia de distribución de datos, que define cómo se asignan a los distintos dispositivos o nodos. De acuerdo con (Zhou y Chen, 2023), esta estrategia debe garantizar decisiones eficientes, minimizar la cantidad de movimiento de datos y mantener un balance de carga adecuado entre los nodos.

Como se menciona en (Zhou y Chen, 2023), algunos sistemas logran mantener un buen equilibrio de datos, de manera que solo se requiere una migración mínima al agregar o eliminar nodos. No obstante, aunque estas estrategias permiten una distribución uniforme de los datos, en muchos casos no contemplan las características específicas de cada dispositivo dentro de entornos heterogéneos, lo que puede repercutir negativamente en la eficiencia del almacenamiento y en el acceso a la información. Investigaciones recientes han intentado superar estas limitaciones mediante enfoques orientados a la gestión de franjas de archivos, la optimización del rendimiento a través del uso de dispositivos de alta velocidad como los SSD, o la aplicación de funciones hash convencionales que presuponen capacidades y propiedades homogéneas entre los nodos, sin considerar sus diferencias particulares.

A esta consideración se debe agregar que los documentos se deberán preservar sobre una escala de tiempo que abarca, incluso, varias décadas. ¿Qué contingencias pueden presentarse en estos plazos en un sistema del tamaño y la complejidad de un Biobanco de datos?, podría tratarse de eventos naturales, apagones, fallas en los dispositivos de almacenamiento, cambios tecnológicos, sólo por citar algunos. Además, dado el aumento de la demanda de soluciones de almacenamiento económico para sistemas de archivo, elegir una tecnología de archivo adecuada puede traducirse en considerables ahorros a largo plazo y en un desempeño superior. De acuerdo con (Byron y Long, 2018), se trata entonces de diseñar y construir un sistema de almacenamiento distribuido, definido por software y tolerante a fallas.

“Distribuido” significa que debe coordinar un número considerable de dispositivos de almacenamiento alojados dentro de equipos de cómputo que se comunican mediante una red de datos. “Definido por software”, significa que por medio de software se crean interfaces de acceso a los dispositivos para ocultar los detalles de la tecnología con que están construidos. Visto de otra forma, los dispositivos físicos se “envuelven” en un software para crear dispositivos virtuales normalizados. Con ello se garantiza la independencia entre las operaciones de lectura y escritura de datos y los dispositivos que hay detrás de esta. Tal como se señala en (Opara-Martins y Sahandi, 2014), en el mediano y largo plazo, esto significa que los dispositivos pueden ser intercambiables y con ello se evita caer en dependencias con tecnológicas o con proveedores, que pueden ser muy costosas.

Además, se abre la puerta a las innovaciones tecnológicas. Por último, la “tolerancia a fallas” significa que el diseño del sistema debe considerar la redundancia

de componentes. Es decir, la disponibilidad de partes de reserva o refacciones, que puedan entrar en operación cuando se detecte la falla de un componente activo. Como se explica en (Opara-Martins y Sahandi, 2016) y (Różańska y Kritikos, 2019), en aplicaciones de uso crítico, este reemplazo debería de realizarse, idealmente, de manera automática para evitar que el sistema salga de operación. A propósito del almacenamiento de grandes volúmenes de datos y su preservación en el largo plazo, existe un debate en curso sobre la manera en que los administradores de esta colección deben abordar su gestión tecnológica.

Por un lado, están quienes abogan por la contratación de un servicio privado, ofrecido por un proveedor de infraestructura “en la nube”. Por otro lado, están quienes se inclinan por la construcción de las capacidades propias a cargo de la misma organización que es “dueña” de la información. Además, de acuerdo con (Raghavan y Schneier, 2023) desde el primer enfoque se dice que con ello se delegan los detalles técnicos a un proveedor que cuenta con el personal especializado para atender estos requerimientos. Relacionado con esto, se encuentra el concepto de soberanía de la información. Este concepto ha sido objeto de debate académico y recientemente ha cobrado relevancia en el ámbito gubernamental. Se entiende como parte de un marco más amplio que incluye la soberanía tecnológica —la capacidad de un Estado para tomar decisiones en tecnología basadas en sus propios valores y normas— y la soberanía digital, que aplica este principio al ciberespacio.

Ambas buscan garantizar la integridad de la infraestructura de datos, el control sobre redes y comunicaciones, y la autonomía en el desarrollo de capacidades tecnológicas. La soberanía de los datos pone el énfasis en el control exclusivo sobre la información almacenada y procesada, así como en la capacidad de decidir quién puede acceder a ella. De esta manera, el control de los datos se convierte en el núcleo de la soberanía digital, y los “sujetos soberanos de datos” son quienes pueden ejercer poder sobre ellos.

Según (Kushwaha y Roguski, 2020), la soberanía de los datos es particularmente relevante en el contexto de la computación en la nube. Como muchos de los mayores proveedores de servicios en la nube (CSP) están ubicados en Estados Unidos y actualmente pueden almacenar datos en centros de datos en el extranjero, los datos sensibles podrían almacenarse en servidores situados en los territorios de varios Estados y quedar sujetos a la jurisdicción —y, por lo tanto, al control soberano— de cada uno de esos Estados. Así, de acuerdo con (Kushwaha y Roguski, 2020) las jurisdicciones en

competencia y el control sobre la infraestructura de red (centros de datos) y los propios CSP desafían directamente el control exclusivo que un gobierno podría esperar tener sobre sus datos y, en general, la soberanía de los Estados sobre la información que les pertenece.

El proveedor por su parte, garantiza la disponibilidad aún bajo contingencias, atiende los problemas derivados del crecimiento de la colección, de manera que resultan transparentes para el contratante. Los costos del servicio pueden ser muy atractivos cuando la colección de documentos es pequeña y su volumen cabe en un disco personal. Como se menciona en (Raghavan y Schneier, 2023), estos precios pueden crecer linealmente (o sublinealmente), siempre que no se considere el crecimiento de las operaciones de carga y descarga de datos. Sin embargo, para cuando el volumen se acerque a la escala de los petabytes se puede comprometer la viabilidad presupuestal del proyecto por los costos del servicio. Para entonces, si no se han construido las capacidades propias, para migrar la colección a un sitio a cargo de la propia organización, el proyecto puede verse limitado por problemas financieros.

El balance entre mantener la soberanía sobre los datos y depender de tecnologías externas representa un reto importante para las instituciones, aunque los propietarios puedan perder cierto control sobre la ubicación y la jurisdicción de su información, los servicios en la nube ofrecen ventajas operativas que resultan difíciles de pasar por alto. La capacidad de los proveedores para asegurar alta disponibilidad, resistencia ante contingencias y una escalabilidad prácticamente ilimitada hace que el crecimiento de los volúmenes de datos sea casi transparente para los usuarios. Por lo tanto, según (Giese y Anderl, 2022) la decisión de desarrollar infraestructura propia o recurrir a servicios externos debe evaluar no solo los aspectos legales y de soberanía, sino también los beneficios en términos de eficiencia, costos iniciales reducidos y facilidad de operación.

Por último, como se cita en (Hsu y Irie, 2018) y (Lantz y Furrer, 2025), la escala del proyecto y los patrones en el uso de la información obliga a pensar también en el desarrollo de una jerarquía de almacenamiento. Esto implica la implementación de distintos dispositivos de almacenamiento que respondan a criterios de uso diferenciados. Por un lado, se contempla un conjunto de unidades, preferentemente de estado sólido, denominado almacenamiento primario, con una capacidad relativamente reducida (del orden de algunos terabytes) pero con tiempos de acceso mínimos, destinado a alojar los documentos de consulta más frecuente. Por otro lado, como complemento del almacenamiento primario, se considera el almacenamiento secundario, caracterizado

por una capacidad significativamente mayor, donde se conserva la mayor parte del acervo documental, priorizando la preservación a largo plazo por encima de la velocidad de acceso (probablemente implementada sobre discos mecánicos e incluso cintas magnéticas).

Puede pensarse también en el uso de una federación de celdas. Estas representan un enfoque arquitectónico para integrar y gestionar múltiples unidades o fragmentos de almacenamiento distribuidos en diversos nodos o ubicaciones, viéndolos como un único sistema coherente. En esencia, la federación implica que cada "celda" o unidad de almacenamiento opera de forma autónoma, manteniendo sus propios datos, reglas y gestión. Sin embargo, mediante una capa de federación—que funciona como un sistema virtual— estas celdas se combinan para ofrecer una visión unificada de los datos al usuario o aplicaciones. Esto permite consultar, almacenar y recuperar información como si fuera un solo stock a pesar de estar distribuido físicamente.

De acuerdo con (Wylot y Marcin, 2018) y (Gadiraju y Naayini, 2025), entre sus principales caracteristicas estan:

- No es necesario copiar o mover físicamente los datos, se accede a ellos en su ubicación original.
- Se proporciona una interfaz unificada que traduce y distribuye las consultas a cada celda según corresponda.
- Cada celda puede operar, administrar y proteger su información independientemente. Facilita la distribución de carga y la redundancia para mejorar la disponibilidad y confiabilidad.
- A diferencia de un sistema tradicional donde todos los datos se consolidan y almacenan en un repositorio único, la federación mantiene los datos en sus ubicaciones originales y maneja su integración a través de software especializado (capa de federación).
- Reducción de costos y mejora de tiempos de acceso sobre bases de datos heterogéneas.

5. La gestión tecnológica con visión de largo plazo

Se reconoce que un Biobanco está concebido para sustentar investigaciones a lo largo de periodos extensos, que pueden abarcar varias décadas. Bajo este principio, la gestión del sistema debe integrar un proceso de planeación estratégica que contemple planes de contingencia y de continuidad operativa, esquemas de mantenimiento

preventivo y correctivo (tanto de hardware como de software), estrategias de sustitución o migración tecnológica y mecanismos de relevo generacional. Asimismo, se debe considerar el establecimiento de alianzas estratégicas entre instituciones, así como una gestión financiera de largo plazo que enfatice el retorno de la inversión, entendido desde una perspectiva social que no solo valore los beneficios en la salud de la población, sino también el potencial de transferencia de conocimiento en los ámbitos de la investigación en ciencia básica y el desarrollo tecnológico.

6. El Sistema de Almacenamiento Distribuido Biobanco Iztapalapa (SADBI)

Los diferentes artículos que se han publicado a partir de la investigación basada en los Biobancos, han puesto en evidencia la importancia de disponer de suficiente información sobre las poblaciones locales ya que, aunque muchas de las conclusiones son generalizables, los hallazgos más relevantes provienen de las particularidades de cada población: su ambiente, sus condiciones y estilo de vida, su composición genética o ancestría. De aquí la importancia de construir Biobancos nacionales.

El Biobanco de la alcaldía Iztapalapa es un proyecto multidisciplinario que integra a profesores de las tres divisiones de la UAMI y un equipo extendido que incluye a profesores de otras unidades académicas e investigadores del sector salud (del INRLGII) y nutrición entre otros. Con un apoyo inicial de la SECTEI del gobierno de la CDMX, se han comenzado a desarrollar las capacidades técnicas para la gestión de la información. A este subproyecto se le conoce como el Sistema de Almacenamiento Distribuido: Biobanco de Iztapalapa (SADBI). En una primera etapa se busca alcanzar una capacidad de almacenamiento de entre 60 y 100 TB. En principio, se considera gestionar imágenes, señales biomédicas, estudios de genómica (RNA-seq) y datos sociodemográficos.

Por principio, se considera que el SADBI será un sistema “alimentado” con los documentos conteniendo datos de salud y sociodemográficos, registrados en los diferentes laboratorios e instituciones de salud asociados al proyecto. Visto de otra forma, serán las instituciones los que proveerán al Biobanco con los estudios que realicen. El planteamiento genera tres requerimientos de operación que deben ser atendidos: 1) por un lado se desea que en el Biobanco ingresen documentos con datos protegidos que no vinculen a los estudios con las personas a las que corresponden. Lo anterior significa que los datos personales, deben pasar por un proceso de anonimización, en el que la información sensible se sustituye por una clave numérica, 2) por otro lado, se desea que esta clave numérica coincida en estudios de la misma persona aún si proceden de

diferentes laboratorios. Finalmente,

3) es deseable que sólo el laboratorio de origen retenga la información necesaria para vincular a la clave numérica con la persona a quien corresponde.

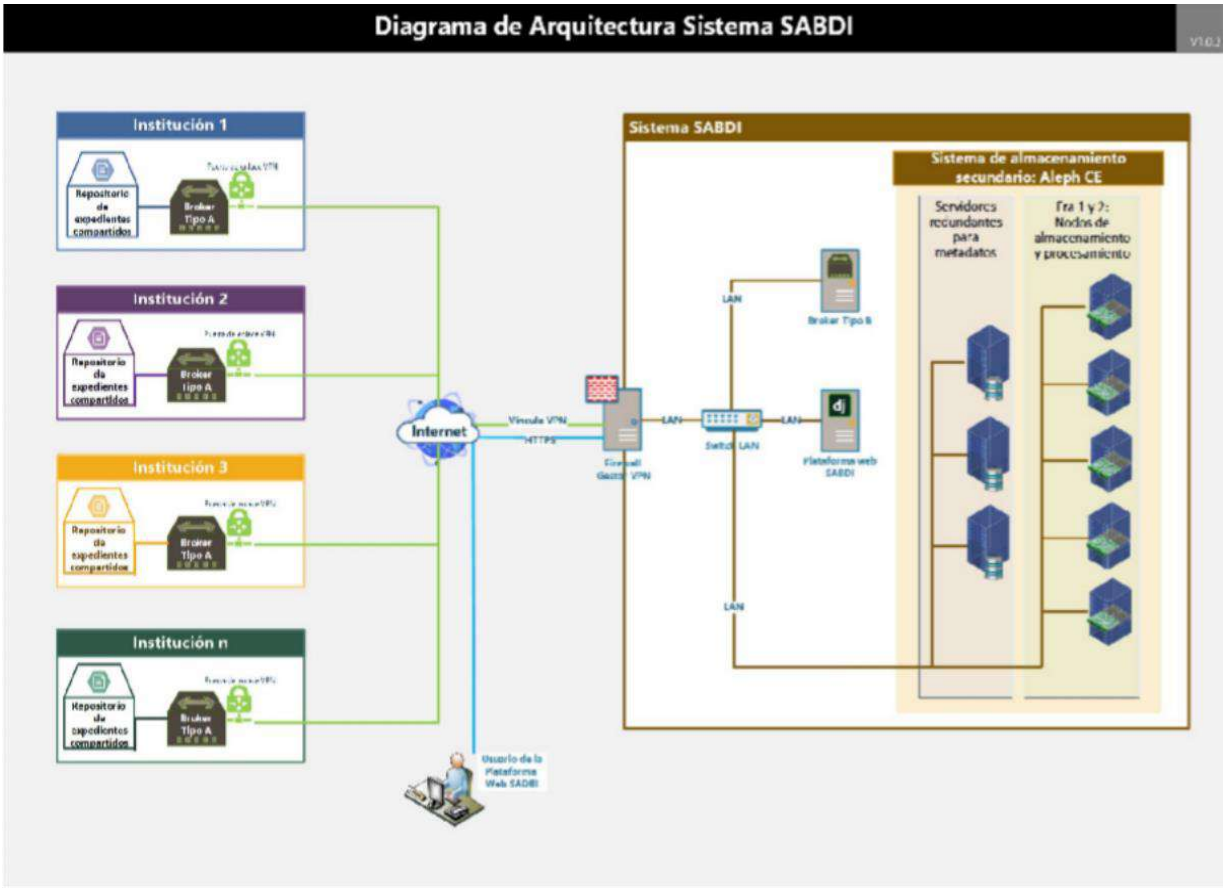
Es importante tomar en cuenta que cada una de las instituciones participantes que genera documentos en el ámbito de sus competencias, es decir, que corresponden a diferentes especialidades, enfoques de estudio y técnicas de análisis: algunos ofrecen imágenes, otras señales, otros datos genómicos y en el curso del tiempo pueden integrarse nuevos tipos de estudios. Cada documento obedece a una norma, posiblemente diferente, que estandariza el intercambio de información, dentro de un contexto específico. Esto es, normas para el intercambio de imágenes, normas para el intercambio de señales, normas para el intercambio de registros de muestras biológicas, etc. Por encima de estas se requiere de una convención de términos llamada "metanorma", que describe el intercambio de todo tipo de documentos de salud, considerando para ello aquellos datos que podría vincular a todos los estudios independientemente de su origen: tales como nombre del paciente (o la clave que lo sustituye), edad, sexo, domicilio (bajo reservas), datos de seguridad social, entre otros.

Bajo estas consideraciones se presenta un caso de uso, a manera de ejemplo, para mostrar un posible flujo de información y la forma como se procesarán los documentos en las diferentes etapas de su recolección, desde su laboratorio de origen hasta su registro y depósito en el SADBI. Para ilustrar este ejemplo, se referirá a la figura 2 donde se observan los componentes que intervienen en el funcionamiento del sistema. Se supone que una institución de salud 1, designa un estudio de imagenología de RX para ser depositado en el SADBI. Se asume que el documento correspondiente se adhiere al estándar DICOM. Este documento será ingresado al bloque denominado "repositorio de expedientes compartidos" en donde debe pasar por el proceso de anonimización de los datos sensibles. Es decir, la protección de los campos o registros del documento que deban sustituirse por cadenas numéricas.

Por los requerimientos previamente considerados, este bloque (a cargo de la institución de salud 1) debe retener la información que haga posible invertir el proceso de anonimización, esto es vincular a un registro numérico con los datos sensibles de una persona. Cuando este proceso termina, el documento se deposita en el "broker tipo A" donde, por su parte, es "ensobretado" dentro de otro documento que obedece las especificaciones de HL7, a la que previemente se definió como la "metanorma". El nuevo documento contendrá la información necesaria para su registro en el SADBI. Enseguida,

el “broker tipo A” envía un mensaje al “broker tipo B” del SADBI, para informarle que está por enviarle por un canal seguro, el documento que ha preparado. Luego de recibirlo, el “broker tipo B” extrae los datos de tipo HL7, que servirán para asociar el documento con una colección dentro del módulo denominado “plataforma web SADBI”.

Figura 2: Arquitectura del sistema de almacenamiento SADBI.



Considerando que el Biobanco SADBI, se enfoca en la población de la alcaldía Iztapalapa de la CDMX, ya que su densidad poblacional es una muestra representativa para estudiar las morbilidades, midiendo la incidencia y prevalencia de enfermedades de dicha población con el objetivo de definir políticas de salud. No obstante, el sistema de Biobanco podría implantarse en otras entidades, obteniendo los beneficios antes mencionados. Los datos extraídos deben incluir información que sirva para establecer que el estudio recibido obedece especificaciones del estándar DICOM. Esto servirá para recuperar los datos que son propios de este tipo de documento y con ello complementar el registro. Finalmente, cuando el estudio haya sido asociado con una colección y almacenado en la “plataforma web SADBI” (a la que se llamará como el almacenamiento

primario) esta misma guardará una copia en el "sistema de almacenamiento secundario". Además, (Núñez-Gaona y Marcelín- Jiménez, 2018), si el documento no es consultado al cabo de un cierto tiempo, entonces sería borrado del almacenamiento primario. Esta decisión de diseño obedece a 2 consideraciones: 1) por un lado se evita el desbordamiento del almacenamiento primario, ya que éste sólo retiene los documentos que hayan sido consultados últimamente, por otro lado 2) se mejora la garantía de disponibilidad del documento pues se dispone de un respaldo.

Recapitulando, el ejemplo presentado muestra las diferentes etapas por las que debe pasar un documento desde su laboratorio de origen hasta su registro y depósito en el almacenamiento secundario. El tratamiento inicial que recibe el documento, en el llamado "repositorio de expedientes compartidos" obedece a la naturaleza particular de cada estudio intercambiado. Esto quiere decir que, si en lugar de imágenes se tratara por ejemplo de datos asociados con muestras biológicas, entonces su procesamiento antes de depositarlo en el "broker tipo A" de la institución, se realizará atendiendo a la norma que corresponda. Luego, se guarda dentro de otro documento con datos que obedecen a la metanorma HL7 y se transfiere a "broker tipo B" el cual debe recuperar los datos necesarios para asociarlo con la colección particular y completar su registro-depósito.

Finalmente, vale la pena mencionar que el "almacenamiento secundario" es un bloque construido con un enfoque de sistema distribuido, tolerante a fallos, escalable y con capacidades definidas por software. Esto último significa que puede ser desplegado sobre cualquier hardware disponible y mediante una interfaz programable. También se habilitan recursos virtuales de almacenamiento que son independientes de los dispositivos físicos que los soportan. Esto quiere decir que se trata de un sistema que evita las dependencias tecnológicas y que puede adaptarse a la disponibilidad de los recursos, al mismo tiempo que puede crecer en la medida que se necesite.

Lo que hace especial al SADBI es que, de inicio, no se considera almacenar datos a partir de tecnologías o servicios contratados, sino que se apuesta por el desarrollo de soluciones propias, en particular, relacionadas con la gestión de la información. Con ello se abre la posibilidad no solo de hacer investigación en temas en ciencias de la salud, sino que permitirá transferir soluciones tecnológicas relacionadas con el manejo de datos para el mismo sector y con ello limitar la posibilidad de dependencias tecnológicas o problemas como el "vendor lockin" (Shaik y Natarajan K., 2024).

El potencial del proyecto para impulsar la transferencia tecnológica incluye desde

las soluciones para el intercambio de expedientes electrónicos, su control de acceso, su preservación en el mediano y largo plazo, apostando por el desarrollo de tecnología propia, libre y abierta, que abone a la soberanía tecnológica.

Conclusiones

La función de los Biobancos es clave para que la investigación científica cuente con las muestras biológicas y los datos necesarios, destinados a estudios, ensayos clínicos y el diseño de nuevos enfoques para diagnosticar y tratar enfermedades. Otra de sus misiones es servir como plataforma para el intercambio de saberes y competencias entre distintas instituciones e investigadores, un factor que acelera el progreso y aumenta la calidad de la investigación. Finalmente, (Bukreeva y Malsagova, 2024) también tienen un importante compromiso con la educación de los profesionales del área biomédica y con la divulgación de los procedimientos y técnicas más eficaces.

Este sistema puede integrarse con otras soluciones de almacenamiento, esta integración depende de los mecanismos de interoperatividad soportados por terceros, sin embargo, el Biobanco define interfaces de comunicación estándar como se describió en la figura 2, cuya funcionalidad puede extenderse al incluir las especificaciones o requerimientos de un tercero, facilitando de esta forma la creación de repositorios extendidos o compartidos.

El éxito de un Biobanco de datos depende de la integración de soluciones tecnológicas avanzadas, desde la preservación de muestras biológicas hasta la seguridad y gestión de grandes volúmenes de datos. La privacidad, calidad y sostenibilidad son aspectos críticos que requieren una planificación meticulosa y el uso de herramientas especializadas, pensado siempre en un horizonte de largo plazo.

Este proyecto establece las bases para integrar modelos de inteligencia artificial especializados en el análisis automatizado de estudios de imagenología y expedientes clínicos. Entre las soluciones más avanzadas destaca Medical Open Network for AI (MONAI) mencionado en (Brudfors y Graham, 2024), el cual es un framework basado en PyTorch desarrollado para imágenes médicas 3D que implementa arquitecturas de vanguardia como nnU-Net para segmentación volumétrica, el sistema Swin-UNETR (Diaz-Pinto y Alle, 2024) para clasificación multi- etiqueta y MONAI Label para anotación semi-supervisada colaborativa, todo ello con fines de investigación. Por último, estan los sistemas Leveraging NVIDIA Clara, que se emplea para la reconstrucción de estudios de tomografía computarizada, resonancia magnética y ultrasonido (Navaneethan y

Hemanth, 2024), y NiftyNet, framework de código abierto diseñado específicamente para análisis de imágenes médicas e intervenciones asistidas por computadora (Gibson y Li, 2018).

Referencias

Aleksandra Mileva, Luca Caviglione (2021). *Risks and Opportunities for Information Hiding in DICOM Standard*, In Proceedings of the 16th International Conference on Availability, Reliability and Security (ARES '21). Association for Computing Machinery, New York, NY, USA, pages. 1–8. https://doi.org/10.1145/3465481.3470072.

Alkhatib R, Gaede KI. (2024). *Data Management in Biobanking: Strategies, Challenges, and Future Directions*, BioTech (Basel) 2;13(3):34. doi: 10.3390/biotech13030034.

Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. (2021). *The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities*,.9(7):e21929. doi: 10.2196/21929.

Bahcall, O.G. (2018). *UK Biobank — a new era in g*enomic medicine, Nat Rev Genet 19, 737, https://doi.org/10.1038/s41576-018-0065-3.

Bakr Kamel, A.A., Belal, N.A., and El-Sonbaty, Y. (2023). *A Novel Computational Method for Predicting DNA Methylation Sites within WGBS of Chromosome Y with WGS Data and Illumina Array Data*, pages. 57-63, doi: 10.1109/ICCTA60978.2023.10969268.

Bomewar, M., Baraskar, T., and Mankar, V. (2015), *DICOM image size reduction and data embedding using randomization technique*, International Conference on Pervasive Computing (ICPC), pages. 1-6, doi: 10.1109/PERVASIVE.2015.7087166.

Brudfors, M., Graham, M,. Ryu, H., and Kutter, O. (2024). *Monai for Deep-Learning Based CBCT Reconstruction*, pages 75-76, doi: 10.1109/ICASSPW62465.2024.10626056.

Bukreeva, Anastasiia & Malsagova, Kristina & Petrovskiy, Denis & Butkova, Tatiana & Nakhod, Valeriya & Rudnev, Vladimir & Izotov, Alexander & Kaysheva, A.. (2024). *Bio*bank Digitalization: From Data Acquisition to Efficient Use, Biology. pages. 13- 957, DOI: 10.3390/biology13120957.

Byron, J., Long, D.D.E., and Miller, E.L. (2018), *Using Simulation to Design Scalable and Cost-Efficient Archival Storage Systems*, pages. 25-39, doi: 10.1109/MASCOTS.2018.00011.

Coughlin, T., y Hoyt, R. (2024). *La Hoja de Ruta del IEEE describe el desarrollo de la tecnología de almacenamiento digital masivo*, vol. 57, n.º 09, págs. 111-116, doi: 10.1109/MC.2024.3416230.

Dandu RV (2008), *Storage media for computers in radiology. Indian J Radiol Imaging*, 18(4): 287-9. doi: 10.4103/0971-3026.43838.

Diaz-Pinto, Andrés, Sachidanand, Alle, Et. al.,(2024), *MONAI Label: A framework for AI-assisted interactive labeling of 3D medical i*mages, Volume 95, https://doi.org/10.1016/j.media.2024.103207.

Eli Gibson; Li, V.; Sudre, Carole Et. Al (2018). *NiftyNet: a deep-learning platform for medical imaging, Computer Methods and Programs in Biomedicine*, Volume 158, Pages 113-122, https//doi.org/10.1016/j.cmpb.2018.01.025.

Fadhil, Jawaher & Zeebaree, Subhi. (2024). *Blockchain for Distributed Systems Security in Cloud Computing: A Review of Applications and Challenges*, vol 13. pages.1576-1605, https://doi.org/10.33022/ijcs.v13i2.3794.

Gadiraju, P., Naayini, P., Somayajula, R., and Aunugu, D.R. (2025). *Federated-Aware Cluster Computing for Resilient and Scalable Machine Learning*, pages. 1-7, doi: 10.1109/ICOCT64433.2025.11118503.

Giese, T., and Anderl, R. (2022), *Maintaining Control over Distributed Data Through a Data Sovereignty Model*, pages. 1-7, doi: 10.1109/ICITDA55840.2022.9971218.

Hernaez, M., Pavlichin, D., Weissman, T. & Ochoa, I. (2019). *Genomic data compression*, Annu. Rev. Biomed. Data Sci. 2, pages. 19–37. https://doi.org/10.1146/annurev-biodatasci-072018-021229.

Hsu, Y.F., Irie, R., Murata, S. and Matsuoka, M. (2018). *A Novel Automated Cloud Storage Tiering System through Hot-Cold Data Classification*", pages. 492-499, doi: 10.1109/CLOUD.2018.00069.

Jonathan M. Mudge, Sílvia Carbonell-Sala (2025), GENCODE 2025: reference gene annotation for human and mouse, Vol 53, pages. D966–D975, DOI: 10.1093/nar/gkae1078.

Juozapaitė, D, et al. (2023). *The COVID-19 pandemic reveals the wide-ranging role of biobanks*, Frontiers in Public Health, pages. 1-11, doi: 10.3389/fpubh.2023.125660.

Kushwaha, N., Roguski, P., and Watson, B.W. (2020. *Up in the Air: Ensuring Government Data Sovereignty in the Cloud*", pages. 43-61, doi: 10.23919/CyCon49761.2020.9131718.

Leila Namvari-Tazehkand, Saeid Pashazadeh, (2021). *Investigating the Reliability in Three RAID Storage Models and Effect of Ordering Replicas on Disks*, https://doi.org/10.48550/arXiv.2104.01238.

Malone, J.H., Oliver, B.(2011). *Microarrays, deep sequencing an*d the true measure of the transcriptome, https://doi.org/10.1186/1741-7007-9-34.

Mark A. Lantz, Simeon Furrer, Martin Petermann, Hugo Rothuizen, Stella Brach, Luzius Kronig, Ilias Iliadis, Beat Weiss, Ed R. Childers, and David Pease. (2025), *Magnetic Tape Storage Technology. ACM Trans. Storage,* Volume 21, pages. 1-70, https://doi.org/10.1145/3708997.

Marta Różańska and Kyriakos Kritikos (2019), Good Bye Vendor Lock-in: Getting your Cloud Applications Multi-Cloud Ready!, pages. 171–173. https://doi.org/10.1145/3368235.3370267.

Müller, Heimo & Reihs, Robert, Et. Al. (2015). *State-of-the-Art and Future Challenges in the Integration of Biobank Catalogues*, pages 3-11,doi: 10.1007/978-3-319-16226- 3_11.

Nadaf, R., Sumangala, N.B., Mandi, M., and Konnur, A. (2023). *Symmetric and Asymmetric Cryptographic Approach based Security Protocol for Key Exchange*", pages. 1-6, doi: 10.1109/ICAISC58445.2023.10200989.

Naranjo-Martínez, Y., Zaldívar-Revé, C., & González-Chaveco. (2022). *Sistema de Gestión de la Información de Laboratorios en BioCubaFarma / Laboratory Information Management System*. 28(4), pages. 1-10,doi: 10.22201/fesc.20075057e.2022.28.4.

Navaneethan, S., Hemanth, G., and Nikilkumar, R. (2024), *Leveraging NVIDIA Clara for Real-Time Cardiac Image Segmentation and Diagnosis*", pages. 1-7, doi: 10.1109/ICSES63760.2024.10910358.

Nicholas Bradley Larson, Ann L. Oberg, Alex A. Adjei, Liguo Wang (2023). *A Clinician's Guide to Bioinformatics for Next-Generation Sequencing*, Journal of Thoracic Oncology, Volume 18, Issue 2, Pages 143-157, ISSN 1556-0864, https://doi.org/10.1016/j.jtho.2022.11.006.

Núñez-Gaona MA, Marcelín-Jiménez R, Gutiérrez-Martínez J, Aguirre-Meneses H, Gonzalez-Compean JL. (2018), *A Dependable Massive Storage Service for Medical Imaging*. (5):628-639. doi: 10.1007/s10278-018-0091-x.

Opara-Martins, J., Sahandi, R., & Tian, F. (2014). *Critical review of vendor lock-in and its impact on adoption of cloud computing*, vol. 5, pages. 92-97, doi: 10.1109/i-Society.2014.7009018.

Opara-Martins, J., Sahandi, R., & Tian, F. (2016). *Critical analysis of vendor lock-in and its impact on cloud computing migration: a business perspective,* 5(1), 4, https://doi.org/10.1186/s13677-016-0054-z.

Păscuțoiu, O., Tache Ungureanu, M.D., Teodor, C.I., Moraru, S. A. (2025). *Enhancing DICOM Security in Medical Imaging Networks Using Software-Defined Networking*, pages. 1-7, doi: 10.1109/ECAI65401.2025.11095443.

Priya, B., Deepan, N. (2023). *A Web-based Dicom Image and Plane Viewer*, pages. 1-7, doi: 10.1109/ICECCT56650.2023.10179664.

Raghavan, B., and Schneier, B. (2023). A Bold New Plan for Preserving Online Privacy and Security: Decouplin*g our identities from our data and actions could safeguard our secrets in the cloud*", vol. 60, pages. 22-29, doi: 10.1109/MSPEC.2023.10352412.

Roberta Gazzarata, Joao Almeida, (2024). *HL7 Fast Healthcare Interoperability Resources (HL7 FHIR) in digital healthcare ecosystems for chronic disease management: Scoping review*, International Journal of Medical Informatics, Vol. 189, 2024, pages. 1-11, https://doi.org/10.1016/j.ijmedinf.2024.105507.

Sankaranarayanan, S. et al. (2024): *Enhancing Healthcare Imaging Security: Color Secret Sharing Protocol for the Secure Transmission of Medical Images*, vol. 12, pages. 100200-100216, doi: 10.1109/ACCESS.2024.3426935.

Scapicchio C, Gabelloni M, Forte SM, Alberich LC, Faggioni L, Borgheresi R, Erba P, Paiar F, Marti-Bonmati L, Neri E.(2021). *DICOM-MIABIS integration model for biobanks: a use case of the EU PRIMAGE project,* Eur Radiol Exp. 12;5(1):20. doi: 10.1186/s41747-021-00214-4. PMID: 33977357; PMCID: PMC8113005.

Seebode, C., Ort, M., Hufnagl, P., and Regenbrecht, C. R. A. (2015). *Next generation biobanks*, pages. 1362-1367, doi: 10.1109/BigData.2015.7363896.

Vaheedbasha Shaik, Natarajan K. (2024). *Cloud databases: A resilient and robust framework to dissolve vendor lock-in, Software Impacts*, Volume 21, pages. 1-5, https://doi.org/10.1016/j.simpa.2024.100680.

Watts, Geoff (2012). *UK Biobank opens its data vaults to researchers*, BMJ. 344: e2459. doi:10.1136/bmj.e2459. ISSN 1756-1833.

Wylot, Marcin & Hauswirth, Manfred & Cudre-Mauroux, Philippe & Sakr, Sherif. (2018). *RDF Data Storage and Query Processing Schemes: A Survey*, Volume 51. pages. 1-36, doi: 10.1145/3177850.

Yujie Lian, Rui Geng, and Deqiang Zhu. (2025). *Whole-Genome Sequencing and Compara*tive *Genomic Analysis of a Selenium-Enriched Saccharomyces cerevisiae Strain*, pages. 235–241, https://doi.org/10.1145/3745034.3745071.

Zhou, J., Chen, Y., Zheng, M., and Wang, W. (2023), *Data Distribution for Heterogeneous Storage Systems*, vol. 72, no. 6, pages. 1747-1762, doi: 10.1109/TC.2022.3223302.

Conflicto de interés

Los autores de este manuscrito declaran no tener ningún conflicto de interés.

Declaración ética

Los autores declaran que el proceso de investigación que dio lugar al presente manuscrito se desarrolló siguiendo criterios éticos, por lo que fueron empleadas en forma racional y profesional las herramientas tecnológicas asociadas a la generación del conocimiento.

Copyright

Licencia CreativeCommons