Normas Internacionales de Anotación de Información de Imagen Médica

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN
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			
Paquete Trabajo	WP4 – Datos Médicos e Imagen			
Tarea	T4.2. Adaptación, instalación y uso de herramientas de software de código abierto para el seguimiento y validación de los procesos de extracción y tratamiento de los datos de imagen médica			
Entregable	E4.4. Normas Internacionales de Anotación de Información de Imagen Médica			
Versión	1.1.1			
Fecha Entrega	31/03/2022		***Fecha Aprobación***	17/05/2023
Responsable	ISS La Fe-GIBI230			
Nivel Diseminación	X	PU	Public	
		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***
ISS La Fe-GIBI230	Luis Martí Bonmatí	Coordinador / Autor
ISS La Fe-GIBI230	Pedro Mallol Roselló	Coordinador / Autor
ISS La Fe-GIBI230	Leonor Cerdá Alberich	Autor
FISABIO-UMIB	Mariam de la Iglesia-Vayá	Autor
FISABIO-UMIB	Silvia Nadal	Autor
INCLIVA	Andrés Cervantes	Revisor
CNIC	Daniel Jiménez Carretero	Revisor
HCB	Santiago Andrés Frid	Autor
HCB	Antonio López	Autor
SAS-HUVR	Carlos Luis Parra Calderón	Autor
SAS-HUVR	Sara González García	Autor
IGTP-HCGB	Lauro Sumoy Van Dyck	Autor

Historial de versiones			
Nro.	***Fecha***	***Descripción***	***Autor***
V 0.1	11/01/2022	Borrador Índice	L. Martí (ISS La Fe-GIBI230) P. Mallol (ISS La Fe-GIBI230)
V 0.2	17/01/2022	Aportes al Índice	M. de la Iglesia (FISABIO-UMIB)
V 0.3	25/01/2022	Modificaciones al Índice	L. Martí (ISS La Fe-GIBI230) P. Mallol (ISS La Fe-GIBI230)
V 0.4	10/02/2022	Aportaciones al entregable (SNOMED, CIE)	S.A. Frid (HCB)
V 0.5	16/02/2022	Aportaciones al entregable (DICOM, NIFTI, Población en repositorios, Características herramientas)	R. Romero García (INiBICA)
V 0.6	17/02/2022	Aportaciones al entregable	L. Martí (ISS La Fe-GIBI230) P. Mallol (ISS La Fe-GIBI230)
V 0.7	22/02/2022	Aportaciones al entregable (Principios FAIR en Imagen Médica y Principios FAIR en Repositorios)	S. González (SAS-HUVR)
V 0.8	27/02/2022	Aportaciones al entregable	L. Martí (ISS La Fe-GIBI230) P. Mallol (ISS La Fe-GIBI230)
V 0.9	28/02/2022	Aportaciones al entregable (DICOM RT, DICOM SR, Selección repositorio caso de uso)	S.A. Frid (HCB) A. López (HCB)
V 0.10	28/02/2022	Aportaciones al entregable (Ontologías en el dominio biomédico, anonimización en imagen médica, UMLS, BIDS, MIDS, Estándares en anonimización, Informe exploratorio, DOI)	Marian de la Iglesia (FISABIO_UMIB) Silvia Nadal (FISBIO-UMIB)
V 0.11	28/02/2022	Aportaciones al entregable (Herramientas de procesado en imagen médica: Definición, Preparación,	L. Cerdá (ISS La Fe-GIBI230)

		Detección y Segmentación, Extracción de Radiómica)	
V 0.12	02/03/2022	Aportaciones finales e Integración con el resto del entregable	M. de la Iglesia (FISABIO-UMIB) P. Mallol (ISS La Fe-GIBI230)
V 0.13	11/03/2022	Revisión general del documento	D. Jimenez (CNIC)
V 0.14	13/03/2022	Revisión general del documento	L. Martí (ISS La Fe-GIBI230)
V 0.15	17/03/2022	Revisión general del documento	P. Mallol (ISS La Fe-GIBI230)
V 0.16	19/03/2022	Revisión general del documento	Marian de la Iglesia (FISABIO_UMIB) Silvia Nadal (FISBIO-UMIB)
V 0.17	21/03/2022	Aporte sobre UMLS	Lauro Sumoy (IGTP-HCGB)
V 1.0	21/03/2022	Revisión general del documento	L. Martí (ISS La Fe-GIBI230) P. Mallol (ISS La Fe-GIBI230)
V 1.1	17/05/2023	Cambio visibilidad a público y aprobado	Comité Dirección
V 1.1.1	14/06/2023	Cambio de formato para publicar en la Web de IMPaCT	David Velasco (ISCIII)

Contenido

Contenido 5
Tablas 8
Figuras 8
Resumen Ejecutivo 9
Introducción 9
Audiencia 9
Ámbito 9
Relación con otros Entregables 9
Estructura Entregable 10
Terminología 11
1.1 Principios FAIR en imagen médica 11
1.2 Definición de conceptos 12
1.2.1 Codificación 12
1.2.2 Ontologías en el dominio biomédico 12
1.2.3 Niveles de Anotación 13
1.2.3.1 Anotación a nivel de Sujeto 13
1.2.3.2 Anotación a nivel de Imagen 13
1.2.3.3 Anotación a nivel de Capa 14
1.2.4 Anonimización en imagen médica 14
1.2.5 Repositorios de imagen médica 14
1.2.6 Herramientas de análisis en imagen 14
2 Estándares de Anotación Imagen Médica 15
2.1 Nivel de Sujeto 15
2.1.1 SNOMED-CT 15
2.1.2 CIE (Clasificación Internacional de Enfermedades) 17
2.1.3 UMLS (Sistema Unificado de Lenguaje Médico) 18
2.1.3.1 Fuentes de conocimiento de la UMLS 19
2.1.3.2 Identificadores Únicos de Concepto (CUI) 20
2.1.3.3 Utilizaciones de UMLS 20
2.2 Nivel de Imagen 21

2.2.1 DICOM 21
2.2.1.1 Estructura de un fichero DICOM 22
2.2.1.2 Elemento de Datos o Atributos 23
2.2.2 Analyze 7.5 25
2.2.3 NifTI 25
2.2.4 BIDS 26
2.2.5 MIDS 27
2.2.6 REMBI 28
2.3 Nivel de Capa 29
2.3.1 DICOM SR 29
2.3.2 DICOM SEG 29
2.3.3 DICOM RWVM 29
2.3.4 DICOM RT 30
2.3.5 Radiomics Ontology 30
2.3.6 NIfTI 30
2.4 Estándares de relación entre la Historia Clínica Electrónica y la Imagen Médica 30
2.5 Estándares de Anonimización de la Imagen Médica 31
2.5.1 De-identificación de meta-información DICOM 31
2.5.1.1 Perfil básico 32
2.5.1.2 Mantener segura la información privada 33
2.5.2 Protocolo adoptado 33
2.5.3 De-identificación de información gráfica 33
2.5.3.1 De-identificación de información facial 33
2.5.3.2 De-identificación de anotaciones de texto superpuestas a la imagen 34
2.5.4 De-identificación en informes radiológicos 35
3 Repositorios de imagen médica 38
3.1 Características y Contenido 38
3.1.1 Identificación del repositorio 39
3.1.2 Arquitecturas de los repositorios 40
3.1.2.1 Repositorios Centralizados 40
3.1.2.2 Repositorios Descentralizados (Federados) 41
3.1.2.3 Repositorios Mixtos 42
3.1.3 Descripción de la población de estudio 42

3.1.4 Acceso a los datos del repositorio 43
3.1.5 Descripción del repositorio 43
3.1.5.1 Anotaciones 44
3.1.5.2 Patrón de referencia 45
3.1.6 Cumplimiento FAIR 45
3.2 Estructura del listado de repositorios 45
4 Herramientas de procesado en imagen médica 46
4.1 Definiciones y Clasificación 46
4.1.1 Preparación de la imagen 47
4.1.1.1 Preprocesado de la imagen 47
4.1.1.2 Registro 49
4.1.2 Detección y Segmentación 50
4.1.3 Extracción de radiómica 51
4.2 Estructura del listado de herramientas 52
5 Caso de Uso 53
5.1 Selección de un repositorio de imagen médica 53
5.2 Informe exploratorio 53
5.3 Análisis del repositorio 55
5.3.1 Población del repositorio 55
5.3.2 Acceso del repositorio 55
5.3.3 Descripción del repositorio 55
6 Conclusiones 56
Referencias 57
Acrónimos y Abreviaturas 61
Anexo A. Estructura Listado Repositorios en Imagen Médica 62
Anexo B. Estructura Listado Herramientas en Imagen Médica 63
Anexo C. Perfil de de-identificación de las cabeceras DICOM propuesto por la RSNA 64
Anexo D. Cabecera archivo NifTI 79

Tablas

Tabla 1.- Data Elements necesarios para la lectura adecuada de una imagen 24

Tabla 2.- Selección de herramientas de segmentación de órganos y tumores. 50

Figuras

Figura 1.- Tres tipos de anotación en imagen: a) Segmentación de región anatómica de interés, b) Segmentación de región patológica de interés, c) Medidas geométricas (6) 13

Figura 2.- Modelo Lógico de SNOMED-CT 16

Figura 3.- Modelo Conceptual de SNOMED-CT 16

Figura 4.- eCIE10 - Edición electrónica de la CIE-10", 2010 17

Figura 5.- "CIE-11 para estadísticas de mortalidad y morbilidad", 2021. 18

Figura 6.- Fuentes de conocimiento de UMLS 19

Figura 7.- Estructura de un archivo DICOM (izquierda) y ejemplo de un archivo DICOM (derecha).....21

Figura 8.- Transformación a un objeto IOD 22

Figura 9.- Jerarquía de un archivo DICOM.....23

Figura 10.- Campos de un Elemento de Datos DICOM.....24

Figura 11.- Estructura de archivos BIDS 27

Figura 12.- Anotación en Imagen médica basada en UNLS. Proyecto PADCHEST.....31

Figura 13.- Esquema conceptual de anonimización facial 34

Figura 14.- Eliminación del texto incrustado en la imagen 34

Figura 15.- Máscara utilizada en la figura 10.....35

Figura 16.- De-identificaión de informes.....37

Figura 17.- Modelo de repositorio centralizado (27) 40

Figura 18.- Modelo de repositorio descentralizado (federado) (27).....41

Figura 19.- Escala MIDaR.....44

Figura 20.- Arquitectura de tipo GAN. Fuente: Medium.....49

Figura 21.- Visualización análisis exploratorio.....54

Resumen Ejecutivo

En este documento se hace una recopilación de los Estándares de Anotación de Información en Imagen Médica más utilizados. De ellos se seleccionarán los más adecuados para la prueba de concepto que se realizará en el entregable 4.6.

Entendiendo que la definición de anotación en imagen médica puede ser un poco confusa, en este trabajo se considera la anotación como el proceso de añadir información adicional a las imágenes médicas ya adquiridas. Se pretende abordar de una manera más amplia la anotación, diferenciando en los siguientes estratos: nivel de Sujeto, nivel de Imagen y nivel de Capa. Existen normas de anotación compartidas con las normas descritas en el entregable 4.1 y de las cuales simplemente se hará una mención en este documento.

Este documento servirá de base para el futuro análisis de los repositorios de imagen médica existentes y de las herramientas de programación para realizar las anotaciones. Se establecerán criterios de calidad para poder valorar los diferentes repositorios y herramientas, y de esta forma finalmente evaluar un repositorio de imagen médica y realizar un informe de su calidad global.

Introducción

Audiencia

Este entregable está dirigido a las personas que pretendan realizar una investigación que requiera del análisis de imágenes médicas, especialmente con imágenes médicas de radiología y medicina nuclear, pero también de anatomía patológica. Será de especial interés para los participantes del proyecto IMPaCT-DATA que estén implicados en el paquete de trabajo 4, en la integración de la imagen que se pretende hacer en el paquete 5, y en los posibles casos de uso que se pretenden realizar en el paquete 6.

Ámbito

Este documento supone un punto de partida para trabajar con el análisis de las imágenes médicas, describiendo un amplio abanico de posibilidades y permitiendo seleccionar los estándares que mejor se ajusten a la investigación que se quiera llevar a cabo.

Relación con otros Entregables

Este entregable forma parte de la Tarea 4.2 en el Paquete de Trabajo 4. Tiene como objetivo tratar los principales estándares de anotación en Datos Médicos y de Imagen. Este entregable guarda relación con el entregable 4.1 que pretende recoger los estándares de anotación de información de la Historia Clínica Electrónica (HCE). También estará relacionado con el entregable 4.5 donde se realizará una selección de estos estándares, e

indirectamente con el entregable 4.6 donde se realizará una prueba de concepto de extracción de información de la imagen médica con los estándares seleccionados en el entregable 4.5.

Este documento también guarda relación con el entregable 5.1.- Técnicas de Integración de Datos Biomédicos; entregable 5.5.- Especificaciones Revisadas para la Inclusión de Marcadores Biomédicos de Imagen Médica; tarea 6.1.- Desarrollos de casos de uso alineados con IMPaCT-DATA.

Estructura Entregable

El entregable tiene la siguiente estructura general: (1) Terminología, (2) Estándares de Anotación en Imagen Médica, (3) Repositorios de Imagen Médica, (4) Herramientas de procesado en Imagen Médica y (5) Caso de uso.

En la primera sección se describen los conceptos más relevantes que se van a tratar durante el entregable; posteriormente se definen los estándares relacionados con la imagen médica; a continuación, se describen los repositorios de imagen médica, analizando su importancia y su necesidad; en el cuarto apartado se introducirán los tipos de herramientas más utilizadas en el procesado de imagen; y por último se pondrán las bases del análisis de un caso de uso de repositorio en imagen médica.

Terminología

1.1 Principios FAIR en imagen médica

En este documento se considera el uso secundario de la imagen médica y su información contenida. Por ello es importante establecer los criterios que deben cumplir los datos para poder ser utilizados en investigación. En el marco de la Ciencia Abierta, los principios FAIR (1), formalmente publicados en 2016 por la comunidad Force11 (2), proveen cuatro guías que deben cumplir estos datos de uso secundario. FAIR viene dado por las siglas en inglés Findable-Accessible-Interoperable-Reusable, lo que indica que los datos deben ser Localizables, Accesibles, Interoperables y Reutilizables. Con el fin de cumplir cada una de estas dimensiones, los datos de imagen médica deben poder ser trazables en estas propiedades:

- **Localizables**: las imágenes deben describirse con metadatos ricos y relacionados entre sí, mediante un identificador único, y estar registradas o indexadas en un recurso que permita realizar búsquedas.
- **Accesibles**: las imágenes y sus metadatos deben poder recuperarse a partir de su identificador, utilizando un protocolo de comunicación abierto, gratuito y de aplicación universal, que permita la autenticación y la autorización si es necesario.
- **Interoperabilidad** de las imágenes médicas: debe utilizarse un lenguaje específico para la representación del conocimiento, mediante el uso de ontologías públicas y otras tecnologías semánticas (3,4), como el Marco de Descripción de Recursos (RDF) (5) para la creación de un modelo de datos semántico que permita representar las relaciones entre los múltiples conceptos de datos y metadatos.
- **Reutilización** de las imágenes médicas: las descripciones de sus atributos deben incluir información de licencia y procedencia, entre otros.

El plan estratégico de IMPaCT-Data establece el siguiente compromiso respecto al cumplimiento de los principios FAIR en la acción nº 2 de la LET1 de Integridad científica, en la que se deben establecer los procedimientos para garantizar los principios FAIR en la organización y gestión de la información en IMPaCT. Esto se materializará en un entregable E1 “Sistema de Evaluación de principios FAIR”.

1.2 Definición de conceptos

1.2.1 Codificación

Codificar es el proceso de transformar un tipo de información en otra. La nueva información debe estar bien delimitada, previamente consensuada y basada en normas internacionales, sin perder nada de la información inicial en esta transformación.

1.2.2 Ontologías en el dominio biomédico

El vocabulario propio de las ciencias de la salud se caracteriza por su riqueza, ambigüedad, dependencia contextual, y uso de jerga, acrónimos y términos altamente especializados. Estos atributos determinan la necesidad de elaborar estrategias de representación del conocimiento que optimicen el rendimiento de los sistemas de información en salud, tanto para fines asistenciales como estadísticos y de investigación. La codificación emerge entonces como una solución para que los sistemas puedan "*comprender*" el lenguaje.

Pueden distinguirse dos clasificaciones para este tipo de estándares: las terminologías y los clasificadores. Las **Terminologías** son conjuntos limitados de términos que tienen por objetivo desambiguar la información transmitida en un ámbito particular. Una vez que la terminología de un dominio específico es revisada, depurada y validada por alguna organización, se convierte en una **Nomenclatura**. Por otra parte, las **Clasificaciones** son sistemas ordenados de conceptos pertenecientes a un dominio y diseñados con un fin particular. Se conforman por listas de categorías, cada una de las cuales alberga conceptos clínicos vinculados por alguna característica común.

En Ciencia y Tecnología de la Información, las ontologías son clasificaciones. Se utilizan como un medio para categorizar o agrupar la información en clases. Las ontologías se aplican también en web semántica y en inteligencia artificial para asimilar y codificar el conocimiento, definiendo las relaciones existentes entre los conceptos de un determinado dominio. Las ontologías en el dominio biomédico proveen un marco organizativo de los conceptos involucrados en entidades y procesos biológicos, en un sistema de relaciones jerárquicas y asociativas para razonar sobre el conocimiento del dominio. Las terminologías biomédicas promueven una manera estándar de nombrar los conceptos del dominio a estudiar.

Las principales ontologías en el dominio biomédico son:

- SNOMED: Systematized Nomenclature of Medicine Clinical Terms.
- UMLS: Unified Medical Language System.
- MeSH: Medical Subject Headings.
- ICD: International Classification of Diseases (en Español, CIE).

1.2.3 Niveles de Anotación

La **Anotación** en la Imagen Médica puede definirse como el proceso de añadir información adicional a las imágenes. En 2017, durante la primera conferencia anual de Inteligencia Artificial en Imagen Médica (6), se definió la anotación como una clase especial de metadatos que pertenecen a los diagnósticos asociados, o a las regiones anatómicas o patológicas contenidas en una determinada imagen (Figura 1)

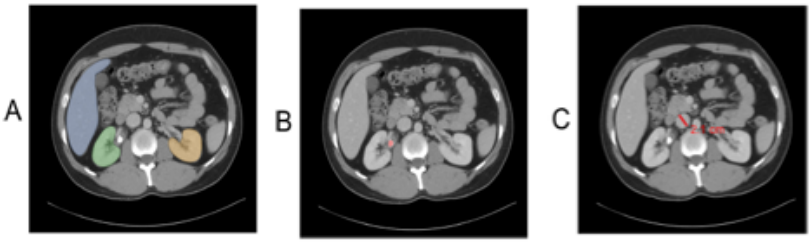

Figura 1.- Tres tipos de anotación en imagen: a) Segmentación de región anatómica de interés, b) Segmentación de región patológica de interés, c) Medidas geométricas (6)

En este trabajo se ha optado por un enfoque más amplio, incluyendo la información imprescindible para desarrollar algoritmos de inteligencia artificial con un interés final hacia la práctica clínica habitual. Así, se considerará la información clínica asociada al sujeto de estudio, la información que se deriva de la propia imagen médica, y la información relacionada con el procesamiento digital de la imagen. Así pues, se consideran los tres niveles de anotación en imagen médica que se describen a continuación.

1.2.3.1 Anotación a nivel de Sujeto

Se corresponde con todas aquellas anotaciones que hacen referencia al sujeto de estudio, tanto si se trata de controles como de pacientes con enfermedades definidas. En este nivel se añade información de la enfermedad que padece el paciente y afecta a la imagen o de cualquier otra variable clínica o de tratamiento que se considere relevante para los análisis que se realicen. Este tipo de anotaciones también pueden servir para identificar grupos de imágenes en función de una determinada característica del sujeto, como cohortes o subpoblaciones de estudio.

1.2.3.2 Anotación a nivel de Imagen

Son aquellas anotaciones que se representan como metadatos contenidos en la misma imagen médica reconstruida o añadidos *a posteriori* y que se relacionados con el estudio y el protocolo de adquisición. Ejemplos relevantes son la dosis del radiofármaco que se le ha administrado a un paciente en un estudio PET o la estandarización del tipo de secuencia de RM adquirida.

1.2.3.3 Anotación a nivel de Capa

Por último, esta anotación se añade como pieza de información solapada a la propia imagen. Como ejemplos están las máscaras de las segmentaciones, el etiquetado de los órganos y estructuras identificados y los marcadores geométricos como las medidas de distancia, superficie y ángulos.

1.2.4 Anonimización en imagen médica

La anonimización en imagen médica consiste en la eliminación de aquellos elementos de la imagen y de sus metadatos que permitan identificar al sujeto al cual se le ha realizado esta imagen. Entre la información identificable que puede aparecer en las imágenes médicas podemos listar principalmente a los identificadores en metadatos, textos superpuestos sobre la imagen, características faciales e información codificada en los datos gráficos.

1.2.5 Repositorios de imagen médica

Un repositorio de imágenes médicas se define como el espacio donde se almacenan, organizan, mantienen y difunden imágenes médicas digitales. En este entregable nos centraremos principalmente en los repositorios que disponen de imágenes médicas radiológicas, incluyendo medicina nuclear.

Según el grupo de trabajo de la Sociedad Europea de Radiología (ESR, 2014) (7), estos repositorios se definen como una base organizada de datos de imágenes médicas y biomarcadores de imagen asociados, compartida entre investigadores y ligada a otros biorepostorios. Un objetivo inmediato de los biobancos de imágenes es la generación de conocimiento y de biomarcadores de imágenes para su uso en estudios de investigación, apoyando la validación biológica de los biomarcadores de imágenes existentes.

1.2.6 Herramientas de análisis en imagen

Las herramientas de análisis en imagen son un conjunto de algoritmos computacionales y aplicaciones o programas informáticos utilizados para procesar imágenes, con el objetivo de mejorar su calidad, o extraer información y medidas que representen características cualitativas y cuantitativas de interés. Estas características pueden ser de utilidad para aplicaciones médicas, para el diagnóstico o para el pronóstico de enfermedades.

2 Estándares de Anotación Imagen Médica

2.1 Nivel de Sujeto

Conocer el entorno clínico del paciente permite orientar la investigación en función del uso clínico por el que se hayan adquirido. Es pues necesario que las imágenes estén anotadas con un mismo vocabulario dentro las diferentes ontologías clínicas de las que disponemos (8). Además, el creciente uso de la IA y el procesamiento automatizado ha ido permitiendo un mejor y más completo control y uso de datos masivos (en este caso imágenes y repositorios de imágenes), en parte gracias a la capacidad de poder utilizar toda la información disponible de los sujetos de estudio. Sin embargo, ello hace imprescindible el uso de una misma anotación común que permita integrar los datos de distintos sujetos e interoperar con otros repositorios de imágenes médicas. Así pues, el uso de ontologías para facilitar la anotación semántica de imágenes médicas se convierte en un proceso vital a la hora de realizar anotaciones en la imagen médica (9).

En este apartado, se nombrarán los estándares considerados más relevantes a la hora de realizar anotaciones a nivel de los sujetos de estudio

2.1.1 SNOMED-CT

Actualmente, la *terminología* clínica de referencia más completa y más utilizada a nivel mundial es **SNOMED-CT**. Está basada en conceptos, sus descripciones y relaciones que describen la vinculación jerárquica y los atributos de dichos conceptos. Representa el conocimiento a través de una ontología médica, lo cual permite una gran riqueza semántica y facilidad de crecimiento evolutivo. De esta forma, con SNOMED-CT se puede desarrollar contenido médico completo y de alta calidad, permitiendo codificar, comunicar, o anotar datos médicos de manera precisa e inequívoca.

El modelo lógico de SNOMED-CT define el modo en el que se relaciona y representa cada tipo de componente de la terminología (conceptos, descripciones y relaciones) y sus derivados (Figura 2).

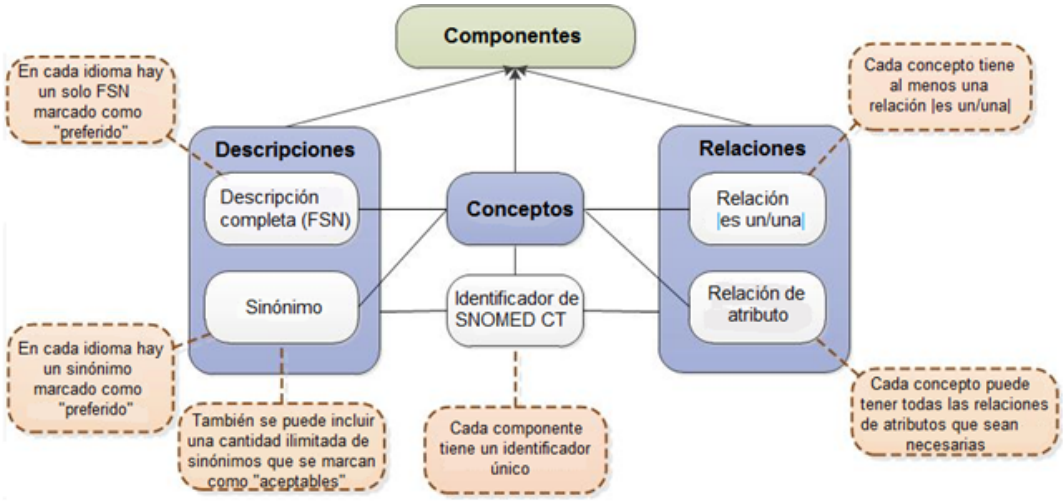

Figura 2.- Modelo Lógico de SNOMED-CT

A su vez, el modelo conceptual de SNOMED-CT es el que especifica cómo se definen los conceptos de la terminología mediante una combinación de lógica formal y reglas editoriales. Los atributos que describen a los conceptos deben tener un dominio (la jerarquía a la que pertenecen) y un rango (los conceptos de SNOMED-CT permitidos como valores del atributo). De las 19 jerarquías que tiene la terminología (conceptos ubicados en el nivel superior), los atributos permiten representar el significado de conceptos de nueve de ellas (Figura 3).

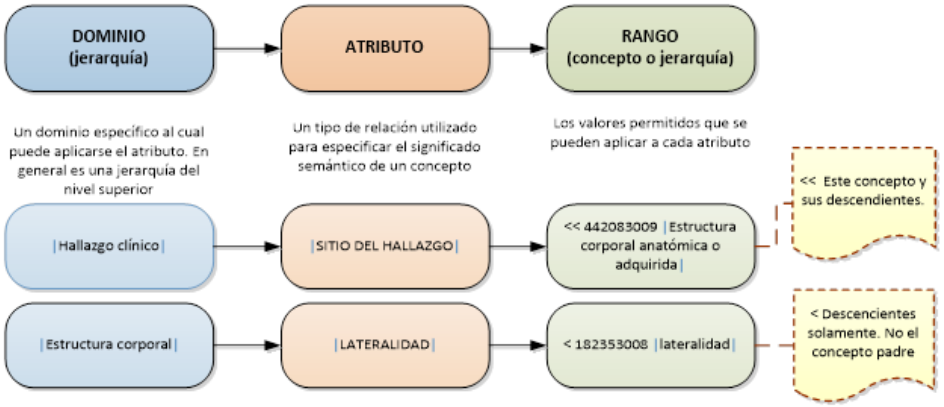

Figura 3.- Modelo Conceptual de SNOMED-CT

Por último, cabe mencionar que, a partir de la post-coordinación, SNOMED-CT permite representar frases clínicas, aunque el concepto exacto no esté definido en la terminología, mediante la construcción de expresiones que contienen dos o más identificadores de concepto.

2.1.2 CIE (Clasificación Internacional de Enfermedades)

La *clasificación* más ampliamente utilizada a nivel nacional e internacional es la **Clasificación Internacional de Enfermedades (CIE),** de la Organización Mundial de la Salud (OMS). Con más de 100 años de evolución, la versión más ampliamente distribuida e implementada es la CIE-10. Se trata de una clasificación mono axial, donde la jerarquía de términos está basada en una raíz común y cada término pertenece únicamente a una clase. Este sistema fue diseñado para informar diagnósticos médicos, aunque las versiones modificadas (CIE-9-CM, CIE-10-CM) también incluyen procedimientos (Figura 4).

Clasificación de Enfermedades y Lesiones

01. (A00-B99) CIERTAS ENFERMEDADES INFECCIOSAS Y PARASITARIAS
02. (C00-D48) TUMORES [NEOPLASIAS]
03. (D50-D89) ENFERMEDADES DE LA SANGRE Y DE LOS ÓRGANOS HEMATOPOYÉTICOS, Y CIERTOS TRASTORNOS QUE AFECTAN EL MECANISMO DE LA INMUNIDAD
04. (E00-E90) ENFERMEDADES ENDOCRINAS, NUTRICIONALES Y METABÓLICAS
05. (F00-F99) TRASTORNOS MENTALES Y DEL COMPORTAMIENTO
06. (G00-G99) ENFERMEDADES DEL SISTEMA NERVIOSO
07. (H00-H59) ENFERMEDADES DEL OJO Y SUS ANEXOS
08. (H60-H95) ENFERMEDADES DEL OÍDO Y DE LA APÓFISIS MASTOIDES
09. (I00-I99) ENFERMEDADES DEL SISTEMA CIRCULATORIO
10. (J00-J99) ENFERMEDADES DEL SISTEMA RESPIRATORIO
11. (K00-K93) ENFERMEDADES DEL SISTEMA DIGESTIVO
12. (L00-L99) ENFERMEDADES DE LA PIEL Y DEL TEJIDO SUBCUTÁNEO
13. (M00-M99) ENFERMEDADES DEL SISTEMA OSTEOMUSCULAR Y DEL TEJIDO CONJUNTIVO
14. (N00-N99) ENFERMEDADES DEL SISTEMA GENITOURINARIO
15. (O00-O99) EMBARAZO, PARTO Y PUERPERIO
16. (P00-P96) CIERTAS AFECCIONES ORIGINADAS EN EL PERÍODO PERINATAL
17. (Q00-Q99) MALFORMACIONES CONGÉNITAS, DEFORMIDADES Y ANOMALÍAS CROMOSÓMICAS
18. (R00-R99) SÍNTOMAS, SIGNOS Y HALLAZGOS ANORMALES CLÍNICOS Y DE LABORATORIO, NO CLASIFICADOS EN OTRA PARTE
19. (S00-T98) TRAUMATISMOS, ENVENENAMIENTOS Y ALGUNAS OTRAS CONSECUENCIAS DE CAUSAS EXTERNAS
20. (V01-Y98) CAUSAS EXTERNAS DE MORBILIDAD Y DE MORTALIDAD
21. (Z00-Z99) FACTORES QUE INFLUYEN EN EL ESTADO DE SALUD Y CONTACTO CON LOS SERVICIOS DE SALUD
22. (U00-U99) CÓDIGOS PARA PROPÓSITOS ESPECIALES

Figura 4.- eCIE10 - Edición electrónica de la CIE-10", 2010

La versión más reciente es la **CIE-11**, aunque aún no se ha desplegado significativamente. Además de agregar nuevos capítulos y modificar algunos antiguos, tiene dos principales cambios con respecto a la CIE-10. Por un lado, ya no es una clasificación estrictamente mono axial, sino que un término puede pertenecer a más de una clase. Además, los códigos principales pueden ser post-coordinados (agregado de códigos de extensión), permitiendo agregar granularidad manteniendo la estandarización (Figura 5).

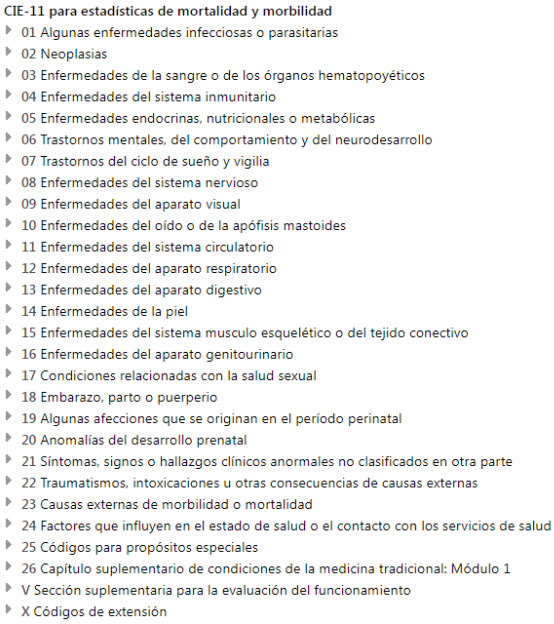

Figura 5.- "CIE-11 para estadísticas de mortalidad y morbilidad", 2021.

Dentro de la clasificación CIE, también podemos encontrar la **Clasificación Internacional de Enfermedades para Oncología (CIE-O)**, que es muy utilizada en registros de tumores para codificar la localización anatómica (topografía) y la histología (morfología) de la neoplasia, generalmente obtenidos del informe de patología. La sección de morfología de CIE-O está incorporada en la jerarquía de morfología de SNOMED CT[1].

2.1.3 UMLS (Sistema Unificado de Lenguaje Médico)

UMLS (*Unified Medical Language System*) es la fuente de conocimiento biomédico más conocida y utilizada en el sector biomédico. Su principal objetivo es asignar un identificador único (CUI) a cada concepto biomédico y relacionarlos formando una estructura de red. Fue desarrollado por la National Library of Medicine (NLM), de los Estados Unidos, y es un sistema que garantiza referencias cruzadas entre más de treinta vocabularios y clasificaciones, incluyendo la CIE, MeSH, CPT, COSTAR, DSM IV, READ 3.1 y SNOMED. Si bien el inglés es el idioma dominante (116 vocabularios de origen), 52 vocabularios del Metathesaurus están en idiomas distintos del inglés (6 en holandés, alemán y español; 5 en francés; 4 en italiano y portugués; 2 en checo, finlandés, húngaro, japonés, coreano, noruego y sueco, y 1 en euskera, croata, danés, hebreo, letón, polaco y ruso). UMLS no reconoce catalán/valenciano, lo que puede presentar un problema para los centros que utilicen este idioma. Aunque existe un estudio de viabilidad[2].

[1] https://apps.who.int/iris/bitstream/handle/10665/96612/9789241548496_eng.pdf

[2]https://plantl.mineco.gob.es/tecnologias-lenguaje/comunicacion-formacion/eventos/eventosinfoday2019/Estudios%20en%20el%20%C3%A1mbito%20del%20Plan%20TL/umls-HSaggion.pdf

NLM no traduce ninguno de estos vocabularios, sino que integra traducciones de vocabularios en inglés cuando están disponibles. Se incluyen los encabezados de temas médicos (MeSH) presentes en el Metathesaurus de 16 idiomas; la Clasificación Internacional de Atención Primaria (ICPC) en 14; el Diccionario Médico de Terminología de Actividades Reguladoras (MedDRA) en 10; y la terminología de reacciones adversas a medicamentos de la OMS en 5. Cabe destacar que SNOMED CT y la terminología de procedimientos médicos actuales (CPT) están presentes en inglés y español, porque estos dos idiomas son los principales idiomas que se usan en los EE. UU. Por último destacar que UMLS proporciona mapeos directos entre todos estos vocabularios.https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MSHSPA/index.html

2.1.3.1 Fuentes de conocimiento de la UMLS

UMLS presenta 3 fuentes de conocimiento (Figura 6): el metatesauro (Metathesaurus), la Red Semántica (Semantic Network) y el léxico especializado y herramientas léxicas (SPECIALIST Lexicon & Lexical Tools).

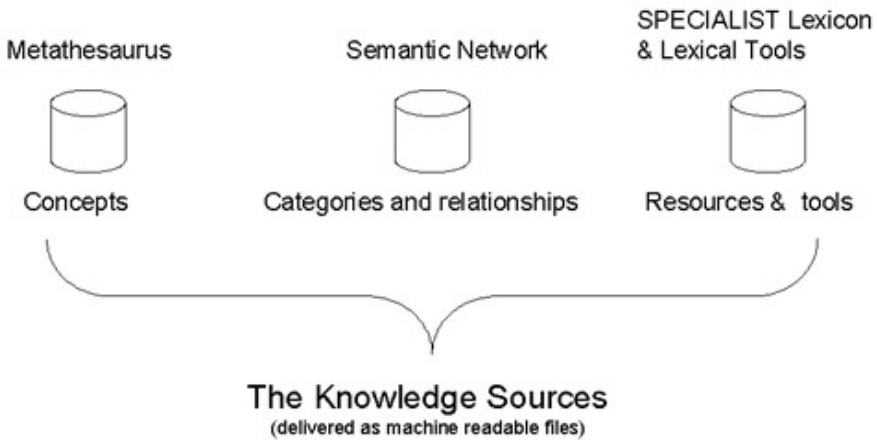

Figura 6.- Fuentes de conocimiento de UMLS[3]

La base de UMLS está formada por un **Metatesauro** organizado por conceptos o significados que enlaza términos sinónimos de más de 200 vocabularios fuente. Además, este vocabulario identifica relaciones útiles entre conceptos y mantiene los significados, los nombres de conceptos y relaciones de cada vocabulario. Esta construido a partir de las versiones electrónicas de diferentes tesaurus, clasificaciones y listas de términos controlados utilizados en el cuidado de pacientes, en la elaboración de estadísticas sobre salud, en el indexado y la catalogación de literatura biomédica y en la investigación clínica. Su propósito es enlazar nombres alternativos y vistas de un mismo concepto, así como identificar relaciones útiles entre diferentes conceptos.

[3] https://www.nlm.nih.gov/research/umls/new_users/online_learning/OVR_001.html

El **léxico especializado**, en lengua inglesa, contiene en su versión actual unos 108.000 informes léxicos y más de 186.000 cadenas de términos. Cada entrada presenta información sintáctica, morfológica y ortográfica. La información léxica incluye la categoría sintáctica, la variación de la inflexión (singular o plural para los sustantivos, conjugación de los verbos, comparativo y superlativo para los adjetivos y adverbios), y posibles patrones de complementación (objetos y otros argumentos que pueden acompañar a los verbos, nombre y adjetivos). El léxico distingue entre once categorías sintácticas: verbos, nombres, adjetivos, adverbios, auxiliares, modales, pronombres, preposiciones, conjunciones y determinantes.

La **red semántica**, presenta 132 tipos semánticos, y garantiza una categorización consistente de todos los conceptos representados en el methasaurus. Los 53 enlaces entre los tipos semánticos establecen la estructura de la red y representan las relaciones más importantes en el dominio biomédico. Los tipos semánticos se clasifican en seis agrupaciones básicas: organismos, estructuras anatómicas, funciones biológicas, productos químicos, eventos, objetos físicos y conceptos o ideas, y permiten la categorización semántica de un amplio abanico de terminología en múltiples dominios de especialidad.

2.1.3.2 Identificadores Únicos de Concepto (CUI)

Un concepto es un significado y un significado puede tener muchos nombres diferentes. Un objetivo clave de la construcción de Metathesaurus es comprender el significado previsto de cada nombre en cada vocabulario de origen y vincular todos los nombres de todos los vocabularios de origen que significan lo mismo (los sinónimos). Los CUI contienen la letra C seguida de siete números. Ej. “C0018681”.

2.1.3.3 Utilizaciones de UMLS

Entre las utilizaciones de UMLS destacamos: Enlace de términos y códigos entre médico, farmacia y compañía de seguros; Coordinar la atención del paciente entre varios departamentos dentro de un hospital; Procesar textos para extraer conceptos, relaciones o conocimientos; Facilitar el mapeo entre terminologías; Desarrollar un sistema de recuperación de información; Extraer terminologías específicas del Metathesaurus; Crear y mantener una terminología local; Desarrollar un servicio de terminología; Investigación de terminologías u ontologías.

2.2 Nivel de Imagen

2.2.1 DICOM

DICOM (*Digital Imaging and Communications in Medicine*) no es sólo un formato de archivo para imágenes médicas, es un estándar que busca garantizar todas las necesidades de un sistema PACs[4] (Pictures Archiving and Communication System) como almacenamiento, transmisión, comunicación e impresión (10). DICOM está universalmente aceptado y proporciona los medios para apoyar intercambio no sólo de imágenes adquiridas, sino también de datos clínicos y diversos tipos de resultados con el objetivo de permitir su intercambio y reutilización. Actualmente se encuentra en su versión 3.0, aunque la versión más ampliamente extendida es la versión 2.0.

De manera general, los ficheros DICOM consisten en dos partes diferenciadas (Figura 7):

- Una *Cabecera* con multitud de campos estandarizados que especifican tantos datos administrativos (datos de paciente o datos del hospital donde se realizó la adquisición), como datos sobre la imagen.
- Cuerpo de la *Imagen*, que puede estar comprimido con distintos estándares.

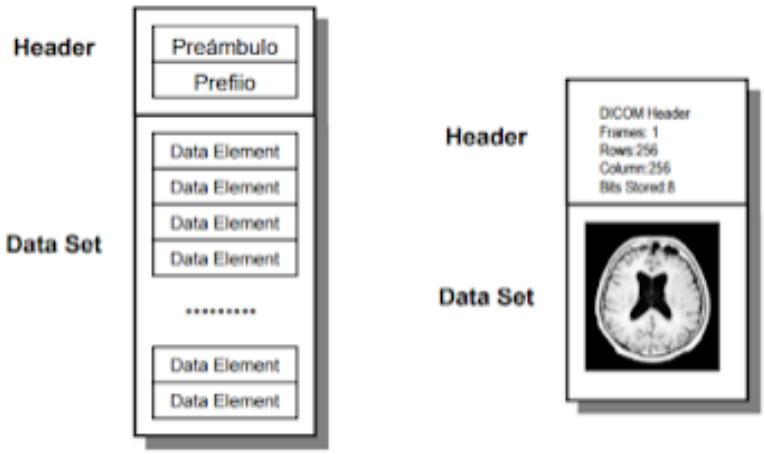

Figura 7.- Estructura de un archivo DICOM (izquierda) y ejemplo de un archivo DICOM (derecha)

En un contexto más técnico, el formato DICOM cuenta con objetos **IOD** (*Information Object Definition*), formados por la imagen y su información asociada (Son una representación lógica de objetos del mundo real) y **DIMSE** (*DICOM Message Service Element*), operaciones que pueden realizarse sobre un objeto. IOD y DICOM forman **SOP** (*Service Object*), la unidad funcional de DICOM.

[4] https://www.postdicom.com/es/services/pacs

Un IOD se compone de **IEs** (*Entidades de información*) (Hay IE de paciente, de estudio, de serie, de equipo, de imagen…) que a su vez se componen de uno o varios módulos que a su vez se contienen varios atributos. Un atributo se define con nombre, etiqueta, tipo y descripción (11)(Figura 8).

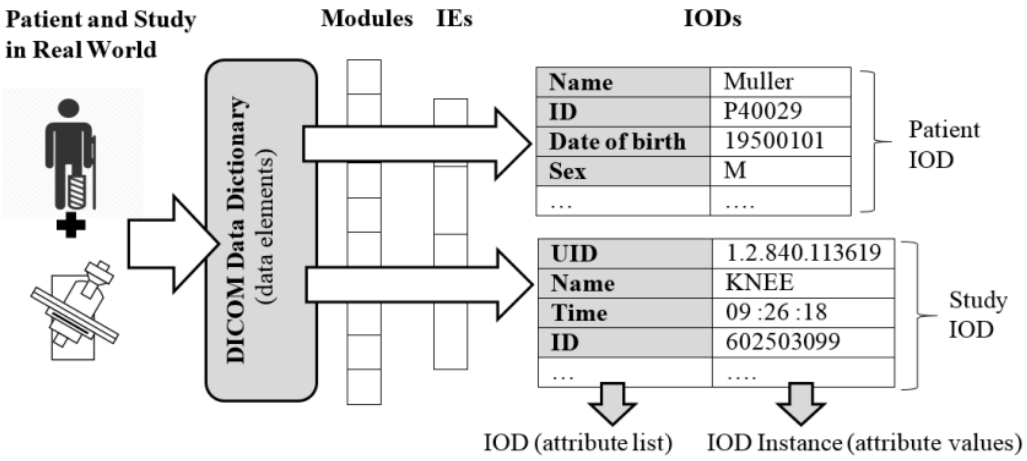

Figura 8.- Transformación a un objeto IOD

2.2.1.1 Estructura de un fichero DICOM

Centrándonos en el punto de vista del implementador, la estructura de un fichero DICOM se puede dividir en las siguientes partes (10) (12) (Figura 7**¡Error! No se encuentra el origen de la referencia.**):

- *Preámbulo:* Es un elemento opcional y tiene un tamaño fijo de 128 bytes, y está pensado para tener un uso definido por la implementación. Por ejemplo, puede contener información sobre el nombre de la aplicación usada para crear el fichero, o puede tener información que permita a aplicaciones acceder directamente a los datos de la imagen almacenada en el fichero.

- *Prefijo:* Es un elemento de 4 bytes que contiene la cadena de caracteres DICM. Esta cadena debe estar codificada siempre con las letras en mayúscula y usando el repertorio de caracteres ISO 8859 GO. Este apartado permite a las implementaciones diferenciar si se trata de un fichero DICOM o no.

- *Cabecera y Meta-cabecera*: Consisten en una serie de campos o atributos con toda la información necesaria sobre la imagen en cuestión, incluyendo la propia imagen. Entre estos campos se encuentran, por ejemplo, datos sobre el paciente (nombre, sexo), sobre el tipo de imagen y muchos más. Así pues, en este apartado está la información necesaria para que se pueda visualizar correctamente la imagen y toda la información asociada a esa imagen que hará que el personal médico pueda interpretar correctamente el estudio.

- *Imagen*: Aunque desde el punto de vista del formato, la imagen es un elemento más de la cabecera.

Además, DICOM establece una jerarquía que permite realizar una clasificación según el contenido de la información por grupos, facilitando la identificación, el acceso a las variables y los parámetros de interés de un mismo archivo (13)(Figura 9Figura 9).

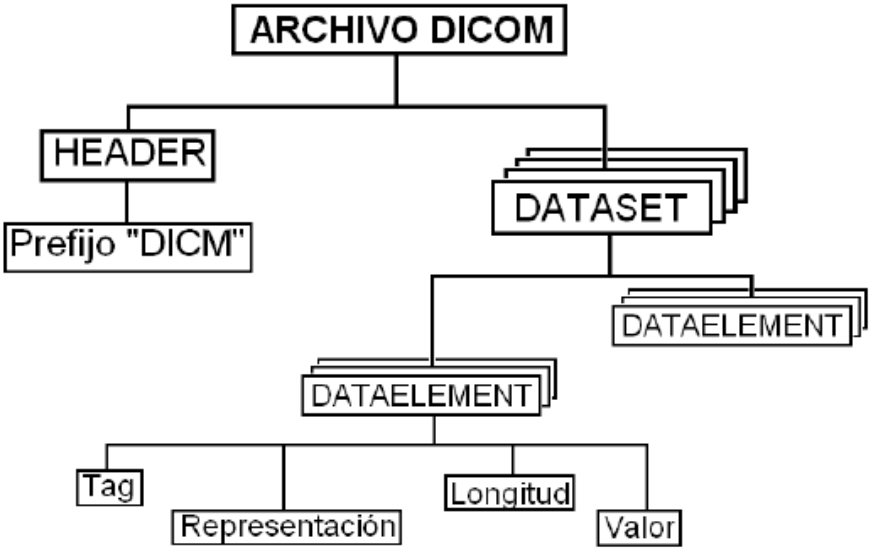

Figura 9.- Jerarquía de un archivo DICOM

2.2.1.2 Elemento de Datos o Atributos

Al conjunto de toda la información codificada sobre un campo se le conoce con el nombre de Elemento de Datos. Así, tanto la cabecera como la meta-cabecera de un fichero DICOM consisten en una sucesión de elementos de datos. Estos elementos de datos están codificados de manera estandarizada y muy detallada según se puede ver en la documentación oficial del estándar DICOM[5]. En el capítulo 6 del estándar podemos observar que existen miles de elementos DICOM, desde atributos más básicos como el nombre del paciente o su fecha de nacimiento, hasta alguno más especializados como el Índice de reconstrucción. Se han definido unos elementos mínimos que deben estar presentes en todos los archivos DICOM para la lectura adecuada de una imagen (13) (Tabla 1). Además, DICOM también permite la creación de atributos privados para favorecer diferentes procesos propios de cada investigación o de cada centro.

[5] https://www.dicomstandard.org/current

Un *Elemento de Datos* está constituido por los siguientes campos (Figura 10):

- *Etiqueta*: Identifica cada Elemento de Datos de forma unívoca. Está constituida por un primer valor hexadecimal de 2 bytes, que corresponde al grupo y un segundo valor, que es el número del elemento. Se suelen representar como un vector de dos dimensiones, Ej.- (0010, 0020) Patient ID
- *Representación del Valor*: Muestra la forma en que se codifica el valor del elemento.
- *Longitud del Valor*: Es el tamaño que ocupa el campo Valor.
- *Valor*: Es el valor del elemento de datos, codificado según el campo "*Representación del Valor*" y con la longitud que se indica en el campo "*Longitud del Valor*".

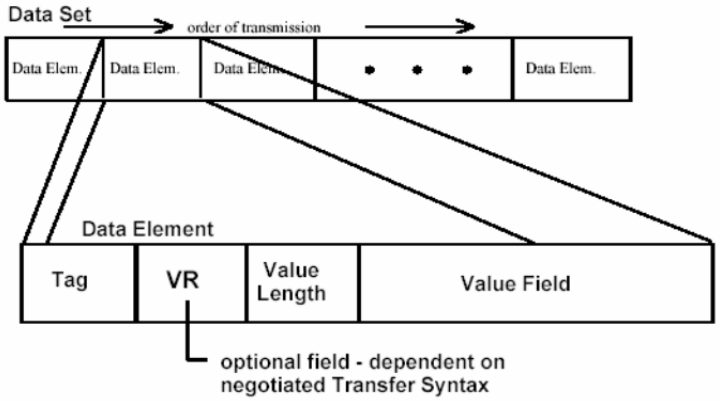

Figura 10.- Campos de un Elemento de Datos DICOM

Tabla 1.- Data Elements necesarios para la lectura adecuada de una imagen

TAG	Descripción	Tipo
(0028,0002)	Samples per Pixel	Int
(0028,0008)	Number of Frames	Int
(0028,0010)	Rows	Int
(0028,0011)	Columns	Int
(0028,0100)	Bits Allocated	Int
(0028,0101)	Bits Stored	Int
(0028,0102)	Hight Bit	Int
(0028,0103)	Pixel Representation	Int
(0028,1050)	Window Center	Int
(0028,1051)	Window Width	Int
(0028,1052)	Rescale Intercept	Int
(0028,1053)	Rescale Slope	Int
(7FE0,0010)	PixelData	Byte[ ] ó Uint16[ ]

2.2.2 Analyze 7.5

Analyze 7.5 es el formato desarrollado por el software Analize[6]. Creado a finales de los años 80 por el grupo "*Biomedical Imaging Resource"* (BIR)[7] en la Clínica Mayo en Rochester (Minnesota, EEUU) y por más de una década se convirtió en el estándar para el post-procesamiento de las imágenes médicas. Actualmente se considera "antiguo", pero se sigue usando en una gran cantidad de aplicaciones hoy en día.

Una imagen con formato Analyze 7.5 consiste en un par de archivos: Un archivo con extensión "*.img"* donde nos encontramos la imagen y otro con extensión *".hdr"* donde se encuentra la cabecera con los metadatos asociados a esa imagen.

Actualmente existe versiones más recientes que Analyze 7.5, pero hemos decidido nombrar esta porque es las más extendida y, además, fue la precursora del estándar NIfTI, mucho más estandarizado en la actualidad.

2.2.3 NifTI

Neuroimaging Informatics Technology Initiative. (NIfTI)[8] surge como una iniciativa de los grupos "*National Institute of Mental Health"* y "*National Institute of Neurological Disorders and Stroke"*, con el objetivo de proporcionar servicios, capacitación e investigación coordinados y específicos para acelerar el desarrollo de los modelos informáticos en el campo de la neuroimagen y más concretamente en el desarrollo e implementación de herramientas para el análisis de la resonancia magnética funcional (fMRI). Con esto se pretendía favorecer la mejora de las herramientas existentes en busca de minimizar la falta de interoperabilidad y maximizar la usabilidad entre la comunidad científica. A día de hoy, se ha convertido en un estándar más allá de la fMRI y es uno de los formatos más utilizados para almacenar y procesar imágenes médicas junto con DICOM.

A través del grupo de trabajo "*Data Format Working Group" (DFWG),* se establecieron las recomendaciones para un nuevo formato de datos común a partir del standard ANALYZE 7.5, NifTI-1, sobre el cual se añadieron nuevas características propias de la imagen médica, entre ellas la capacidad de relacionar las coordenadas de la imagen con la posición espacial del mundo real. En el formato NIfTI, las tres primeras dimensiones se reservan para definir las tres dimensiones espaciales (*x,y,z*), mientras que la cuarta dimensión se reserva para definir los puntos de tiempo (*t*). Las dimensiones restantes, de la quinta a la séptima, son para otros usos. La quinta dimensión, sin embargo, puede seguir teniendo algunos usos predefinidos, como almacenar parámetros de distribución específicos de los vóxeles o mantener datos basados en vectores.

[6] https://analyzedirect.com/

[7] https://www.mayo.edu/research/core-resources/biomedical-imaging-resource-core/overview

[8] https://nifti.nimh.nih.gov/

Aunque este estándar tiene la capacidad de almacenar las adquisiciones en dos ficheros como Analyze 7.5, la práctica más usada es guardarlas en un solo fichero con extensión "*.nii",* donde la imagen y los metadatos de la cabecera están fusionados. Además, este único fichero contiene todo el estudio, a diferencia del formato DICOM, en el que cada imagen corresponde a un fichero (cada secuencia o adquisición tiene entre 100 y 300 ficheros dependiendo del tipo de estudio).

La cabecera de un archivo NIfTI se ha se ha mantenido como la utilizada en el modelo Analyze 7.5 para mantener la compatibilidad con dicho formato (Anexo D). Esta cabecera es mucho menos completa que la utilizada en los ficheros DICOM, y por ello, es práctica habitual a la hora de convertir ficheros DICOM a NIfTI, el archivar un fichero de texto auxiliar con toda la información que se encontraba en la cabecera DICOM. NIfTI también ofrece la posibilidad de extender la cabecera y poder almacenar más atributos, pero hay que ir con cuidado porque no todas las herramientas pueden acceder correctamente a esta extensión de la cabecera.

Existe en la actualidad el formato NIfTI-2, desarrollado en 2011 para gestionar grandes cantidades de datos, conserva, en general, la misma cantidad de información que la anterior, con la diferencia crucial de que permite un número mucho mayor de datos en cada dimensión, lo que permite utilizar la misma estructura general de archivos para almacenar, por ejemplo, datos escalares basados en superficies o grandes matrices de conectividad.

2.2.4 BIDS

Es de interés primordial para todos los científicos que los resultados sean reproducibles. Si bien dicha reproducibilidad se consideró de suma importancia en el campo de la tomografía por emisión de positrones (PET) (14), la evaluación cuantitativa de la reproducibilidad se ha descuidado en gran medida en la comunidad de la Resonancia Magnética funcional (fMRI): como lo describieron Bennett y Miller: "La confiabilidad no es un tema típico de conversación" entre los investigadores de fMRI (15)

Debido a que históricamente los experimentos de neuroimagen producen resultados complejos y heterogéneos dependiendo de las herramientas utilizadas, se realizaron múltiples propuestas para obtener un acuerdo en cuanto a realizar un formato válido y de consenso. En 2016 BIDS se presentó como una primera propuesta de organización de los datos de neuroimagen por parte del Comité de Mejores Prácticas en Análisis e Intercambio de Datos (COBIDAS). COBIDAS[9] es parte de la sociedad líder en neuroimagen, OHBM (Organización para el Mapeo del Cerebro Humano).

Después de diversas reuniones, BIDS se estableció como el estándar de facto en neuroimagen, posicionándose dentro de múltiples plataformas bien conocidas, destacando entre otras la plataforma EBRAINS, creada por Human Brain Project.

[9] www.humanbrainmapping.org/cobidas

BIDS (Brain Imaging Data Structure) (16) es un estándar que permite la organización, descripción y anotación de datos obtenidos durante pruebas de neuroimagen. Este estándar se basa en una estructura definida de archivos y carpetas e incluye, además, documentación de los metadatos DICOM correspondientes en ficheros JSON. La documentación sobre los metadatos correspondientes a las pruebas se organiza en archivos tabulares. Entre otros, BIDS incluye archivos de información sobre el paciente y también sobre la modalidad de imagen que se ha adquirido; esta estructura de datos jerárquica lo convierte en idóneo para organizar colecciones y realizar experimentos con IA. En resumen, BIDS describe una forma simple y fácil para organizar datos de neuroimagen (Figura 11). No obstante, y hasta donde sabemos, el estándar BIDS todavía está restringido a imágenes cerebrales, lo que limita sus ventajas solo a los investigadores de esta área.

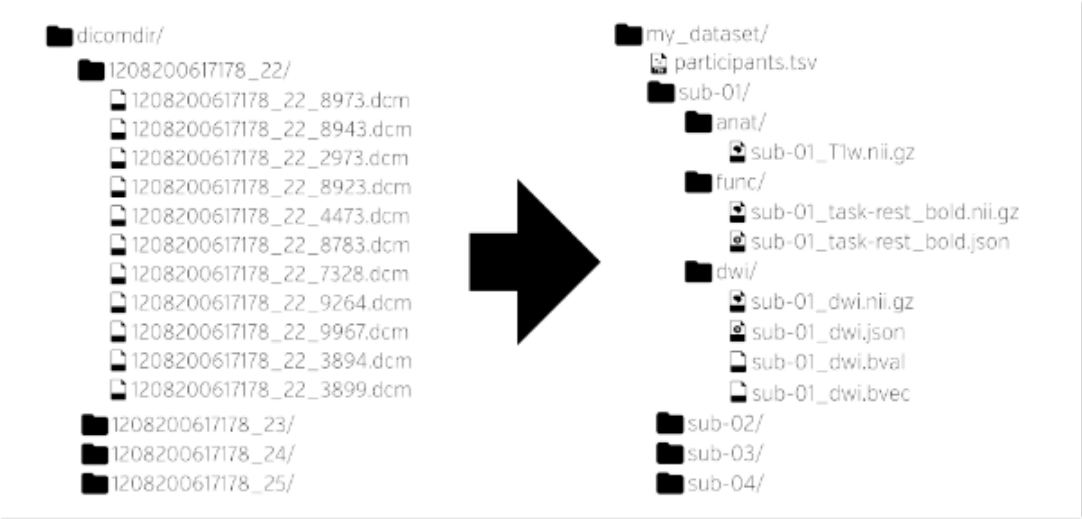

Figura 11.- Estructura de archivos BIDS

2.2.5 MIDS

MIDS (Medical Imaging Data Structure) (17) se crea con el objetivo de extender el estándar BIDS al resto de regiones anatómicas y técnicas de imagen. MIDS se concibe como una extensión de BIDS, concretamente BEP025[10] siguiendo la misma estructura para datos y metadatos, pero adaptándolo a múltiples modalidades de técnicas de imagen médica (resonancia magnética, tomografía computarizada, ecografía/ultrasonido, entre otras).

[10] https://bids.neuroimaging.io/get_involved.html

MIDS añade a la jerarquía de directorios de BIDS un nuevo nivel que describe el tipo de imágenes médicas utilizadas para una sesión en particular. Esta información se organiza en un tercer archivo tabular, que proporciona información referente a las sesiones del paciente, tales como edad en la sesión, procedimientos o diagnósticos.

Por último, destacar que se está desarrollando en diversos proyectos nacionales la integración de OMOP[11] y MIDS. El modelo de datos comunes de OMOP permite el análisis sistemático de bases de datos observacionales dispares. El concepto detrás de este enfoque es transformar los datos contenidos dentro de esas bases de datos en un formato común (modelo de datos), así como una representación común (terminologías, vocabularios, esquemas de codificación), y luego realizar análisis sistemáticos utilizando una biblioteca de rutinas analíticas estándar que tienen sido escrito basado en el formato común. OMOP no se desarrolla en este documento al estar más relacionado con los datos de la historia clínica electrónica, por lo que se puede encontrar una descripción más detallada en el entregable 4.1.

2.2.6 REMBI

REMBI (*Recommended Metadata for Biological Images*) como estándar. Se desarrolló con colaboradores de más de 30 instituciones de todo el mundo con el objetivo de maximizar el potencial de reutilización de los datos de bioimagen (18). Para que los datos de bioimagen desarrollen todo su potencial, es esencial que se archiven sistemáticamente en bases de datos públicas y que vayan acompañados de metadatos normalizados. REMBI propone directrices para dichos metadatos, con el objetivo de atender las necesidades de todas las comunidades de imagen biológica que utilizan la microscopía óptica y electrónica. Los metadatos cubiertos por REMBI se han dividido en ocho categorías, entre las que se incluyen la información del estudio; los componentes del estudio (por ejemplo, el método de obtención de imágenes); la información de la biomuestra (por ejemplo, las especies de las que se obtienen las imágenes y las variables experimentales); los datos del espécimen (por ejemplo, el método de preparación); la adquisición de imágenes (por ejemplo, los parámetros de adquisición); los datos de las imágenes (por ejemplo, el formato y la compresión); la correlación de las imágenes (por ejemplo, las imágenes relacionadas y su relación) y los datos analizados (por ejemplo, el método de análisis y los detalles).

[11] https://ohdsi.github.io/CommonDataModel/index.html

2.3 Nivel de Capa

A la hora de trabajar con imágenes y procesarlas para realizar cualquier procedimiento posterior, se añaden capas de información que nos permiten segmentar ciertas estructuras (órganos o lesiones), etiquetarlas e incluso añadir marcadores o medidas como ángulos o distancias.

El estándar DICOM proporciona una gran variedad de objetos que se pueden usar para almacenar la información derivada del procesamiento de las imágenes. En los siguientes apartados, vamos a describir algunos de estos objetos

2.3.1 DICOM SR

DICOM SR (DICOM Structured Reporting) es un documento estructurado que contiene texto con enlaces a otros datos, como imágenes y coordenadas espaciales o temporales. Su estructura, junto con su amplio uso de información codificada, permite la comprensión semántica de los datos que son esenciales para la HCE (19).

El uso de un código permite que el documento de diagnóstico atraviese las definiciones locales y las barreras del idioma. Además, la codificación y la estructura de SR permiten consultas sofisticadas y operaciones de extracción de datos. También los enlaces a otros datos (p. ej., imágenes) proporcionan una forma de integrar la imagen (es decir, la entrada de interpretación) con el resultado de la interpretación en el mismo documento. (20) (21) (22).

Los objetos DICOM SR pueden utilizarse para texto, contenido codificado, medidas numéricas, referencias espaciales y temporales de ROI, resultados de análisis y referencias a imágenes clave, información de la segmentación.

2.3.2 DICOM SEG

DICOM SEG *(DICOM Segmentation).* Además de poder almacenar las segmentaciones en los objetos DICOM SR, también se pueden almacenar en un objeto DICOM SEG, que define la información gráfica ya sea como objetos rasterizados o de malla de superficie. Representan una clasificación de píxeles ya sea binaria o fraccionada.

2.3.3 DICOM RWVM

DICOM RWVM *(DICOM Real World Value Mapping).* Este objeto proporciona un mecanismo para describir el cálculo que se utilizó (y puede reutilizarse) para crear "valores del mundo real" (como el SUV) a partir de los valores de datos de píxeles almacenados. Un RWVM puede estar incrustado dentro de otro objeto DICOM (como una imagen adquirida o derivada), o puede ser codificado como un objeto independiente, que a su vez puede ser referenciado desde otros objetos, como los SR, o reconocido como relevante a partir de lo común de los identificadores del paciente y del estudio. Estos pueden incluir parámetros de

difusión, valores normalizados de algunas modalidades, volumen sanguíneo cerebral regional (RCBV) o flujo (RCBF), etc.

2.3.4 DICOM RT

DICOM RT (DICOM Radiation Therapy) fue creado para la integración de información de radioterapia para un registro electrónico de pacientes, facilitando así la interoperabilidad de diferentes sistemas de radioterapia, haciendo posible así el intercambio de información de diferentes sistemas. Estos objetos incluyen información sobre el plan de radioterapia, el registro del tratamiento, dosis, etc.

2.3.5 Radiomics Ontology

RO (Radiomics Ontology) es otro enfoque con el propósito de modelar el dominio de las características radiómicas y documentar los algoritmos de segmentación y los filtros de imagen que se han utilizado. Esto se completa con la *Iniciativa de Estandarización de Biomarcadores de Imagen* (IBSI): una nomenclatura de biomarcadores de imagen para el análisis cuantitativo de imágenes de alto rendimiento.

2.3.6 NIfTI

Aunque este estándar ya se ha definido en el apartado anterior de estándares de anotación, cuándo hemos hablado de la anotación a nivel de imagen médica. Cabe resaltar que este NIfTI también se puede utilizar para almacenar segmentaciones. Generalmente, la segmentación se realiza mediante la creación de máscaras, las cuales contienen el valor 1 en aquellos vóxeles pertenecientes a la región de interés, dejando a 0 todos los demás, de modo que la región de interés pueda obtenerse fácilmente multiplicando la imagen original por la máscara. Otra opción menos frecuente cuando se segmenta es la de generar una nueva imagen en la que solo los vóxeles correspondientes a la región de interés, tienen valor diferente a 0, en este caso, tomarían el valor de la imagen original. Para la creación de estas segmentaciones, bien sea a través de máscaras o de imágenes segmentadas, así como para todos los diferentes pasos que puedan existir en el preproceso, postproceso y generación de resultados, es habitual emplear el formato NIfTI.

2.4 Estándares de relación entre la Historia Clínica Electrónica y la Imagen Médica

En los apartados anteriores se ha hecho referencia a los estándares de anotación en imagen médica en función de los diferentes niveles que hemos establecido, pero no se puede pasar por alto la relación que nos encontramos con ciertos estándares tratados en el entregable 4.1. Nos referimos a los estándares terminológicos y de clasificación que son imprescindibles en la anotación de imagen médica para favorecer la interoperabilidad entre centros.

En imagen médica, entre las ontologías más ampliamente aceptadas para realizar anotaciones semánticas nos encontramos con SNOMED-CT. La norma DICOM contempla el uso de SNOMED-CT para normalizar conceptos como los de preparación de la muestra, y algunos autores han propuesto utilizar SNOMED-CT para normalizar la "codificación de las anotaciones" en imágenes microscópicas para facilitar su integración con informes sinópticos o registro de enfermedades (23)(24).

Algunos proyectos internacionales conocidos de anotación de imagen médica, como CaBIG[12], BIMCV-COVID-19[13], y PADCHEST[14] contemplan el uso de UMLS (Unified Medical Language System) (Figura 12), junto a SNOMED-CT y RadLex, para el vocabulario controlado de regiones anatómicas (25).

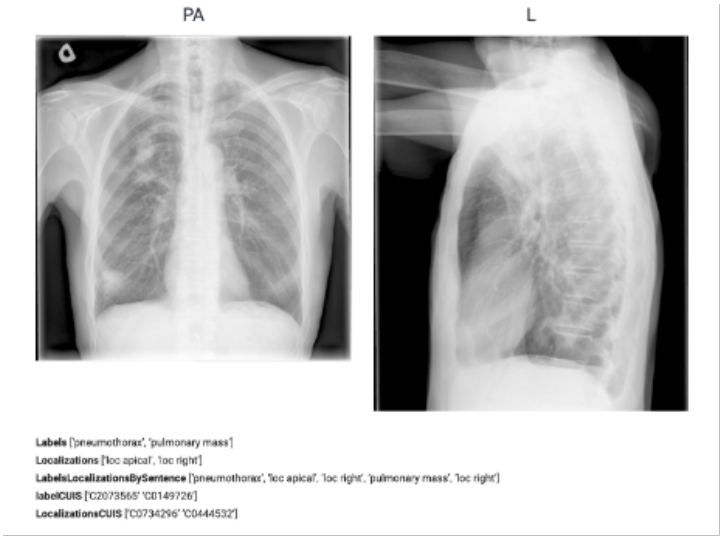

Figura 12.- Anotación en Imagen médica basada en UNLS. Proyecto PADCHEST

2.5 Estándares de Anonimización de la Imagen Médica

2.5.1 De-identificación de meta-información DICOM

El estándar DICOM especifica un conjunto de reglas de de-identificación para ser utilizados, más detalle[15]. En el Anexo C se proporciona el detalle sobre el perfil de de-identificación de las cabeceras DICOM propuesto por la RSNA.

Las reglas se agrupan en un Perfil de Confidencialidad de Aplicación Básica, que elimina toda la Información Sanitaria Protegida (PHI, Protected Health Information), y un conjunto de opciones que prescinden de restricciones en ciertas clases de PHI que se pueden mantener en casos específicos de interés.

[12] https://biospecimens.cancer.gov/relatedinitiatives/overview/caBig.asp

[13] https://b2drop.bsc.es/index.php/s/BIMCV-COVID19-cIter_1_2

[14] https://b2drop.bsc.es/index.php/s/BIMCV-PadChest-FULL

[15] http://dicom.nema.org/medical/dicom/current/output/html/part15.html#chapter_E.

Para evitar comprometer la privacidad, es crucial implementar módulos o scripts seguros y robustos que desarrollen la infraestructura de protección de datos personales desde el diseño. Dentro del estándar DICOM, se definen los 10 perfiles listados a continuación:

- Perfil básico
- Mantener segura la información privada
- Mantener el Unified Identifier DICOM (UIDs[16])
- Mantener la identidad del aparato
- Mantener las características de los pacientes
- Mantener las fechas completas
- Mantener las fechas modificadas completas
- Limpiar descriptores
- Limpiar el contenido estructurado
- Limpiar gráficos

A continuación, se describen los perfiles de de-identificación más complejos. El resto de perfiles se autodescriben por su título.

2.5.1.1 Perfil básico

Este es el perfil más estricto y elimina toda la información relativa a:

- La identidad además de las características identificativas y demográficas del paciente
- La identidad de los autores, personas responsables o familiares
- La identidad de todo el personal involucrado en ordenar o llevar a término el método
- La información (no relativa al paciente) que pueda ser usada para identificar los archivos originales no anonimizados (p. ej. UID, fecha y marca de tiempo).
- Los atributos privados (que no son parte del estándar). Algunos fabricantes mantienen información esencial para la imagen, p. ej. gradientes usados en Imágenes de Tensor de Difusión (ITD)

Este perfil es obligatorio y siempre está activo. La activación de otros perfiles modifica el perfil básico de anonimización.

[16] Este es un número único para identificar, por ejemplo, imágenes, informes, etc. Algunos UID comunes que se usan en imágenes son: UID de clase de implementación: identifica un tipo de dispositivo particular, intercambiado durante la negociación de la Asociación DICOM.

2.5.1.2 Mantener segura la información privada

Si se especifica esta opción, junto con un perfil, se retienen los atributos privados que se sabe que son seguros y que no permiten la identificación. Al preservar estas etiquetas (supuestamente seguras), el sistema debe despejarlas, es decir, ya sea eliminando la información o reemplazándolas con valores con significado similar (pero que no permitan la identificación) y consistentes con su Valor de Representación (VR).

2.5.2 Protocolo adoptado

La RSNA ha trabajado de la mano de expertos DICOM para desarrollar una Guía y un Protocolo de De-identificación de Datos basada en estándares y buenas prácticas reconocidos. Dentro del ámbito de metainformación es necesario asegurar que los procedimientos de de-identificación sigan este protocolo. (Ver anexo C). El procedimiento requiere que se cree un identificador único (seudónimo) para cada sujeto.

2.5.3 De-identificación de información gráfica

La información de identificación puede quedar almacenada en capas y anotaciones gráficas que luego se superponen a la imagen. La anonimización de imágenes gráficas debe validarse muchas veces mediante inspección visual; es un trabajo arduo y tedioso, pero es crucial porque nos ayuda a ser tan rigurosos como exige la legislación vigente.

Las imágenes médicas son un campo amplio que incluye todas las técnicas y modalidades de imágenes médicas existentes. Un método genérico de anonimización de imágenes médicas no puede llevarse a cabo de forma sistemática, dado que cada modalidad tiene sus propias características distintivas. Sin embargo, se pueden realizar aproximaciones. Los dos principales desafíos para abordar la de-identificación de las imágenes son:

- **Información facial**: Puede identificar a un paciente en imágenes de resonancia magnética (MRI) de alta calidad.
- **Anotaciones de texto superpuestas a la imagen que contienen información identificativa**: Estos textos se pueden encontrar en imágenes de rayos X y otras modalidades como la tomografía por emisión de positrones.

2.5.3.1 De-identificación de información facial

La tecnología asociada a la anonimización dentro de la propia imagen es muy diversa y depende del contexto de actuación y de la modalidad de imagen médica a la que se aplica.

En la resonancia magnética de alta calidad, las características faciales de un paciente se pueden reconstruir y utilizar para el reconocimiento facial. Por lo tanto, los píxeles que contengan esta información deben borrarse (Figura 13), en particular aquellos que contengan información de la nariz y rasgos faciales como la boca, la cuenca del ojo, las cejas y la barbilla.

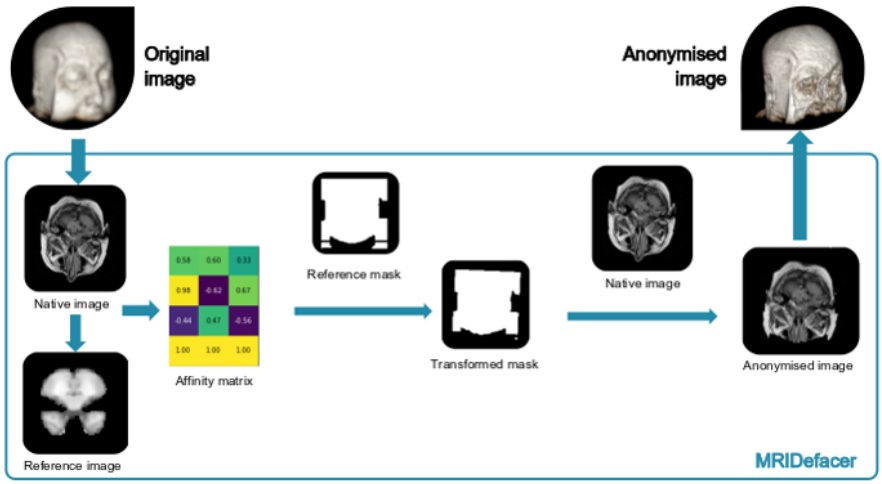

Figura 13.- Esquema conceptual de anonimización facial

2.5.3.2 De-identificación de anotaciones de texto superpuestas a la imagen

Los textos se incrustan en la imagen fuera de las capas especificadas por el estándar para este fin. Estos textos suelen contener información relevante para el médico, pero no es extraño encontrar datos identificativos del paciente, como la edad o los apellidos.

Es común encontrar esta información en las regiones superior e inferior de una imagen, que normalmente no contiene datos relevantes para la patología. Estas regiones están formadas por píxeles que contienen metadatos grabados, pero no partes del escaneo médico. En estos casos, eliminar las regiones superior e inferior de una imagen es más que suficiente para garantizar la de-identificación del texto incrustado en la imagen (Figura 14). En este caso se recomienda aplicar máscaras como las que se muestran en la Figura 15, teniendo en cuenta que la máscara a aplicar debe estar diseñada específicamente para cada imagen en particular.

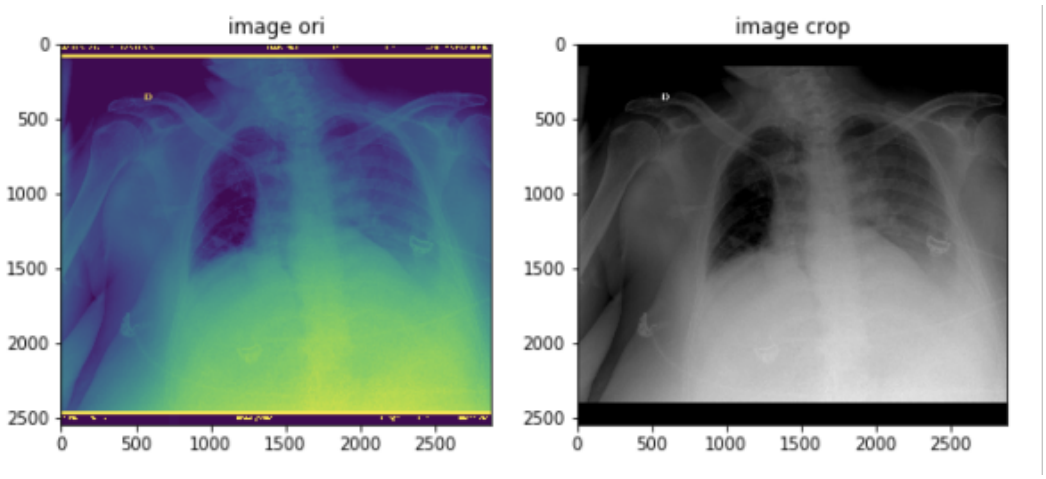

Figura 14.- Eliminación del texto incrustado en la imagen

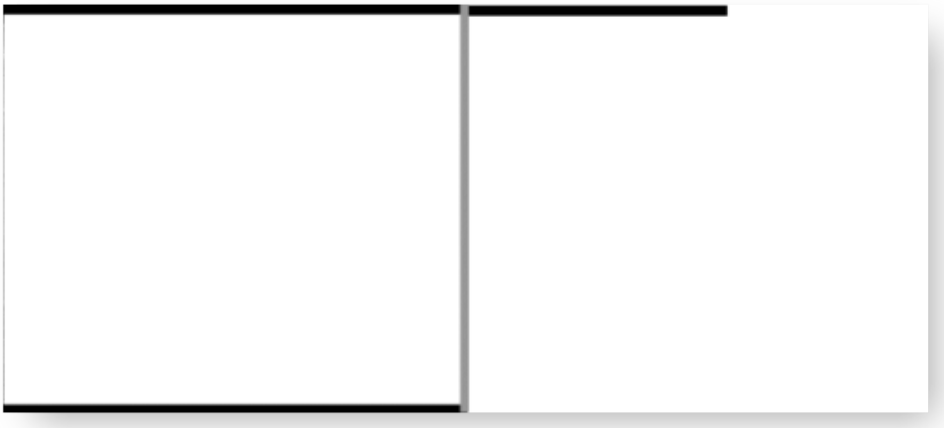

Figura 15.- Máscara utilizada en la figura 10

El primer paso del proceso es, por tanto, el etiquetado de las imágenes. Esto generará una lista de imágenes que incluyen información de identificación y dónde se ubica espacialmente esta información en la imagen. El segundo paso del proceso es la eliminación automática de los datos confidenciales mediante la aplicación de máscaras a las imágenes según el etiquetado.

En casos de mayor complejidad, la imagen es descartada y tiene que ser procesada de forma diferente. En casos más complejos, el texto se ubicará en regiones donde el recorte implica perder parte del escaneo médico. En estas situaciones, puede ser necesario recurrir a redes neuronales que puedan detectar texto en una escena. Es el caso de EAST (Efficient and Accurate Scene Text Detector, arXiv:1704.03155), que es una red neuronal ya implementada y diseñada para esta tarea.

En términos generales, para las anotaciones de texto incrustadas en una imagen, la variabilidad de estructuras es tan grande como la variedad de modalidades de imágenes médicas, lo que hace imposible establecer un método generalizado de de-identificación aplicable a cada caso.

2.5.4 De-identificación en informes radiológicos

Actualmente, los métodos de de-identificación empleados en informes radiológicos se basan en inteligencia artificial y, concretamente, en técnicas de procesamiento de lenguaje natural (NLP). Sin embargo, como ya se ha introducido en el apartado 10.2 del entregable 4.1, el principal inconveniente de NLP es que requiere un corpus relativamente grande de textos anotados, entre 500 y 1000 textos, para entrenar las redes neuronales. Este primer paso es esencial y debe realizarse con cuidado y minuciosidad. La calidad de la anotación tiene un impacto directo en la precisión y sensibilidad de los modelos entrenados y, por lo tanto, tiene un efecto en el resultado de la de-identificación. Es necesario inspeccionar los textos y definir las categorías a anotar, así como acordar qué reglas y metodología se van a aplicar a la anotación entre todos los anotadores involucrados.

Un buen criterio para definir categorías son las categorías de PHI definidas por HIPAA[17]. No obstante, es recomendable evaluar la adecuación de estas categorías a los textos. El conocimiento de la estructura de los informes radiológicos y qué información confidencial contiene, combinado con la lista de PHI, conducirá a la definición de categorías más adecuadas y mejorará el resultado de la de-identificación. Las categorías definidas son de amplio alcance, y podrían aplicarse a otros casos con pocas o ninguna modificación, ver Figura 16 basada en la metodología DISMED (26).

- **Nombres**: Esta categoría incluye tanto nombres como apellidos. Si los informes radiológicos están muy estructurados, puede ser útil diferenciar los nombres de los pacientes de los del personal médico.
- **Ubicaciones**: ciudades, tanto dentro como fuera de las direcciones.
- **Direcciones**: calles, números y códigos postales. Si los informes radiológicos están muy estructurados y bien formateados, el nombre de la ciudad se puede incluir en esta categoría en lugar de la ubicación de la categoría.
- **Números**: números o cadenas alfanuméricas que pueden identificar a alguien, incluidas firmas digitales, números de teléfono, números de registros médicos, números de licencia médica, números de seguro social y números de licencia médica, entre otros.
- **Fechas**: esta categoría incluye fechas en cualquier formato. La lista de PHI incluye todos los elementos de fechas (incluido el año) que indican la edad del paciente y todos los elementos de fechas (excepto el año) para cualquier otra fecha directamente relacionada con una persona. Para ser aún más restrictivo, es recomendable incluir el año en todos los casos. No obstante, en caso de que se quieran conservar años (salvo los indicativos de edad), esta categoría deberá dividirse en dos, incluyendo una de ellas las fechas sin los años.
- **Instituciones**: nombres de hospitales, centros de salud o cualquier otra institución que pueda indicar la ubicación de alguien. Aunque esta categoría no está incluida en la lista de PHI por HIPAA, las instalaciones hospitalarias y los centros de atención médica señalan la ubicación geográfica de la residencia del paciente. Eliminar estas palabras reduce el riesgo de fuga de información.

Si bien estas bases de datos se pueden aplicar para textos en español, para cualquier otro idioma es necesario recuperar bases de datos similares. Debido a la Directiva Europea (UE) 2019/1024 sobre datos abiertos y reutilización de la información del sector público, esta tarea no debería ser un problema en ningún país europeo.

Una vez que las bases de datos requeridas están disponibles y las entidades nombradas se reconocen en el texto para de-identificarlas, se puede aplicar el paso de aleatorización. Consiste en la sustitución de la palabra sensible por una nueva palabra extraída de la base de datos correspondiente a la misma categoría de la palabra detectada. Por lo tanto, un nombre sería sustituido por una nueva palabra de la base de datos de nombres. En el caso

[17] https://www.hhs.gov/hipaa/for-professionals/index.html

de los municipios, los nombres y apellidos, la población y el número de personas registradas con un nombre específico pueden utilizarse para ponderar la selección aleatoria, es decir, disminuyendo la probabilidad de sustituir un nombre por otros menos comunes en la población. El resultado de aplicar estas ponderaciones es un informe de aspecto más natural, lo que reduce la probabilidad de fuga de información confidencial. También es recomendable valorar que la palabra o número original no sea el mismo que el seleccionado al azar.

En caso de que se quiera aplicar la aleatorización de entidades nombradas, el resultado será el archivo de texto original con toda la información sensible detectada sustituida por palabras sintéticas elegidas aleatoriamente de las bases de datos. Por otro lado, si se prefiere la eliminación de palabras en lugar de la sustitución, todas las palabras sensibles detectadas por el modelo entrenado pueden simplemente borrarse.

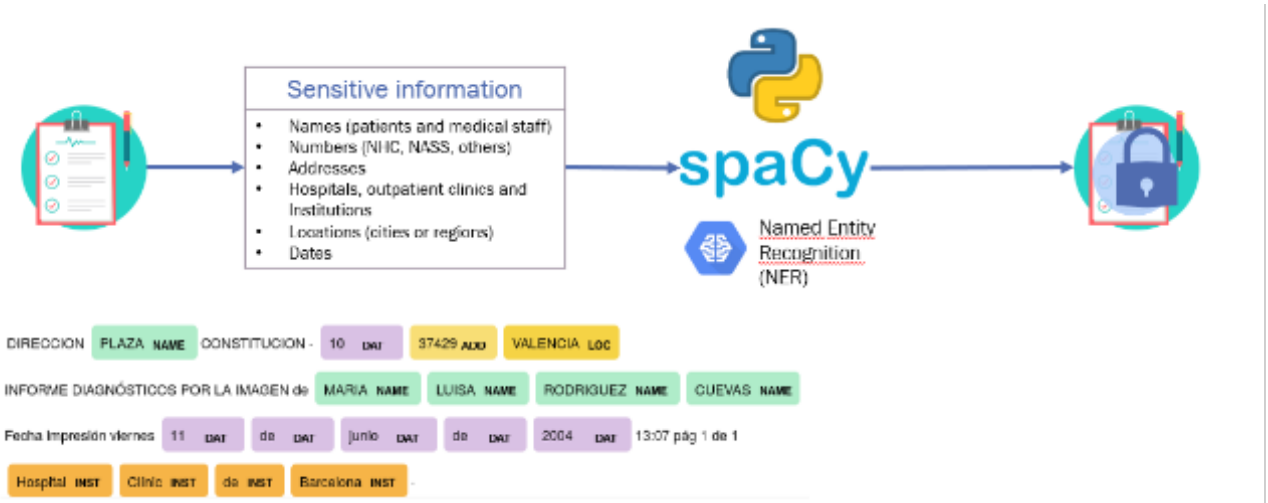

Figura 16.- De-identificaión de informes

3 Repositorios de imagen médica

Desde hace unos años y gracias en parte a la aparición de la Inteligencia Artificial (IA) y los modelos de Machine Learning y Deep Learning, las imágenes médicas se han convertido en un recurso muy apreciado y buscado, sin embargo, no es fácil encontrar imágenes de calidad estables en el tiempo y que se puedan acceder en abierto sin necesidad de pago.

Para realizar modelos de inteligencia artificial se requiere una gran cantidad de información, que depende de muchos factores, como la complejidad del modelo que se intenta construir, el rendimiento que se pretende obtener o, a veces, incluso puede deberse al tiempo del que se dispone.

En este documento se definen las características más importantes y el contenido que debería de tener un repositorio de imagen médica para trabajar adecuadamente con modelos de IA. Propondremos una estructura para albergar la información de los repositorios, y será tarea del entregable 4.5 seleccionar repositorios de imagen médica relevante y compararlos en base a lo que vamos a definir en este apartado.

3.1 Características y Contenido

En repositorios de imagen médica se debe establecer una jerarquía de la información que identifique la población o grupo de estudio, las modalidades de imagen que contiene y las anotaciones sobre la imagen médica.

A día de hoy no existe un consenso claro sobre cómo debería de estar estructurado un repositorio de imagen médica, qué información debería de tener ni de cómo debería de ser su acceso e intercambio de información con otros biobancos. Así pues, podemos encontrarnos diferentes tipos de repositorios de imagen médica en función de sus diferentes características. Según la Sociedad Europea de Radiología (7), nos podemos encontrar tres escenarios posibles:

- **Repositorio para investigación clínica**: En este escenario, los repositorios de imágenes se conciben como una infraestructura para archivar, compartir y difundir (para uso secundario) los datos de imágenes que se utilizaron originalmente en el contexto de proyectos de investigación clínica, como los ensayos clínicos. En este caso, las imágenes pueden haber sido sometidas a algún tipo de procesamiento para extraer uno o más biomarcadores de imagen relevantes para la cuestión científica que motivó el estudio. Es natural almacenar dichos biomarcadores, así como los metadatos de procedencia relacionados, en la base de datos del biobanco de imágenes.

- **Repositorio para una enfermedad específica**: En este escenario, los repositorios de imágenes se contemplan como recursos para recibir, archivar, compartir y difundir imágenes en ámbitos clínicos específicos, por ejemplo, la enfermedad de Alzheimer o la esclerosis múltiple. Dichos sistemas tienen como objetivo recopilar las imágenes clínicas de pacientes con una determinada patología a gran escala (por ejemplo, a escala nacional). Suelen abordar protocolos específicos comunes de adquisición de imágenes y garantizan el control de calidad en el momento de la importación de las imágenes para asegurar su uso óptimo. Estos biobancos de imágenes también pueden basarse en una iniciativa de cribado regional o nacional que recoja datos de imágenes de un grupo de personas con características específicas, como hábitos, enfermedades previas o un perfil genético. Los cribados de cáncer de mama, cáncer de pulmón y cáncer de colon son ejemplos de cribados basados en imágenes con posterior almacenamiento de las mismas en un biobanco de imágenes
- **Repositorios con datos de población general**: En este escenario, los repositorios de imágenes se componen de colecciones de datos obtenidos de la población asintomática. Los datos pueden ser longitudinales o transversales, y estos repositorios cuentan con la mayor cantidad de datos recogidos del mayor número de sujetos posible. La configuración técnica de este escenario es similar a la del segundo escenario, y la recogida de datos se realizan usualmente con fines epidemiológicos

3.1.1 Identificación del repositorio

A la hora de trabajar con repositorios de imagen médica, es de vital importancia poder identificarlos de manera inequívoca si queremos asegurar la reproducibilidad y autoría de las investigaciones. Para conseguir tal propósito, contamos con el **DOI** (Identificador de Objeto Digital - *Digital Object Identifier*). DOI es un enlace permanente en forma de código alfanumérico que identifica de forma única un contenido electrónico. Esta pieza específica de contenido intelectual puede ser un artículo científico, una imagen, un libro, una canción u otro, siempre que se trate de un objeto en el ambiente digital, inclusive una colección de datos de imagen médica.

Una forma común de emplear el sistema DOI es dar a las publicaciones científicas un número específico que cualquiera puede utilizar para localizar a través de la Red el citado contenido digital. A diferencia del sistema URL, usado en las páginas web, el sistema DOI no cambia con el paso del tiempo, aunque el contenido digital sea reubicado en una dirección distinta puesto que lleva la información incorporada en forma de metadatos.

Debido al arduo trabajo que implica la curación de cohorte de datos de imagen biomédica, consideramos necesaria la asignación de un data set DOI a cada una de las colecciones que se compartan. Se considera de interés la creación de una agencia de registro estatal

que asuma estas competencias. De momento solo se puede optar por utilizar algunas agencias internacionales para tal propósito, por ejemplo, IEEE[18].

Otro punto de vital importancia a la hora de identificar el repositorio es contar con un control de versiones del mismo. El control de versiones permite la gestión de los cambios realizados en el repositorio, y nos identifica el conjunto de datos concreto con el que hemos realizado nuestra investigación para asegurar así la reproducibilidad de nuestro trabajo.

3.1.2 Arquitecturas de los repositorios

A la hora de almacenar los datos en un repositorio podemos encontrarnos con repositorios centralizados, donde los datos se almacenan en un mismo lugar o repositorios descentralizados (o federados), donde los datos se almacenan localmente en cada centro.

3.1.2.1 Repositorios Centralizados

En estos repositorios, los datos están almacenados en un lugar único, lo que da el beneficio de una mayor eficiencia a la hora de trabajar con los datos. Además, constituye una integración de datos de diferentes procedencias, lo que permite una homogenización de los mismos en el caso de que no se hayan adoptado estándares en los diferentes lugares de origen. La parte negativa de esta aproximación es la transformación previa que se realiza para homogeneizar la información y para poder cumplir con las regulaciones pertinentes en cuestión de protección de datos que apliquen en cada caso, y para el correcto almacenamiento de la información (Figura 17).

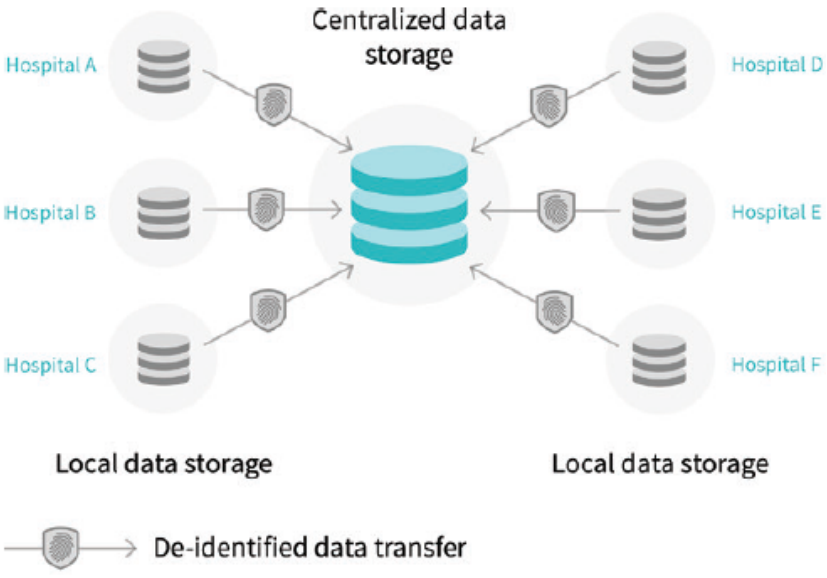

Figura 17.- Modelo de repositorio centralizado (27)

[18] https://ieee-dataport.org/open-access/bimcv-covid-19-large-annotated-dataset-rx-and-ct-images-covid-19-patients

3.1.2.2 Repositorios Descentralizados (Federados)

Un repositorio de imagen médica descentralizada o federada (Figura 18), es un sistema en el cual varias bases de datos trabajan conjuntamente entre sí como una sola, pero mantienen su propia autonomía (28). No existe una única localización donde todas las imágenes se almacena juntas como en el anterior caso. En esta arquitectura, cada centro actuaría como un nodo funcional independiente donde se almacenarían todas las imágenes. Cuando se hace una consulta a un repositorio federado, el sistema busca que nodo contiene los datos que se le solicitan y manda una solicitud a cada uno de ellos.

La ventaja de las arquitecturas federadas frente a las centralizadas es que las fuentes de datos son controladas localmente por sus propietarios y las normas y restricciones políticas pueden aplicarse adecuadamente sin que los datos salgan de las instalaciones locales, lo que en muchos casos exigen los hospitales y los propietarios de los datos.

Sin embargo, las arquitecturas federadas complican el acceso y la recuperación de los datos, ya que éstos no están disponibles en un punto central y deben consultarse de forma distribuida. Además, esto requiere que los nodos distribuidos tengan una potencia de procesamiento y almacenamiento eficiente, ya que los datos deben ser almacenados y servidos localmente. Otro punto a tener en cuenta es que, al actuar cada nodo de manera independiente, no existe un punto común donde se pueda armonizar las imágenes, por ello, otro posible problema que nos podemos encontrar es la heterogeneidad de las imágenes a la hora de trabajar.

La Sociedad Europea de Radiología creó en 2014 un grupo de trabajo (ESR WG on Imaging Biobank) (7) para monitorizar los diferentes biobancos de imagen médica en Europa y promover este tipo de arquitectura.

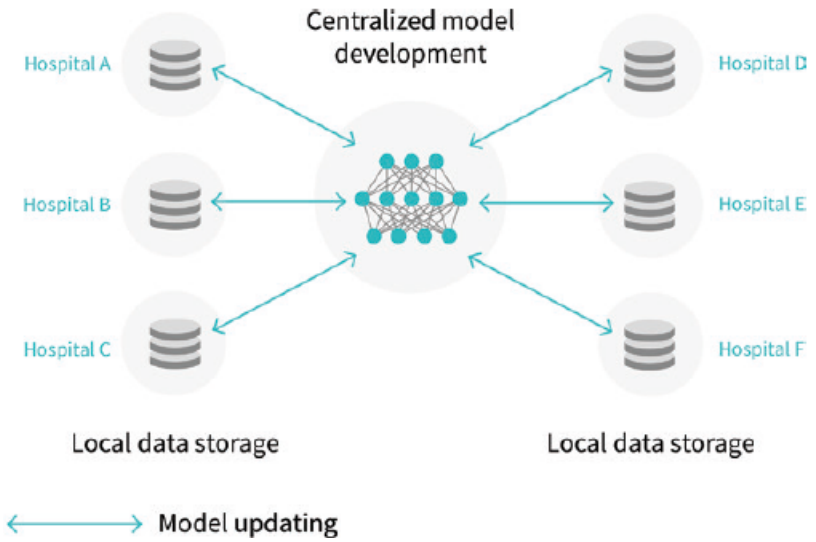

Figura 18.- Modelo de repositorio descentralizado (federado) (27)

3.1.2.3 Repositorios Mixtos

Queremos hacer mención también a repositorios que usan las dos aproximaciones anteriores. Hemos de tener en cuenta que algunos procesamientos de la imagen médica requieren de una gran capacidad de computación que no se encuentra disponible en los centros, y es por ello que deciden usar este enfoque mixto, donde los datos se almacena de manera local, pero para determinados usos específicos que requieran de esa gran capacidad de computación, se han de enviar a una plataforma centralizada que si disponga de esa capacidad.

3.1.3 Descripción de la población de estudio

Los repositorios de imagen médica contienen cierta información clínica de los pacientes, pero suele ser una información limitada y centrada básicamente en la descripción de la población para la creación de cohortes de estudio (la enfermedad estudiada, el sexo del paciente, franja de edad o alguna característica relevante concreta). Esta información suele ser insuficiente para investigaciones más avanzadas, como en la medicina de precisión o en la creación de modelos de IA, donde la integración de los datos clínicos obtenidos de la historia clínica del paciente, junto con la información extraída de la imagen médica, se vuelve imprescindible.

Uno de los grandes retos a día de hoy, es conectar de forma fiable los biobancos de imágenes con los biobancos de tejidos para explorar posibles biomarcadores de imágenes y proporcionar acceso a fenotipos (29). De hecho, la integración de datos moleculares y de imagen es necesaria para un enfoque radio-genómico de los pacientes en un entorno de medicina personalizada (30). La vinculación e integración de datos de imagen existentes, así como la conexión entre los biobancos de imágenes y los biobancos tradicionales es el factor clave para desarrollar y validar nuevos biomarcadores de imágenes, así como para mejorar la comprensión general de su de su importancia biológica (31).

Para ayudar a realizar esta integración, se hace necesario el uso de estándares que faciliten la integración de los datos clínicos y su posterior explotación. En los diferentes modelos de datos común usados en biobancos, como MIABIS (32), se puede almacenar información básica que nos permite realizar una selección de cohortes, pero esta información es insuficiente en otros escenarios, como en el entrenamiento de ciertos modelos de IA, donde puede ser relevante saber los tratamientos previos del paciente, sus comorbilidades o sus antecedentes familiares.

Existen algunas propuestas de modelos como la que realiza el proyecto europeo PRIMAGE[19], donde generan un modelo teórico híbrido entre MIABIS y DICOM. Su objetivo es desarrollar un modelo de metadatos que permita describir los datos de imagen y facilitar la interoperabilidad entre los biomarcadores de imagen y los biobancos tradicionales. La estrategia se basa en mantener el núcleo fundamental del modelo de datos MIABIS como punto de partida, y se crea una nueva subcategoría que se amplía en dos partes

[19] https://www.primageproject.eu/

principales. Por un lado, está la parte relativa a los metadatos de imagen más específicos para el tipo concreto de datos adquiridos y basados en las etiquetas DICOM. Por otro lado, hay una parte dependiente del proceso de análisis de imágenes realizado para extraer biomarcadores de imagen y datos radiómicos (33). Actualmente, el equipo responsable del modelo está trabajando en la integración de ese modelo híbrido con los datos clínicos del paciente en formato OMOP[20].

De las variables clínicas más importantes a tener en cuenta, destacan las relacionadas con el paciente, como la edad, estado funcional, sexo; variables relacionadas con la enfermedad, como los antecedentes personales y familiares, enfermedad actual, comorbilidades, tratamientos farmacológicos y físicos; variables de alteración morfológica, como las variables topográficas, morfológicas (medidas, diagnóstico de sospecha, tipo histológico, necrosis, calcificación), cuantificación biomarcadores; variables de localización específica, como la relación entre compartimentos (estroma/tumor), detectar alteraciones funcionales (mitosis).

3.1.4 Acceso a los datos del repositorio

El acceso a los datos de los repositorios es diferente en función de la gobernanza que se haya definido previamente en ellos. Así pues, nos encontramos con diferentes escenarios, como tener un acceso abierto a los datos sin necesidad de realizar ningún tipo de pago ni aprobación previa, puede que necesitemos ser aprobados por un comité de acceso antes de proceder a descargar las imágenes o también es posible que necesitemos realizar algún tipo de aportación para poder acceder a los datos del repositorio.

3.1.5 Descripción del repositorio

A la hora de conectar múltiples repositorios de imágenes, nos encontramos con una gran variabilidad de información debido a la variedad de métodos y secuencias de adquisición de las mismas. Es por ello imprescindible almacenar los parámetros de adquisición y toda la información que podamos considerar importante junto con las imágenes (7).

Se consideran parámetros relevantes a la hora de describir un repositorio de imagen médica los relacionados con: modalidad del estudio; protocolo de adquisición; área analizada; enfermedad analizada; fecha creación/actualización del repositorio; número de casos; cantidad de estudios, series e imágenes disponibles; tamaño del repositorio; existencia de patrón de referencia; existencia de imágenes anotadas; realización de análisis de imagen (como biomarcadores o máscaras de segmentación); conocer el número de centros que han aportado casos al repositorio; saber si disponemos de imágenes de seguimiento de los casos; ver si el repositorio cuenta con capacidad de computación; si existe una armonización previa de las imágenes.

[20] https://www.ohdsi.org/data-standardization/the-common-data-model/

3.1.5.1 Anotaciones

Los repositorios de imagen médica pueden limitarse estrictamente a los estudios con la información generada durante la adquisición de la imagen, pero existe un valor añadido muy importante que encontramos en los repositorios de imagen médica, y es la posibilidad de que tengan ya incorporadas ciertas anotaciones relacionadas con el tipo de estudio que estamos buscando, como por ejemplo segmentaciones, anotaciones semánticas, marcadores geométricos o biomarcadores.

Es tan importante la existencia de estas anotaciones, como la necesidad de que haya una calidad y estandarización en las mismas. Errores en este tipo de información pueden provocar problemas muy serios en nuestros experimentos (34).

Para minimizar este problema, se propuso en 2019 el uso de una escala que integra los principios FAIR y nos informa de lo preparado que puede estar nuestro repositorio para realizar tareas de Machine Learning, la escala *MIDaR* (35) (Figura 19).

En esta escala nos encontramos con distintos niveles:

- *Nivel A*: Nuestro repositorio se encuentra estructurado, completamente anotado, con ruido mínimo y contextualizado para tareas específicas de Machine Learning.
- *Nivel B*: La cantidad y calidad de los conjuntos de datos del repositorio están totalmente contabilizadas, y se han resuelto los errores a gran escala en la estructura y el formato de los datos. Los datos que se van a usar están separados de los datos que con mala calidad.
- *Nivel C*: Los datos se han anonimizado y son accesibles. Sin embargo, los datos siguen estando sujetos a errores, omisiones, ruido y artefactos, que afectan a la calidad de las imágenes y los metadatos; Un *Nivel D*, donde los datos están presentes en el sistema y son inaccesibles para los investigadores. La cantidad o calidad exacta de los datos es desconocida o difícil de estimar fácilmente.

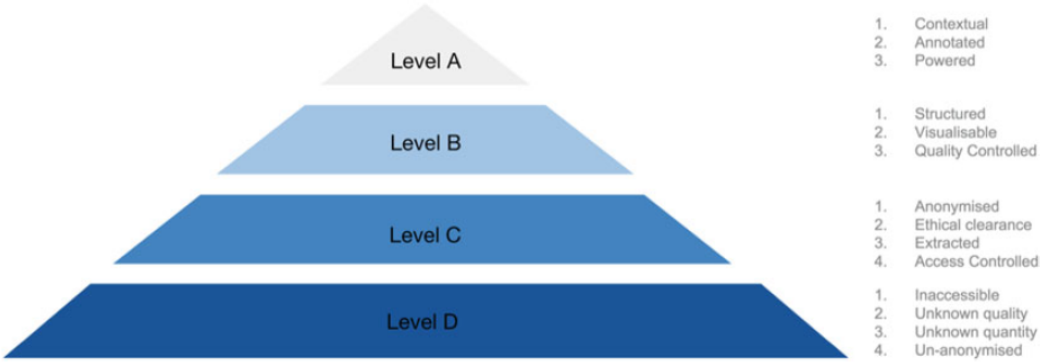

Figura 19.- Escala MIDaR

Por último, resaltar que existe un documento guía creado en la primera conferencia anual de IA en Imagen médica (C-MIMI) (6), que específica los metadatos que deberían de existir en los repositorios de imagen médica si queremos trabajar con modelos de IA.

3.1.5.2 Patrón de referencia

En los algoritmos actuales de IA para tareas de clasificación de imágenes médicas suelen basarse en un enfoque de aprendizaje supervisado. Esto significa que antes de poder entrenar y probar un algoritmo de IA, es necesario definir el patrón de referencia y vincularlo a la imagen. El término *patrón de referencia* suele referirse a la información adquirida por la observación directa (como los resultados de una biopsia, un parámetro de laboratorio, estado clínico del paciente o un seguimiento de la enfermedad) (27).

Las imágenes médicas pueden considerarse como patrón de referencia en algunos casos, como en las fracturas o hemorragias, pero muchas veces, la imagen por sí sola no es suficiente y necesita de confirmación con un seguimiento, un informe patológico o incluso con el uso de otra modalidad de imagen. En cualquier caso, para que una imagen se pueda considerar como patrón de referencia, deben de estar etiquetada con una información específica, como puede ser un diagnóstico, una anotación en la imagen o una segmentación (36).

3.1.6 Cumplimiento FAIR

La adopción de los principios FAIR en el ámbito de la imagen médica está siendo más lento que sobre conjuntos de datos clínicos procedentes de las HCE, siendo los estudios de radiómica los que mayor avance han presentado actualmente en el proceso de "fairificación". Un ejemplo representativo es la aplicación de los principios FAIR sobre 4 conjuntos de datos de radioterapia, proporcionado por Kalendralis et al. (37), y alojados en el Archivo de Imágenes de Cáncer (TCIA, por sus siglas en inglés), donde se proporcionan los metadatos clínicos, las características cuantitativas de las imágenes y los metadatos DICOM en formato RDF (Resource Description Frameworks)[21] mediante el uso de ontologías públicas.

3.2 Estructura del listado de repositorios

En el apartado anterior se describen las características que se han de tener en cuenta a la hora de describir un repositorio de imagen médica. Dichas características nos ayudarán a seleccionar el repositorio indicado para cada tipo de investigación que queramos llevar a cabo.

En el Anexo A se encuentra la estructura que tendrá el listado de repositorios en imagen médica que con el que trabajaremos en el próximo entregable (entregable 4.5) y con el que nos apoyaremos para seleccionar el repositorio que usaremos en la prueba de concepto del entregable 4.6.

[21] https://www.w3.org/RDF/

4 Herramientas de procesado en imagen médica

En el procesado digital de la imagen médica existen gran número de herramientas accesibles. Analizar todas las herramientas disponibles sería una tarea que va más allá de los objetivos de este trabajo. En este documento se definirán las características más relevantes de cada aplicación para poder hacer una selección adecuada de las soluciones que vayan a usarse en la investigación, así como establecer una clasificación de las herramientas en función del uso que les podamos dar.

4.1 Definiciones y Clasificación

El procesado de imagen médica hace referencia al uso de algoritmos computacionales para mejorar la calidad de las imágenes o para extraer información que puede ser de utilidad para aplicaciones médicas como el diagnóstico de enfermedades o la realización de ciertos procedimientos.

El flujo de procesado de imagen médica se compone, por tanto, de una serie de procesos algorítmicos secuenciales que son fundamentales para la extracción de biomarcadores de imagen reproducibles y la predicción robusta de eventos clínicos finales. Una posible clasificación de estos procesos se basa en la confección de 3 grupos diferenciados basados en la naturaleza y la secuencialidad temporal de las tareas a realizar. El primero de ellos es la preparación de la imagen, que incluye aspectos como la reducción de ruido, la armonización y el registro de datos de las imágenes médicas. A continuación, tienen lugar los procesos de detección y segmentación de órganos y tumores, que pueden variar en función del grado de automatización de los mismos. La última tarea del flujo de procesado de imagen propuesto es la extracción de características de radiómica, que se define como una extensión del diagnóstico asistido por computadora y se refiere a la cuantificación integral de, por ejemplo, los fenotipos tumorales mediante la extracción de un gran número de características de imagen cuantitativas, para realizar minería de datos y medicina de precisión.

A continuación, de describen cada uno de estos procesos con mayor detalle y se enumeran algunas de las herramientas de procesado de imagen médica más empleadas por radiólogos e investigadores en imagen médica.

4.1.1 Preparación de la imagen

4.1.1.1 Preprocesado de la imagen

Cuando se trata de datos multicéntricos del mundo real, se ven grandes variabilidades en las imágenes procedentes de los diferentes centros clínicos. Diferentes fabricantes, escáneres y protocolos de adquisición producen variabilidades intrínsecas sobre los valores de intensidad de imagen que afectan al desarrollo de modelos predictivos generalizables y reproducibles, afectando, al mismo tiempo, a la reproducibilidad de los biomarcadores de imagen cuantitativos, de especial relevancia en el desarrollo de herramientas de apoyo a la decisión clínica. Estas variabilidades son especialmente evidentes en el caso de imágenes de RM, ya que los valores de intensidad no se normalizan a unidades estándar como sucede en el caso de las imágenes de TAC, donde los valores de intensidad se estandarizan a través de las unidades Hounsfield.

En los últimos años, se han propuesto diferentes soluciones para resolver el problema de la armonización de la imagen. Estas soluciones van desde el uso de algoritmos tradicionales de visión artificial para la normalización y reducción de ruido de imágenes hasta soluciones más avanzadas basadas en Inteligencia Artificial (IA) y, en concreto, arquitecturas de Redes Neuronales Convolucionales (CNN) como las Redes Generativas Adversariales (GAN).

Las técnicas tradicionales de visión por computadora se basan principalmente en el uso de métodos de normalización y estandarización. Los más comunes son los siguientes:

- **Normalización**: el rango de intensidad se cambia a un rango común. Los rangos típicos son [0-1] o [-1 – 1].

$$I'_{i,j} = \frac{I_{i,j} - \min(I)}{\max(I) - \min(I)}$$

  Donde $I'_{i,j}$ es el valor de intensidad del píxel (i,j) después de la normalización, $I_{i,j}$ es la intensidad de píxel original. min(I) es el valor de intensidad mínima de la imagen y max(I) es el valor de intensidad máxima de la imagen.

- **Centrado**: los valores de intensidad se escalan para tener media cero. El centrado se puede aplicar empleando dos estrategias distintas. Por un lado, en una aproximación muestral (“samplewise”), cada imagen se centra de forma independiente restando a cada intensidad de píxel el valor de intensidad media de la imagen. Por otro lado, en una aproximación característica (“featurewise”), cada imagen se centra utilizando el valor medio del conjunto de datos.

$$Samplewise: I'_{i,j} = I_{i,j} - \mu(I)$$

$$Featurewise: I'_{i,j} = I_{i,j} - \mu(D)$$

Donde $I'_{i,j}$ es el valor de intensidad del píxel (i,j) después de centrar, $I_{i,j}$ es la intensidad de píxel original. Y μ se refiere al valor medio de la imagen (I) o del conjunto de datos (D).

- **Estandarización**: los valores de intensidad se escalan para tener media cero y varianza unitaria. En cuanto al centrado, la estandarización se puede hacer tanto siguiendo una aproximación muestral como una aproximación característica.

$$Samplewise: I'_{i,j} = \frac{I_{i,j} - \mu(I)}{\sigma(I)}$$

$$Featurewise: I'_{i,j} = \frac{I_{i,j} - \mu(D)}{\sigma(D)}$$

Donde $I'_{i,j}$ es el valor de intensidad del píxel (i,j) después de centrar, $I_{i,j}$ es la intensidad de píxel original. μ se refiere al valor medio de la imagen (I) o del conjunto de datos (D). Y σ se refiere al valor de desviación estándar de la imagen (I) o del conjunto de datos (D).

Sin embargo, todos estos métodos no han conseguido demostrar un rendimiento de generalización suficientemente bueno en el dominio de la imagen médica del mundo real, lo que hace necesaria la recogida de una gran cantidad de datos multicéntricos para utilizarlos durante el proceso de entrenamiento de modelos.

Asimismo, para mejorar la calidad de la imagen, tradicionalmente se han utilizado algoritmos de **reducción de ruido** como filtros gaussianos, filtros anisotrópicos de difusión, Non-local Means, o filtros de corrección de inhomogeneidades de campo (como el N4), entre otros (38). Sin embargo, su generalización a cada fuente de ruido y el alto coste computacional que introducen, limitan su aplicación.

Para la extracción de biomarcadores de imágenes cuantitativas, otra aproximación común es utilizar los valores de intensidad de una región homogénea de tejido sano (por ejemplo, músculo) como referencia, contra la cual los valores se normalizan. Sin embargo, este proceso requiere la delineación manual de este tejido de referencia.

Una de las herramientas de armonización de imágenes más empleadas en la actualidad es **ComBat** (39), que ha demostrado un funcionamiento robusto al utilizarse sobre imágenes multicéntricas y de gran variabilidad, y que también se puede aplicar de forma satisfactoria después del procesado de imagen para la armonización de características de radiómica. No obstante, se han detectado dificultades de generalizabilidad al enfrentarse a imágenes adquiridas en centros, equipos y protocolos nuevos (no vistos durante la fase de entrenamiento).

Por todas estas razones, se están explorando técnicas de **Deep Learning** para hacer frente a estas heterogeneidades en el campo de la imagen médica a través de la creación de soluciones de estandarización de imágenes. Sin embargo, desde nuestro conocimiento, todavía no existe una solución general que cubra la enorme variabilidad presente a lo largo de los diferentes centros de imagen.

En este sentido, el método actual que está consiguiendo los mejores resultados a la hora de reconstruir en profundidad una imagen son las llamadas arquitecturas **GAN** (Figura 20).

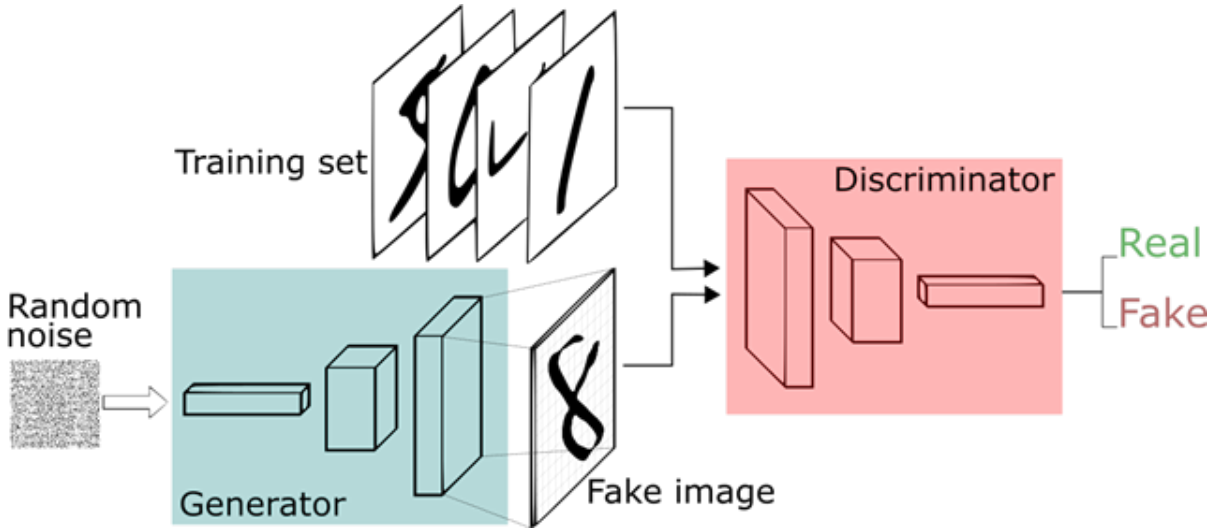

Figura 20.- Arquitectura de tipo GAN. Fuente: Medium

Esta red fue introducida por Ian Goodfellow en 2014 (40) y fue diseñada para crear imágenes sintéticas de una manera no supervisada. La arquitectura se basa en dos redes principales: un generador, que genera datos sintéticos, y un discriminador, que discrimina entre datos reales y datos sintéticos producidos por el generador. El objetivo es aumentar la tasa de error del discriminador, momento en el que el generador crea nuevas imágenes que el discriminador piensa que no son sintéticas. De esta manera, las imágenes generadas serán similares a las reales.

Esta técnica ha sido muy utilizada con fines de aumento de datos, siendo de gran interés en conjuntos de datos desequilibrados. Para la armonización de datos, el enfoque de **CycleGAN** (41) se ha utilizado ampliamente para convertir imágenes de un contexto original de gran variabilidad a un contexto de referencia. De esta forma, armonizando los escáneres, las imágenes de diferentes fabricantes se pueden convertir con el objetivo de incrementar su similitud a otra imagen tomada como referencia.

4.1.1.2 Registro

El registro de imágenes es una técnica de uso frecuente en el procesamiento de imágenes médicas multimodales que se refiere al proceso de alineación automática de datos de imágenes. Consiste en transformar imágenes en un sistema de coordenadas común para que los píxeles correspondientes representen puntos biológicos homólogos. En los últimos

años ha aumentado el interés por relacionar información en diferentes imágenes médicas urgido por una creciente disponibilidad de escáneres, modalidades de imagen y potencia de cómputo. El registro de imágenes multimodales constituye un paso esencial en el objetivo de extraer biomarcadores de imágenes reproducibles y construir modelos computacionales para la predicción de resultados clínicos a partir de datos de imágenes altamente heterogéneos. Dos de las herramientas de registro de imágenes de última generación más populares y competitivas son **Elastix** (42) y **Advanced Normalization Tools** (ANT) (43).

4.1.2 Detección y Segmentación

Detectar es identificar una lesión o anormalidad sobre un órgano o tejido. Se describirán herramientas de detección que en general dependen de la modalidad que empleemos (por ejemplo, para la detección automática de pólipos en estudios de TC-colonografía o de lesiones isquémicas en RM cerebral).
La *segmentación* de órganos y tumores, entendido como el delineamiento de estas estructuras, se divide en 3 tipos, principalmente, dependiendo del grado de automatización del proceso e intervención del usuario clínico: manual, semiautomática y automática. La Tabla 2.- Selección de herramientas de segmentación de órganos y tumores., detalla las herramientas de software más utilizadas en la comunidad de imagen médica, en particular, por los radiólogos. En el caso de las segmentaciones manuales, solo se incluyen las herramientas de software equipadas con herramientas de segmentación y de renderizado 3D que sean utilizables por un no experto en procesamiento de imágenes. Todas las herramientas de software están disponibles gratuitamente en Internet.

Tabla 2.- Selección de herramientas de segmentación de órganos y tumores.

Herramienta	Tipo	Acceso (link)
3D Slicer	Manual / Semiautomática	http://www.slicer.org/
Anatomist	Manual / Automática (cerebro)	http://brainvisa.info/web/anatomist.html
Freesurfer	Manual / Automática (cerebro)	http://surfer.nmr.mgh.harvard.edu/
FSL	Manual / Automática (cerebro)	http://www.fmrib.ox.ac.uk/fsl
ImageJ	Manual / Semiautomática	http://imagej.net/
ITK-SNAP	Manual / Semiautomática	http://www.itksnap.org
Mango	Manual / Semiautomática	http://ric.uthscsa.edu/mango/
MedInria	Manual	http://med.inria.fr/
MIPAV	Manual / Semiautomática	http://mipav.cit.nih.gov/
Myrian Studio	Manual / Semiautomática / Automática (IA)	http://studio.myrian.fr/
OsiriX	Manual / Semiautomática	http://www.osirix-viewer.com/
Seg3D	Manual / Semiautomática	http://www.sci.utah.edu/software/seg3d.html
nnU-Net	Automática (IA)	https://github.com/MIC-DKFZ/nnUNet

4.1.3 Extracción de radiómica

La radiómica (44) es un campo de investigación traslacional emergente que se caracteriza por la extracción, el almacenamiento y el análisis de alto rendimiento de una gran cantidad de características de imágenes cuantitativas (que podrían constituir biomarcadores de imágenes) que proporcionan información cuantitativa (biopsias virtuales) para el diagnóstico temprano de enfermedades, fenotipado de enfermedades, clasificación de enfermedades, terapias dirigidas y evaluación de la respuesta de la enfermedad al tratamiento. Los descriptores cuantitativos exhiben diferentes niveles de complejidad y expresan propiedades relacionadas con la morfología, la intensidad y la textura de la región de interés:

- Las características morfológicas describen propiedades semánticas, como la forma de la región de interés trazada y sus propiedades geométricas, como el volumen, el diámetro máximo a lo largo de diferentes direcciones ortogonales, la superficie máxima, la compacidad del tumor y la esfericidad.
- Las características estadísticas de primer orden describen la distribución de los valores de las intensidades de los vóxeles individuales, sin preocuparse por las relaciones espaciales. Estas propiedades están basadas en distribuciones de histograma, y proporcionan información sobre la media, la mediana, el máximo y el mínimo de las intensidades de los vóxeles en la imagen, así como de su asimetría, curtosis, uniformidad y aleatoriedad (entropía). Dependiendo de la modalidad de imagen empleada, estas características adquieren un significado concreto que permiten aumentar la explicabilidad de la radiómica. Por ejemplo, el contraste radiográfico es la diferencia de densidad entre regiones vecinas en una radiografía simple; en imagen PET, los valores de intensidad adquiridos permiten determinar la captación de glucosa en la región de interés.
- Las características estadísticas de segundo orden se obtienen calculando las interrelaciones estadísticas entre vóxeles vecinos. Proporcionan una medida de la disposición espacial de las intensidades de vóxel y, por lo tanto, de la heterogeneidad dentro de la región de interés. Tales características pueden derivarse, por ejemplo, de la matriz de co-ocurrencia de nivel de gris (GLCM), que cuantifica la incidencia de vóxeles con las mismas intensidades a una distancia predeterminada a lo largo de una dirección fija, o de la matriz de longitud de ejecución de nivel de gris (GLRLM), que cuantifica vóxeles consecutivos con la misma intensidad a lo largo de direcciones fijas.

Estas características se pueden extraer directamente en las imágenes originales o después de aplicar diferentes filtros o transformaciones matemáticas a las imágenes. Estas últimas reciben el nombre de características de orden superior. Se obtienen, por ejemplo, con el objetivo de identificar patrones repetitivos, suprimir ruido o resaltar detalles. Estos incluyen métodos como el análisis fractal, el funcional de Minkowski, la transformada de tipo “wavelet” y la transformada laplaciana aplicada a imágenes procesadas por un filtro gaussiano, a partir de los que se pueden extraer regiones con patrones de textura cada vez más complicados.

Una de las herramientas más populares para la extracción de características es el software **Pyradiomics** (45), que permite la extracción de miles de características radiómicas cuantitativas, normalmente de imágenes de RM, TC y PET, así como mapas paramétricos. Otras herramientas de acceso abierto ampliamente utilizadas por la comunidad de imagen médica son: **RaCaT** (46), **Moddicom** (47), **CERR** (48), **MaZda** (49) e **IBEX** (50).

4.2 Estructura del listado de herramientas

En el apartado anterior hemos descrito las herramientas más utilizadas a día de hoy para el procesado de la imagen médica en función de sus diferentes usos. Al igual que hemos hecho con los repositorios de imagen médica, hemos creado el Anexo B, donde vamos a definir la estructura que tendrá el listado de herramientas en imagen médica que con el que trabajaremos en el próximo entregable (entregable 4.5) y con el que nos apoyaremos para seleccionar las herramientas que usaremos en la prueba de concepto del entregable 4.6.

5 Caso de Uso

5.1 Selección de un repositorio de imagen médica

Para finalizar el documento, vamos a proceder a seleccionar un repositorio de imagen médica en abierto para poder analizarlo en función de las características que hemos definido como importantes en el apartado 3.

Para la selección del repositorio, se ha usado la plataforma OpneNEURO[22] por la facilidad de uso que tiene y la gran cantidad de repositorios que nos podemos encontrar en ella. A día del presente trabajo, la plataforma cuenta con más de 23000 participantes y más de 650 repositorios de imagen médica en Resonancia, tomografía de emisión de positrones (PET), magnetoencefalografía (MEG), electroencefalograma (EEG) y electroencefalograma intracraneal (iEEG). La plataforma está centrada en imágenes de neurociencia y por ello utiliza el estándar BIDS que hemos definido previamente.

El repositorio seleccionado tiene como nombre "Lausanne_TOF-MRA_Aneurysm_Cohort" (doi:10.18112/openneuro.ds003949.v1.0.0).

5.2 Informe exploratorio

El análisis exploratorio de los datos (AED o EDA, Exploratory Data Analysis) o Informe exploratorio consiste en la aplicación de técnicas estadísticas que permiten explorar, describir y resumir la naturaleza de los datos para poder comprender qué relaciones existen entre las variables de interés y así maximizar la comprensión del conjunto de datos. La exploración minuciosa de los datos permite detectar posibles errores tales como datos introducidos de forma incorrecta, revelar la presencia de valores atípicos o perdidos, o realizar un análisis descriptivo mediante representaciones gráficas y resúmenes de los aspectos más significativos. Estamos convencidos de que se trata de una buena práctica y se recomienda como previo a cualquier análisis de IA en imagen poblacional y/o repositorios de estudios retrospectivos.

El EDA debe contener los siguientes pasos:

- Análisis descriptivo de las variables
- Re-ajuste de los tipos de las variables
- Detección y tratamiento de datos ausentes
- Identificación de datos atípicos y su tratamiento

[22] https://openneuro.org/

- Examen numérico y gráfico de las relaciones entre las variables analizadas para determinar el grado de correlación entre ellas.

En la Figura 21 se muestra una vista previa del análisis exploratorio de los datos del repositorio elegido, "**Lausanne_TOF-MRA_Aneurysm_Cohort**". Para visualizar el análisis completo del repositorio de imagen médica seleccionado, se puede visitar el siguiente enlace en: https://ceib.cipf.es/reports/impact_data/bimcv_eda_impactdata_openneuro.html

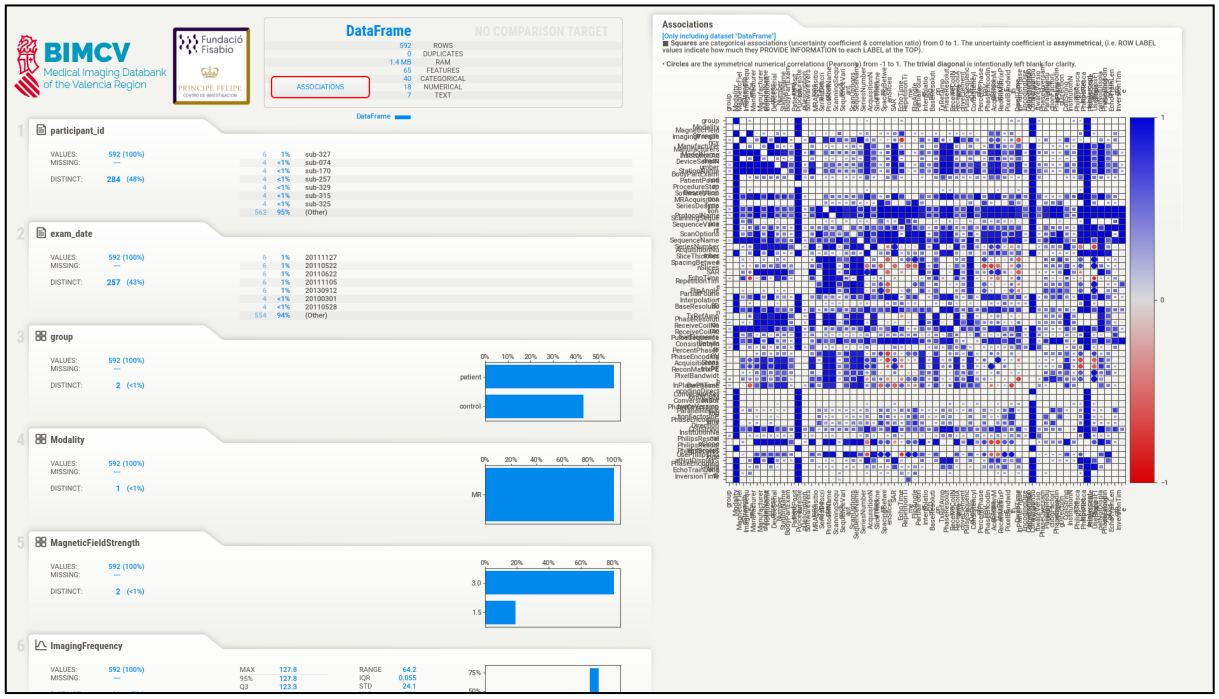

Figura 21.- Visualización análisis exploratorio

5.3 Análisis del repositorio

5.3.1 Población del repositorio

Las imágenes incluidas en este repositorio provienen del Hospital “Lausanne University Hospital”, en Suiza. Los casos son pacientes que ha sufrido de aneurisma(s) y sujetos sanos que no han tenido ningún episodio de aneurisma. Este repositorio no contiene información clínica de los pacientes, más allá de la fecha de realización de la prueba y si pertenece al grupo control o no.

5.3.2 Acceso del repositorio

El repositorio se encuentra perfectamente identificado con su correspondiente DOI (doi:10.18112/openneuro.ds003949.v1.0.0) y con un control de versiones perfectamente detallado (actualmente se encuentra en su versión 1.0.0). Los autores del repositorio están identificados y existe una política de citación del repositorio muy detallada y accesible. Nos encontramos con un repositorio unicéntrico y con una arquitectura centralizada en la plataforma OpenNEURO y con un acceso abierto de las imágenes.

5.3.3 Descripción del repositorio

El repositorio son 13.46 GB y consta de 284 Angiografías por Resonancia Magnética realizadas mediante la técnica TOF (*Time of light*), de las cuales, 127 corresponden a sujetos sanos y 157 corresponden a pacientes con aneurisma(s). Las imágenes se han convertido a formato NIfTI y junto con las mismas, se adjunta un archivo *json* que aporta todas las etiquetas DICOM que se habrían perdido durante el proceso de conversión. Nos encontramos con un total de 568 series (dos por sujeto) y no está definido el número total de imágenes. Además de las resonancias, se aporta también: El volumen del cráneo, y en los pacientes, se aporta también una máscara manual del aneurisma(s); Correcciones del volumen; Métricas de calidad del registro; Parámetros utilizados para cada registro; Atlas probabilístico de vasos. El repositorio no tiene capacidad de procesamiento de imágenes médicas, por lo que necesitas descargar las imágenes para poder realizar cualquier procesamiento de las mismas o para poder entrenar los modelos de IA.

6 Conclusiones

Como se ha podido comprobar a lo largo de este documento, para realizar una investigación sobre datos es de vital importancia que estos datos sigan los principios FAIR. En un campo como la imagen médica se vuelve aún más relevante cumplir con estos principios y ser capaces de reutilizar nuestros datos y sacarles el máximo partido posible.

Un punto crítico para conseguir ese cumplimiento FAIR es el uso de estándares para facilitar la interoperabilidad de los datos. Estos estándares están definidos tanto a nivel de sujeto, que ayudarán a identificar enfermedades, estructuras, lesiones o cualquier información relevante de manera común entre los investigadores; a nivel de Imagen, donde el uso una estructura común de datos facilita el intercambio de información entre sistemas; como a nivel de Capas, que permite compartir con los demás colaboradores nuestros análisis y procesamientos de las imágenes.

En este trabajo se presentan los estándares más importantes que se utilizan para facilitar esta interoperabilidad, pero este ámbito está en continuo desarrollo y a día de hoy no paran de salir nuevas necesidades y estándares o combinaciones de estándares que pretenden cubrirlas.

Es imprescindible que los datos estén almacenados de manera efectiva, para realizar búsquedas personalizadas y acceder a ellos en el caso que sea posible. Para este punto, se describen los repositorios de imagen médica y se definen unos puntos básicos que deberían de tener si queremos utilizarlos para nuestras investigaciones. Todo ello cumpliendo con los principios FAIR y con las correspondientes leyes en materia de protección de datos pertinentes, y por ello, hemos definido los estándares usados para la de-identificación de la imagen médica.

No se puede pasar por alto la importancia creciente de la inteligencia artificial en nuestro día a día y la gran aplicación que encuentra en las imágenes médicas. Se hace especialmente relevante poder acceder a toda la cantidad de información disponible de los sujetos de estudio, así como de las técnicas utilizadas durante las adquisiciones. Se introduce en este documento las herramientas más comúnmente usadas para procesar las imágenes.

Finalmente, en un caso de uso, podemos ver cuál es la realidad actual de trabajar con un repositorio de imagen médica, como encontrar un conjunto de datos para su investigación, que herramientas tiene disponibles, y que estándares se pueden usar. Es tarea ahora del investigador decidir que repositorios seleccionar para encontrar sus datos, o que tipo de repositorio puede utilizar para alojar sus datos en el caso de que sea una investigación propia; que estándares quiere adoptar, y que herramientas necesita utilizar.

Referencias

1. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 15 de marzo de 2016;3(1):160018.

2. The FAIR Data Principles [Internet]. FORCE11. 2014 [citado 23 de febrero de 2022]. Disponible en: https://force11.org/info/the-fair-data-principles/

3. Alonso-Calvo R, Perez-Rey D, Paraiso-Medina S, Claerhout B, Hennebert P, Bucur A. Enabling semantic interoperability in multi-centric clinical trials on breast cancer. Comput Methods Programs Biomed. marzo de 2015;118(3):322-9.

4. VAN SOEST J, LUSTBERG T, GRITTNER D, MARSHALL MS, PERSOON L, NIJSTEN B, et al. Towards a semantic PACS: Using Semantic Web technology to represent imaging data. Stud Health Technol Inform. 2014;205:166-70.

5. RDF - Semantic Web Standards [Internet]. [citado 23 de febrero de 2022]. Disponible en: https://www.w3.org/RDF/

6. Kohli MD, Summers RM, Geis JR. Medical Image Data and Datasets in the Era of Machine Learning—Whitepaper from the 2016 C-MIMI Meeting Dataset Session. J Digit Imaging. agosto de 2017;30(4):392-9.

7. European Society of Radiology (ESR). ESR Position Paper on Imaging Biobanks. Insights Imaging. agosto de 2015;6(4):403-10.

8. Seifert S, Kelm B, Moeller M, Mukherjee S, Cavallaro A, Huber M, et al. Semantic Annotation of Medical Images. En 2010.

9. Ceballos O, Garcia A, García-Castro L, Millan M, de Cali S. ANOTACIÓN SEMÁNTICA DE IMÁGENES MÉDICAS. :16.

10. Martínez SMR, Vido SABG, Jorente MJV. Representación conceptual de imágenes médicas digitales: Integración de Contexto y Contenido Visual. Rev Gen Inf Doc. 2016;26(2):651-72.

11. Nguyen C-D. Workload- and Data-based Automated Design for a Hybrid Row-Column Storage Model and Bloom Filter-Based Query Processing for Large-Scale DICOM Data Management. :217.

12. Cabecera Del DICOM - [PDF Document] [Internet]. [citado 27 de febrero de 2022]. Disponible en: https://cupdf.com/document/cabecera-del-dicom.html

13. Serna W, Trujillo JP, Hernando J. DESCRIPCIÓN DEL ESTÁNDAR DICOM PARA UN ACCESO CONFIABLE A LA INFORMACIÓN DE LAS IMÁGENES MÉDICAS. Sci Tech Año XVI No 45. agosto de 2010;45.

14. Adams MC, Turkington TG, Wilson JM, Wong TZ. A systematic review of the factors affecting accuracy of SUV measurements. AJR Am J Roentgenol. agosto de 2010;195(2):310-20.

15. Bennett CM, Miller MB. How reliable are the results from functional magnetic resonance imaging? Ann N Y Acad Sci. marzo de 2010;1191(1):133-55.

16. Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, Das S, Duff EP, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci Data. 21 de junio de 2016;3:160044.

17. Saborit-Torres JM, Saenz-Gamboa JJ, Montell JÀ, Salinas JM, Gómez JA, Stefan I, et al. Medical imaging data structure extended to multiple modalities and anatomical regions. ArXiv201000434 Phys [Internet]. 30 de septiembre de 2020 [citado 2 de marzo de 2022]; Disponible en: http://arxiv.org/abs/2010.00434

18. Sarkans U, Chiu W, Collinson L, Darrow MC, Ellenberg J, Grunwald D, et al. REMBI: Recommended Metadata for Biological Images—enabling reuse of microscopy data in biology. Nat Methods. diciembre de 2021;18(12):1418-22.

19. Noumeir R. DICOM structured report document type definition. IEEE Trans Inf Technol Biomed. diciembre de 2003;7(4):318-28.

20. Clunie DA. DICOM structured reporting: an object model as an implementation boundary. En: Siegel EL, Huang HK, editores. San Diego, CA; 2001 [citado 17 de febrero de 2022]. p. 207-15. Disponible en: http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=906985

21. Noumeir R. Multimedia medical diagnostic document. En: Proceedings IEEE International Conference on Multimedia and Expo [Internet]. Lausanne, Switzerland: IEEE; 2002 [citado 17 de febrero de 2022]. p. 325-8. Disponible en: http://ieeexplore.ieee.org/document/1035594/

22. Noumeir R. Benefits of the DICOM Structured Report. J Digit Imaging. diciembre de 2006;19(4):295-306.

23. Campbell W, Foster K, Hinrichs S. Application of whole slide image markup and annotation for pathologist knowledge capture. J Pathol Inform. 2013;4(1):2.

24. Lindman K, Rose JF, Lindvall M, Lundström C, Treanor D. Annotations, Ontologies, and Whole Slide Images – Development of an Annotated Ontology-Driven Whole Slide Image Library of Normal and Abnormal Human Tissue. J Pathol Inform. 23 de julio de 2019;10:22.

25. Channin DS, Mongkolwat P, Kleper V, Sepukar K, Rubin DL. The caBIG™ Annotation and Image Markup Project. J Digit Imaging Off J Soc Comput Appl Radiol. abril de 2010;23(2):217-25.

26. Pérez-Díez I, Pérez-Moraga R, López-Cerdán A, Salinas-Serrano J-M, la Iglesia-Vayá M de. De-identifying Spanish medical texts - named entity recognition applied to radiology reports. J Biomed Semant. diciembre de 2021;12(1):6.

27. Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing Medical Imaging Data for Machine Learning. Radiology. abril de 2020;295(1):4-15.

28. Sheth AP, Larson JA. Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Comput Surv. septiembre de 1990;22(3):183-236.

29. Prescott JW. Quantitative Imaging Biomarkers: The Application of Advanced Image Processing and Analysis to Clinical and Preclinical Decision Making. J Digit Imaging. febrero de 2013;26(1):97-108.

30. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. diciembre de 2017;14(12):749-62.

31. Alberich-Bayarri Á, Hernández-Navarro R, Ruiz-Martínez E, García-Castro F, García-Juan D, Martí-Bonmatí L. Development of imaging biomarkers and generation of big data. Radiol Med (Torino). 1 de junio de 2017;122(6):444-8.

32. Norlin L, Fransson MN, Eriksson M, Merino-Martinez R, Anderberg M, Kurtovic S, et al. A Minimum Data Set for Sharing Biobank Samples, Information, and Data: MIABIS. Biopreservation Biobanking. agosto de 2012;10(4):343-8.

33. Scapicchio C, Gabelloni M, Forte SM, Alberich LC, Faggioni L, Borgheresi R, et al. DICOM-MIABIS integration model for biobanks: a use case of the EU PRIMAGE project. Eur Radiol Exp. 12 de mayo de 2021;5:20.

34. Suman G, Patra A, Korfiatis P, Majumder S, Chari ST, Truty MJ, et al. Quality gaps in public pancreas imaging datasets: Implications & challenges for AI applications. Pancreatology. agosto de 2021;21(5):1001-8.

35. Harvey H, Glocker B. A Standardised Approach for Preparing Imaging Data for Machine Learning Tasks in Radiology. En: Ranschaert ER, Morozov S, Algra PR, editores. Artificial Intelligence in Medical Imaging [Internet]. Cham: Springer International Publishing; 2019 [citado 18 de febrero de 2022]. p. 61-72. Disponible en: http://link.springer.com/10.1007/978-3-319-94878-2_6

36. Langlotz CP, Allen B, Erickson BJ, Kalpathy-Cramer J, Bigelow K, Cook TS, et al. A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop. Radiology. junio de 2019;291(3):781-91.

37. Kalendralis P, Shi Z, Traverso A, Choudhury A, Sloep M, Zhovannik I, et al. FAIR-compliant clinical, radiomics and DICOM metadata of RIDER, interobserver, Lung1 and head-Neck1 TCIA collections. Med Phys. noviembre de 2020;47(11):5931-40.

38. Fan L, Zhang F, Fan H, Zhang C. Brief review of image denoising techniques. Vis Comput Ind Biomed Art. diciembre de 2019;2(1):7.

39. Fortin J-P, Parker D, Tunç B, Watanabe T, Elliott MA, Ruparel K, et al. Harmonization of multi-site diffusion tensor imaging data. NeuroImage. 1 de noviembre de 2017;161:149-70.

40. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Networks. ArXiv14062661 Cs Stat [Internet]. 10 de junio de 2014 [citado 21 de marzo de 2022]; Disponible en: http://arxiv.org/abs/1406.2661

41. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. En: 2017 IEEE International Conference on Computer Vision (ICCV). 2017. p. 2242-51.

42. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. elastix: A Toolbox for Intensity-Based Medical Image Registration. IEEE Trans Med Imaging. enero de 2010;29(1):196-205.

43. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage. febrero de 2011;54(3):2033-44.

44. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. febrero de 2016;278(2):563-77.

45. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 1 de noviembre de 2017;77(21):e104-7.

46. Pfaehler E, Zwanenburg A, de Jong JR, Boellaard R. RaCaT: An open source and easy to use radiomics calculator tool. Wang Y, editor. PLOS ONE. 20 de febrero de 2019;14(2):e0212223.

47. Dinapoli N, Alitto AR, Vallati M, Gatta R, Autorino R, Boldrini L, et al. Moddicom: a complete and easily accessible library for prognostic evaluations relying on image features. En: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) [Internet]. Milan: IEEE; 2015 [citado 2 de marzo de 2022]. p. 771-4. Disponible en: https://ieeexplore.ieee.org/document/7318476/

48. Apte AP, Iyer A, Crispin-Ortuzar M, Pandya R, van Dijk LV, Spezi E, et al. Technical Note: Extension of CERR for computational radiomics: A comprehensive MATLAB platform for reproducible radiomics research. Med Phys. agosto de 2018;45(8):3713-20.

49. Szczypiński PM, Strzelecki M, Materka A, Klepaczko A. MaZda—A software package for image texture analysis. Comput Methods Programs Biomed. 1 de abril de 2009;94(1):66-76.

50. Zhang L, Fried DV, Fave XJ, Hunter LA, Yang J, Court LE. ibex: An open infrastructure software platform to facilitate collaborative work in radiomics. Med Phys. marzo de 2015;42(3):1341-53.

Acrónimos y Abreviaturas

IMPaCT	Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología
IMPaCT-Data	Programa de ciencia de datos de IMPaCT
IA	Inteligencia Artificial
HCE	Historia Clínica Electrónica
fMRI	Resonancia Magnética funcional
PHI	Información Sanitaria Protegida (Protected Health Information)
RSNA	Radiological Society of North America
DOI	Digital Object Identifier
NIfTI	Neuroimaging Informatics Technology Initiative
RWVM	DICOM Real World Value Mapping
BIDS	Brain Imaging Data Structure
MIDS	Medical Imaging Data Structure
REMBI	Recommended Metadata for Biological Images)
DICOM SR	DICOM Structured Reporting
DICOM SEG	DICOM Segmentation
RO	Radiomics Ontology
UMLS	Unified Medical Language System
CIE	Clasificación Internacional de Enfermedades
NLM	National Library of Medicine

Anexo A. Estructura Listado Repositorios en Imagen Médica

			Descripción Población		
Nombre	URL	DOI	Descripción del episodio	Datos Clínicos-Analíticos asociados	Información de Tratamientos realizados al paciente

			Descripción de acceso						
Nombre	URL	DOI	Identificador único del repositorio	Control de versiones	Información sobre la política de citación del repositorio	Información de los autores	Nº Centros	Arquitectura	Tipo de acceso

			Descripción imágenes								
Nombre	URL	DOI	Tipo Imágenes	Protocolo de aquisición	Nº Pacientes	Nº Estudios	Nº Series	Nº Imágenes	Exitencia de Anotaciones	Patrón de Referencia	Otros descriptores

Anexo B. Estructura Listado Herramientas en Imagen Médica

Modelo estándar para herramienta

Nombre	Clasificación	Dirección web	Documentación	Descripcíon

Modelo específico para Herramienta de Segmentación

Nombre herramienta	Grado automatismo	Modalidad	Repositorio	Código abierto	Órganos	Preparación de imagen	Validación Interna	Validación Externa	Métricas de Calidad	Uso comercial	Publicaciones

Anexo C. Perfil de de-identificación de las cabeceras DICOM propuesto por la RSNA

Data Element	Description	De-identification
(0008,0014)	InstanceCreatorUID	@hashuid(@UIDROOT,this)
(0008,0015)	InstanceCoercionDateTime	@hashdate(this,PatientID)
(0008,0018)	SOPInstanceUID	@hashuid(@UIDROOT,this)
(0008,0020)	StudyDate	@hashdate(this,PatientID)
(0008,0021)	SeriesDate	@hashdate(this,PatientID)
(0008,0022)	AcquisitionDate	@hashdate(this,PatientID)
(0008,0023)	ContentDate	@hashdate(this,PatientID)
(0008,0024)	OverlayDate	@hashdate(this,PatientID)
(0008,0025)	CurveDate	@hashdate(this,PatientID)
(0008,002A)	AcquisitionDatetime	@hashdate(this,PatientID)
(0008,0050)	AccessionNumber	@hash(this,8)
(0008,0058)	FailedSOPInstanceUIDList	@remove()
(0008,0080)	InstitutionName	@empty()
(0008,0081)	InstitutionAddress	@remove()
(0008,0082)	InstitutionCodeSeq	@remove()
(0008,0090)	ReferringPhysicianName	@empty()
(0008,0092)	ReferringPhysicianAddress	@remove()
(0008,0094)	ReferringPhysicianPhoneNumbers	@remove()
(0008,0096)	ReferringPhysicianIdentificationSeq	@remove()
(0008,009C)	ConsultingPhysicianName	@empty()
(0008,009D)	ConsultingPhysicianIdentificationSeq	@remove()
(0008,1010)	StationName	@empty()

(0008,1040)	InstitutionalDepartmentName	@remove()
(0008,1041)	InstitutionalDepartmentTypeCodeSeq	@remove()
(0008,1048)	PhysicianOfRecord	@remove()
(0008,1049)	PhysicianOfRecordIdentificationSeq	@remove()
(0008,1050)	PerformingPhysicianName	@remove()
(0008,1052)	PerformingPhysicianIdentificationSeq	@remove()
(0008,1060)	NameOfPhysicianReadingStudy	@remove()
(0008,1062)	PhysicianReadingStudyIdentificationSeq	@remove()
(0008,1070)	OperatorName	@empty()
(0008,1072)	OperatorIdentificationSeq	@remove()
(0008,1080)	AdmittingDiagnosisDescription	@remove()
(0008,1084)	AdmittingDiagnosisCodeSeq	@remove()
(0008,1110)	RefStudySeq	@remove()
(0008,1111)	RefPPSSeq	@remove()
(0008,1120)	RefPatientSeq	@remove()
(0008,1140)	RefImageSeq	@remove()
(0008,1155)	RefSOPInstanceUID	@remove()
(0008,1195)	TransactionUID	@remove()
(0008,2111)	DerivationDescription	@remove()
(0008,2112)	SourceImageSeq	@remove()
(0008,3010)	IrradiationEventUID	@remove()
(0008,4000)	IdentifyingCommentsRetired	@remove()
(0010,0010)	PatientName	@param(@SITEID)-@integer(PatientID,"ptid",6)
(0010,0020)	PatientID	
(0010,0021)	IssuerOfPatientID	@param(@SITEID)-@integer(PatientID,"ptid",6)
(0010,0030)	PatientBirthDate	@remove()
(0010,0050)	PatientInsurancePlanCodeSeq	@hashdate(this,PatientID)
(0010,0101)	PatientPrimaryLanguageCodeSeq	@remove()
(0010,0102)	PatientPrimaryLanguageModifierCodeSeq	@remove()

(0010,1000)	OtherPatientIDs	@remove()
(0010,1001)	OtherPatientNames	@remove()
(0010,1002)	OtherPatientIDSeq	@remove()
(0010,1005)	PatientBirthName	@remove()
(0010,1040)	PatientAddress	@remove()
(0010,1050)	InsurancePlanIdentificationRetired	@remove()
(0010,1060)	PatientMotherBirthName	@remove()
(0010,1080)	MilitaryRank	@remove()
(0010,1081)	BranchOfService	@remove()
(0010,1090)	MedicalRecordLocator	@remove()
(0010,1100)	RefPatientPhotoSeq	@remove()
(0010,2000)	MedicalAlerts	@remove()
(0010,2110)	ContrastAllergies	@remove()
(0010,2150)	CountryOfResidence	@remove()
(0010,2152)	RegionOfResidence	@remove()
(0010,2154)	PatientPhoneNumbers	@remove()
(0010,2155)	PatientTelecomInformation	@remove()
(0010,2180)	Occupation	@remove()
(0010,21B0)	AdditionalPatientHistory	@remove()
(0010,21D0)	LastMenstrualDate	@remove()
(0010,21F0)	PatientReligiousPreference	@hashdate(this,PatientID)
(0010,2297)	ResponsiblePerson	@remove()
(0010,2299)	ResponsibleOrganization	@remove()
(0010,4000)	PatientComments	@remove()
(0012,0010)	ClinicalTrialSponsorName	@remove()
(0012,0020)	ClinicalTrialProtocolID	@empty()
(0012,0021)	ClinicalTrialProtocolName	@empty()
(0012,0030)	ClinicalTrialSiteID	@empty()
(0012,0031)	ClinicalTrialSiteName	@empty()

(0012,0040)	ClinicalTrialSubjectID	@empty()
(0012,0042)	ClinicalTrialSubjectReadingID	@empty()
(0012,0050)	ClinicalTrialTimePointID	@empty()
(0012,0051)	ClinicalTrialTimePointDescription	@empty()
(0012,0060)	CoordinatingCenterName	@remove()
(0012,0062)	PatientIdentityRemoved	@empty()
(0012,0064)	DeIdentificationMethodCodeSeq (see table)	YES
(0012,0071)	ClinicalTrialSeriesID	113100/113101/113102/113103
(0012,0072)	ClinicalTrialSeriesDescription	@remove()
(0012,0081)	ClinicalTrialProtocolEthicsCommitteeName	@remove()
(0012,0082)	ClinicalTrialProtocolEthicsCommitteeApprovalNum	@empty()
(0016,002B)	MakerNote	@remove()
(0016,004B)	DeviceSettingDescription	@remove()
(0016,004D)	CameraOwnerName	@remove()
(0016,004E)	LensSpecification	@remove()
(0016,004F)	LensMake	@remove()
(0016,0050)	LensModel	@remove()
(0016,0051)	LensSerialNumber	@remove()
(0016,0070)	GPSVersionID	@remove()
(0016,0071)	GPSLatitudeRef	@remove()
(0016,0072)	GPSLatitude	@remove()
(0016,0073)	GPSLongitudeRef	@remove()
(0016,0074)	GPSLongitude	@remove()
(0016,0075)	GPSAltitudeRef	@remove()
(0016,0076)	GPSAltitude	@remove()
(0016,0077)	GPSTimeStamp	@remove()
(0016,0078)	GPSSatellites	@remove()
(0016,0079)	GPSStatus	@remove()
(0016,007A)	GPSMeasureMode	@remove()

(0016,007B)	GPSDOP	@remove()
(0016,007C)	GPSSpeedRef	@remove()
(0016,007D)	GPSSpeed	@remove()
(0016,007E)	GPSTrackRef	@remove()
(0016,007F)	GPSTrack	@remove()
(0016,0080)	GPSImgDirectionRef	@remove()
(0016,0081)	GPSImgDirection	@remove()
(0016,0082)	GPSMapDatum	@remove()
(0016,0083)	GPSDestLatitudeRef	@remove()
(0016,0084)	GPSDestLatitude	@remove()
(0016,0085)	GPSDestLongitudeRef	@remove()
(0016,0086)	GPSDestLongitude	@remove()
(0016,0087)	GPSDestBearingRef	@remove()
(0016,0088)	GPSDestBearing	@remove()
(0016,0089)	GPSDestDistanceRef	@remove()
(0016,008A)	GPSDestDistance	@remove()
(0016,008B)	GPSProcessingMethod	@remove()
(0016,008C)	GPSAreaInformation	@remove()
(0016,008D)	GPSDateStamp	@remove()
(0016,008E)	GPSDifferential	@remove()
(0018,1000)	DeviceSerialNumber	@remove()
(0018,1002)	DeviceUID	@empty()
(0018,1004)	PlateID	@hashuid(@UIDROOT,this)
(0018,1005)	GeneratorID	@remove()
(0018,1007)	CassetteID	@remove()
(0018,1008)	GantyID	@remove()
(0018,1009)	UniqueDeviceIdentifier	@remove()
(0018,100A)	UDISeq	@remove()
(0018,100B)	ManufacturerDeviceClassUID	@remove()

(0018,1030)	ProtocolName	@hashuid(@UIDROOT,this)
(0018,1400)	AcquisitionDeviceProcessingDescription	@remove()
(0018,2042)	TargetUID	@remove()
(0018,4000)	SeriesCommentsRetired	@hashuid(@UIDROOT,this)
(0018,700A)	DetectorID	@remove()
(0018,9185)	RespiratoryMotionCompensationTechniqueDescrip	@remove()
(0018,9367)	XRaySourceID	@remove()
(0018,9369)	SourceStartDateTime	@empty()
(0018,936A)	SourceEndDateTime	@hashdate(this,PatientID
(0018,9371)	XRayDetectorID	@hashdate(this,PatientID)
(0018,9373)	XRayDetectorLabel	@empty()
(0018,937B)	MultienergyAcquisitionDescription	@remove()
(0018,937F)	DecompositionDescription	@remove()
(0018,9424)	AcquisitionProtocolDescription	@remove()
(0018,9516)	StartAcquisitionDateTime	@remove()
(0018,9517)	EndAcquisitionDateTime	@hashdate(this,PatientID)
(0018,A003)	ContributionDescription	@hashdate(this,PatientID)
(0020,000D)	StudyInstanceUID	@remove()
(0020,000E)	SeriesInstanceUID	@hashuid(@UIDROOT,this)
(0020,0010)	StudyID	@hashuid(@UIDROOT,this)
(0020,0052)	FrameOfReferenceUID	@empty()
(0020,0200)	SynchronizationFrameOfReferenceUID	@hashuid(@UIDROOT,this)
(0020,3401)	ModifyingDeviceIDRetired	@hashuid(@UIDROOT,this)
(0020,3406)	ModifiedImageDescriptionRetired	@remove()
(0020,4000)	ImageComments	@remove()
(0020,9158)	FrameComments	@remove()
(0020,9161)	ConcatenationUID	@remove()
(0020,9164)	DimensionOrganizationUID	@hashuid(@UIDROOT,this)
(0028,1199)	PaletteColorLUTUID	@hashuid(@UIDROOT,this)

(0028,1214)	LargePaletteColorLUTUID	@hashuid(@UIDROOT,this)
(0028,4000)	PixelCommentsRetired	@hashuid(@UIDROOT,this)
(0032,0012)	StudyIDIssuer	@remove()
(0032,1020)	ScheduledStudyLocation	@remove()
(0032,1021)	ScheduledStudyLocationAET	@remove()
(0032,1030)	ReasonforStudy	@remove()
(0032,1032)	RequestingPhysician	@remove()
(0032,1033)	RequestingService	@remove()
(0032,1060)	RequestedProcedureDescription	@remove()
(0032,1066)	ReasonForVisit	@empty()
(0032,1067)	ReasonForVisitCodeSeq	@remove()
(0032,1070)	RequestedContrastAgent	@remove()
(0032,4000)	StudyComments	@remove()
(0034,0001)	FlowIdentifierSeq	@remove()
(0034,0002)	FlowIdentifier	@remove()
(0034,0005)	SourceIdentifier	@empty()
(0034,0007)	FrameOriginTimestamp	@empty()
(0038,0004)	RefPatientAliasSeq	@remove()
(0038,0010)	AdmissionID	@remove()
(0038,0011)	IssuerOfAdmissionIDRetired	@remove()
(0038,0014)	IssuerOfAdmissionIDSeq	@remove()
(0038,001E)	ScheduledPatientInstitutionResidence	@remove()
(0038,0020)	AdmittingDate	@remove()
(0038,0021)	AdmittingTime	@remove()
(0038,0040)	DischargeDiagnosisDescription	@remove()
(0038,0050)	SpecialNeeds	@remove()
(0038,0060)	ServiceEpisodeID	@remove()
(0038,0061)	IssuerOfServiceEpisodeIDRET	@remove()
(0038,0062)	ServiceEpisodeDescription	@remove()

(0038,0064)	IssuerOfServiceEpisodeIDSeq	@remove()
(0038,0300)	CurrentPatientLocation	@remove()
(0038,0400)	PatientInstitutionResidence	@remove()
(0038,0500)	PatientState	@remove()
(0038,4000)	VisitComments	@remove()
(0040,0001)	ScheduledStationAET	@remove()
(0040,0002)	SPSStartDate	@remove()
(0040,0003)	SPSStartTime	@remove()
(0040,0004)	SPSEndDate	@remove()
(0040,0005)	SPSEndTime	@remove()
(0040,0006)	ScheduledPerformingPhysicianName	@remove()
(0040,0007)	SPSDescription	@remove()
(0040,000B)	ScheduledPerformingPhysicianIdentificationSeq	@remove()
(0040,0010)	ScheduledStationName	@remove()
(0040,0011)	SPSLocation	@remove()
(0040,0012)	PreMedication	@remove()
(0040,0241)	PerformedStationAET	@remove()
(0040,0242)	PerformedStationName	@remove()
(0040,0243)	PerformedLocation	@remove()
(0040,0244)	PPSStartDate	@remove()
(0040,0245)	PPSStartTime	@remove()
(0040,0250)	PPSEndDate	@remove()
(0040,0251)	PPSEndTime	@remove()
(0040,0253)	PPSID	@remove()
(0040,0254)	PPSDescription	@remove()
(0040,0275)	RequestAttributesSeq	@remove()
(0040,0280)	PPSComments	@remove()
(0040,050A)	SpecimenAccessionNumber	@remove()
(0040,0512)	ContainerIdentifier	@remove()

(0040,0513)	IssuerOfTheContainerIdentifierSeq	@empty()
(0040,051A)	ContainerDescription	@remove()
(0040,0551)	SpecimenIdentifier	@remove()
(0040,0554)	SpecimenUID	@empty()
(0040,0555)	AcquisitionContextSeq	@remove()
(0040,0562)	IssuerOfTheSpecimenIdentifierSeq	@remove()
(0040,0600)	SpecimenShortDescription	@remove()
(0040,0602)	SpecimenDetailedDescription	@remove()
(0040,0610)	SpecimenPreparationSeq	@remove()
(0040,06FA)	SlideIdentifier	@remove()
(0040,1001)	RequestedProcedureID	@remove()
(0040,1002)	ReasonForTheRequestedProcedure	@remove()
(0040,1004)	PatientTransportArrangements	@remove()
(0040,1005)	RequestedProcedureLocation	@remove()
(0040,100A)	ReasonforRequestedProcedureCodeSeq	@remove()
(0040,1010)	NamesOfIntendedRecipientsOfResults	@remove()
(0040,1011)	IntendedRecipientsOfResultsIdentificationSeq	@remove()
(0040,1101)	PersonIdentificationCodeSeq	@remove()
(0040,1102)	PersonAddress	@remove()
(0040,1103)	PersonTelephoneNumbers	@remove()
(0040,1104)	PersonTelecomInformation	@remove()
(0040,1400)	RequestedProcedureComments	@remove()
(0040,2001)	ReasonForTheImagingServiceRequest	@remove()
(0040,2008)	OrderEnteredBy	@remove()
(0040,2009)	OrderEntererLocation	@remove()
(0040,2010)	OrderCallbackPhoneNumber	@remove()
(0040,2011)	OrderCallbackTelecomInformation	@remove()
(0040,2016)	PlacerOrderNumber	@remove()
(0040,2017)	FillerOrderNumber	@empty()

(0040,2400)	ImagingServiceRequestComments	@empty()
(0040,3001)	ConfidentialityPatientData	@remove()
(0040,4005)	SPSStartDateAndTime	@remove()
(0040,4008)	ScheduledProcedureStepExpirationDateTime	@remove()
(0040,4010)	SPSModificationDateandTime	@remove()
(0040,4011)	ExpectedCompletionDateAndTime	@remove()
(0040,4023)	RefGPSPSTransactionUID	@remove()
(0040,4025)	ScheduledStationNameCodeSeq	@remove()
(0040,4027)	ScheduledStationGeographicLocationCodeSeq	@remove()
(0040,4028)	PerformedStationNameCodeSeq	@remove()
(0040,4030)	PerformedStationGeographicLocationCodeSeq	@remove()
(0040,4034)	ScheduledHumanPerformersSeq	@remove()
(0040,4035)	ActualHumanPerformersSeq	@remove()
(0040,4036)	HumanPerformerOrganization	@remove()
(0040,4037)	HumanPerformerName	@remove()
(0040,4050)	PerformedProcedureStepStartDateTime	@remove()
(0040,4051)	PerformedProcedureStepEndDateTime	@remove()
(0040,4052)	ProcedureStepCancellationDateTime	@remove()
(0040,A027)	VerifyingOrganization	@remove()
(0040,A073)	VerifyingObserverSeq	@empty()
(0040,A075)	VerifyingObserverName	@remove()
(0040,A078)	AuthorObserverSeq	@empty()
(0040,A07A)	ParticipantSeq	@remove()
(0040,A07C)	CustodialOrganizationSeq	@remove()
(0040,A088)	VerifyingObserverIdentificationCodeSeq	@remove()
(0040,A123)	PersonName	@remove()
(0040,A124)	UID	@empty()
(0040,A171)	ObservationUID	@remove()
(0040,A172)	RefObservationUIDTrial	@remove()

(0040,A192)	ObservationDateTrial	@remove()
(0040,A193)	ObservationTimeTrial	@remove()
(0040,A307)	CurrentObserverTrial	@remove()
(0040,A352)	VerbalSourceTrial	@remove()
(0040,A353)	AddressTrial	@remove()
(0040,A354)	TelephoneNumberTrial	@remove()
(0040,A358)	VerbalSourceIdentifierCodeSequenceTrial	@remove()
(0040,A402)	ObservationSubjectUIDTrial	@remove()
(0040,A730)	ContentSeq	@remove()
(0040,DB0C)	TemplateExtensionOrganizationUID	@remove()
(0040,DB0D)	TemplateExtensionCreatorUID	@remove()
	ContainerComponentID	@remove()
(0050,001B)	DeviceDescription	@remove()
(0050,0020)	LongDeviceDescription	@remove()
(0050,0021)	TrackingUID	@remove()
(0062,0021)	GraphicAnnotationSeq	@remove()
(0070,0001)	PresentationCreatorName	@remove()
(0070,0084)	ContentCreatorIdentificationSeq	@empty()
(0070,0086)	FiducialUID	@remove()
(0070,031A)	PresentationDisplayCollectionUID	@remove()
(0070,1101)	PresentationSequenceCollectionUID	@remove()
(0070,1102)	StorageMediaFileSetUID	@remove()
(0088,0140)	IconImageSeq	@remove()
(0088,0200)	TopicTitle	@remove()
(0088,0904)	TopicSubject	@remove()
(0088,0906)	TopicAuthor	@remove()
(0088,0910)	TopicKeyWords	@remove()
(0088,0912)	DigitalSignatureUID	@remove()
(0400,0100)	RefDigitalSignatureSeq	@remove()

(0400,0402)	RefSOPInstanceMACSeq	@remove()
(0400,0403)	MAC	@remove()
(0400,0404)	ModifiedAttributesSeq	@remove()
(0400,0550)	OriginalAttributesSeq	@remove()
(0400,0561)	InstanceOriginStatus	@remove()
(0400,0600)	TextString	@remove()
(2030,0020)	LabelText	@remove()
(2200,0002)	BarcodeValue	@empty()
(2200,0005)	RefFrameOfReferenceUID	@empty()
(3006,0024)	RelatedFrameOfReferenceUID	@remove()
(3006,00C2)	FirstTreatmentDate	@remove()
(3008,0054)	MostRecentTreatmentDate	@remove()
(3008,0056)	SourceSerialNumber	@remove()
(3008,0105)	TreatmentDate	@remove()
(3008,0250)	TreatmentTime	@remove()
(3008,0251)	RTPlanLabel	@remove()
(300A,0002)	RTPlanName	@remove()
(300A,0003)	RTPlanDescription	@remove()
(300A,0004)	RTPlanDate	@remove()
(300A,0006)	RTPlanTime	@remove()
(300A,0007)	PrescriptionDescription	@remove()
(300A,000E)	DoseReferenceUID	@remove()
(300A,0013)	DoseReferenceDescription	@remove()
(300A,0016)	FractionGroupDescription	@remove()
(300A,0072)	RefDoseReferenceUID	@remove()
(300A,0083)	TreatmentMachineName	@remove()
(300A,00B2)	BeamDescription	@remove()
(300A,00C3)	BolusDescription	@remove()
(300A,00DD)	FixationDeviceDescription	@remove()

(300A,0196)	ShieldingDeviceDescription	@remove()
(300A,01A6)	SetupTechniqueDescription	@remove()
(300A,01B2)	SourceManufacturer	@remove()
(300A,0216)	CompensatorDescription	@remove()
(300A,02EB)	TreatmentPositionGroupLabel	@remove()
(300A,0608)	TreatmentPositionGroupUID	@remove()
(300A,0609)	RTAccessoryHolderSlotID	@remove()
(300A,0611)	RTAccessoryDeviceSlotID	@remove()
(300A,0615)	RadiationDoseIdentificationLabel	@remove()
(300A,0619)	RadiationDoseInVivoMeasurementLabel	@remove()
(300A,0623)	RTToleranceSetLabel	@remove()
(300A,062A)	PatientSetupUID	@remove()
(300A,0650)	EquipmentFrameOfReferenceDescription	@remove()
(300A,0676)	RadiationGenerationModeLabel	@remove()
(300A,067C)	RadiationGenerationModeDescription	@remove()
(300A,067D)	ReasonForOmissionDescription	@remove()
(300C,0113)	ReviewerName	@remove()
(300E,0008)	ConceptualVolumeUID	@remove()
(3010,0006)	RefConceptualVolumeUID	@remove()
(3010,000B)	ConceptualVolumeCombinationDescription	@remove()
(3010,000F)	ConstituentConceptualVolumeUID	@remove()
(3010,0013)	SourceConceptualVolumeUID	@remove()
(3010,0015)	ConceptualVolumeDescription	@remove()
(3010,0017)	DeviceAlternateIdentifier	@remove()
(3010,001B)	DeviceLabel	@remove()
(3010,002D)	RefFiducialsUID	@remove()
(3010,0031)	UserContentLabel	@remove()
(3010,0033)	UserContentLongLabel	@remove()
(3010,0034)	EntityLabel	@remove()

(3010,0035)	EntityName	@remove()
(3010,0036)	EntityDescription	@remove()
(3010,0037)	EntityLongLabel	@remove()
(3010,0038)	RTTreatmentPhaseUID	@remove()
(3010,003B)	ManufacturerDeviceIdentifier	@remove()
(3010,0043)	IntendedPhaseStartDate	@remove()
(3010,004C)	IntendedPhaseEndDate	@remove()
(3010,004D)	RTPrescriptionLabel	@remove()
(3010,0054)	RTTreatmentApproachLabel	@remove()
(3010,0056)	RTPhysicianIntentNarrative	@remove()
(3010,005A)	ReasonForSuperseding	@remove()
(3010,005C)	PriorTreatmentDoseDescription	@remove()
(3010,0061)	DosimetricObjectiveUID	@remove()
(3010,006E)	RefDosimetricObjectiveUID	@remove()
(3010,006F)	TreatmentSite	@remove()
(3010,0077)	TreatmentTechniqueNotes	@remove()
(3010,007A)	PrescriptionNotes	@remove()
(3010,007B)	FractionationNotes	@remove()
(3010,007F)	PrescriptionNotesSeq	@remove()
(3010,0081)	ArbitraryRetired	@remove()
(4000,0010)	ArbitraryCommentsRetired	@remove()
(4000,4000)	ResultsIDIssuer	@remove()
(4008,0042)	InterpretationRecorder	@remove()
(4008,0102)	InterpretationTranscriber	@remove()
(4008,010A)	InterpretationText	@remove()
(4008,010B)	InterpretationAuthor	@remove()
(4008,010C)	InterpretationApproverSeq	@remove()
(4008,0111)	PhysicianApprovingInterpretation	@remove()
(4008,0114)	InterpretationDiagnosisDescription	@remove()

(4008,0115)	ResultsDistributionListSeq	@remove()
(4008,0118)	DistributionName	@remove()
(4008,0119)	DistributionAddress	@remove()
(4008,011A)	InterpretationIDIssuer	@remove()
(4008,0202)	Impressions	@remove()
(4008,0300)	ResultsComments	@remove()
(4008,4000)	DigitalSignaturesSeq	@remove()
(FFFA,FFFA)	DataSetTrailingPadding	@remove()
		@remove()

Anexo D. Cabecera archivo NifTI

TYPE	NAME	OFFSET	SIZE	DESCRIPTION
int	sizeof_hdr	0B	4B	Size of the header. Must be 348 (bytes).
char	data_type[10]	4B	10B	Not used; compatibility with ANALYZE.
char	db_name[18]	14B	18B	Not used; compatibility with ANALYZE.
int	extents	32B	4B	Not used; compatibility with ANALYZE.
short	session_error	36B	2B	Not used; compatibility with ANALYZE.
char	regular	38B	1B	Not used; compatibility with ANALYZE.
char	dim_info	39B	1B	Encoding directions (phase, frequency, slice).
short	dim[8]	40B	16B	Data array dimensions.
float	intent_p1	56B	4B	1st intent parameter.
float	intent_p2	60B	4B	2nd intent parameter.
float	intent_p3	64B	4B	3rd intent parameter.
short	intent_code	68B	2B	NIFTI intent.
short	datatype	70B	2B	Data type.
short	bitpix	72B	2B	Number of bits per voxel.
short	slice_start	74B	2B	First slice index.
float	pixdim[8]	76B	32B	Grid spacings (unit per dimension).
float	vox_offset	108B	4B	Offset into a .nii file.
float	scl_slope	112B	4B	Data scaling, slope.
float	scl_inter	116B	4B	Data scaling, offset.
short	slice_end	120B	2B	Last slice index.
char	slice_code	122B	1B	Slice timing order.
char	xyzt_units	123B	1B	Units of pixdim[1..4].
float	cal_max	124B	4B	Maximum display intensity.

TYPE	NAME	OFFSET	SIZE	DESCRIPTION
float	cal_min	128B	4B	Minimum display intensity.
float	slice_duration	132B	4B	Time for one slice.
float	toffset	136B	4B	Time axis shift.
int	glmax	140B	4B	Not used; compatibility with ANALYZE.
int	glmin	144B	4B	Not used; compatibility with ANALYZE.
char	descrip[80]	148B	80B	Any text.
char	aux_file[24]	228B	24B	Auxiliary filename.
short	qform_code	252B	2B	Use the quaternion fields.
short	sform_code	254B	2B	Use of the affine fields.
float	quatern_b	256B	4B	Quaternion b parameter.
float	quatern_c	260B	4B	Quaternion c parameter.
float	quatern_d	264B	4B	Quaternion d parameter.
float	qoffset_x	268B	4B	Quaternion x shift.
float	qoffset_y	272B	4B	Quaternion y shift.
float	qoffset_z	276B	4B	Quaternion z shift.
float	srow_x[4]	280B	16B	1st row affine transform
float	srow_y[4]	296B	16B	2nd row affine transform.
float	srow_z[4]	312B	16B	3rd row affine transform.
char	intent_name[16]	328B	16B	Name or meaning of the data.
char	magic[4]	344B	4B	Magic string.
Total size			348B	