Project title: European Federation for Cancer Images

Project acronym: EUCAIM

Grant Agreement: 101100633

Call identifier: DIGITAL-2022-CLOUD-AI-02

D5.6. Minimum Data Federation and Interoperability Framework

Author(s): Mirna El Ghosh (LIMICS), Christel Daniel (LIMICS), Catherine Duclos (LIMICS), Xavier Tannier (LIMICS), Ferdinand Dhombres (LIMICS), Valia Kalokyri (FORTH), Manolis Tsiknakis (FORTH), Laure Saint Aubert (MEDEX), Celia Martin Vicario (QUIBIM), Alejandro Vergara (QUIBIM), Jose Munuera (QUIBIM), Eirini Kaldeli (MAG), Gianna Taskou (MAG), Irene Marin (HULAFE), Pedro Miguel Martínez-Gironés (HULAFE), Olga Giraldo (DKFZ), Wahyu Wijaya Hadiwikarta (DKFZ), Hanna Leisz (DKFZ), Clara Meinzer (DKFZ), Ignacio Blanquer (UPV), Esther Bron (Health-RI, Erasmus MC), Alexander Harms (Health-RI), Carles Hernandez-Ferrer (BSC), David Rodriguez Gonzalez (CSIC-IFCA), Marcel Koek (Erasmus MC)

Contributor(s): Marco Aiello (SYNLAB), Victor Sóñora Pombo (BAHIA), Alexandra Kosvyra (AUTH), Ioanna Chouvarda (AUTH), Dimitris Filos (AUTH), Dimitris Fotopoulos (AUTH), Alexandra Groth (Philips), Federica Cruciani (IFOM)**,** Nuno Cruz, Santiago Frid, Sebastiaan Huntjens, Katerina Nikiforaki, Maciej Bobowicz (GuMed), Michal Konso (GuMed), Jose Alejandro Matute Flores, Dario Livio Longo, Heimo Müller (BBMRI), Kurt Maicen (BBMRI)

Reviewer(s): Enola Knezevic (DKFZ), Tobias Kussel (DKFZ)

Date of delivery: 31/1/2025

Version: 1.0

Table of contents

1. Introduction

1.1 Document Purpose

This deliverable introduces the EUCAIM minimum data Federation and Interoperability Framework (min-FIF). It demonstrates its potential to support the efficient, secure, and standardized exchange of clinical and imaging data provided by disparate data holders, repositories, or infrastructures. EUCAIM min-FIF helps to establish baseline standards and guidelines that permit different types of federated data to be integrated while ensuring data privacy, protection, and usability. Thus, a minimal set of clinical and imaging data and the associated semantic and technical requirements, necessary for effective interoperability and collaboration, are specified. By focusing on the essentials, min-FIF ensures that data infrastructures can achieve seamless data exchange without significant complexity.

To maintain and support interoperability, the EUCAIM framework adheres to the European Interoperability Framework (EIF) interoperability model, where different interoperability layers have been covered, including legal, organizational, semantic, and technical. This deliverable focuses on the semantic and technical levels. While semantic interoperability ensures that the precise format (syntactic aspect) and meaning (semantic aspect) of exchanged data and information are preserved and understood throughout exchanges between parties, technical interoperability covers the applications and infrastructures linking systems and services. Different international standards and terminologies have been used in the context of EUCAIM to maintain the semantic level. Besides, common semantic models have been developed in EUCAIM, such as the EUCAIM hyper-ontology and Common Data Model (CDM), aiming to maximize semantic interoperability and support data harmonization and integration across disparate sources. Data preprocessing and interoperability tools, such as FAIR EVA[1] for data FAIRification and MITK Workbench[2] for data annotation , have also been employed to support technical interoperability. Moreover, the technical specifications for local node setup and dataset transfer to reference nodes have been defined. To ensure the establishment of EUCAIM data repositories and facilitate the integration of new data holders, the technical requirements are provided as guidelines covering the tools and services to be used and the workflows to be followed.

1.2 Document Scope

This deliverable outlines min-FIF's technical requirements and specifications, which are crucial to ensuring seamless data sharing and usability across disparate infrastructures and supporting data holders' integration into EUCAIM. These requirements are provided as guidelines covering the different levels of compliance with the EUCAIM data federation and interoperability framework (FIF): dataset cataloging, federated querying, and processing. The scope of this document is to provide a clear, concise set of minimum technical requirements for achieving interoperability and federation among disparate data holders/repositories/infrastructures and acts as a complementary document to the deliverable D4.4 “Final rules of participation”. Specifically, it addresses the following key aspects:

[1] https://github.com/IFCA-Advanced-Computing/FAIR_eva
[2] https://www.mitk.org/wiki/Downloads

- *Minimum interoperability requirements for the dataset metadata and guidelines for dataset preparation*: a set of fundamental standards crucial for ensuring that datasets can be seamlessly discovered, accessed, and integrated across diverse repositories and infrastructures.

- *Minimum interoperability requirements for the clinical and imaging data*: a set of essential guidelines that specify the mandatory clinical and imaging information provided by data holders (at the record/patient level) required for federated querying and processing. These requirements focus on essential data formatting, considering a limited set of standard formats to support basic federated queries and processing for data aggregation and retrieval from disparate sources.

- *Guidelines for supporting federated query and processing*: involve a series of steps and workflows planned to ensure data can be accurately and consistently prepared for joint analysis across disparate infrastructures/repositories. Data preparation includes data cleaning, quality, FAIRification, harmonization, annotation, de-identification, and transformation (ETL (Extract, Transform, Load) and mapping processes), with the required preprocessing tools or services, permitting seamless integration and usability of data.

By adhering to these minimum requirements and specifications, interoperability is ensured among diverse data holders, repositories, or infrastructures in the context of the EUCAIM framework. Another level of interoperability, which spans across the EUCAIM data federation and other data infrastructures, such as the EHDS (European Health Data Space), will be maintained in further works.

Based on the minimum data Federation and Interoperability Framework (min-FIF), we will continue to evolve and enrich the framework to fulfill the needs of oncology researchers and clinicians. We will gradually move towards the maximum data Federation and Interoperability Framework (max-FIF). Transitioning from min-FIF to max-FIF involves expanding the framework to support more complex and broader requirements and specifications for data integration, querying, processing, and sharing across diverse and disparate data holders, repositories, and infrastructures. This transition will support the capabilities of the data federation to meet the requirements of more diverse data sources and holders, including complex querying and processing, such as Machine Learning (ML) models on federated data.

2 The EUCAIM Data Federation and Interoperability Framework

2.1 Overview

This section outlines the EUCAIM data federation and interoperability framework (FIF) defined as a structured approach comprising standards and practices that promote integrating, sharing, and querying data across different sources, repositories, or infrastructures while ensuring data accessibility, standardization, and consistent usability. Establishing this framework is critical regarding the complexity and heterogeneity of data sources in the context of the EUCAIM platform, where heterogeneous clinical and imaging information needs to be harmonized, integrated, and aggregated to ensure the exploration of data collections, federated querying, and processing. The data federation and interoperability framework involves two key aspects: data

federation focuses on linking disparate sources, while interoperability permits seamless integration and communication of heterogeneous data across these sources.

Interoperability, the "I" in FAIR (Findability, Accessibility, Interoperability, Reusability) principles[3], crucial for overcoming various challenges of data integration in healthcare, aims to ensure unambiguous communication among various heterogeneous healthcare systems and efficient sharing of essential patient data in a standardized and meaningful way. In EUCAIM, which deals with a considerable amount of diverse data from different repositories/sources, two interoperability levels are distinguished: 1) interoperability among data holders, repositories, or infrastructures and 2) interoperability among the EUCAIM data overall federation and the EHDS. Maintaining interoperability across diverse data holders/repositories is essential but challenging in EUCAIM, which deals with a considerable amount of diverse data from different repositories/sources, requiring the definition of standards and structures and semantics to handle how the data are modeled and stored to avoid ambiguity and allows machines, artificial intelligence (AI) systems, or any information tools to deal with the data and metadata. Additionally, EUCAIM intends to incorporate diverse new data holders/repositories from different environments, such as research (e.g., secondary use data repositories) or real-world data (e.g., hospitals).

Thereby, the implementation of the EUCAIM framework faces different challenges due to the complexity of dealing with heterogeneous, complex, and disparate data sources, formats, and standards and the necessity to ensure data privacy, quality, and security while permitting federated querying and processing. Adhering to the European Interoperability Framework (EIF), it is essential to overcome those challenges in a structured approach. In EUCAIM, the different interoperability layers provided by the EIF interoperability model[4] (Figure 1) are covered, supporting seamless exchange and integration of different types of data across heterogeneous sources. The interoperability layers (Legal, Organisational, Semantic, and Technical) represent different levels or stages of integration among different systems, ensuring that data from heterogeneous sources can be integrated, exchanged, protected, and managed efficiently. While the *Legal* and *Organisational* interoperability layers focus on the legal value of exchanged information complying with European and national laws and the coordinating processes and shared responsibilities, *Semantic* and *technical* layers address the techniques and protocols used to ensure data are exchanged and shared unambiguously and efficiently across information systems. This deliverable focuses on the alignment with the semantic and technical layers. The legal and organisational interoperability requirements are addressed in D4.4 (Participation Rules).

[3] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
[4] https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf

Figure 1. Interoperability model adapted from the European Interoperability Framework (EIF).

Lastly, to be aligned with the policies and strategies of the European Health Data Space (EHDS), the EUCAIM Interoperability Framework will take into account the important steps towards the adoption of the European EHRxF Standards, resulting in better and safer healthcare, more resilient health systems, and a more competitive European digital health industry in the EHDS. EUCAIM will be especially aware of the efforts of the XSHARE project[5], aiming to expand the European EHRxF to share and effectively use health data within the EHDS. The EUCAIM Interoperability Framework will also take into account the HL7 Vulcan Retrieval of real-world data for clinical research[6] aiming to facilitate the integration of care and research activities through the use of HL7 FHIR interoperability standards such as the minimal Common Oncology Data Elements (mCODE)[7] for the cancer domain. The EUCAIM Interoperability Framework will also consider the specifications of the QUANTUM project[8] and develop a data quality and utility label for the EHDS.

2.2. Interoperability Layers

This section highlights two main interoperability layers (semantic and technical) covered within EUCAIM, complying with the European Interoperability Framework (EIF) interoperability model (Figure 1).

2.2.1 Semantic Interoperability

Semantic interoperability ensures that the precise format or structure (syntactic) and meaning (semantic) of exchanged data and information are unambiguously maintained and understood throughout exchanges between all parties. While the syntactic aspect describes the exact format and structure of the information to be exchanged, the semantic aspect refers to the meaning of data elements and their interactions (relationships). Common healthcare data standards, such as OHDSI-OMOP and HL7-FHIR, enable syntactic and semantic interoperability. They permit standardization and structuring of healthcare data, ensuring that information systems can

[5] https://xshare-project.eu/
[6] http://hl7.org/fhir/uv/vulcan-rwd/#overview
[7] https://build.fhir.org/ig/HL7/fhir-mCODE-ig/
[8] https://quantumproject.eu/

effectively communicate and exchange data. While OMOP focuses on data standardization, FHIR focuses on data exchange. FAIR-compliant terminologies and vocabularies ensure the data is structured, standardized, and accessible, enabling its use in diverse applications while maintaining consistency, accuracy, and reusability. Finally, ontologies are developed to support the unambiguous description of domain entities and their interactions, ensuring that the meanings of concepts are understandable by machines and humans, maximizing semantic interoperability. In what follows, we provide an overview of the healthcare data standards, terminologies and semantic models used in the EUCAIM framework to maintain semantic interoperability.

2.2.1.1 Common Healthcare Data Standards

In EUCAIM, two primary common healthcare data standards, OHDSI-OMOP and HL7-FHIR, which concern data storage and exchange, have been used to support syntactic and semantic interoperability. These standards have been adopted in the AI4HI network: ProCAncer-I and ChAImeleon have used OMOP, and INCISIVE and EuCanImage have adopted FHIR.

OHDSI-OMOP

The Observational Medical Outcomes Partnership Common Data Model (OMOP CDM)[9] facilitates the analysis of diverse observational datasets by standardizing their structure and representation. It provides a unified data model and employs standard terminologies, vocabularies, and coding schemes. This standardization enables consistent and systematic analyses using pre-built libraries of analytic tools that leverage the common data structure. Due to its flexibility and widespread adoption, the OMOP CDM is often regarded as a robust framework for harmonizing and sharing data, particularly in studies leveraging longitudinal data from electronic health records (EHRs).

The OMOP CDM organizes clinical data into various domains relevant to healthcare and research. Key domains include conditions, procedures, drug exposures, measurements, observations, and visits. Each domain corresponds to dedicated tables in the CDM, such as the CONDITION_OCCURRENCE table for recording patient diagnoses, the DRUG_EXPOSURE table for medication data, the MEASUREMENT table for laboratory results and vital signs, and the VISIT_OCCURRENCE table for patient encounters. These structured tables ensure clinical data is categorized consistently, supporting seamless integration and analysis across different datasets and institutions. Furthermore, the use of closed, standardized dictionaries for clinical concepts avoids the need for sites to explicitly share information to define a conceptual framework.

HL7 FHIR

Health Level 7[10] is a non-profit organization that develops health data interoperability and standardizes clinical, financial, and administrative information exchange between hospital information systems (HIS). HL7 defines a set of technical specifications integrated into the corpus

[9] https://www.ohdsi.org/data-standardization
[10] HL7, https://www.hl7.org/

of American (ANSI) and International (ISO) formal standards. Initially American, these specifications are exported and become an international standard for this type of application. Fast Healthcare Interoperability Resources[11] is a standard developed by HL7 describing data formats and elements (called “Resource”) as well as an application programming interface (API) for the exchange of information in the field of healthcare.

Several HL7 FHIR initiatives focus on enabling the secondary use of EHR data for research and public health through implementation guides such as the European XSHARE implementation guide (IG), HL7 Vulcan Retrieval of real-world data (RWD) for clinical research IG. The FAIR4Health project aims to promote the use of HL7 FHIR to support the health datasets FAIRification process. A recent scoping review stated that the most prominent domain in HL7 FHIR IG registries is the cancer (40 %) and that frameworks or platforms in this domain use HL7 FHIR IG to facilitate secondary data use in public health registries and clinical trials [12,13,14,15]. The most cited FHIR IGs are produced by HL7 International: HL7 Genomics Reporting followed by Common Oncology Data Elements (mCODE) IG.

a - **The XSHARE IG**: The XSHARE project aims to expand the European EHRxF to share and effectively use health data within the EHDS to empower individuals, health systems, and businesses. The XSHARE will promote the development of the XSHARE Yellow Button and toolbox for data portability under GDPR and support early adopters in getting the XSHARE label. In this context, the XSHARE project is building a harmonized core data set across health care, population health and clinical research relying on several international initiatives such as the ISO/TC 215 Health Informatics - International Patient Summary (IPS) document, the International Patient Access IG providing additional HL7 FHIR profiles with data elements and value sets and the US Core IG providing a set of HL7 FHIR profiles in support of the US Core Data for Interoperability (USCDI). Building on all three HL7 FHIR IGs (IPS, IPA, and US Core), the XSHARE IG provides access to the artifacts produced by the XSHARE project for supporting the Yellow Button (aka xShare button) capabilities.

b - **The HL7 Vulcan Retrieval of real-world data for clinical research IG**: HL7 Vulcan-RWD aims to facilitate integrating care and research activities to improve patient lives, reduce costs, and improve efficiency, using HL7 FHIR interoperability standards. This FHIR Acceleration Program develops FHIR resources needed to execute prioritized use cases of secondary use of “real-world” data, especially EHR data. The main goal of the HL7 Vulcan FHIR implementation

[11] FHIR, https://hl7.org/fhir/

[12] L. A. Pollack, S. Jones, W. Blumenthal, T. O. Alimi, D. E. Jones, J. D. Rogers. Population Health Informatics Can Advance Interoperability: National Program of Cancer Registries Electronic Pathology Reporting Project, JCO Clinical Cancer Informatics (4) (2020) 985–92.

[13] M. Murugan, L.J. Babb, C.O. Taylor, L.V. Rasmussen, R.R. Freimuth, E. Venner, et al., Genomic considerations for FHIR®; eMERGE implementation lessons, Journal of Biomedical Informatics 118 (2021) 103795.

[14] N. Zong, N. Ngo, D.J. Stone, A. Wen, Y. Zhao, Y. Yu, Leveraging Genetic Reports and Electronic Health Records for the Prediction of Primary Cancers: Algorithm Development and Validation Study, JMIR Med Inform 9 (5) (2021) 23586.

[15] M. Lambarki, J. Kern, D. Croft, C. Engels, N. Deppenwiese, A. Kerscher, et al., Oncology on FHIR: A Data Model for Distributed Cancer Research, Stud Health Technol Inform. 24 (278) (2021 May) 203–210, https://doi.org/10.3233/ SHTI210070. PMID: 34042895.

guide (IG) for retrieval of real-world data for clinical research is to help define a minimal set of clinical research FHIR resources and elements in an EHR that can be utilized in an interoperable and consistent manner for research or innovation purposes. The profiles detail the data elements needed to convey data of interest in clinical research. The guide defines the FHIR building blocks to meet use cases, which will eventually mature the minimal set of common resources and elements. It is being developed using an iterative use case approach. The Vulcan accelerator promotes additional projects exploring mappings needed to achieve different outcomes (e.g., FHIR to CDISC, FHIR to OMOP, etc.). Complementary initiatives provide additional specifications in particular domains of interest, e.g., mCODE in oncology.

c - **FHIR4FAIR IG**[16] : The FAIR4Health project aims to facilitate and encourage the health research community to reuse datasets derived from publicly funded research initiatives using the FAIR principles. The 'FAIRness for FHIR' project aims to provide guidance on how HL7 FHIR could be utilized as a common data model to support the health datasets FAIRification process. This first expected result is an HL7 FHIR Implementation Guide (IG) called FHIR4FAIR, covering how FHIR can be used to cover FAIRification in different scenarios. This IG aims to provide practical underpinnings for the FAIR4Health FAIRification workflow as a domain-specific extension of the GoFAIR process, while simplifying curation, advancing interoperability, and providing insights into a roadmap for health datasets FAIR certification

d - **Minimum Common Oncology Data Elements (mCODE) IG**: mCODE is an initiative to increase interoperability by assembling a core set of structured data elements for oncology electronic health records (EHRs). mCODE is a step towards capturing research-quality data from the treatment of all cancer patients. This would enable the treatment of every cancer patient to contribute to comparative effectiveness analysis of cancer treatments by allowing for easier methods of data exchange between health systems. The mCODE core model explicitly defines the real-world entities of the oncology domain and their semantic relations. This approach has helped to clarify or overcome the ambiguity and heterogeneity of how well-known terminologies/ontologies define essential clinical concepts, such as Disease and Morphology. In the context of EUCAIM, mCODE is the basis of the conceptual model for representing various cancer types, cancer stages, performance status metrics, and scales, as well as assessments (e.g., radiological assessments (ACR Reporting and Data Systems (RADS)).

2.2.1.2 Terminologies/Vocabularies

Standardized terminologies or vocabularies are crucial to maintain semantic interoperability in healthcare. They ensure seamless communication and exchange of information between healthcare professionals, researchers, and information systems. Moreover, they are essential for data consistency, research integrity, and healthcare quality and outcome improvement. In EUCAIM, various FAIR-compliant terminologies have been used to map healthcare information to standard resources, supporting semantic interoperability. These resources, previously defined in the AI4HI network, are determined depending on different clinical and imaging data types. For instance, the terminologies used for cancer conditions and histological types are SNOMED-CT

[16] https://confluence.hl7.org/pages/viewpage.action?pageId=91991234: https://hl7.org/fhir/uv/fhir-for-fair/

and ICD-O-3, and those used for procedures are SNOMED-CT, CPT4, and ICD10PCS, depending on the projects.

- **SNOMED-CT** (Systematized Nomenclature of Medicine Clinical Terms)[17] is an internationally recognized clinical terminology and coding system. It organizes medical concepts hierarchically, providing precise and standardized descriptions for healthcare information.
- **LOINC** (Logical Observation Identifiers Names and Codes)[18] is a standardized system for identifying and naming laboratory tests, observations, and clinical measurements.
- **ICD10** is the 10th revision of the International Classification of Diseases (ICD)[19], a medical classification list by the World Health Organization (WHO). It contains codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases.
- **ICD10PCS** (International Classification of Diseases, Tenth Revision, Procedure Coding System) is a standardized medical coding system that classifies procedures and surgeries.
- **CPT4** (Current Procedural Terminology)[20] is a uniform coding system that identifies medical services and procedures provided by physicians and other healthcare professionals.
- **ICD-O-3**, the WHO International Classification of Diseases for Oncology, 3rd edition[21], is an international standard classification of tumors and related diseases. It provides a specialized representation of cancer histology, topography, and behavior.
- **NAACCR** (North American Association of Central Cancer Registries)[22] is a data standard used to code data in the US Cancer Registries. It covers most cancer types and includes critical diagnostic features and high-level treatment classification used in cancer epidemiology.
- **Cancer Modifier** is generated by OMOP to describe the Diagnostic Modifiers of Cancer.
- **OMOP Genomic** is generated by OMOP to describe the genomic variants and mutations.

2.2.1.3 EUCAIM Hyper-Ontology, CDM, and DCAT-AP

In the EUCAIM framework, although common healthcare standards and models, such as OMOP/FHIR, have been adopted to standardize and structure clinical and biological information, semantic interoperability remained challenging due to the complexities of dealing with heterogeneous formats, standards, terminologies, and interpretations. An ontology integration approach is followed to maximize semantic interoperability and ensure seamless integration of heterogeneous information, including different types of data (clinical, biological, and imaging) in

[17] SNOMED Clinical Terms. Available from URL: https://www.snomed.org/
[18] LOINC, the international standard for identifying health measurements, observations, and documents. Available from URL: https://loinc.org/.
[19] https://icd.who.int/browse10/2019/en
[20] https://www.ama-assn.org/practice-management/cpt-current-procedural-terminology
[21] https://seer.cancer.gov/icd-o-3/
[22] https://www.naaccr.org/data-standards-data-dictionary/

a common semantic meta-model called EUCAIM hyper-ontology[23,24]. Using the hyper-ontology for the integration of heterogeneous data is advantageous in different aspects:

- **A common understanding of data**: The EUCAIM hyper-ontology is a structured and standardized vocabulary that ensures a consistent understanding of terms and concepts. This has reduced ambiguity and resolved the issue of co-existing terminologies, achieved by establishing syntactic and semantic mappings among various terminologies and vocabularies. Besides, the hyper-ontology is grounded on the mCODE specifications, supporting a conceptual clarity of oncology. The oncology essential entities and relationships have been analyzed and explicitly and formally represented in the ontology model.

- **Mapping to healthcare and imaging standards**: The hyper-ontology ensures syntactic mappings with OMOP/FHIR considering the OMOP domain and FHIR resource components for clinical and biological data. For imaging data, mappings with DICOM are maintained, linking the concepts to their respective DICOM tags and names. These mappings align the hyper-ontology concepts with healthcare and imaging standards and support interoperability, consistency, and effective data exchange across healthcare repositories. The mapping process has helped bridge the gap between healthcare standards and ensured that data is interpreted consistently across various systems.

- **Supporting semantic search**: The hyper-ontology organizes data at different granularity levels in structured semantic model rich in axiomatizations, reducing ambiguity and supporting precise semantic search and querying of information from disparate and heterogeneous sources. It also supports complex queries involving heterogeneous data by aggregating disparate knowledge collected from various sources while ensuring that the meaning of information is preserved across them.

Moreover, the EUCAIM framework defines a standardized Common Data Model (CDM), supporting consistent interpretation of clinical, biological, and imaging data. A terminology-binding process is in progress to ensure that the data elements represented in the EUCAIM CDM are semantically aligned with the knowledge (concepts and object/data properties) described in the hyper-ontology. This ensures a coherent interpretation and understanding of data between the hyper-ontology and CDM. For instance, the fields or attributes "*birthSex*" and "AgeAtDiagnosis" defined in the tables "*Patient*" and "*Primary Cancer Condition*", respectively, in the CDM are aligned with the object properties "*hasBirthSex*"/"*hasAgeAtDiagnosis*", which link the concepts "*Sex assigned at birth"* (eucaim:COM1001396)/"*Age at diagnosis"* (eucaim:COM1000131) with the classes "*Patient*" (eucaim:COM1001047) and "*Cancer Patient*" (eucaim:COM1001051), respectively, in the ontology model. Although the binding process is challenging due to the complexity of representing different oncology aspects and the associated semantic relations at various granularity levels in the hyper-ontology, the mapping of the minimum clinical and imaging information between the hyper-ontology and the CDM has been ensured, supporting semantic

[23] EUCAIM Hyper-Ontology, Accessed via Zenodo, https://doi.org/10.5281/zenodo.10777925.
[24] El Ghosh, et al. Towards Semantic Interoperability among Heterogeneous Cancer Data Models using a Layered Modular Hyper-Ontology. FOIS 2024.

interoperability and seamless integration of data in the context of the min-FIF. Tables 19, 20, and 21 (Section 6) highlight various examples of terminology binding for the mandatory or minimum clinical and imaging information. Mapping broader and more complex elements and structures will be considered in further works to maintain semantic compatibility and alignment between the hyper-ontology and CDM towards the max-FIF.

Additionally, DCAT-AP[25], a widely recognized metadata standard, has been adopted, with HealthDCAT-AP[26], a health-related extension of DCAT-AP, to specify the EUCAIM DCAT-AP to describe health-related datasets. EUCAIM DCAT-AP is an application profile that re-uses terms from base standards, but also adds more specificity by identifying mandatory, recommended and optional elements, and requires the use of specific controlled vocabularies to guarantee interoperability. In this deliverable, a set of mandatory dataset metadata properties have been provided in the context of the EUCAIM DCAT-AP (see Tables 2 to 6, Section 4).

More details about the EUCAIM DCAT-AP, CDM, and hyper-ontology are provided in deliverable D5.2.

2.2.2 Technical Interoperability

Technical interoperability covers the applications and infrastructures linking systems and services. Aspects of technical interoperability include interface specifications, interconnection services, data integration services, data presentation and exchange, and secure communication protocols. Technical interoperability is maintained by the use of formal technical specifications. For instance, the structured exchange of health data is supported by international standards development organizations (SDOs) such as Health Level Seven International (HL7) or Digital Imaging and Communications in Medicine (DICOM). HL7's Fast Healthcare Interoperability Resources (FHIR) not only provides semantic interoperability (see Section 2.2.1.1), but also specifies APIs for technical interoperability.

DICOM

The Digital Imaging and Communications in Medicine (DICOM) standard is a crucial framework for seamlessly exchanging and managing medical image data and associated information across different healthcare systems and devices. DICOM standardizes the structure and format of various elements within medical imaging, such as MRI, CT scans, X-rays, and ultrasounds. This standardization helps medical professionals to view, share, and analyze images accurately, regardless of the manufacturer, modality, or software used to acquire or process the images. DICOM permits this interoperability by combining defined data structures, encoding rules, and network communication protocols. Central to the standard is a DICOM Information Object, which stores pertinent information like patient demographics, acquisition parameters, and image pixel data. These attributes are structured consistently, making the data understandable and usable across platforms. Furthermore, DICOM specifies communication protocols that allow devices to connect, exchange messages, and transmit image data over networks. This is essential for systems like Picture Archiving and Communication Systems (PACS) to interact seamlessly. Over

[25] https://semiceu.github.io/DCAT-AP/releases/3.0.0/
[26] https://healthdcat-ap.github.io/

time, DICOM has evolved by adding new modules and supplements to incorporate technological advancements and changes in medical practices. These updates accommodate emerging imaging modalities, enhance security, and support the integration of medical imaging into the broader electronic health record (EHR) systems.

DICOM-SEG

DICOM-SEG is a crucial aspect of the DICOM standard, specifically for medical image segmentation. Segmentation involves precisely outlining regions of interest within medical images, such as tumors or organs, and DICOM-SEG provides a standardized format for storing and sharing this data. It allows medical professionals to delineate and quantify structures, aiding in treatment planning, disease diagnosis, and research. DICOM-SEG's interoperability ensures that segmentation data can be seamlessly exchanged between imaging devices and healthcare institutions, facilitating multi-center studies. Its utility lies in improving collaboration, research, and patient care by standardizing how segmented regions are represented within DICOM images.

NIfTI

The Neuroimaging Informatics Technology Initiative (NIfTI) standard is a fundamental framework for representing and sharing neuroimaging data and information. NIfTI revolves around a standardized file format, known as the NIfTI-1 format, which stores neuroimaging data such as MRI, fMRI, and PET scans and uses a single file with header and image data, making it more straightforward to manage and share data between different systems. While initially geared towards neuroimaging, NIfTI's adaptability extends to oncology research, particularly in scenarios involving brain tumors where it is leveraged to store annotations, such as tumor segmentations. These annotations are pivotal for delineating specific regions of interest, a crucial aspect of oncology tasks such as treatment planning and monitoring. However, although the NIfTI standard offers significant advantages for imaging data representation and sharing, DICOM-SEG is preferred over NIfTI for medical image segmentation due to its dedicated and standardized format, ensuring precise representation and interoperability of segmented regions within DICOM images[27].

3 Architecture of the EUCAIM Data Federation

EUCAIM focuses on a federated model in which data holders can contribute by connecting their sites to the federation services of the central hub of EUCAIM. EUCAIM offers two reference nodes that can host data from data holders that cannot provide the required service level needed for the access and processing of their data. Reference Nodes have Secure Processing Environments where data can be processed safely.

The central hub features a Dashboard that gives access to the Catalogue, the Distributed Processing, the Access Negotiation, the Helpdesk and the Federated Search services. Figure 2 shows a high level schema that denotes all the different entities, including Data Holders that contribute with Real World Data (RWD) through observational studies.

[27] Aiello M, Esposito G, Pagliari G, Borrelli P, Brancato V, Salvatore M. How does DICOM support big data management? Investigating its use in medical imaging community. Insights Imaging. 2021 Nov 8;12(1):164. doi: 10.1186/s13244-021-01081-8. PMID: 34748101; PMCID: PMC8574146

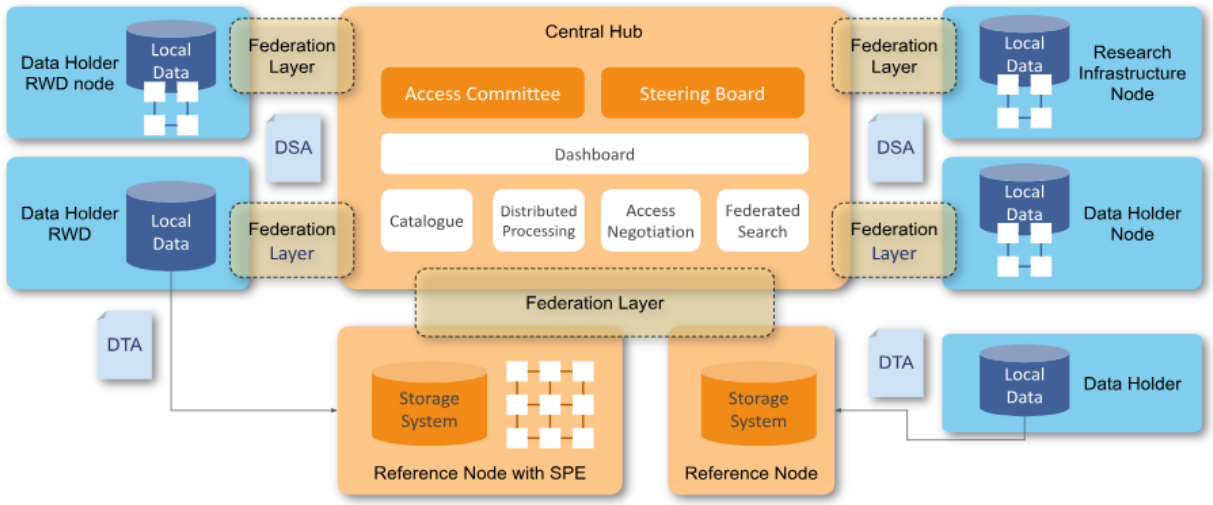


Figure 2: EUCAIM federation model. Central services are depicted in orange and data providers in blue. DTA stands for Data Transfer Agreement and DSA for Data Sharing Agreement.

The federation model of EUCAIM implies three different tier levels, already described in several documents, which define:

- Tier 1: Interoperability at the level of the collection/dataset metadata. If the data holder owns a catalogue, this catalogue should be searchable and accessible via FAIR Data Points[28] and DCAT. Data could be registered manually in the general catalogue otherwise.
- Tier 2: Interoperability at the level of the federated search. The data should be searchable so aggregated results can be retrieved according to searching criteria. The node should provide a query mediator to adapt the standard queries of the federated search to the local searching service format.
- Tier 3: Interoperability at the level of the distributed processing. The data must comply with the data model of EUCAIM and should be made available in an execution environment by means of a materializator component.

To federate a node those components described above have to be developed and deployed at the provider's side to achieve interoperability at the different tier levels. Figure 3 shows a schema of the EUCAIM central services, the data access flows, and the required federation tier.

[28] https://oldcatalogue-eucaim.grycap.i3m.upv.es/api/fdp/fdp_Catalog/aaaadedy5am5bvxematnxliaae/

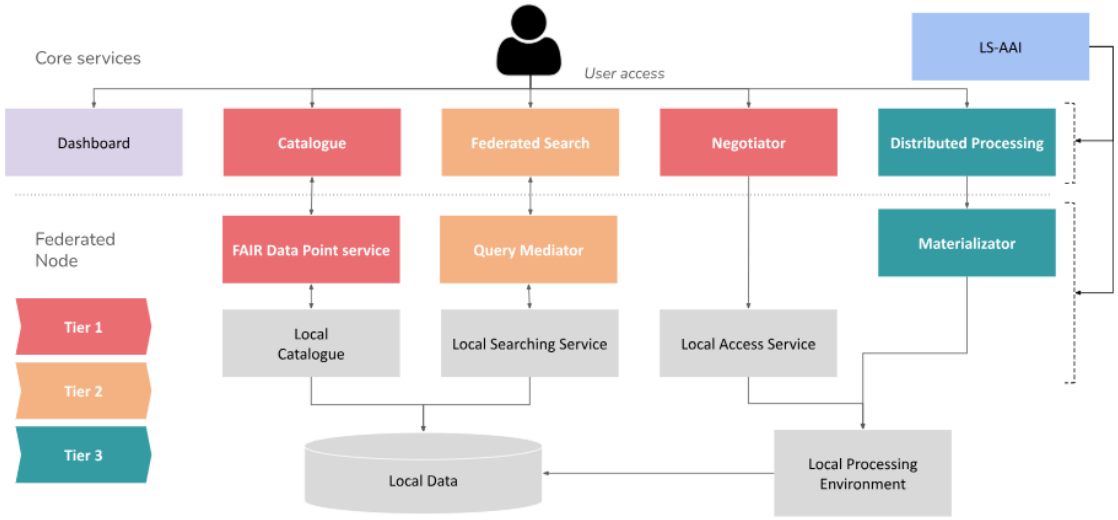


Figure 3: Components required at the federation node to interact with the federation services. Colours define the components involved in the interoperability at the different tiers (red for tier 1, orange for tier 2 and green for tier 3).

The following subsections describe in detail the current version of the architecture of EUCAIM and its main components.

3.1 Authentication and Authorization Infrastructure (AAI)

Only the Dashboard and the Catalogue in EUCAIM allow anonymous access, as they provide access to general information, onboarding processes and aggregated data. The rest of the services are only accessible for duly authenticated and authorised users, and the dashboard and the catalogue expose additional features to authenticated users.

Authentication and Authorisation in EUCAIM services is performed through the Life Science AAI (LS-AAI)[29]. Every core service is registered as one LS-AAI service, so the management of the authorisation can be centrally applied by means of Virtual Organization (VO) groups. As depicted in Figure 4, the researchers can use their own Identity Provider (IdP) credentials (if included in the eduGAIN Federation[30]) to authenticate themselves. Only users of the EUCAIM VO Group[31] are authorised to access the services and EUCAIM VO membership is manually validated.

[29] https://lifescience-ri.eu/ls-login/
[30] https://edugain.org/
[31] Enrollment URL for the EUCAIM VO Group https://signup.aai.lifescience-ri.eu/fed/registrar/?vo=lifescience&group=communities_and_projects:EUCAIM

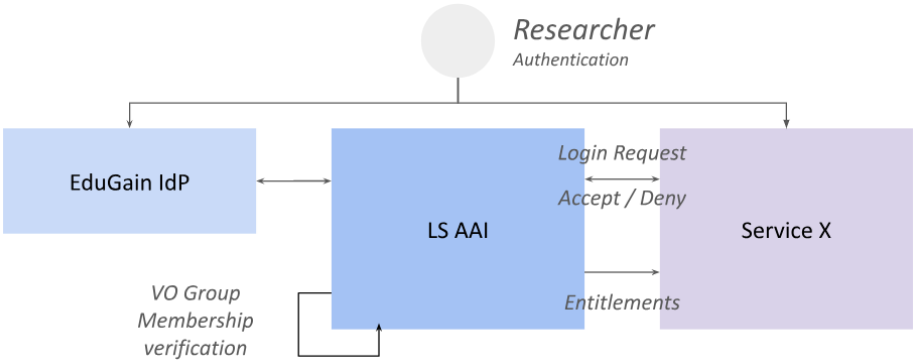


Figure 4: Basic interaction of the Core Services with the LS-AAI.

The following services are currently registered in the LS-AAI:

- Dashboard, accessible under dashboard.eucaim.cancerimage.eu, which verifies that the attribute urn:geant:lifescience-ri.eu:group:lifescience:communities_and_projects has the value EUCAIM#aai.lifescience-ri.eu. This membership attribute is also verified in other services.
- Federated Search, accessible under explorer.eucaim.cancerimage.eu, which is directly restricted to users in the EUCAIM VO group at the authentication on the LS-AAI service, using an OAuth Proxy in front of the Lens-based[32] exploration service.
- Negotiator UI, accessible under negotiator.eucaim.cancerimage.eu, and Negotiator backend, accessible under negotiator.eucaim.cancerimage.eu/api, which retrieve not only the membership attribute but also the roles with respect to the datasets. The responsible person for each dataset is registered in LS-AAI, as well as the negotiator admin role.
- Helpdesk, accessible under helpdesk.eucaim.cancerimage.eu, retrieving the membership attribute.
- Reference Node at UPV, accessible under eucaim-node.i3m.upv.es, through a federated authentication service (Keycloak in the case of this deployment) that interacts with the LS-AAI and retrieves the membership attribute. Additional authorisation configurations are defined at the level of the node Keycloak. It is important to outline that the reference node is a canonical implementation for other nodes in the federation.

3.2 Dashboard

The Dashboard is a web application that integrates the Graphical User Interfaces of the different components in a seamless environment with a common design. This website links to the core services as described in deliverable *D4.5 First Federated Core Services*: The Public Catalogue, Federated Search, Negotiator, the Authentication and Authorisation Infrastructure (AAI), Data Population Monitoring and the Helpdesk.

The architecture of the Dashboard application is simple, comprising two components, a NodeJs server with the application and the database for the persistence layer of the application. The

[32] https://github.com/samply/lens

application interacts with other services in EUCAIM. Figure 5 describes the components and the interactions.

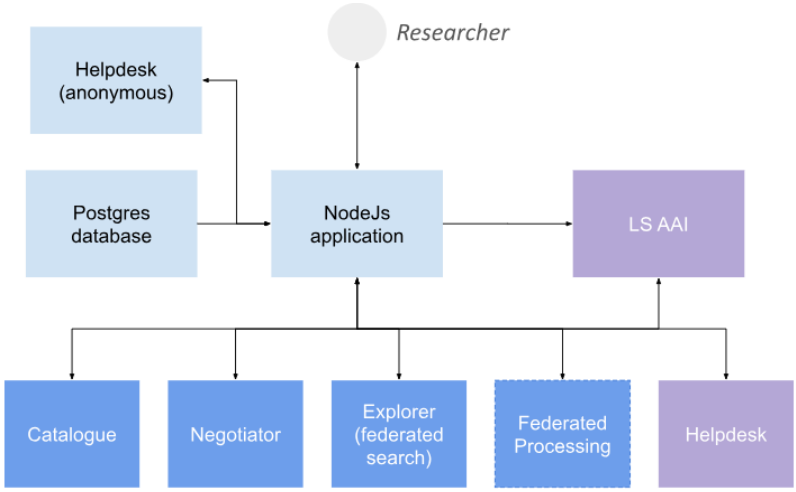


Figure 5: Dashboard architecture.

3.3 Public Catalogue

The public catalogue stores the metadata of data sets, and offers the researchers descriptive information about the available datasets, while displaying data characteristics as well as access conditions.

The metadata catalogue consists of the Molgenis emx2 platform[33] as a back-end service with a custom Javascript front-end which is based on prior catalogues[34]. The catalogue lists the datasets registered in the platform grouped into dataset series. Figure 6 shows the four components involved: the Molgenis front-end and backend, an Elasticsearch component for indexing data and a Postgres database for the persistence of all the information.

The metadata catalogue offers an API[35] through the Molgenis platform that facilitates the querying of the metadata in the catalogue. Just like the metadata which are made publicly accessible in the GUI, this information is also made accessible through the API.

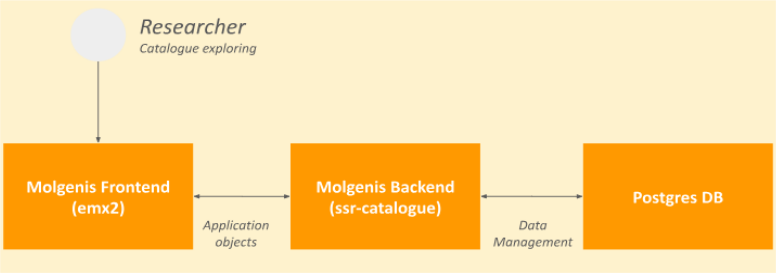


Figure 6: Molgenis emx2 architecture.

[33]van der Velde, K. Joeri, et al. "MOLGENIS research: advanced bioinformatics data software for non-bioinformaticians." Bioinformatics 35.6 (2019): 1076-1078. (https://doi.org/10.1093/bioinformatics/bty742)
[34]EIBIR catalogue: https://molgenis.eibir-edc.org/#/, BBMRI catalogue: https://directory.bbmri-eric.eu/#/catalogue
[35] API: https://catalogue.eucaim.cancerimage.eu/api/v2

To allow the dissemination of dataset metadata into multiple catalogues, without going through the trouble of repeatedly registering the datasets, the FAIR Data Point (FDP) protocol[36] is used to connect the catalogues. The FDP protocol uses the DCAT vocabulary and the DCAT-AP Health application profile, plus some additional fields defined for EUCAIM. Through an FDP endpoint that is exposed from the metadata catalogue, other catalogues can harvest the metadata in a standardised format. Figure 7 shows the architecture.keycloa

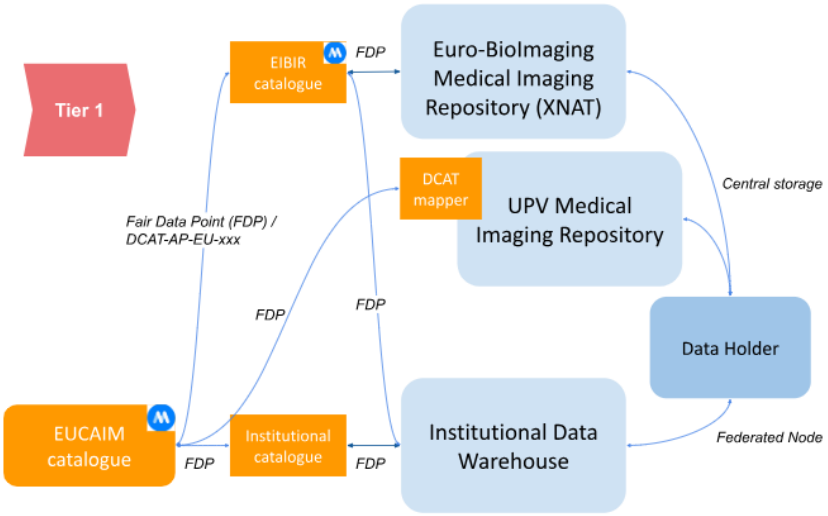


Figure 7: Catalogue harvesting architecture.

3.4 Federated Search

The Federated Search enables users to retrieve the number of subjects that fulfil specific criteria. The federated search is deployed in two different environments: The central core services, which consist of the front-end, its back-end, the federated query brokering system, and the certificate storage; and, on each providers' side, the query dispatcher, the store, and the data holders customised components to translate the query into the local format. To ensure secure and efficient communication between the central components and components in each providers' networks, the Samply.Beam[37] network communication middleware is employed.

The central core services are the following:

- **Lens:** The front-end application allowing researchers to construct complex search queries, and explore the search results.
- **Spot:** Lens backend, creating a task containing the query in abstracted form from Lens and sending it to the sites for execution using the Beam Proxy.
- **Beam Proxy:** Handling communication with the central Beam Broker, taking care of authentication, encryption, and signatures.
- **Beam Broker**: The central task broker.
- **Vault:** Used to store the certificates for each data holder's Beam Proxy registered in the system. The certificates are required for service authentication, encryption, and signing of the tasks.

[36] Fair Data Point specification: https://specs.fairdatapoint.org/fdp-specs-v1.2.html
[37] https://github.com/samply/beam

The data-holder side integrates the following components:

- **Beam Proxy:** Handling communication with the central Beam Broker, taking care of authentication, encryption, and signatures.
- **Focus:** The query dispatcher receiving Beam tasks using the local Beam Proxy, translating queries depending on the types of endpoints, executing them, and returning the results to Lens via Beam.
- **Local Data Management:** Different stores (DBMS) and custom components translating the queries (mediator).

An implementation of the mediator component for connecting the CHAIMELEON data holder has been integrated in the CHAIMELEON Dataset service. The implementation can be found in the service's Github repository[38]. ProCancer-I has also implemented a mediator for the datasets available from this initiative.

Figure 8 shows the architecture diagram of the above components and the interactions with the data holders.

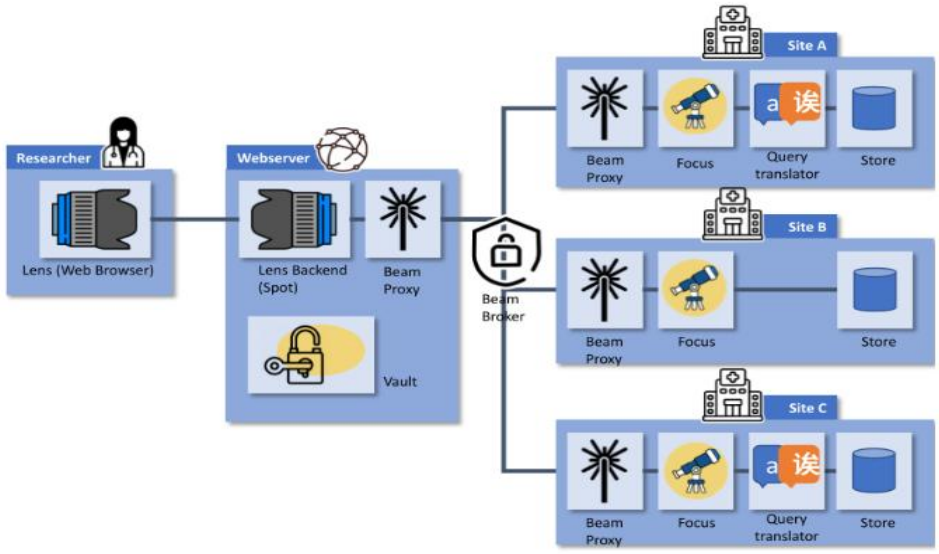


Figure *8: Architecture of the Federated Search system.*

3.5 Federated Access

Access requests to datasets are collected through the Negotiator. This is a service that collects the information about a data access request, which should be evaluated by the Access Committee. Access requests are currently triggered through the Catalogue, in the future it should be possible to directly request access through the Federated search as well. The user selects the set of datasets of interest and sends the request to the Negotiator. Then, the Negotiator presents a dynamic form that could depend on the type of the dataset and creates the full request. Further interactions with the data access process can be performed directly on the Negotiator to follow-on (or evaluate) the requests. In the case of observational studies, a special dataset ("Build an observational study with RWD") should be selected, leading to a specific access form and triggering a procedure of contacting the data holders.

[38] https://github.com/chaimeleon-eu/dataset-service?tab=readme-ov-file#integration-with-eucaim-federated-search

The Negotiator comprises three services: the front-end, which builds user interfaces according to the specifics of the requests, the backend service, which exposes the API of the Negotiator's functionality. Finally, a Postgres database persists the specifications of the access form and the information regarding special privileged roles for each dataset. The Negotiator periodically collects this information from the Catalogue. Figure 9 shows the interaction among the different components. Currently, all negotiations start from the catalogue which interacts through the "export" method with the Negotiator under the user "directory".

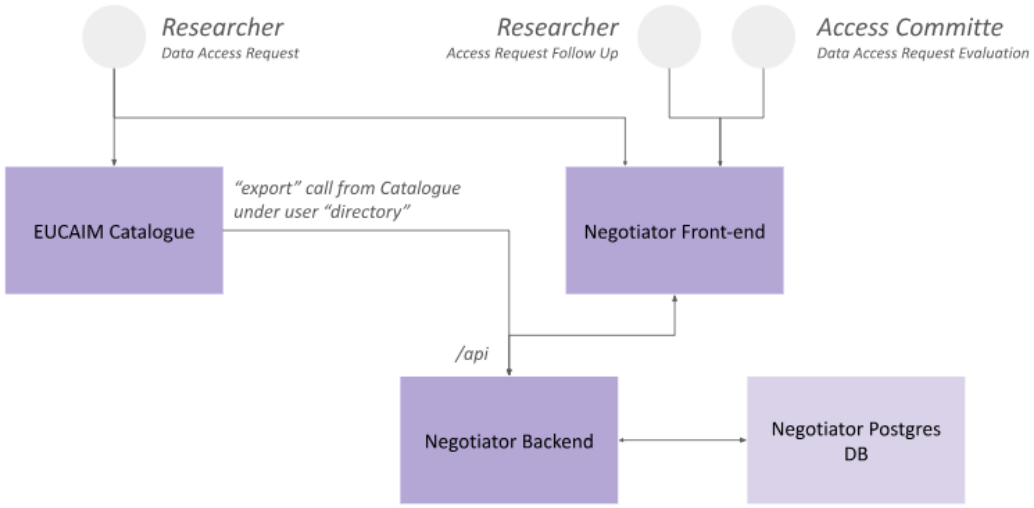


Figure 9*:* Architecture of the Negotiator Service.

3.6 Federated Processing

Depending on the workload and the data holding sites' compute capabilities, data processing and data analysis can be performed either purely locally, or using the Federated Processing service. This section briefly describes the architecture of the Federated Processing capabilities in EUCAIM.

The Federated Processing' architecture uses a pull model to fit restricted environments in which services are not exposed to the outside world (incoming connections), but can connect to external services (outgoing connections). The central services for Federated Processing manages an execution queue that is populated, and managed, by the Federated Execution Manager (FEM) backend but it is consumed by the clients (the nodes in the federation). The processing job description includes details regarding the execution environment (the local environment of a node), the reference to the datasets, and the execution parameters. Jobs are pulled by the clients at each processing node and run locally. The required data is made available to the job in a sandbox environment through the Data Materializator Tool (DMT), which is configured locally to adapt to local capabilities and which can call local tools to transform the data to the EUCAIM expected format.

Figure 10 presents the architecture of the Federated Processing.

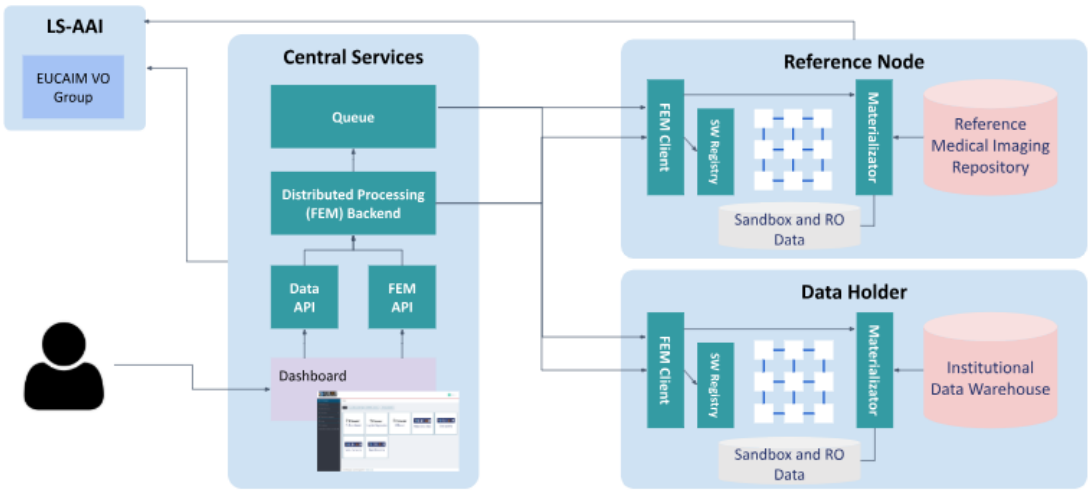


Figure 10: Federated Processing service.

As previously stated in deliverable D2.3, the minimum hardware requirement for federated nodes are presented in Table 1.

***Table 1**: Minimum hardware requirement for federated nodes*

Hardware	Option 1	Option 2	Notes
CPU	*Minimum Cores*: 16 >=1.8GHZ	*Minimum Cores*: 12 >=3.0Ghz	• If a GPU is not present, a server-grade, high core-count CPU is necessary for the Second Prototype. • If not comparable by cores, the ideal thread count is 24+.
RAM	64GB		• DDR5 is ideal. • ECC memory is highly recommended for stability.
Motherboard	4+ RAM Slot		• Make sure to double check the compatibility of selected CPUs with the Chipset of the motherboard. • In the case of DDR5, double check motherboard compatibility with DDR5.
Storage	• 521 GB SSD Drive for Operating System (Either NVMe M.2 PCI Gen4 or SATA III) • 1TB++ SATA III Drive (SSD or HDD) for local storage of medical data		• M.2, NVMe, Gen4 Drives are suggested for the OS • For data storage size, Data Holders (DH) are expected to plan their purchase depending on the size of the Data they will provide. 1TB is a minimum, with some DHs already planning for 2 TB + datasets. • For data storage, SSD are preferred for speed but are not mandatory.
Graphics card	NVIDIA Quadro	NVIDIA RTX 3XXX	• 12GB RAM+ is preferred. • Maximizing the amount of Tensor Cores is a priority, most recent GPUs will generally have higher Tensor Core counts. • Ampere and Volta architectures are preferred.

Operating System	Linux	● The latest version of any mainstream Linux distribution is acceptable: Ubuntu, Alpine or other. ● Windows is NOT acceptable, unless absolutely impossible for a DP to setup a Linux environment
Power Supply	**-**	● Each DH must make calculations depending on the hardware setup that will be selected to make sure that needed Wattage is covered and ideally exceeded to prepare for any future upgrades to the machine.
Internet	100mbps (baseline)	● Each DH must make best efforts to provide the best possible connection to their Node. Network performance will directly affect node stability and can invalidate AI training or prevent successful demonstrations of the platform.

These are the real minimum requirements necessary to use most of the software that will be on boarded. However, it is important to note that, over the lifespan of the project, new software and models may emerge requiring increased computational resources.

We recommend starting by querying the federated nodes to specify their technical capabilities and classify them accordingly. Some initial classification ideas include: no-GPU, GPU-low, GPU-mid, or GPU-high. These categories can encompass broader specifications such as the number of CPUs, RAM memory, and hard disk space to ensure nodes are appropriately categorized and utilized.

3.7 Local Data Nodes (Architectural Representation)

Figure 11 provides a high-level overview of the Local Nodes' architectural schema, showcasing its main local components and their interactions with the EUCAIM central services. Local data, including image data and, if relevant, clinical data should be processed following the workflows depicted in Figures 13, 15, and 19 in subsequent Sections of this document, depending on the Tier of the local node and whether it chooses to transfer its data to a Reference Node. After the data are processed so that they are in line with the EUCAIM standards, they have to be registered to the Public Catalogue (see Section 4.3). The local node may choose to transfer their datasets to the EUCAIM Reference Node (this possibility is shown with a dashed line in the Figure below).

Tier 2 nodes need to connect with the Federated Search Beam broker to support federated queries. To this end, they need to deploy the Beam Proxy and Focus Docker components (see Section 5.2). Moreover, they potentially need to implement a Mediator component that maps the queries and the responses between the formats expected by the federated search and the ones produced by the local search implementation. In addition to the Tier 2 requirements mentioned above, Tier 3 nodes must deploy and configure the FEM's client, that connects to the FEM manager located at the EUCAIM's central node, as well as the DMT, so that FEM can allocate the right data (see Section 6.2).

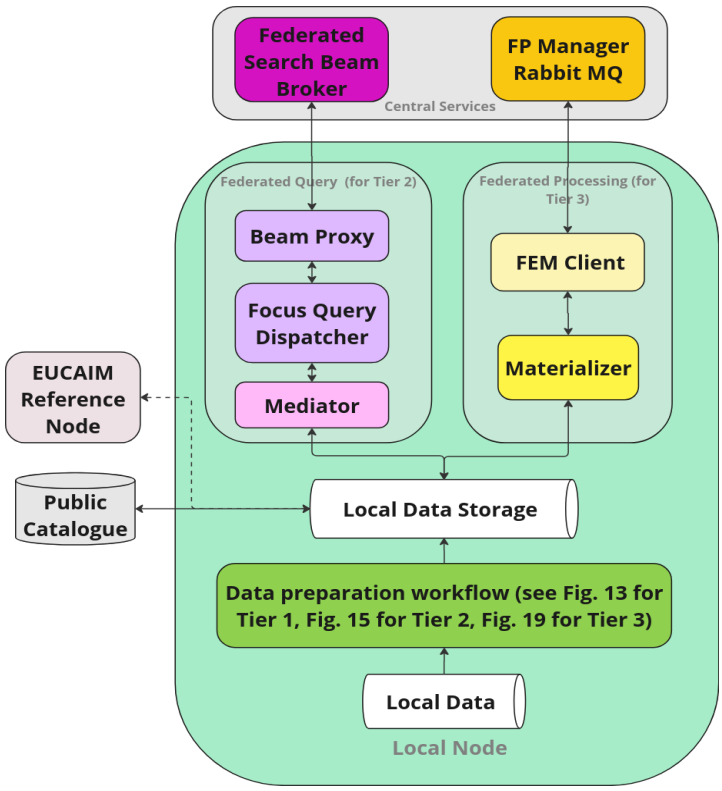


Figure 11: *Architectural overview of a Local Node*

4 Minimum Technical Requirements for Tier 1 Data Federation and Interoperability Framework (Dataset Cataloguing)

Tier 1 represents the lowest level of compliance with the EUCAIM Data Federation and Interoperability Framework. While the EUCAIM platform offers limited capabilities at this level, it serves as a critical starting point for all data holders to properly catalogue their datasets and make them accessible for discovery and exploration.

4.1 Minimum Interoperability Requirements for the Dataset Metadata (aggregated level metadata)

Tier 1 achieves interoperability at a dataset cataloguing level. Its main purpose is to standardize the definition, documentation and exchange of aggregated dataset metadata across the EUCAIM federation. This is achieved by adopting widely recognized metadata standards (i.e. DCAT-AP), use controlled vocabularies to prevent ambiguity, and facilitate automatic metadata exchange through FAIR data point services.

As described both in deliverables D5.1 and D5.2, an initial version of the EUCAIM DCAT Application Profile has been defined. The EUCAIM DCAT-AP is a specification based on DCAT-AP, as well as the recently published HealthDCAT-AP for describing health-related datasets that also comply with the European Health Data Space Regulation. EUCAIM DCAT-AP, as an application profile, re-uses terms from one or more base standards, adds more specificity by identifying mandatory, recommended and optional elements, and requires the use of specific controlled vocabularies to guarantee interoperability. In the context of this deliverable only the

minimum/mandatory elements will be outlined, and as such the following general requirements must be met:

- The mandatory requirements defined in the DCAT-AP v3.0 must be respected.
- The controlled vocabularies of the DCAT-AP v3.0 specification must be respected.
- The mandatory requirements defined in the HealthDCAT-AP specification should be respected. Please note that certain attributes not applicable in the EUCAIM context will be neglected (such as the HDAB). In addition, as HealthDCAT-AP is still not finalized and cardinalities are subject to change, all changes will get reflected in future deliverables.
- New EUCAIM domain-specific controlled vocabularies must be adopted based on the initial EUCAIM hyper-ontology specification defined in D5.2.

The following section describes the mandatory dataset metadata properties split into five distinct categories:

1. **General/Discovery metadata**
2. **Contact Details**
3. **Domain-specific metadata**
4. **Distribution metadata**
5. **Technical metadata**

All participating nodes in the EUCAIM data federation must provide the following information:

1. **General/Discovery:** A set of metadata properties that describe essential attributes of the dataset, facilitating its relevance, and applicability for various use cases. These properties provide insights into the content, scope, context, etc.

***Table 2**: EUCAIM DCAT Application Profile - Dataset Discovery*

Property	Description	Property IRI	Range	Cardinality	Usage Note	Example
title	A clear and concise name for the dataset.	`dct:title`	`rdfs:Literal`	1..n	This property can be repeated for multiple language versions of the title. The English version is mandatory.	`dct:title "Open Challenge Prostate Cancer V1"@en;`
description	A detailed description of the dataset's content, purpose, and scope.	`dct:description`	`rdfs:Literal`	1..n	This property can be repeated for multiple language versions of the description. The English version is mandatory.	`dct:description "This ProCAncer-I project imaging dataset contains a collection of patients with mpMRI examinations (T2ax, DWI, ADC) who have confirmed PCa at biopsy and/or prostatectomy."@en`

provenance	A statement about the lineage of a Dataset.	dct:provenance	dct:ProvenanceStatement	0..n	Information about how the data was created, or processed, including methodologies, tools, and protocols used.	dct:provenance [a dct:ProvenanceStatement; rdfs:label "This data is sourced from several existing datasets, including the Duke dataset, ParcTauli and TCGA datasets."@en];
intended Purpose	The primary objective for which the dataset was created.	dpv:hasPurpose	dpv:Purpose	1..n	A free text statement of the purpose of the processing of data or personal data.	dpv:hasPurpose[a dpv:Purpose ; dct:description " The primary objective of this dataset is the detection of prostate cancer with high accuracy both in peripheral and transitional zones to identify which men have cancer and those with no cancer."@en;] ;
imageCreationYear	A temporal period that the dataset covers. This corresponds to the year range that the actual (DICOM) images were created/acquired.	dct:temporal	dct:PeriodOfTime	1..n	This can be extracted from the DICOM acquisition date (0008,0022), if this has not been changed/removed in the anonymization process. If this is not available, an approximation should be added.	dct:temporal [a dct:PeriodOfTime; dcat:endDate "2023-12-31"^^<http://www.w3.org/2001/XMLSchema#date>; dcat:startDate "2021-01-01"^^<http://www.w3.org/2001/XMLSchema#date>];
geographicalCoverage	A geographic region that is covered by the Dataset.	dct:spatial	dct:Location	1..n	The EU country vocabulary is recommended for this attribute: https://publications.europa.eu/resource/authority/country, or alternatively the Geonames https://sws.geonames.org/	dct:spatial <http://publications.europa.eu/resource/authority/country/GRC>;

2. **Contact Details:** Contact details related to the dataset.

***Table 3**: EUCAIM DCAT Application Profile - Contacts*

Property	Description	Property IRI	Range	Cardinality	Usage Note	Example
contact Point	Contact information of the individual/managing organization of the Dataset for sending comments about the Dataset.	`dcat:contactPoint`	`vcard:Kind`	1..n	Contact information is limited to the contact email and/or the contact page. At least one of the two MUST be provided. In case the dataset is transferred to one of the reference nodes, the Data Access Committee will be designated as the contact point.	`dcat:contactPoint [ a vcard:Organization; vcard:fn "FORTH"; vcard:hasEmail <mailto:access-commitee@procancer-i.com> ];`
publisher	An entity (organisation) responsible for making the Dataset available. (Name and URL (landing page) of the organisation should be given)	`dct:publisher`	`foaf:Organization`	1..n	An entity (organisation) responsible for making the Dataset available.	`dct:publisher [ a foaf:Organization; locn:address [ a locn:Address; foaf:name "FORTH"; foaf:mbox <mailto:access-commitee@procancer-i.com>; foaf:homepage <https://forth.ics.gr>; ];];`
publisher Type	A type of organisation that makes the Dataset available.	`healthdcatap:publishertype`	`skos:Concept`	1..n	One of: Research Institute, Hospital or Healthcare System Repository, European project, Cancer screening program, Patient association, Data altruism organization, ERIC and EDIC.	`healthdcatap:publisherType <.../authority-table/publisher-type/ResearchInstitute>;`

3. **Domain-specific metadata:** A set of metadata properties that classify and describe the key domain-specific characteristics and compliance aspects of the dataset.

***Table 4**: EUCAIM DCAT Application Profile - Domain-specific metadata*

Prefix eucaim: <https://cancerimage.eu/ontology/EUCAIM#> (Hyper-Ontology IRI)

Property	Description	Property IRI	Range	Cardinality	Usage Note	Example
applicableLegislation	The legislation that mandates the creation or management of the Dataset.	`dcatap:applicableLegislation`	`rdfs:Resource`	1..n	The value must include the ELI of the EHDS Regulation. Multiple legislations may apply to the dataset.	`dcatap:applicableLegislation <http://data.europa.eu/eli/reg/2022/868/oj>;`
theme	A category of the dataset.	`dcat:theme`	`skos:Concept`	1..1	fixed to: http://publications.europa.eu/resource/authority/data-theme/HEAL	`dcat:theme <http://publications.europa.eu/resource/authority/data-theme/HEAL>;`
type	A type of the Dataset.	`dct:type`	`skos:Concept`	1..n	One of Original Dataset, Annotated Dataset, Processed Dataset. The value “PersonalData” will also be registered by default.	`dct:type a skos:Concept ; skos:prefLabel "Annotated Dataset"@en .`
age low	The minimum age of subjects within the dataset.	`eucaim:ageLow`	`rdfs:Integer`	1..1		`eucaim:ageLow "18" ^^xsd:int ;`
age high	The maximum age of subjects within the dataset.	`eucaim:ageHigh`	`rdfs:Integer`	1..1		`eucaim:ageHigh "18" ^^xsd:int ;`
birthsex	BirthSex of subjects in the dataset.	`eucaim:hasBirthSex`	`skos:Concept`	1..*	EUCAIM Controlled vocabulary: **Subclasses of “Sex assigned at birth”**	`eucaim:hasBirthSex eucaim:COM1001370 (Female)` `eucaim:COM1001366 (Male)`

						eucaim:COM1001288 (Unspecified)
number of studies	Total count of DICOM studies.	eucaim:nbrOfStudies	rdfs:Integer	1..1		eucaim:nbrOfStudies "8789" ^^xsd:int ;
number of subjects	Total count of unique individuals in the dataset.	eucaim:nbrOfSubjects	rdfs:Integer	1..1		eucaim:nbrOfSubjects "8237" ^^xsd:int ;
collection method	This attribute defines the scope of data aggregation within the dataset. It specifies how data records are organized based on different criteria, allowing users to understand the context in which the data was collected.	eucaim:collectionMethod	subproperty of dct:subject skos:Concept (fixed to a predefined set of values presented in D4.4)	1..n	EUCAIM Controlled vocabulary	eucaim:collectionMethod a skos:Concept ; skos:prefLabel "Only-Image"@en.
quality label	A statement related to quality of the Dataset, including rating, quality certificate as per the EHDS requirements.	dqv:hasQualityAnnotation	dqv:QualityCertificate	1..1	The set of certificate values are still in progress (to align with EHDS and QUANTUM initiatives).	dqv:hasQualityAnnotation [a dqv:QualityCertificate ; oa:hasTarget <https://…/dataset/123>; oa:hasBody <https://…/certificate>; oa:motivatedBy dqv:qualityAssessment];
legal basis	Legal basis used to justify processing of data or use of technology in accordance with a law.	dpv:hasLegalBasis	dpv:LegalBasis	1..n		dpv:hasLegalBasis [a dpv:LegalBasis ; dct:description "Deliberation no. 21/028 of february 18, 2021, last amended on june 18, 2021, relating to the communication of data to pseudonymized personal

						character relating to the healthdata of.. , as part of the EUCAIM project and the subsequent processing of personal data pseudonymised by…"@en;] ;
condition	The primary cancer condition of individuals in the dataset.	eucaim:hasCondition	skos:Concept	1..1	EUCAIM controlled vocabulary based on ICD-10 **subclasses of "Malignant neoplastic disease"**	eucaim:CLIN1000075 (Malignant neoplasm of prostate)
image modality	The set of modalities for the images in the dataset.	eucaim:hasImageModality	skos:Concept	1..n	EUCAIM controlled vocabulary based on Radlex: **subclasses of "Imaging Modality"**	eucaim:IMG1000022 (Magnetic Resonance Imaging)
image equipmentManufacturer	Manufacturer of the imaging device as it is defined in DICOM tag (0008,0070).	eucaim:hasEquipmentManufacturer	skos:Concept	1..n	EUCAIM controlled vocabulary based on Birnlex: **subclasses of "Manufacturer"**	eucaim:IMG1000047 (General Electric)
image body part	Anatomical areas captured in the images.	eucaim:hasImageBodyPart	skos:Concept	1..n	EUCAIM controlled vocabulary based on ICD-O3: **subclasses of "Body structure"**	eucaim:BP1000233 (Neck and chest)

4. **Dataset Distribution:** A set of metadata properties that describe various ways to access and interact with the dataset:
    a. Dataset Distribution: Describes how the dataset is made accessible.
    b. Dataset Sample: Provides subsets or representative examples of the dataset to facilitate evaluation and understanding. These samples could be a synthetic subset or solely exhibit the dataset's structure, i.e. human-readable structural metadata providing the properties or columns of the dataset schema. A sample distribution of the dataset is mandatory in case the dataset's access conditions is "Authorization to remotely process the datasets without the ability to access and visualise data, even remotely." Providing such a sample as a downloadable file can offer insights into the data's format, structure and possible set of values, aiding in understanding and utilisation while ensuring privacy and security.

***Table 5**: EUCAIM DCAT Application Profile - Dataset distribution*

Property	Description	Property IRI	Range	Cardinality	Usage Note	Example
accessURL	A URL that gives information about accessing the dataset.	`dcat:accessURL`	rdfs:Resource	1..1	In EUCAIM, this is the URL of the negotiator service for the specific dataset.	dcat:accessURL <https://negotiator.eucaim.cancerimage.eu/collection/a96b56cd-59d4-444a-8e59-32a7fb0d7dea> ;
accessRights	The accessRights of the dataset.	`dct:accessRights`	dct:RightsStatement	1..1	one of the public, non-public, restricted: https://publications.europa.eu/resource/authority/access-right	dcterms:accessRights <http://publications.europa.eu/resource/authority/access-right/NON-PUBLIC> ;
accessConditions	A statement about the conditions of access and usage of the dataset.	`dct:rights`	dct:RightsStatement	1..1	fixed to a predefined set of values: "Authorisation to download the datasets" "Authorisation to access, view and process in-situ the datasets" "Authorisation to remotely process the datasets without the ability to access and visualise data, even remotely."	dct:rights [a dct:RightsStatement; rdfs:label "Authorisation to access, view and process in-situ the datasets"@en];

imageSize (in GB)	The total size of all Distributions in the dataset, which is mainly the image size.	`dcat:byteSize`	xsd:decimal	0..1		dcat:byteSize "325"^^xsd:decimal
format	The file format of the Distributions included in the Dataset.	`dct:format`	dct:MediaTypeOrExtent (IANA Media Types)	0..n	Imaging data: the imaging format of the images in your dataset (e.g. DICOM, Nifti), Annotation data: the format of the annotations (e.g. DICOM-SEG, Nifti), if available. Clinical data: the format of the available clinical data (e.g. CSV, XLS, JSON, parquet).	dct:format <https://www.iana.org/assignments/media-types/application/dicom>;
sample	A sample distribution of the dataset.	`adms:sample`	dcat:Distribution	0..n	Mandatory in case the access condition of the dataset is: “Authorisation to remotely process the datasets without the ability to access and visualise data, even remotely”	adms:sample [a dcat:Distribution ; dct:description "Synthetic data of the X Dataset"@en; dcat:downloadURL <https://github.com/CAVDgit/EHDS2_UC_Scensano/blob/main/use_case_1_synthetic_data_10K_individuals.csv>; dcat:mediaType <http://www.iana.org/assignments/media-types/text/tab-separated-values> ;];

5. **Technical Metadata:** A set of metadata properties used primarily for metadata management within the EUCAIM federation.

***Table 6**: EUCAIM DCAT Application Profile - Technical Metadata*

Property	Description	Property IRI	Range	Cardinality	Usage Note	Example
identifier	A unique identifier for the dataset, i.e. the URI in the	`dct:identifier`	rdfs:Literal	1..1	The use of persistent dereferenceable	dct:identifier "https://catalogue.eucaim.cancerimage.eu/#/co

	context of the EUCAIM Public Catalogue. ((in compliance with the findability aspect of the FAIR principles))				URIs is mandatory.	llection/1a1a6653-975a-4a0a-a79b-b2bfc7317119"^^<http://www.w3.org/2001/XMLSchema#anyURI>;
version	The version of the dataset.	`dcat:version`	rdfs:Literal	1..1	in SemVer or CalVer format	dcat:version "20231122"
interoperabilityTier	The EUCAIM data federation and interoperability tier the specific dataset belongs to.	`adms:interoperabilityLevel`	skos:concept (EUCAIM controlled vocabulary)	1..1	One of “Tier 1”, “Tier 2”, “Tier 3”, “Tier 1A+”, “Tier 1C+”, “Tier 2A+”,”Tier 2C+”, “Tier 3A+”, “Tier3C+“.	adms:interoperabilityLevel a skos:Concept ; skos:prefLabel "Tier 1"@en.

4.2 Minimum Requirements for the Clinical and Imaging Data (at record level/patient level)

Although Tier 1 only requires the publication and visualization of datasets - mandating the standardization of the dataset descriptions - there are specific requirements for the imaging and clinical data at a patient level (whose dataset descriptions are registered in the EUCAIM Catalogue) that must be met to successfully transition to a higher level/tier in the EUCAIM Data Federation.

Imaging Data

- *Imaging data must be in DICOM format, and associated annotations and segmentations (when available) desired in DICOM-SEG format.*
- *At least one DICOM study per patient must be present. For positive or diagnostic cases, the image scan (DICOM study) where the tumor (or the lesion) was first detected must be present. For negative screening and control groups, the image scan (DICOM study) corresponding to the time of screening or control group participation must be provided.*
- *Any image modality is accepted.*

Imaging data must be provided in DICOM format, as it is the interoperable standard within the EUCAIM infrastructure. In exceptional cases where NIfTI format images are provided, these will only be considered if the Data Holder (DH) is unable to retrieve the original DICOM files from which the NIfTI images were derived. In such cases, DHs are responsible for ensuring the minimum requirements for anonymization, risk analysis, and quality through their own tools and procedures. However, NIfTI-based datasets will be deprioritized during ingestion due to their limited interoperability.

If the DH is able to convert NIfTI images back to valid DICOM format images and supplement them with all minimum and relevant clinical information that would have been extracted from the original DICOM files (as encouraged by EUCAIM), the dataset can be upgraded to higher

interoperability tiers. This enhancement will align the dataset with the other DICOM-based datasets within the EUCAIM federation. If uploading non-DICOM formatted data is unavoidable, all the required DICOM-compatible imaging metadata must be supplied in DICOM JSON format as specified by the DICOM standard[39]. Although it is not a requirement, datasets with associated annotations are highly desired and recommended into the EUCAIM federation as they enhance the datasets' value and usability. Datasets with associated annotations will get prioritized in their inclusion to the federation and receive a higher quality stamp/label (Tier 1A+). However, only manual and semi-automatic annotations will be accepted in the federation.

When the annotations provided are segmentations, they are desired in DICOM SEG format. If they are in other formats (as NIfTI or RTSTRUCT) and the related original imaging data in DICOM format are available, the EUCAIM Converter tool will be provided by EUCAIM (see section 4.3.2) to convert them to DICOM SEG. This is especially relevant when datasets are to be transferred to any of the Reference Nodes.

Imaging Metadata

- *Imaging data must be accompanied by a set of minimum imaging metadata.*

All images must include a minimum set of imaging metadata to ensure compliance and interoperability. This minimum set of metadata includes:

- **Mandatory DICOM tags:** All required DICOM tags, as defined by the DICOM standard, must be present at all times.
- **Additional imaging metadata:** An additional set of metadata, detailed in Table 7, must also be present. This metadata must be extracted from the individual DICOM images, aggregated, and populated in the EUCAIM public catalogue. EUCAIM will provide a tool to facilitate this process (see section 4.3.2). Once aggregated, the metadata values must be standardized and comply with the EUCAIM DCAT-AP specification detailed in Section 4.1.

Table 7: Minimum imaging metadata

Attribute	DICOM tag	Requirement	Example
Patient ID	(0010,0020)	Mandatory	X123456
Image modality	(0008,0060)	Mandatory	CT
Image body part	(0018,0015)	Mandatory	Chest
Image manufacturer	(0008,0070)	Mandatory	Siemens
Date of image acquisition (YYYYMMDD)	(0008,0022)	Mandatory	20240101

[39] https://dicom.nema.org/medical/dicom/current/output/chtml/part18/sect_f.2.html

The patient's age at the time of each imaging study must be provided, either:

- Directly in the PatientAge DICOM tag (0010,1010);
- Or indirectly, by ensuring it can be calculated using the 'Age at diagnosis' (from the clinical attributes) and the 'Date of image acquisition'.

Clinical Metadata

- *Imaging data must be accompanied by a set of minimum clinical metadata.*

All imaging data must be accompanied by a minimum set of clinical metadata. Exceptions may be made for imaging-only datasets that include only imaging metadata. These cases will be evaluated on a case-by-case basis and accepted on the platform only when the associated clinical metadata is deemed irrelevant to the intended purpose for which the dataset was created.

Table 8 presents the minimum clinical metadata required for positive or diagnostic cases:

- **Positive Cases:** Patients that have been confirmed to have cancer through diagnostic testing, such as imaging studies, biopsies, or other pathological examinations. These cases show evidence of cancer presence, such as tumors, malignant cells, or other markers specific to the disease.
- **Diagnostic Cases:** Cases where diagnostic evaluations were performed to confirm, rule out, or further characterize a cancer diagnosis. This could include patients undergoing imaging studies (e.g., CT, MRI, PET scans) or laboratory tests (e.g., tumor markers) even if the final outcome is negative. Often emphasizes the data collected during the diagnostic process, such as imaging or histopathological data, regardless of whether the result confirmed cancer.

Table 8: Minimum clinical metadata for positive or diagnostic cases

Attribute	Description	Requirement	Examples
Patient ID	A unique identifier for the patient. This should match the patient ID DICOM tag (0010,0020) and the anonymization processes.	Mandatory	X123456
Population	Categorization of the subjects in the dataset based on their status.	Mandatory	Patient with Cancer; Patient with lesion not being a malignant tumor.
Sex	Biological sex at Birth	Mandatory	Female, Male, Unspecified
Date of radiology detection*	Date when the tumor or the lesion was first detected by an imaging study (or the nearest study to the diagnosis confirmation).	Mandatory if available	January 1, 2024

Attribute	Description	Requirement	Examples
Date of pathology confirmation / diagnosis date *	Date when the tumor is histologically confirmed (or confirmed by an imaging study if histology was not performed, in specific cases such as HCC)	Mandatory if available	February 1, 2024
Age at diagnosis	Age of the patient at the time the tumor or lesion was confirmed. The age must be provided in years with one decimal.	Mandatory	45,5
Pathology confirmation	Method used to confirm the pathology (histological (surgery, biopsy) or by imaging in specific cases such as HCC). The method used before the treatment decision will be considered.	Mandatory if available	Biopsy
Topography	Location of the lesion, stratified in three steps: organ, region, and laterality	Mandatory only for the organ	Lung, Upper Lobe, Right
Imaging procedure protocol	Specific protocol applied to obtain the diagnostic image	Mandatory if available	CT of thorax with contrast
Pathology	Histology and histological subtype of the lesion (in ICDO-3, if available)	Mandatory if available	Adenocarcinoma / Papilar
Treatment	Type of treatment received by the patient	Mandatory if available	Chemotherapy followed by surgery
Date of first treatment*	Date when first treatment occurred	Mandatory if available	March 1, 2024

*IMPORTANT NOTE: If dates are not available in the dataset, or have been altered due to anonymization purposes, relative days to a given baseline time point should be available according to the dataset purposes.

Similarly, Table 9 presents the minimum clinical metadata required for negative screening and control groups:

- **Negative Screening:** Patients who have undergone cancer screening tests and received a negative result, indicating no evidence of cancer or abnormalities suggestive of cancer at the time of the screening.
- **Control Groups:** Groups of individuals used as a baseline or comparison in clinical studies or research. These individuals typically do not have cancer and may not have undergone any diagnostic or screening tests prior to the study. Instead, they are matched to the study population in terms of age, gender, or other characteristics to provide a reliable point of comparison.

Table 9: Minimum clinical metadata for negative screening and control groups

Attribute	Description	Requirement	Examples
Patient ID	A unique identifier for the patient. This should match the patient ID DICOM tag (0010,0020) and the anonymization processes	Mandatory	X123456
Population	The categorization of the subjects in the dataset based on their status	Mandatory	Subject on Screening with a negative result; Subject on a Control group.
Sex	Biological sex at Birth	Mandatory	Female, Male, Unspecified
Date of imaging acquisition	Date when imaging study occurred for screening or control group	Mandatory if available	January 1, 2024
Age at diagnosis	Age of the subject when the imaging study was acquired. Provided in years with one decimal.	Mandatory	45,5
Topography	Area exam with the imaging modality: organ	In negative screening and control group cases, region and laterality are not mandatory.	Lung

Beyond the above clinical metadata, providing extended clinical information significantly enhances the value and usability of imaging datasets, allowing them to support various purposes and use cases. For detailed information about the clinical data currently supported by the EUCAIM CDM, please refer to D5.2: The EUCAIM CDM and Hyper-Ontology for Data Interoperability: Initial Version.

Annotation Metadata

- *Annotations must be accompanied by a set of minimum annotation metadata.*

If annotations are present in the dataset, all annotations must be associated with a set of mandatory annotation metadata. Table 10 presents a set of mandatory minimum annotation metadata required to comply with the DICOM-SEG standard defined in EUCAIM, along with the essential information needed to better describe the annotation dataset. Additionally, all required DICOM tags (type 1), as defined by the DICOM standard, must be present at all times in the DICOM SEG files.

It is worth mentioning that we are currently focusing specifically on segmentation. It is planned to support other types of annotations too, such as rectangular/circular bounding boxes or centroids for detection-related tasks, as well as rulers, arrows, and angles for measurement-related annotations.

Table 10: *Minimum annotation metadata required for conversion from non-standard format to DICOM SEG*

Name	Description	Level	DICOM tag	Type	Example
Segment number	Unique identification number of the segment	Imaging	Segment number (0062, 0004)	Mandatory	1,2,...
Segment label	User-defined or ontology-defined label identifying the segment.	Imaging/ Dataset*	Segment Label (0062, 0005)	Mandatory	“PZ (peripheral zone of prostate)”, or “CZ (central zone of prostate)”
Segment description	User-defined or ontology-defined description for the segment.	Imaging/ Dataset*	Segment Description (0062, 0006)	Mandatory. In the case the segmentations are made in the context of EUCAIM, the Segment Description should have specific terms from the ontology.	“Prostate Central Zone”, or “Prostate Peripheral Zone”
Segmentation method	Type of algorithm used to generate the segment	Imaging/ Dataset*	Segment Algorithm Type (0062, 0008)	Mandatory	Manual, semiautomatic, automatic
Annotation type	Type of annotation	Imaging/ Dataset*		Mandatory if available	Bounding box, Mask, Mesh etc.
Image coordinate system	Defines if the annotation tool does not	Imaging		Mandatory if available	Physical, Pixel

	use the same coordinate system as the image.				
Algorithm name	The name(s) and version of the algorithm(s) used to generate the segment.	Imaging/ Dataset*	Segment Algorithm Name (0062, 0009)	Mandatory if Segment algorithm type (0062, 0008) is not Manual.	“Prostate segmentation Tool v1.0.0”
Number of annotators	Number of annotators involved in the annotation process	Dataset		Mandatory	1,2,...
Annotator type	Specific role of the annotator.	Dataset		Mandatory	Radiologist, imaging technician, etc.
Experience	Years of experience of the annotator as a whole number (integer) without decimals.	Dataset		Mandatory	1,2,3,5,10…
Sequence(s) used for segmentation	Modality(s)/Submodality(s) used to perform the segmentation.	Dataset		Mandatory	T2w, ADC, CT, CT+PET

* It is preferred to provide them at the imaging level in the corresponding DICOM tag. However, if the value is the same for all studies within the dataset, it can be provided once at the dataset level, if necessary.

It is worth considering that it may be necessary to enter information related to the same metadata multiple times, for instance, either because there are several annotated regions, meaning multiple segments to reference, or because there are multiple annotators and information about all of them needs to be included.

4.3 Guidelines for Dataset Preparation

The guidelines for dataset preparation outline a series of steps and tools that a DH can use to ensure data is prepared accurately and consistently for use within each tier. The steps and tools that should be executed will largely depend on the state of the data and the intended tier. For example, in Tier 1, only the publication and visualization of public metadata are supported. However, minimum requirements for both imaging and clinical data must be met to ensure dataset

security and facilitate the eventual upgrade to higher tiers with greater interoperability within the EUCAIM framework. Tier 2 and 3 will require higher efforts to prepare the data. Before data preparation, DH must complete an initial legal and ethical assessment to ensure compliance and ethical adherence (ref).

The processes and tools provided here are primarily designed to support datasets in the DICOM format, as the standard imaging format within EUCAIM, which is strongly encouraged. While accepted for Tier 1, if NIfTI-based datasets are submitted, Data Holders are responsible for using their own tools and procedures to meet the minimum requirements.

By following these guidelines, DH can ensure their datasets align with EUCAIM standards, promoting consistency and compatibility across the infrastructure.

4.3.1 Overview (generic sequential diagram)

Datasets intended for ingestion into the EUCAIM framework exhibit significant heterogeneity, varying significantly in their levels of standardization, quality, or anonymization. Consequently, the preprocessing steps can differ substantially depending on datasets' initial state and the target Tier within EUCAIM. Figure 12 outlines a high level, sequential diagram of the steps a DH should follow, organized by tiers. A DH may opt to prepare the dataset sequentially as it is shown in the figure, progressively transforming the data to achieve higher levels of interoperability. In this approach, datasets will fully meet all the requirements of a given tier before being upgraded to the next. However, it is important to notice that raw data can be directly transformed to achieve higher levels of interoperability (Tier 2&3).

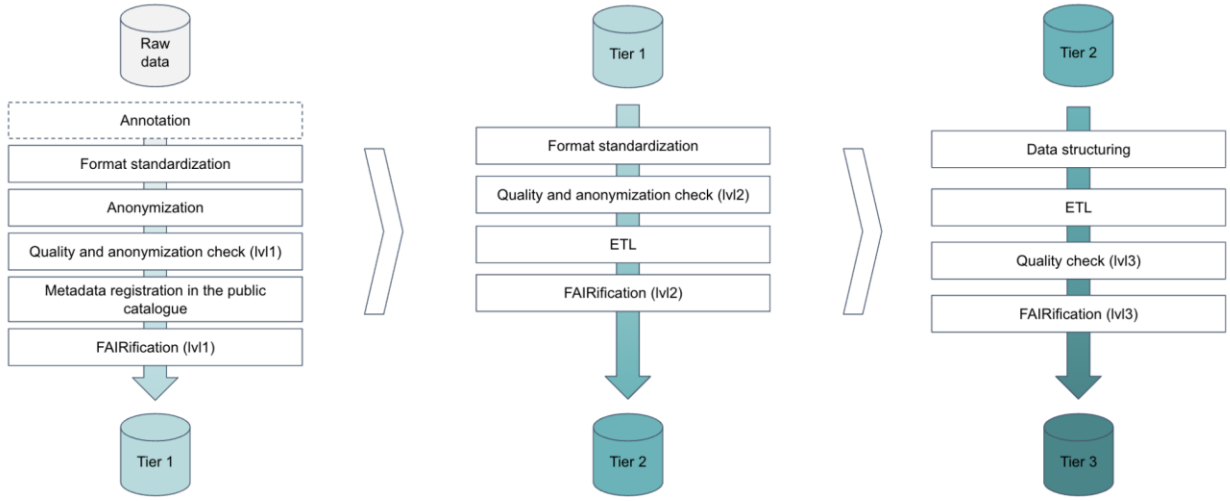


***Figure 12**. The sequential diagram illustrates the steps involved in dataset preparation, organized in a tiered structure. Annotation is represented by a dashed line, indicating that while it is highly recommended, it is not mandatory for dataset preparation. The quality and anonymization check, along with the FAIRification processes, are divided across different levels to accommodate the varying requirements of each tier.*

4.3.2 Dataset to Remain in a Local Node

Workflow and Data Preprocessing/Interoperability Tools

For Tier 1 datasets, data quality and annotation tools may be used by DHs for their imaging dataset preparation (Figure 13), providing the data are in DICOM. The minimum metadata must be manually completed in the EUCAIM catalogue. In-house tools must be used for data de-identification (see below the guidelines for data de-identification).

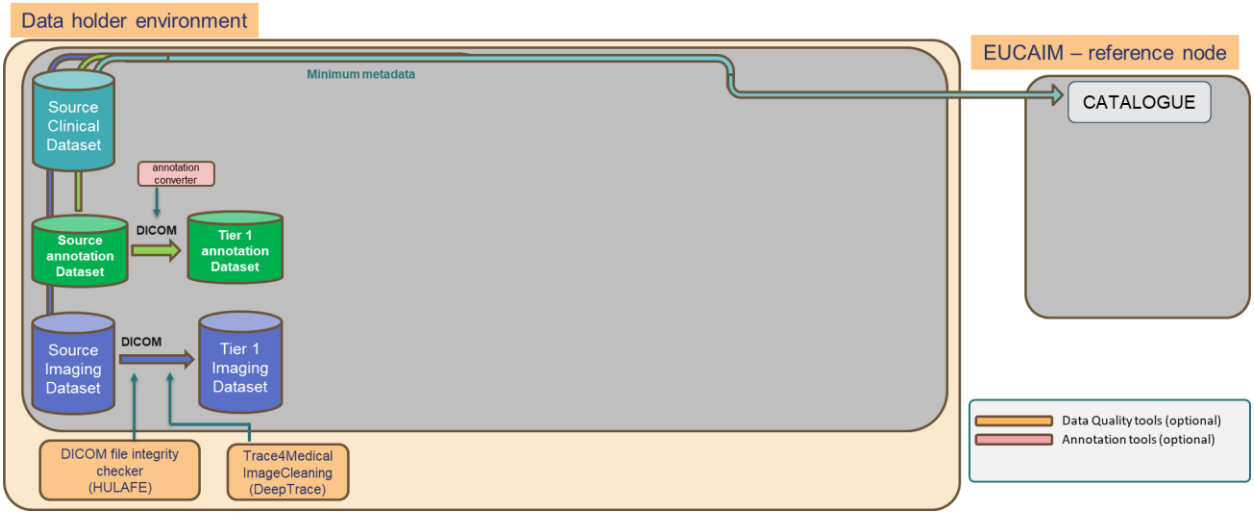


Figure 13. Workflow for dataset preparation (Tier1)

Guidelines for Data Annotation

To enhance quality and utility of the datasets, even if not mandatory, it is highly recommended that datasets are enriched with annotations. If annotations were not previously generated, EUCAIM has a set of tools and guidelines to assist with this task:

- **Tools to annotate**
  - i. Use your own tool. The annotation process can be performed using the annotator's software of choice. The only requirement is to comply with the specifications outlined in deliverable D4.4, primarily related to segmentations in DICOM SEG format and the minimum annotation metadata described in section 4.2 above.
  - ii. MITK workbench:
    The MITK Workbench is a powerful and free application that is part of the Medical Imaging Interaction Toolkit (MITK)[40]. It provides a user-friendly interface for viewing, processing, and segmenting medical images. MITK provides precompiled binaries for Windows, Linux and macOS operating systems. The appropriate version should be selected and downloaded based on the installed operating system.

    Segmentation is one among the many processing options in MITK for medical images. The Segmentation plugin allows for creating multilabel segmentations of anatomical

[40] https://www.mitk.org/wiki/Downloads

structures. The Workbench allows for loading and viewing various medical image formats, such as DICOM and NIfTI.

MITK offers a comprehensive set of slice-based 2D and (semi-)automated 3D segmentation tools including AI tools like Totalsegmentator[41].

For creating a simple 2D segmentation, users should select the “Add” tool and manually draw a contour over the region of interest, or use semi-automatic tools like the “Threshold” tool to select a range of intensities to include in the segmentation.

Once the segmentation is deemed complete, the mask image can be saved in any of the accepted formats, such as DICOM SEG.

Extended MITK Workbench documentation can be found in Annex 1.

- **Harmonization of annotation: Guidelines on how to annotate the data**

To achieve standardized annotations, the development of annotation guidelines is underway. These guidelines are based on establishing a common process and criteria across centres, with a focus on standardization and homogenization. The goal is to ensure consistent annotation criteria for the same use case—considering pathology, task, organ, and modality—while reducing variability both within and between radiologists.

Given that EUCAIM gathers data from diverse sources, such as centres, countries, and annotation protocols, these efforts aim to produce comparable annotation results. Furthermore, high-quality and unbiased annotations are critical for training AI models, as biased or low-quality annotations can lead to poor model performance. Models trained with quality annotations not only deliver better outcomes but also reduce future manual annotation workload through pre-segmentations.

The aim is to prepare two reference use cases: prostate segmentation and glioblastoma segmentation. These examples will include the steps to follow depending on the annotation task and how to proceed based on tumor location and modality (e.g., MRI sequence to use, CT windowing values, contrast/no-contrast, if parametric maps are required, etc.). When ready, the annotation guidelines will be found as training materials in the EUCAIM Moodle.

- **Annotation format standardization: from non-standard segmentations to DICOM SEG**

To avoid limiting the project's annotation collection to those in DICOM SEG format while still maintaining a common standard for segmentation throughout the project, a converter has been developed to enable the conversion of non-standard formats to DICOM SEG format. Currently, the tool enables bidirectional conversion between NIfTI and DICOM SEG formats. It handles four scenarios (each scenario assumes the availability of the original DICOM images):

1. One NIfTI to One SEG: A single NIfTI file generates one SEG file.
2. Multiple NIfTIs to Multiple SEGs: Each NIfTI file generates its own SEG file.
3. Non-Overlapping SEG to NIfTI: One NIfTI file is generated if segments do not overlap.

[41] https://docs.mitk.org/nightly/org_mitk_views_segmentation.html#org_mitk_views_segmentationTotalSegmentator

4. Overlapping SEG to NIfTI: Each segment produces a separate NIfTI file.

The tool can be executed via Docker. All of these functions are containerized and pushed on DockerHub: https://hub.docker.com/r/mariov687/dicomseg

An extension of the tool is currently under development to include support for the DICOM RT STRUCT format as a non-standard format that can be converted to DICOM SEG. DICOM RT STRUCT was chosen due to its widespread use in radiotherapy, where a significant number of annotations are expected to adhere to this standard.

Guidelines for format standardization

DICOM is the standard imaging format within EUCAIM and it is recommended to DHs to convert NIfTI or other formats back to a valid DICOM format if possible. This will align the dataset with the rest of DICOM based EUCAIM datasets, facilitate the dataset preparation process and ease its future upgrade to higher Tiers. However, some datasets in NIfTI format might be accepted for Tier 1, please refer to section 4.2 for further details. Annotations are also preferred in DICOM SEG format, however this is not a requirement for Tier 1. The converter tool provided by EUCAIM to a DH allows the conversion of EUCAIM non-standard formats to DICOM Seg.

Clinical data is accepted in different data formats (JSON, CSV, XLS, Avro, Parquet…), however the use of JSON and CSV formats is encouraged.

Guidelines for Data De-Identification

De-identification requirements

To ensure the privacy and protection of patients in the data present in EUCAIM, it is imperative that the data undergo rigorous processing to eliminate any elements that could potentially lead to patient identification.

It must be ensured that no identifiable information is present in the data shared with EUCAIM. To achieve this, data must be thoroughly processed to modify or remove direct and indirect identifiers. Direct identifiers are elements of the original data that are explicitly unique to a patient, such as their name, social security number, or original patient ID. These elements should either be entirely removed or replaced by pseudonyms before the data is shared.

In addition to direct identifiers, indirect identifiers must also be addressed. Indirect identifiers refer to data points that, while not explicitly unique, can still be used in combination with other information to re-identify an individual. For example, attributes like a patient's date of birth or detailed clinical measurements could, when combined, uniquely identify a person. To prevent this, indirect identifiers should be generalized, aggregated, suppressed, substituted with a dummy value consistent within the dataset, or anonymized in a way that reduces the risk of linkage to a specific patient. In order to achieve that, the data holder may use its own resources to remove or alter the identifiers. However, for the metadata present in the DICOM tags, EUCAIM provides the EUCAIM Anonymizer tool with a pre-configured de-identification profile that describes the actions taken in each of the tags containing sensible information. When applied to the DICOM files, the tool processes the data according to a predefined de-identification profile that could be modified according to the data needs. In addition, if meeting the data requirements, a hashing tool could

be used for anonymizing the ids present in the clinical dataset, preserving the linkage between imaging and the corresponding clinical data.

It is also important to highlight that data privacy concerns extend beyond the individual patient level. Privacy risks can arise at a dataset level due to unique combinations of attributes that create outliers. These outliers, defined by their distinctive characteristics, can potentially expose individual identities when analyzed. For instance, a dataset might include a rare combination of medical conditions or treatment outcomes that uniquely identify a single patient. To mitigate this risk, datasets should be carefully reviewed for such combinations and appropriately processed by the DH. Risk minimization at the dataset level can be achieved through techniques such as data generalization and suppression. Generalization involves grouping or categorizing sensitive attributes, reducing their granularity while retaining their usefulness for analysis. For instance, instead of including exact ages, age rounding (e.g., "30 years instead of 32") could be used. Suppression, on the other hand, entails the removal of particularly sensitive or high-risk data elements entirely, especially if their presence significantly increases the likelihood of re-identification. Techniques such as k-anonymity, k-map or l-diversity can be employed to ensure that no individual stands out within the dataset. For this purpose, EUCAIM provides a tool called Wizard tool that identifies potential risks of data re-identification and proposes ways to mitigate them. In Tier 1 data, this tool can be applied to analyze the metadata present in the DICOM files.

Furthermore, EUCAIM hosts a wide array of medical imaging data, which presents unique privacy risks that must be carefully managed to preserve patient confidentiality. Medical imaging techniques, such as ultrasound, often embed textual information directly into the images. This embedded text may include sensitive patient details, such as names, dates of birth, or medical record numbers. To ensure robust de-identification, this text must be thoroughly removed during the preprocessing phase before the images are shared in EUCAIM.

Beyond textual information, medical images can pose additional risks of re-identification due to their inherent uniqueness. For example, certain anatomical features or abnormalities may be so distinct that they could inadvertently identify an individual. This is especially true for rare medical conditions or unique anatomical variations that act as natural identifiers. Proper risk assessment is crucial to address these challenges. Moreover, medical images that include the head or facial structures also introduce privacy concerns. Advanced technologies, such as three-dimensional imaging and reconstruction techniques, can enable the creation of highly detailed representations of a patient's face. These reconstructions can sometimes be used to identify individuals, particularly when combined with other datasets or biometric tools. Such risks necessitate rigorous de-identification measures taken by the data holder, including the removal or blurring of facial features to ensure that the data remains non-identifiable.

In addition to imaging data, EUCAIM also includes clinical data from patients, making it essential to preserve the link between these datasets during the de-identification process. Specifically, if a patient ID in the DICOM metadata is replaced, the same new ID must be consistently applied to the corresponding clinical data to maintain the integrity and usability of the datasets. If any clinical data is also linked at a study level, the same procedure should be performed to guarantee that no relation is lost. The data holder must ensure the preservation of this linkage as it is crucial for enabling meaningful analyses that integrate clinical and imaging information, such as correlating medical information or test results with imaging findings.

Tools to anonymize

- Use your own tool: The de-identification process can be performed using any software as long as the direct and indirect identifiers are removed or substituted to minimize the risks of patient re-identification.
- DICOM Anonymizer and Hashing: If the data meet the requirements described in the tool's documentation, they can be used for the anonymization purposes. The DICOM Anonymizer will be in charge of the DICOM metadata de-identification while the Hashing function will de-identify the ids of the clinical data preserving the linkage between image and clinical data.
- Trace4medical/skull stripping tools: In particular cases where the images contain the skull of the patient, EUCAIM can provide tools for defacing, adding an extra layer of security to prevent potential re-identification risks coming from face reconstruction.
- Wizard tool: The Wizard tool, powered by the ARX data anonymization framework[42], is available for datasets that meet the project's specific requirements. This tool automatically evaluates the data to identify any risks related to patient uniqueness or unique data attributes that could result in a high re-identification risk. By analyzing both direct and indirect identifiers, it applies appropriate anonymization strategies and provides a comprehensive risk report based on recognized privacy models (e.g., k-anonymity). Through this process, data holders can confidently address GDPR requirements while preserving maximum data utility. The Wizard tool is especially recommended for Tier 1 nodes whose datasets already satisfy the project's minimum data preparation standards. The role of the wizard in this case will be twofold, as it efficiently detects any remaining privacy gaps and suggests necessary adjustments, but it will also provide some descriptive characteristics of the patient cohort and acquisition conditions. The latter can support users to identify possible use cases regarding their usage but also to investigate the possibility of data merging to support the needs of a specific use case under the hypothesis of inadequate data from a single source.

Guidelines for Data Quality

All datasets added to the EUCAIM federation must comply with a data quality standard. This applies to all tiers, including Tier 1 data.

The EUCAIM data quality framework will align with the QUANTUM[43] initiative, a European-funded project under the HORIZON-HLTH-2023-TOOL-05-09 call on Tools and technologies for a healthy society[44], that has started in January 2024 and aims to **develop a Data Quality and Utility Label for the European Health Data Space.** EUCAIM will leverage on the QUANTUM framework by applying it both at the dataset and the data levels, whenever possible.

[42] https://arx.deidentifier.org/
[43] https://quantumproject.eu/
[44] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/calls-for-proposals?callIdentifier=HORIZON-HLTH-2023-TOOL-05

As a start, EUCAIM data quality framework will follow the recommendations from the TEHDAS and TEHDAS2 projects[45] on a data quality framework for the European health data space for secondary use[46].

Article 78 of the European Health Data Space (EHDS) (corrigendum from European Parliament on 27/11/2024), on Data quality and utility label[47] defines six dimensions that the data quality and utility label shall cover, where applicable:

a. for data documentation: meta-data, support documentation, data dictionary, format and standards used, provenance, and when applicable, data model;

b. for assessment of technical quality: **completeness, uniqueness, accuracy, validity, timeliness and consistency of the data**;

c. for data quality management processes: level of maturity of the data quality management processes, including review and audit processes, biases examination;

d. for assessment of coverage: time period, population coverage and, when applicable, representativity of population sampled, and average timeframe in which a natural person appears in a dataset;

e. for information on access and provision: time between the collection of the electronic health data and their addition to the dataset, time to provide electronic health data following an electronic health data access application approval;

f. for information on data modifications: merging and adding data to an existing dataset, including links with other datasets;

g. a. where a data quality and utility label accompanies the dataset pursuant to Article 56, the health data holder shall provide sufficient documentation to the health data access body for that body to confirm the accuracy of the label.

TEHDAS partners rated “Data documentation” and “Technical quality” as *highly relevant* elements, which we will mostly focus on at this stage.

Technical data quality dimensions for the EUCAIM framework

The EUCAIM data quality framework, applicable to all tiers including tier 1, will address the dimensions recommended for the EHDS, listed in the table below. Integrity (in blue in the Table 11) will also be addressed, as it deems very relevant to EUCAIM.

Table 11: *Technical data quality dimensions*

Dimension	Definition (from DAMA’s dimensions of data quality 2020)	Recommended for the EHDS	EUCAIM framework
Completeness	The degree to which all required data values are present	☑	☑

[45] https://tehdas.eu/
[46] https://tehdas.eu/app/uploads/2023/09/tehdas-recommendations-on-a-data-quality-framework.pdf
[47] https://www.europarl.europa.eu/meetdocs/2024_2029/plmrep/COMMITTEES/ENVI/DV/2024/12-04/2022_0140COR01_EN.pdf

Uniqueness	No entity exists more than once within the dataset	☑	☑
Validity	The degree to which dataset values comply with rules	☑	☑
Timeliness	The degree to which the period between the time of creation of the real value and the time that the dataset is available is appropriate [Not assessable on anonymized datasets]	☑	X
Consistency	The degree to which dataset values of two sets of attributes - within a record, - within a data file, - between data files, - within a record at different points in time comply with a rule	☑	☑
Accuracy	The degree of correspondence between dataset values to real values	☑	☑
Integrity	The degree of absence of data value loss or corruption	X	☑

Based on the nature and type of data (also scalar or vector), these data quality dimensions are calculated in a different way. As an example, description of the methodology to measure the various dimensions using the Data Integration Quality Check Tool (DIQCT) is provided in Annex 2.

Tier 1 dataset must comply with the following:

- **Completeness**: DICOM files must be available to data holder for all registered cases, and all associated metadata must be provided;
- **Uniqueness**: no duplicates must be present in the datasets;
- **Validity**: all data must comply with the rules from Tier 1 EUCAIM framework, in terms of data format and de-identification;
- **Consistency**: data and metadata from multiple series, studies, timepoints, must be consistent with one another (eg : a follow-up examination must have a date posterior to the baseline examination);
- **Accuracy**: values from data and metadata of a dataset must conform to their associated labels (eg: values for age must be a numerical value, expected to be comprised between 0 and 110);
- **Integrity**: no corrupted file should be present in the dataset.

Data quality metrics

The EUCAIM data quality framework will rely on robust data quality metrics and establish an error-proof data quality strategy to ensure the quality of data.

The set of tools will vary based on the level of compliance and type of data. Tier 1 data holders will be provided with a set of optional data quality tools to address:

- Validity
- Accuracy
- Integrity

Table 12 lists the tools that may be used by Tier 1 data holders.

Table 12: *Optional data quality tools for Tier 1*

Dimension	Tool	Description	Quality level	Output / Metrics
Validity	**Trace4MedicalImage**	For 2D ultrasound and mammography DICOM files: detect and remove encapsulated text, as they may contain potentially identifying information	Data	List of processed files and corrected DICOM files
Accuracy	**DICOM File Integrity checker**	The tool performs a quality check in terms of the correct number of files per each serie	Data & Dataset	Comparison between the expected quantity of images by serie and existing quantity when they do not match
Integrity	**DICOM File integrity checker**	This tool looks for corrupted files to identify	Data & Dataset	List of the corrupted DICOM imaging files

Guidelines for metadata extraction and registration to catalogue

Within the EUCAIM platform there are two ways to register your dataset in the catalogue. The first way is to use a FAIR Data Point (FDP), from which the metadata is then harvested and added to the catalogue. More information on this can be found in the section 'Data registration in the public catalogue via FDP' below. The second way is to fill the Excel template (see example in [48]). Experts in the technical committee of EUCAIM will check the metadata and manually register them in the public catalogue using the Molgenis user interface. This superuser has the permissions to add entries to the tables for datasets, dataset series and persons, and can use the Molgenis graphical user interface to add catalogue entries manually. During the registration process the property values are validated against the configured schema, matching the EUCAIM metadata model. It is recommended to register the dataset in the catalogue once all the steps of the data preprocessing are completed.

[48] https://dashboard.eucaim.cancerimage.eu/EUCAIM-ingestion-sample.xlsx

Guidelines for Data FAIRification and registration in the public catalogue via FDP

Data FAIRification

The FAIR principles establish a framework of guidelines and best practices designed to facilitate the discovery, access, interoperability, and reuse of data and metadata by both machines and humans. However, within the realm of medical imaging research, particularly in the context of EUCAIM, it is impractical to anticipate the rigorous implementation of the 41 indicators outlined by the Research Data Alliance (RDA) Maturity Model Specification and guidelines across all data holders. This is due to their varied circumstances and backgrounds, as well as the sensitive nature of the data involved.

Instead, before being able to publish a dataset to the EUCAIM catalogue, data providers will ensure such dataset complies with the EUCAIM FAIR compliance level corresponding to the Tier as defined in deliverable D4.4 Annex 5. Those definitions specify the mandatory indicators that are required for the dataset to be accepted at a given Tier. Regardless of this, DHs should be encouraged to make an effort for further FAIRification, as this would be beneficial for users.

Among the Tier 1 mandatory indicators, RDA-F1-01M and RDA-F1-01D are linked to the assignment of persistent identifiers for both data and metadata. EUCAIM can't provide those identifiers and thus the DHs must ensure their proper assignment before submitting the metadata to the EUCAIM catalogue. Further requirements for these identifiers are set for Tier 2.

In order to check a dataset's compliance to a certain level, the FAIR EVA[49] tool will be used. This automatic evaluation tool helps to monitor FAIR compliance following the RDA indicators. On top of this, it provides a plugin mechanism that allows it to extend its functionality for specific use cases. A EUCAIM plugin for FAIR EVA is under development; on top of checking for the RDA indicators it checks for the presence of EUCAIM defined mandatory metadata attributes, and it will incorporate EUCAIM Tier levels compliance checks.

This plugin is meant to interact with FDPs, like the one at the central EUCAIM Catalogue. This will be the main point of FAIR compliance checking, as it will have the metadata for all the datasets in EUCAIM.

However, for Data Holders that install a FDP for facilitating the registration of datasets to the EUCAIM catalogue, FAIR EVA (with the EUCAIM plugin) can be used to test the datasets FAIR compliance in their FDP before importing them into the EUCAIM Catalogue.

Data registration in the public catalogue via FDP

The public Catalogue has been designed and developed to harvest information from other catalogues, and to allow its metadata to be harvested by other catalogues as well. This will eventually enable metadata entered into the EUCAIM Catalogue to automatically become available in other catalogues, preventing a DH from having to enter the metadata manually in every individual catalogue. This implementation is being designed around a FAIR data point (FDP[50]), which will act as an intermediary to manage and share metadata between catalogues.

[49] https://github.com/IFCA-Advanced-Computing/FAIR_eva Retrieved 8th of January 2025
[50] https://www.fairdatapoint.org/

The full functionality of the EUCAIM FDP is expected to be realized and presented in future deliverable D4.6 Final Core Services.

After Data FAIRification, three steps are needed to register them in the public catalogue via FDP, which are detailed below:

1. Implementing a metadata expose pipeline
2. Implementing a FAIR Data Point
3. Harvesting the FAIR Data Point by the EUCAIM public catalogue

1. Implementing a metadata expose pipeline

Metadata can be exported automatically from the local system with an exposed pipeline. Some systems directly expose metadata. In those cases, no additional pipeline is needed.

If the feature to directly expose metadata is not supported by your local system, the following approach can be used as an alternative. This method involves automated exports being conducted from your institute's local system through a script or workflow. By doing so, the export logic is moved outside of the source application. This approach requires the knowledge of a data steward in collaboration with the local IT department.

The data to be exposed is stored within a specific application (e.g., CHAIMELEON, XNAT, Grand-Challenge, Castor) used within the institute. To automate the export process, the raw data that lies beneath this application needs to be accessed. This can typically be done either by accessing the data through an API provided by the application, or by directly connecting to the database where the data is stored (with read access). This allows the data to be retrieved automatically without manual intervention.

A script should be written that queries the local system, aggregates the data if necessary, and generates EUCAIM FDP compliant metadata in RDF format. To ensure up to date representation of the metadata in the catalogue, this script to expose the local metadata should either run when the data in the local system is updated, or on a regular basis, e.g., weekly or daily. This interval can be determined by the data holder and will depend on the expected update frequency of the data.

Example for an XNAT imaging data repository:

An example application that allows the automatic export of metadata from your local system to a FAIR Data Point is img2catalog, developed by Health-RI. This tool queries an XNAT instance and generates DCAT-AP 3.0 metadata. It can register image collections directly at a FAIR Data Point which can then be linked to the EUCAIM catalogue. For details, see https://github.com/Health-RI/img2catalog.

2. Implementation of a FAIR Data Point

There are several approaches for implementation of a FAIR Data Point:

a. Exposing the local system

Enabling the institute’s local system to directly expose metadata as an FDP creates the strongest connection between the original data and publicly available information. These systems establish a direct pathway between the original data and public information, ensuring that the data holder is in charge of managing the data within their institute and how its metadata is exposed. This

setup means that responsibility for maintaining this metadata is kept at the source, whether handled by software or people. Achieving this requires either a system already supporting FAIR Data Points or deep knowledge of the source system and the availability of software engineering capacity to extend the functionality of the system. The systems in the institute must support functionality for displaying metadata from the existing data sources, in order for this approach to be used.

Example for MOLGENIS catalogues:
MOLGENIS EMX2 contains a Fair Data Point (FDP) implementation which complies with the latest v1.1 specification of the FAIR Data Point. This means that MOLGENIS can expose metadata through a FAIR Data Point from the MOLGENIS instance, when configured to do so.

Other systems, such as Castor[51] and LOVD[52], also offer the option to expose the metadata as a Fair Data Point.

b. Implementing a FAIR Data Point using FDP in a box

Running a standalone FAIR Data Point to expose your metadata is an option, if the necessary resources or capacity and knowledge are present in the institute.

The FAIR Data Point (FDP) reference implementation is available as a Docker Compose distribution. To set up an FDP using this software, a (virtual) machine with Docker installed is needed. Detailed instructions on how to use the reference implementation can be found in the official documentation at https://fairdatapoint.readthedocs.io.

c. Publishing dataset metadata via a Central FDP

If the previously suggested methods cannot be implemented, there is the option to use an already existing FDP, e.g., a national FDP. After acquiring an account, the proper permissions on the FDP, metadata can be exported to it automatically, e.g., by using img2catalog to submit metadata from an XNAT image data repository, or by manually adding it through the user interface.

3. Harvesting the FAIR Data Point by the EUCAIM public catalogue

To harvest the exposed metadata, the data holder contacts the EUCAIM service desk with an onboarding request and includes the details of the FAIR Data Point to harvest. EUCAIM then performs the harvesting. The metadata is harvested into a testing environment where a check of the data is performed by EUCAIM and the data holder, before reaching the Catalogue. *N.B. The harvester for the EUCAIM catalogue will be implemented in 2025.*

Technical Requirements for Local Node Setup

Network requirements

There are no strict network requirements for Tier 1 nodes that maintain their data locally. A symmetrical bandwidth of approximately 100 Mbps is recommended, so as to allow access to data within a reasonable time via the means supported by the node itself.

[51] https://www.castoredc.com/
[52] https://www.lovd.nl/

To allow access to their datasets, all local nodes are advised to conduct thorough testing to confirm that the network connection is active and stable. This can involve pinging external servers, performing bandwidth tests, and checking for any packet loss. If additional security protocols such as VPNs or reverse proxy networks are used by the local nodes, compatibility and security compliance should be ensured in collaboration with the Project's Technical Support team, as these cases require specific configurations, and thorough testing before finalizing the installation.

Hardware requirements

Tier 1 local nodes do not need to integrate to the EUCAIM Federated Search and Processing components, but are rather expected to serve requests for data access in-situ, via the services (e.g. for download, visualisation, etc.) they provide independently of the EUCAIM infrastructures. Therefore, there are no particular hardware and storage requirements that they need to comply with, besides the ones following from the needs of their local services.

Software requirements

As of writing, there are no particular requirements for Tier 1 local data nodes to operate any particular software.

Tier 1 local nodes should ensure that their datasets are registered in the EUCAIM central catalogue following the EUCAIM instance of the DCAT-AP model for describing their datasets.

It is expected that Tier 1 Data Holder's data may not already comply with the respective minimum requirements concerning data formats, anonymization, fairness and quality. To facilitate alignment with the data preparation requirements, Tier 1 nodes may opt to set up and deploy locally the EUCAIM software components described in previous sections. The use of the Wizard tool in particular, which checks whether the local node's data is compatible with the EUCAIM's minimum data requirements and identifies relevant risks, is highly recommended. The instructions about how to install and make use of the Wizard tool are provided in Annex 3.

4.3.3 Dataset to be transferred to the EUCAIM Reference Node

EUCAIM has set up two reference nodes to host data transferred from the data holders who choose this approach. These two reference nodes are complementary and use compatible but different technologies.

- **The UPV node**[53] uses an open-source platform developed in the CHAIMELEON project[54] for providing a fully integrated Data Lake, a Registry and a Virtual Research Environment powered by 10 dedicated physical nodes, with a total of 960 cores, 7,5TB of RAM and 30 NVIDIA GPUs with 24GB RAM each. The data ingestion component and the DICOM viewer of this node is a proprietary technology from QUIBIM, the QP-Insights platform. QP-Insights is a web-based cloud platform designed to provide a solution for storing, managing and analyzing large-scale data with an image-centric approach, seamlessly interoperable with other registries. QP-Insights supports the upload of DICOM studies via a DICOMWeb API (DICOMweb). DICOMWeb is a web-based interface for interacting with DICOM services over standard HTTP/HTTPS, enabling web applications to query, retrieve, and store DICOM objects using RESTful services. This approach offers greater efficiency and accessibility compared to traditional DICOM over TCP/IP protocols.

[53] https://eucaim-node.i3m.upv.es/
[54] https://github.com/chaimeleon-eu

Additionally the QP-Insights REST API allows for other operations such as the upload of clinical data in JSON files and the ingestion or retrieval of subject, study and analysis information. Support of non-DICOM images is possible but they are not supported through the API. A request could be done through the issue tracker of the node.

- **The Euro-BioImaging Medical Imaging Repository**[55] is a platform for storing and managing imaging data and is provided as a service through the Euro-BioImaging ERIC. XNAT is an extensible open-source imaging platform that simplifies common tasks in imaging data management. The Euro-BioImaging Medical Imaging Repository service is an XNAT instance operated by Health-RI[56]. The Euro-BioImaging Medical Imaging Repository is also integrated with other core services (see Table 13).

The Imaging Data should be stored in DICOM format if that is available, but can be also stored in other formats like NIfTI (for interoperability purposes the platform requires DICOM compatible metadata in JSON format following the DICOM JSON formatting specified in the DICOM standard). Alongside the imaging data, derived data and clinical data can also be stored in appropriate file formats as described in the Data Management Plan.

Data can be ingested through the XNAT API and by using the Clinical Trail Processor (CTP[57]). Processing capacity can be provided through the resources of SURF, although it is not standardised yet. In addition, data can be accessed from other processing environments.

In both cases, data holders have some degree of control over the data deposited and permissions are assigned at the level of the projects. The choice of provider can be framed according to the immediate need of processing resources, the need to support the ingestion of non-DICOM data, the management of clinical and imaging data in the same data lake, existing cooperation activities, the availability of VREs and Data annotation tools integrated in the node, etc.

Table 13: *A comparison between two EUCAIM reference nodes*

	UPV node	**Euro-BioImaging node**
Service provider	UPV, Valencia, Spain	Health-RI, Utrecht, the Netherlands
End point	https://eucaim-node.i3m.upv.es	https://xnat.health-ri.nl/
Technology	CHAIMELEON (https://github.com/chaimeleon-eu)	Open-source, XNAT: https://www.health-ri.nl/en/services/xnat
Data Ingestion protocol	DICOM Web, QP-Insights REST API	XNAT DICOM Receiver and CTP, XNAT API, XNATpy python library
Imaging Data Formats	DICOM. Support of NIfTI through manual ingestion by UPV staff.	DICOM, NIfTI
Clinical Data Formats	JSON. Support of CSV through a conversion tool.	JSON, CSV

[55] https://xnat.health-ri.nl
[56] https://www.health-ri.nl/en/services/xnat
[57] https://mircwiki.rsna.org/index.php?title=MIRC_CTP

Authentication	LS-AAI and local users	Local users (SRAM connection available in 2025, https://www.surf.nl/en/services/surf-research-access-management)
Authorisation model	Project-based, access control at the level of the dataset.	Project-based, access control at the level of the dataset
Data Lake	Distributed filesystem based on an object-store, both for Images and clinical data	Scalable network attached file storage, snapshotted and replicated for integrity.
Data model	The data lake uses a hierarchical model of folders (Patient/Study/Series) that contains all the image files (mainly DICOM, including DICOM Seg). Clinical data is stored as collections following a predefined schema and coded into JSON.	XNAT provides a tree structure for data management. Hence, the imaging data is organized in projects, under which there are subjects, under which there are scan sessions, under which there are scans and annotations (see also section 4.9 in deliverable D4.13 End-user guide: https://drive.google.com/drive/folders/1dn1xQB9K7Fn3WzzqN5HRiQ7NiVwYt0yy).
Image viewer	QP-Insights from QUIBIM	XNAT OHIF Viewer is built-in. An external viewer, like fsleyes, can also be used.
Secure processing environments	Integrated through VREs in which data cannot be downloaded	Secure processing environment (SPE) is currently being developed as a service. It could be enabled for specific users (through SURF - the Dutch national computer infrastructure provider, https://www.surf.nl/en), but is not standard yet.
Federated processing	Currently being set-up for use within EUCAIM.	Currently being set-up for use within EUCAIM.
Processing capacity	960 cores, 7,5TB of RAM and 30 NVIDIA GPUs with 24GB RAM each.	The data in this node can be used in a processing environment of choice. For instance, resources available at the Dutch national computer infrastructure provider SURF.
Security	https://u.i3m.upv.es/vuhpb	General: https://www.health-ri.nl/quality-and-security-services-provided-health-ri XNAT specific: https://www.health-ri.nl/en/services/xnat#tab4
DPA between node and EUCAIM	HULAFE is currently reviewing the data processing agreement. We can include this in documentation once signed by both parties.	
SLA	https://u.i3m.upv.es/pauzn	HULAFE is currently reviewing the SLA. We can include this in documentation once signed by both parties.

In case a DH chooses to transfer their datasets into one of the reference nodes, e.g., due to project finalization and lack of resources, the high level workflow to be followed is the following:

Workflow and Data Preprocessing/Interoperability Tools

DHs may opt for ingesting the datasets in Tier 1 to one of the two reference nodes described above. For transferring imaging data to the UPV reference node, QP-Insights supports the upload of DICOM studies via a DICOMWeb API (DICOMweb). Additionally, a graphic interface may be used for a case-by-case upload. For transfering of data to the Euro-BioImaging Medical Imaging Repository data can be ingested through the XNAT API and by using the Clinical Trail Processor (CTP[58]).

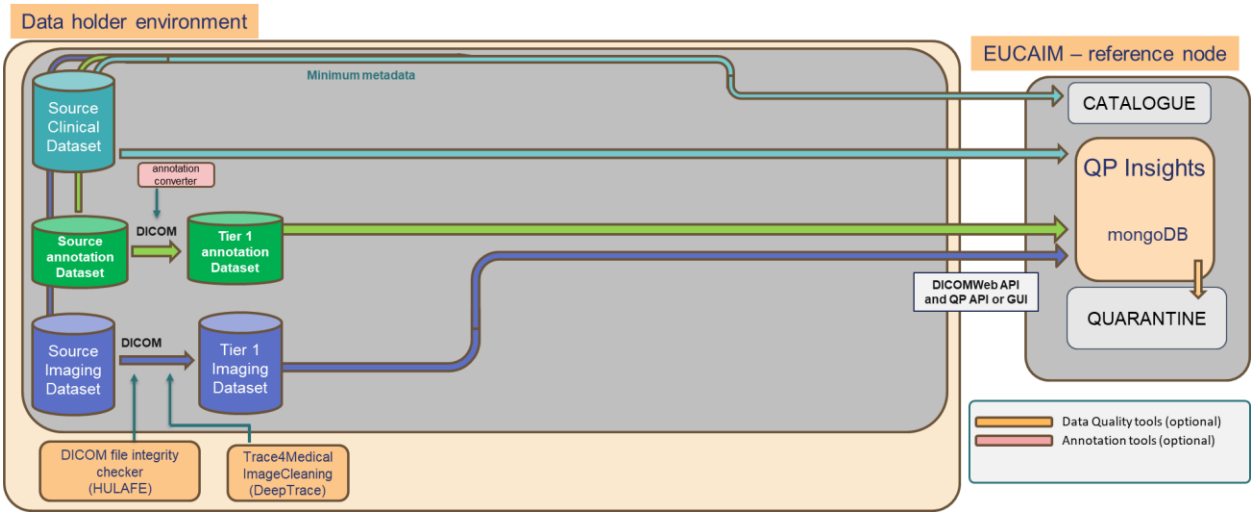


Figure 14. Workflow for Tier 1 data transfer

Guidelines for format standardization

In the reference nodes, clinical data is accepted in JSON and CSV formats, although it is preferable in JSON format. It is encouraged to provide image data in DICOM format, however NIfTI images are also supported in both reference nodes.

FAIRification when dataset are transferred to a Reference Node

While Tier FAIRification requirements remain the same, the movement to the central storage could be used for further FAIRification of the dataset. The check of the FAIRification level will be done using FAIR EVA via the EUCAIM Catalogue FDP.

Guidelines for Data Annotation in the Reference Nodes

Data Holders can take advantage of the viewers provided by reference nodes to annotate their data.

[58] https://mircwiki.rsna.org/index.php?title=MIRC_CTP

Quibim DICOM Web Viewer

Quibim DICOM Web Viewer is integrated into the UPV reference node as the annotation environment, maintaining a “DICOM-in - DICOM-out” approach. This viewer comprises a user interface with tools for image manipulation and manual annotation, as well as a backend that provides access to images and metadata, and handles security issues. Additionally, the annotation tools provided by EUCAIM partners will be integrated into the viewer, allowing for automatic annotation of the images.

XNAT-OHIF Viewer

The Euro-BioImaging Medical Imaging Storage Service provides the XNAT-OHIF viewer for creating annotations on the stored Medical Imaging data. The data needs to be stored as DICOM for this viewer. It offers a range of tools to create Regions-of-Interest (ROI’s), both contour and mask based. It can visualize the data using overlays, fractional segmentation mappings and surface meshes. Furthermore, it is able to present the clinical data stored in eCRF’s inside panel.

Technical Requirements for Local Node Setup

- For supporting the uploading of data (e.g. network etc) and FDP service

A minimum symmetrical bandwidth of 200 Mbps is recommended for Tier 1 nodes who wish to transfer their data to a reference node, to avoid performance bottlenecks during the data transfer procedure.

Local nodes that wish to transfer their data to the UPV reference node should make use of QPInsights, following the instructions that have been provided earlier in this section.

Local nodes that wish to transfer their data to the Euro-BioImaging node need to ingest their data using one of the options. Depending on the specific capabilities of the Data Holder and the data format, the XNAT DICOM Receiver or the A/IO API can be used to upload data to the XNAT repository.

- Clinical Trail Processor[59]
- XNATpy[60]

The Clinical Trail Processor (CTP) is an advanced DICOM processing tool. This is the preferred tool to upload DICOM data to the Euro-BioImaging XNAT. In the Euro-BioImaging Medical Imaging Repository, a CTP is placed between the XNAT built-in DICOM receiver and the public internet. Because the communication protocols of DICOM are not designed for public facing networks, this has to be protected. Two CTP’s can act as a bridge between a Data Holder and the Euro-BioImaging XNAT using an encrypted connection. CTP is able to process DICOM headers and can be used for anonymizing the data. It is fully compliant with the DICOM standard.

XNATpy is a python client, developed by Erasmus MC, which interacts with XNAT through the XNAT API. It offers programmatic access as well as Command Line Interface (CLI) interaction.

[59] https://mircwiki.rsna.org/index.php?title=MIRC_CTP
[60] https://xnat.readthedocs.io/

XNATpy adapts itself to the XNAT instance it communicates with through the XNAT data model XSD definitions. XNATpy offers query and search functionality. It is recommended by the international XNAT community as the tool to use for programmatic access to XNAT.

5 Minimum Technical Requirements for Tier 2 Data Federation and Interoperability Framework (Federated Query)

The implementation of a federated query system across the EUCAIM data federation requires all nodes to adhere to specific interoperability requirements. These requirements ensure that the EUCAIM federated query service can seamlessly query, retrieve and process data from diverse repositories, ranging from fully established repositories to environments without pre-existing infrastructures.

5.1 Minimum Interoperability Requirements for the Clinical and Imaging Data (at record level/patient level)

Executing queries relies on the operations of several local components. *Beam Proxy* is a component used to connect to the central infrastructure (*Beam Broker*), which handles network access through proxy systems and firewalls, as well as cryptographic operations, such as encrypting, decrypting, signing of messages, and validating the signatures. By using *Beam*, the applications themselves do not need to deal with those tasks individually. *Focus* is a query dispatcher which connects to the central infrastructure using *Beam Proxy*, retrieves tasks for that provider, transforms the data structure containing the query, and sends the query to a defined endpoint (the *Mediator*, or directly to the data store). The *Mediator* translates the abstract syntax tree containing the query into the query language of the data store. Even though the queries are executed on a record-level locally, they return only aggregate results (collection based). *Focus* then optionally obfuscates the results, and returns them to the central components using *Beam Proxy*, to be displayed in the Federated Search.

If a federated node is an already established repository with its own data model, it can qualify as a Tier 2 node without requiring a complete transformation of its dataset to the EUCAIM CDM. Instead, the following requirements must be met:

- **Implementation of a Mapping Component:** A mapping component must be implemented, to translate the requested minimum set of clinical and imaging attributes (described in section 4.2) into the EUCAIM concepts. This mapping ensures that queries expressed using the EUCAIM concepts (e.g. “imaging modality”, “age at diagnosis”) can be executed seamlessly across all participating nodes.
- **Query Alignment:** In addition, this mapping component must be responsible for translating the search query (AST syntax), into the local node’s query language.

For nodes lacking a pre-existing database or infrastructure, EUCAIM encourages that datasets are directly transformed to meet the EUCAIM CDM specification:

- **Direct CDM Transformation:** All datasets can be transformed directly into the EUCAIM CDM specification through an ETL (Extract, Transform, Load) process. This transformation ensures automatic compliance with Tier 3 data requirements.
- **No Mapping Component Needed:** Since the datasets adhere to the CDM structure, the federated query can be directly executed through *Focus* without additional mapping.

Tables 14 and 15 present examples of "mandatory" and "mandatory if available" query criteria (defined in the hyper-ontology version 1.2[61]), respectively, used for basic federated queries. Examples of additional attributes, which will be considered for the max-FIF, are presented in Table 16.

***Table 14**. Examples of "mandatory" query criteria (min-FIF) defined in the EUCAIM hyper-ontology*

Category	**Query Criteria**		**Standard Source**	**EUCAIM ID**
Patient Demographics	**Sex**	Male	LOINC:LA2-8	COM1000180
		Female	LOINC:LA3-6	COM1000177
		Unknown	LOINC:LA4489-6	COM1001289
	Age at diagnosis		SNOMEDCT:423493009	COM1000131
Clinical Parameters	**Diagnosis**	Malignant neoplasm of prostate	ICD10:C61	CLIN1000075
		Malignant neoplasm of breast	ICD10:C50	CLIN1000060
		Malignant neoplasm of colon	ICD10:C18	CLIN1000057
		Malignant neoplasm of rectum	ICD10:C20	CLIN1000185
Image Parameters	**Date of image acquisition**			
	Modality	MR	RadLex:RID10312	IMG1000038
		CT	RadLex:RID10321	IMG1000042
		PET	RadLex:RID10337	IMG1000062

[61] https://doi.org/10.5281/zenodo.14765570

	Body part	Prostate	ICDO3:C61	BP1000021
		Breast	ICDO3:C50	CLIN1063727
		Colon	ICDO3:C18	CLIN1063722
	Manufacturer	General Electric (GE)	birnlex_12833	IMG1000047
		Philips	birnlex_3065	IMG1000046
		Siemens	birnlex_3066	IMG1000044

***Table 15**. Examples of "mandatory if available" query criteria (min-FIF) defined in the EUCAIM hyper-ontology version 1.2*

Category	**Query Criteria**	**Standard Source**	**EUCAIM ID**
Clinical Parameters	Year of diagnosis (extracted from date of diagnosis)		
	Date of first treatment		
Pathology confirmation	Biopsy	SNOMEDCT:86273004	CLIN1001712
	Excision	SNOMEDCT:65801008	CLIN1004598
Pathology	Adenocarcinoma	ICDO3:8140/3	CLIN1047138
	Papillary carcinoma	ICDO3:8050/3	CLIN1063544
Imaging procedure protocol	CT of thorax with contrast	SNOMEDCT:75385009	IMG1000076
	CT of breast	SNOMEDCT:241539009	IMG1000078
Treatment	Prostatectomy	SNOMEDCT:90470006	CLIN1000248
	Chemotherapy	SNOMEDCT:367336001	CLIN1024528
	External Beam Radiation Therapy (EBRT)	SNOMEDCT:33195004	CLIN1005277

***Table 16**. Examples of strongly recommended query criteria (max-FIF) defined in the EUCAIM hyper-ontology version 1.2*

Category	**Query Criteria**		**Standard Source**	**EUCAIM ID**
Tumor Marker Test	Estrogen receptor Ag [Presence] in Breast cancer specimen by Immune stain		LOINC:85337-4	CLIN1045815
	HER2 [Presence] in Breast cancer specimen by Immune stain		LOINC:85319-2	CLIN1045851
	Prostate specific Ag [Mass/Volume] in Serum or Plasma		LOINC:2857-1	CLIN1033410
Tumor Marker Test Result	Positive		LOINC:LA6576-8	COM1001310
	Negative		LOINC:LA6577-6	COM1001332
Cancer stage	**cM category**	cM0	NAACCR:960@c0	COM1001860
		cM1	NAACCR:960@c1	COM1000246
		cM1a	NAACCR:960@c1A	COM1000258
		cM1b	NAACCR:960@c1B	COM1000271
		cM1c	NAACCR:900@c1C	COM1001876
	pM category	pM0	NCIT:C48740	COM1001845
		pM1	NAACCR:960@p1	COM1000531
		pM1a	NAACCR:960@p1A	COM1000558
		pM1b	NAACCR:960@p1B	COM1000586
		pM1c	NAACCR:960@p1C	COM1000615
Histologic Grade	**Gleason scoring system for malignant neoplasm of**	Gleason grade score 2 out of 10	SNOMEDCT:49878003	CLIN1025818
		Gleason grade score 3 out of 10	SNOMEDCT:46677009	CLIN1025912

	prostate	Gleason grade score 4 out of 10	SNOMEDCT:18430005	CLIN1026007
	International Society of Urologic Pathology prostate cancer staging system	International Society of Pathology histologic grade group 1	SNOMEDCT:1525761000004109	COM1001362
		International Society of Pathology histologic grade group 2	SNOMEDCT:1525771000004101	COM1001365
		International Society of Pathology histologic grade group 3	SNOMEDCT:1525781000004103	COM1001367
Tumor	**BI-RADS assessment**	BI-RADS 0	RadLex:RID36036	IMG1005470
		BI-RADS 1	RadLex:RID36028	IMG1005469
		BI-RADS 2	RadLex:RID36029	IMG1005468
		BI-RADS 3	RadLex:RID36041	IMG1005466
	PI-RADS assessment (Lesion)	PI-RADS 1	RadLex:RID50296	IMG1005487
		PI-RADS 2	RadLex:RID50297	IMG1005486
		PI-RADS 3	RadLex:RID50298	IMG1005485
		PI-RADS 4	RadLex:RID50299	IMG1005484

5.2 Guidelines for Federated Query support

5.2.1 Dataset in a Federated Node

Guidelines for installing the Mediator Service

The local components are designed to be easy to deploy and easy to maintain with Kubernetes and Docker-Compose based deployment packages provided.

Here's the template of the docker-compose.yml file for running Focus and Beam proxy (alongside the provider's store and optionally a query translating Mediator, which need to be added as separate containers):

```
version: '3.8'
services:
```

```
  beam-proxy:
    image: samply/beam-proxy:main # use tag develop for Test environment
    environment:
      BROKER_URL: ${BROKER_URL}
      PROXY_ID: ${PROXY_ID}
      APP_focus_KEY: ${APP1_KEY}
      PRIVKEY_FILE: /run/secrets/proxy.pem
      BIND_ADDR: 0.0.0.0:8081
    secrets:
      - proxy.pem
      - root.crt.pem
   focus:
      image: samply/focus:main # use tag develop for Test environment
      environment:
        BEAM_PROXY_URL: http://beam-proxy:8081
        BEAM_APP_ID_LONG: focus.${PROXY_ID}
        API_KEY: ${APP1_KEY}
        RETRY_COUNT: 10
        ENDPOINT_URL:  [endpoint to query - root of the API]
        ENDPOINT_TYPE: omop # regardless of the type of store
        PROVIDER: [the name of the provider as it should appear in the
results table]
        PROVIDER_ICON: [the logo of the provider, base64 encoded]
secrets:
  proxy.pem:
    file: ./${PROXY1_ID_SHORT}.priv.pem
  root.crt.pem:
    file: ./root.crt.pem
```

BROKER_URL is:

- **Test:** https://broker-eucaim.grycap.i3m.upv.es
- **Production:** https://broker.eucaim.cancerimage.eu

PROXY_ID_SHORT is the name of the provider, lowercase, and without spaces, as registered with the Beam Broker.

PROXY_ID is ${PROXY_ID_SHORT}.{BROKER_ID}

APP1_KEY is a string the provider sets themself (a password generator, such as pwgen, is recommended).

The commands to generate a private key and corresponding CSR for the Production environment:

openssl genrsa --out [PROXY1_ID_SHORT].priv.pem 2048

openssl req -key [PROXY1_ID_SHORT].priv.pem -new -subj "/CN=[PROXY_ID]/C=[2 letter ISO country code]/L=[city]" -out [PROXY1_ID_SHORT].csr

The generated CSR needs to be sent to the admins of the central components for enrollment in the Beam Broker.

Guidelines for creating a mapping component

Central components send the selected search criteria in an abstract syntax tree with format:

```
AST:
{
       ast: OPERATION,
       id: "uuid__search__uuid"
}

OPERATION:
{
       operand: "AND"|"OR"
     children: [OPERATION|CONDITION]
}

CONDITION:
{
      key: string,
       type:
"EQUALS"|"NOT_EQUALS"|"IN"|"BETWEEN"|"LOWER_THAN"|"GREATER_THAN"|"CONTAINS",
       value: string | [string] | boolean | number | NUM_RANGE | DATE_RANGE | date
}

NUM_RANGE:
{
       min: number,
       max: number
}

DATE_RANGE:
{
              min: date | undefined
              max: date | undefined
}
```

Here's an example of such an AST containing search criteria:

```
{
       "ast": {
       "children": [
              {
              "children": [
                     {
                            "children": [
                            {
                                   "key": "SNOMEDCT263495000",
                                   "system": "",
                                   "type": "EQUALS",
                                   "value": "SNOMEDCT248153007"
                            }
                            ],
                            "operand": "OR"
                     },
                     {
                            "children": [
```

```
                        {
                                "key": "SNOMEDCT439401001",
                                "system": "urn:snomed-org/sct",
                                "type": "EQUALS",
                                "value": "SNOMEDCT363406005"
                        }
                        ],
                        "operand": "OR"
                }
        ],
        "operand": "AND"
        },
        {
        "children": [
                {
                        "children": [
                        {
                                "key": "SNOMEDCT263495000",
                                "system": "",
                                "type": "EQUALS",
                                "value": "SNOMEDCT248152002"
                        }
                        ],
                        "operand": "OR"
                }
        ],
                {
                        "children": [
                        {
                                "key": "SNOMEDCT439401001",
                                "system": "urn:snomed-org/sct",
                                "type": "EQUALS",
                                "value": "SNOMEDCT254837009"
                        }
                        ],
                        "operand": "OR"
                },
        "operand": "AND"
        }
    ],
    "operand": "OR"
    },
    "id":    "4e9cce7d-f544-452e-b588-ae21f5130280__search__4e9cce7d-f544-452e-b588-
ae21f5130280"
}
```


This abstract syntax tree needs to be translated into the query language that the local store supports, for example SQL.

For instance, a query like "finding patients diagnosed with prostate cancer, aged between 50 and 60 at the time of diagnosis, and who have undergone an MR scan" would be translated into an SQL query on an OMOP-CDM PostgreSQL database schema (with a radiology extension) as follows:

```
WITH prostate_cancer_concept AS (
SELECT concept_id
FROM concept
WHERE concept_code ='93974005' -- "Primary malignant neoplasm of prostate"
AND vocabulary_id = 'SNOMED'
AND domain_id = 'Condition'
AND standard_concept = 'S'
),
```

```
descendant_concepts AS (
SELECT DISTINCT ca.descendant_concept_id
FROM concept_ancestor ca
JOIN prostate_cancer_concept pcc ON ca.ancestor_concept_id = pcc.concept_id
WHERE ca.min_levels_of_seperation>0
),
patients_with_prostate_cancer AS (
SELECT DISTINCT co.person_id
FROM condition_oocurrence co
JOIN descendant_concepts dc ON co.condition_concept_id = dc.descendant_concept_id
JOIN person p ON co.person_id = p.person_id
WHERE (EXTRACT(YEAR FROM co.condition_start_date) - p.year_of_birth BETWEEN 50 AND 60
),
patients_with_mr_imaging AS (
SELECT DISTINCT ist.person_id, ist.imaging_study_id
FROM imaging_study ist
JOIN imaging_series ise ON ist.imaging_study_id = ise.imaging_study_id
WHERE ise.modality_concept_id = 'RID10312'
)

SELECT
      COUNT (DISTINCT mris.person_id) AS num_patients, --  returns #patients
      COUNT (DISTINCT mris.imaging_study_id) AS num_studies -returns #studies
FROM patients_with_mr_imaging mris
JOIN patients_with_prostate_cancer pwp ON mris.person_id = pwp.person_id;
```

The return type is a JSON containing following fields:

```
{
      "collections": [
            {
                  "name": "Collection Name",
                  "id": "Collection ID",
                  "studies_count": <result of query>,
                  "subjects_count": <result_of_query>,
                  "age_range": {"min": , "max"},
                  "gender": ["male", "unknown"],
                  "modality": ["MRI", "PET-CT"],
                  "body_parts": ["RECTUM", "PROSTATE", "PELVIS"],
                  "description": "Short description of collection"
            },
            …
      ]
}
```

Workflow and Data Preprocessing/Interoperability Tools (QUIBIM, MEDEX)

Additional pre-processing tools are available to DHs to prepare their Tier 2 datasets. Correct de-identification of imaging data and metadata is possible using dedicated EUCAIM tools, while further tools for data quality checking may be used. The EUCAIM ETL serves to ingest and transform both the clinical data and imaging metadata to query.

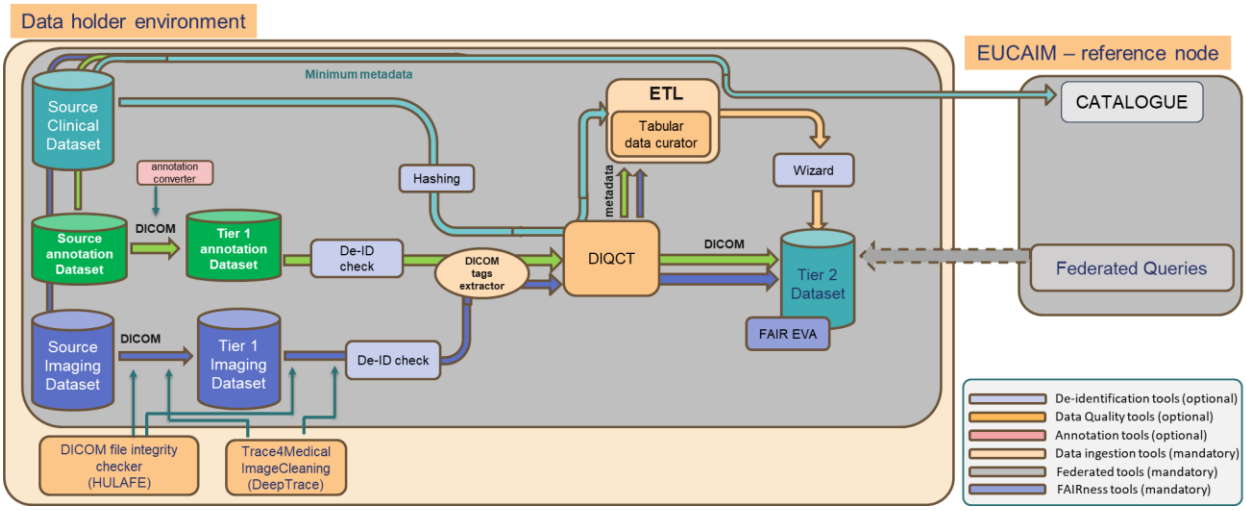


Figure 15. Workflow and pre-processing tools for Tier2

Format standardization

To allow the integration into the Federated Query, at least the minimum imaging and clinical attributes must be structured in a way that supports querying (e.g. JSON in a MongoDB/PostgreSQL database). Image Data should be in DICOM format and annotation in DICOM-SEG.

Guidelines for Data Quality

The same data quality framework as for Tier 1 applies to Tier 2 datasets, which must comply with the same rules (see section 4.2).

Tier 2 data holders will be provided with a set of optional data quality tools to address the following:

- Completeness
- Uniqueness
- Validity
- Consistency
- Accuracy
- Integrity

Table 17 lists the EUCAIM tools that may be used by Tier 2 data holders.

***Table 17**: EUCAIM tools for Tier 2*

Dimension	Tool	Description	Quality level	Output / Metrics
Completeness	**ETL**	The ETL will exclude cases with missing mandatory metadata	Data	only completed cases processed
	Tabular Data Curator	The tool identifies missing data	Data	

Dimension	Tool	Description	Quality level	Output / Metrics
	DIQCT	The tool identifies data that fulfil the pre-defined criteria for mandatory information	Data & Dataset	the percentage of cases with complete entries in the clinical and imaging dataset
Uniqueness	**DIQCT**	The tool identifies possible duplicate records in the dataset (clinical and imaging data)	Dataset	The total number of identified duplicate image files or series within the dataset The absolute number of duplicate cases identified for the clinical dataset.
Validity	**ETL**	The tool will return an error if the clinical data file is not valid	Dataset	only valid cases processed
	Dicom tag extractor	The tool will only process DICOM files	Data	only DICOM files processed
	Wizard	The tool will support the identification of risks and propose ways to mitigate them.	Data & dataset	Modified metadata (in CSV format)
	Trace4MedicalImage	For 2D ultrasound and mammography DICOM files: detect and remove encapsulated text, as they may contain potentially identifying information	Data	
	DIQCT	The tool checks both clinical and imaging data for valid entries based on the predefined validity requirements.	Data	- the percentage of images that do not require any further de-identification - the percentage of series that do not include any image that is not de-identified correctly - the percentage of patients that are correctly de-identified and thus they do not include any identified image - the percentage of clinical values that are valid according to pre-defined format, type and range - percentage of imaging annotations valid according to EUCAIM guidelines

Dimension	Tool	Description	Quality level	Output / Metrics
Consistency	**DIQCT**	The tool checks the link between images and clinical metadata	Dataset	The percentage of imaging modalities properly integrated in the dataset
	Tabular Data Curator	The tool identifies data inconsistencies	Data	
Accuracy	**DICOM File Integrity checker**	The tool performs a quality check in terms of the correct number of files per each serie	Data & Dataset	Comparison between the expected quantity of images by serie and existing quantity when they do not match
	DIQCT	The tool measures how well the clinical data structure conforms to a specific template	Dataset	The absolute number of inaccuracies in the clinical dataset
Integrity	**DICOM File integrity checker**	This tool looks for corrupted files to identify	Data & Dataset	List of the corrupted DICOM imaging files
	DIQCT	This tool checks the series in which a specific segmentation mask can be applied.	Data level	the percentage of annotation series in which the segmentation mask is relevant, is extracted.

Guidelines for Data Annotation Quality

In Tier 2, the capabilities of DHs with respect to annotation tools are extended in the context of their quality evaluation. This is achieved through the DIQCT tool, in particular its module dedicated to annotation. This module provides a command line tool to validate DICOM SEG files against predefined requirements specified in an Excel file. It contains components for finding relevant DICOM files, loading and parsing validation requests and applying validation rules. The main validation process checks each DICOM file for compliance with the Type 1, 1C, 2, 2C and 3 attributes specified in the requirements file. A detailed report is generated, highlighting issues such as missing, invalid or conditionally required attributes, including file paths and affected DICOM tags. The tool is designed to ensure data integrity and compliance with DICOM standards as described in the Guidelines for Data Quality of section 4.3.1.

Guidelines for ETL

In order to allow queries on Tier 2 data and metadata, their harmonised transformation is required. This is achieved by the EUCAIM extract-transform-load (ETL) tool, that is able to ingest, transform, and map the data to a standard. Any type of files can be ingested as source data. The ETL works as a series of processors, easily configurable, that allows a step-by-step approach. The mappings are encapsulated in a single processor to minimize the effort of data holders. The tool converts the entry files and runs quality checks to ensure proper transformation. It allows data

holder users to fully accept the automatic quality checks, review them, or correct them manually. Additional information is provided in deliverable D5.5, with a demonstration video of the ETL.

Guidelines for Data FAIRification

For Tier 2, further FAIRness requirements have been set as explained in deliverable D4.4 Annex 6. On top of having to meet several new RDA indicators, information that will allow users to localise datasets with data that would be relevant to their research questions using the 27 mandatory attributes defined for the EUCAIM metadata catalogue must be included.

As explained for Tier 1 (section 4.3), FAIR EVA will be used for testing the compliance with these requirements accessing the metadata through a FDP. If the dataset complies with them, it will be awarded an approval for Tier 2.

Technical Requirements for Federated Node Setup

Hardware setup

The hardware requirements for Tier 2 local nodes are described below. These requirements were selected according to the expected workload demands by the federated data search.

***Table 18**: Hardware requirements for Tier 2 local nodes*

Hardware	Minimum
CPU	4 Cores /8 Threads
RAM	32 GB
Operating System Drive	160+ GB SSD
Data Storage	1x (Dataset size) Drives

A RAID 1 Configuration that mirrors the data on a second disk is highly recommended for improved data protection and fault tolerance.

Network requirements

Each local node must be connected to the public internet via a stable (outgoing) connection, with a minimum symmetrical bandwidth of 200 Mbps to avoid performance bottlenecks.

Firewall and network policy exceptions must be made to allow the local node access to the following online resources (required for Tier 2 nodes and above):

- https://github.com (for access to code projects)
- https://docker.verbis.dkfz.de or the official docker hub (for access to pre-built docker images). If you choose the latter, you should make sure that the appropriate URLs are added to your firewall allowlist to ensure Docker images can be pulled properly from Docker Hub within your organization (see https://docs.docker.com/desktop/setup/allow-list/).
- https://broker.eucaim.cancerimage.eu (for access to the EUCAIM federation Beam broker)

Relevant network infrastructural adjustments, such as firewall configurations, must be made to allow outbound access on port 443 (HTTPS) to allow for the interactions required by the federated search component.

Organisations which use added security protocols (e.g.,VPN, packet monitoring), must notify and collaborate with EUCAIM's Technical Support Team via Helpdesk.

Software requirements

To install, interact and configure the EUCAIM software components required to provide access to the data stored in Tier 2 local nodes, data holders are required to use an operating system that is compatible with EUCAIM's software stack. This includes stable Linux distributions such as Ubuntu, CentOS, or Debian for Tier 2 nodes.

To provide data access through Tier 2 participation, some software packages are required to allow interaction with the EUCAIM federated search component. The software requirements include:

a) Docker Engine must be installed, version >=20.10.1 (https://docs.docker.com/engine/)
b) Docker Compose must be installed >=2.21 (https://docs.docker.com/compose/)
c) Git must be installed, version >=2.0 (available through the built-in OS package manager)
d) Systemd must be installed (available through the built-in OS package manager)
e) Curl must be installed (available through the built-in OS package manager)

After ensuring the installation of the aforementioned dependencies, Tier 2 nodes can proceed with the installation of following modules which enable interaction with the EUCAIM federated search:

- The Focus query dispatcher component, offered as a Docker container (https://github.com/samply/focus/)
- The Beam proxy component, also offered as a Docker container (https://github.com/samply/beam/).

More information about the aforementioned modules can be found in Sections 3.4 and 3.4 above.

It must be noted that depending on the storage systems and structure of their data, Data Holders may be required to implement a local Mediator service that transforms the queries and data formats expected by the federated search component to the ones corresponding to the local storage system and passes the response in the appropriate format. Instructions about how to configure the Mediator service are provided in Section 5.2 above.

5.2.2 Dataset in the EUCAIM Reference Node

The reference nodes implement a Mediator component to transform the queries of the federated search into the specific syntax of the searching endpoints. Therefore, the main requirement for the Data Holders is to adapt or describe clearly the transformation of the searchable data fields. In particular, the reference nodes require:

- **UPV node**. A common mediator service is available. This service queries the databases (DICOM tags and clinical data in the NoSQL database), retrieves the information in the UPV node native format and transforms it to the format expected by the federated search. In this case, it is desirable that the 27 searchable fields described in this section are coded

in the standard format. Eventually, the mediator can be customized to have different transformations for different providers (which are coded as different projects in the UPV node), but this requires software code development at the UPV side. This will be performed if necessary.

- **Euro-BioImaging Node**. Similarly to the UPV Node, the mandatory searchable fields for supporting federated queries need to be supplied with the data. An adaptor between the XNAT search API and Beacon v2 will be released in 2025. Based on an opt-in mechanism, data holders can indicate if they want their data to be searchable and thus indicate if their data can comply with the EUCAIM tier 2 specifications.

Workflow for Data Preprocessing and Interoperability Tools

The process to transfer Tier 2 datasets to EUCAIM reference nodes is similar to what was described for Tier 1. Similar tools for data preparation as described above are available, although the ETL will not be used before data transfer.
Note that the same workflow applies to Tier 3 data that are to be transferred.

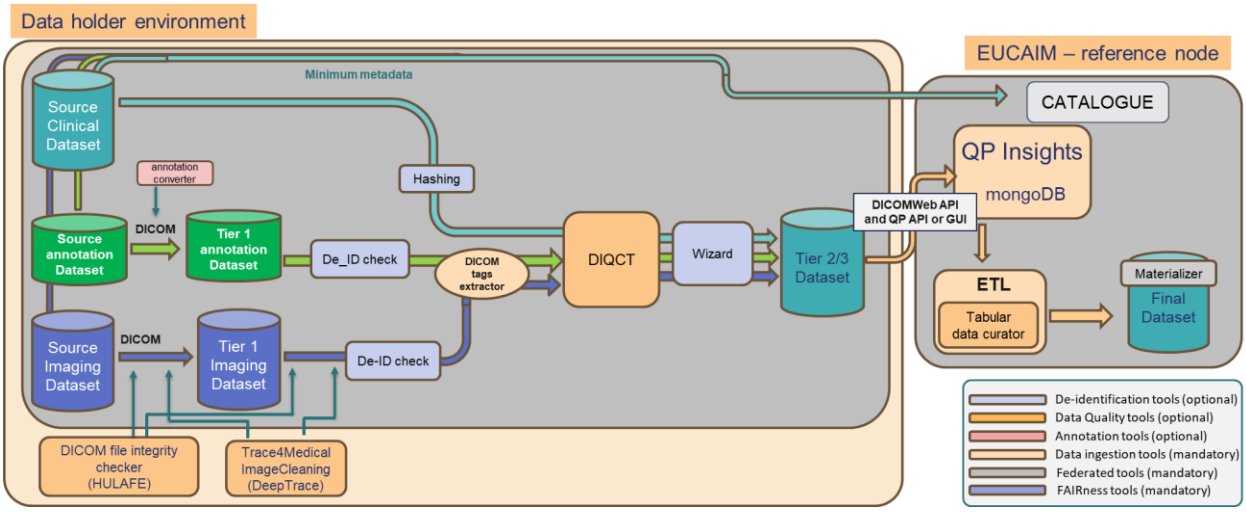


Figure 16. Workflow and data preparation tools for Tier 2 data transfer

6 Minimum Technical Requirements for Tier 3 Data Federation and Interoperability Framework (Federated Processing)

Tier 3 achieves interoperability at the level of federated processing. Its main purpose is to standardise data (clinical and imaging) complying with the EUCAIM data model, including the EUCAIM CDM and hyper-ontology and to ensure technical support for seamless data integration. Semantically, interoperability is maintained by adopting FAIR-compliant standard terminologies to map healthcare information to standard resources and ensuring semantic mapping or alignment with the CDM and hyper-ontology. Technically, a set of guidelines, including software and hardware requirements, is defined to support the federated processing.

6.1 Minimum Interoperability Requirements for the Clinical and Imaging Data (at record level/patient level)

Minimum interoperability requirements for the clinical and imaging data are defined at the patient level to ensure a consistent data exchange and integration across data repositories. These requirements specify a minimum set of clinical and imaging metadata reflecting the essential information required for basic federated processing and ensure that these metadata are structured, standardized, and interoperable, complying with the EUCAIM CDM and hyper-ontology (for the first version please refer to D5.2). A terminology-binding process is performed to support a coherent interpretation and understanding of essential information between the CDM and hyper-ontology.

Minimum Imaging Metadata

At Tier 3, a minimum set of imaging metadata must be extracted and standardized as represented in Table 19. Figure 17 depicts a Unified Modeling Language (UML)[62] model or diagram that clarifies the structure of the minimum imaging metadata and the semantic interactions and properties based on the hyper-ontology (v1.2) semantic content (concepts, object and data properties), which is semantically aligned with the CDM.

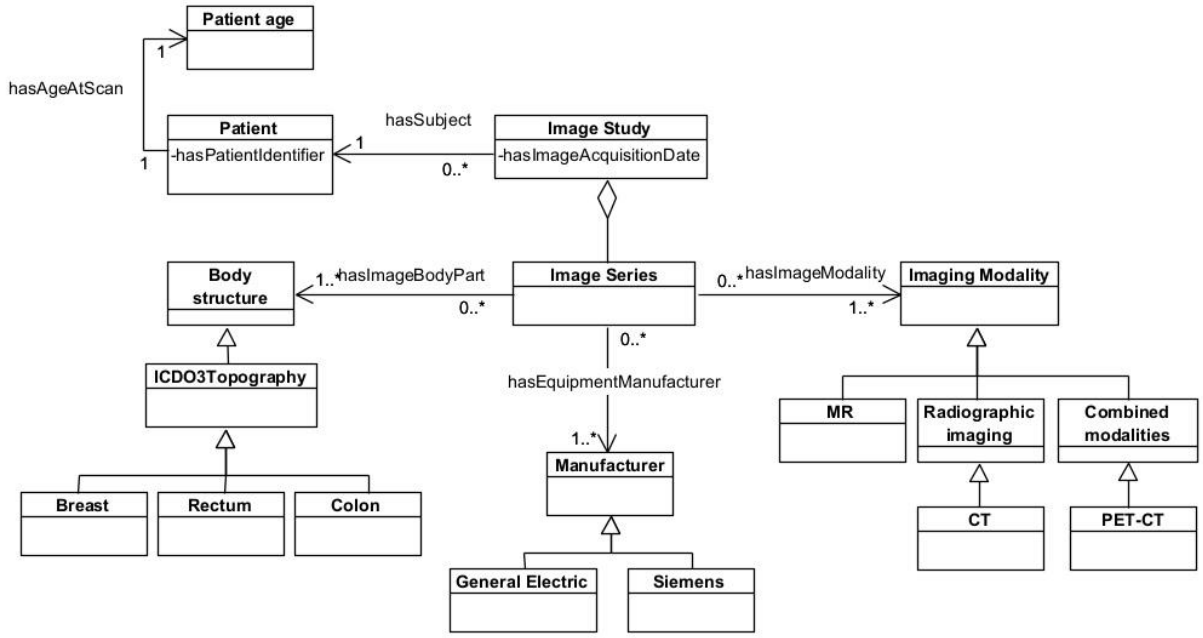


Figure 17. A UML diagram representing part of the structure of the minimum imaging metadata and their semantic interactions as defined in the hyper-ontology.

Prefix eucaim: <https://cancerimage.eu/ontology/EUCAIM#> (Hyper-Ontology IRI)

Table 19: *The minimum imaging metadata and their mapping to EUCAIM CDM and hyper-ontology*

[62] https://www.uml.org/

Variable	CDM attribute	Hyper-ontology property	Vocabulary used (object properties) /Data type (data properties)	Examples of possible values
Patient ID **(**DICOM tag : 0010,0020**)**	Patient.identifier	eucaim:hasPatientIdentifier	xsd:Literal/String	"ECI_125HV8234BD7"
Image modality **(**DICOM tag : 0008,0060**)**	ImageSeries.modality	eucaim:hasImageModality	RadLex, SNOMED-CT	**subclasses of "Imaging Modality"** - Magnetic resonance imaging (MR) (eucaim:IMG1000038) - Computed tomography imaging (CT) (eucaim:IMG1000042) - Positron emission tomographic imaging (PET) (eucaim:IMG1000062) - Single photon emission computed tomography (SPECT) (eucaim:IMG1000061) - PET-CT (eucaim:IMG1004451) - PET-MR (eucaim:IMG1004452) - Mammography (eucaim:IMG1004455)
Image body part **(**DICOM tag : 0018,0015**)**	ImageSeries.bodySite	eucaim:hasImageBodyPart	ICD-O-3, SNOMED-CT, RadLex, OMOP (Cancer Modifier)	**subclasses of "Body structure"** - Breast (eucaim:CLIN1063727) - Colon (eucaim:CLIN1063722) - Cecum (eucaim:CLIN1063731) - Rectum (eucaim:CLIN1063724) - Prostate (eucaim:BP1000021) - Neck and chest (eucaim:BP1000233) - Lung (eucaim:BP1000113) - Adrenal gland (eucaim:BP1000084) - Axillary lymph node (eucaim:BP1000158)
Image manufacturer **(**DICOM tag : 0008,0070**)**	ImageSeries.equipmentManufacturer	eucaim:hasEquipmentManufacturer	Birnlex	**subclasses of "Manufacturer"** - General electric (GE) (eucaim:IMG1000047) - Siemens (eucaim:IMG1000044) - Philips (eucaim:IMG1000046) - Toshiba (eucaim:IMG1000045) - Agfa (eucaim:IMG1000051) - Fujifilm (eucaim:IMG1000060)

Variable	CDM attribute	Hyper-ontology property	Vocabulary used (object properties) /Data type (data properties)	Examples of possible values
Date of image acquisition (DICOM tag : 0008,0022)	ImageStudy.acquisitionDate	eucaim:hasImageAcquisitionDate	xsd:dateTime/dateTime	2023-01-01

Minimum Clinical Metadata

All imaging data must be accompanied by a set of minimum clinical metadata. At this Tier, the following clinical metadata shall be extracted, aggregated and populated in the EUCAIM public catalogue, based on the EUCAIM DCAT-AP specification described in section 4.1. Once extracted they should get standardized as presented in the following tables. To clarify the structure and semantics of the minimum clinical metadata and the associated possible values, as defined in the hyper-ontology (v1.2), a UML diagram is provided (Figure 18). The mappings of these semantics with the CDM are given in the following tables.

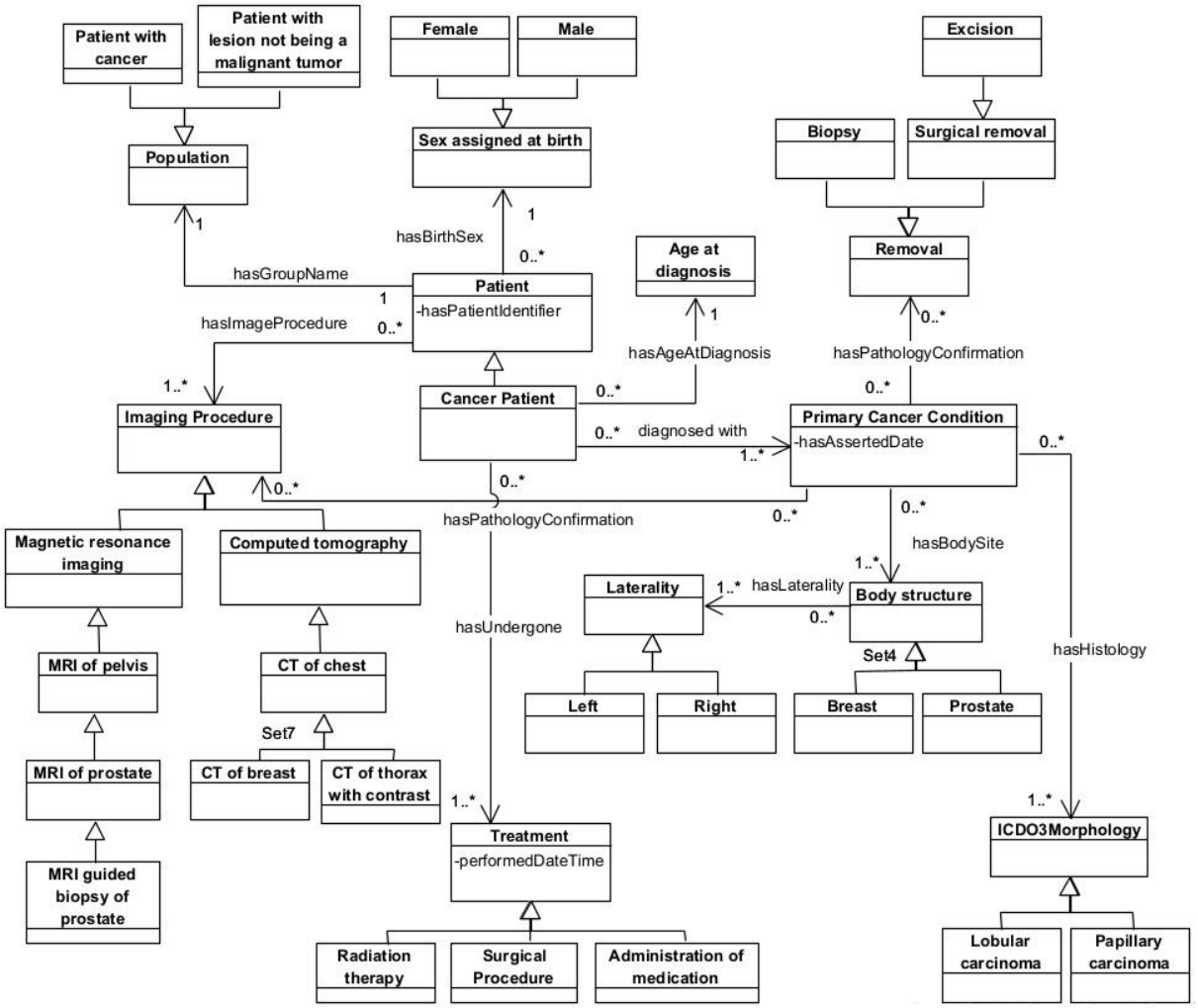


Figure 18. A UML diagram representing part of the structure of the minimum clinical metadata and their semantic interactions as defined in the hyper-ontology.

For confirmed positive or diagnostic patient cases, the clinical data presented in Table 20 must be available.

***Table 20**: The minimum clinical metadata for confirmed positive or diagnostic patient cases and their mapping to EUCAIM CDM and hyper-ontology*

Variable Type	Variable	CDM attribute	Hyper-ontology property	Vocabulary used (object properties) /Data type (data properties)	Possible values -Example Value
Mandatory	**Patient ID**	Patient.identifier	eucaim:hasPatientIdentifier	Literal/String	"ECI_125HV8234BD7" X123456

Variable Type	Variable	CDM attribute	Hyper-ontology property	Vocabulary used (object properties) /Data type (data properties)	Possible values -Example Value
Mandatory	**Population** (Categorization of the subjects in the dataset based on their status.)	Group.name	eucaim:hasGroupName	EUCAIM	**subclasses of "Population"** - Patient with Cancer" (eucaim:COM1001087) - Patient with lesion not being a malignant tumor" (eucaim:COM1001086)
Mandatory	**Sex**	Patient.birthSex	eucaim:hasBirthSex	LOINC, SNOMED-CT	**subclasses of "Sex assigned at birth"** - Female (eucaim:COM1001370) - Male (eucaim:COM1001366) - Unspecified (eucaim:COM1001288)
Mandatory if available	**Date of radiology detection*** Date when the tumor or the lesion was first detected by an imaging study (or the nearest study to the diagnosis confirmation).	This is equivalent to the "Date of image acquisition" in the Imaging metadata table for the imaging study in which the tumor or the lesion was first detected.			
Mandatory if available	**Date of pathology confirmation / diagnosis date*** Date when the tumor is histologically confirmed (or confirmed by an imaging study if histology was not performed, in specific cases such as	PrimaryCancerCondition.assertedDate	eucaim:hasAssertedDate	xsd:dateTime	01-01-2024

Variable Type	Variable	CDM attribute	Hyper-ontology property	Vocabulary used (object properties) /Data type (data properties)	Possible values -Example Value
	HCC)				
Mandatory	**Age at diagnosis (years, with one decimal)**	PrimaryCancerCondition.AgeAtDiagnosis PrimaryCancerCondition.AgeUnitConcept	eucaim:hasAgeAtDiagnosis	Decimal, UCUM	45,5 -Year (eucaim:COM1000151)
Mandatory if available	**Pathology Confirmation** (Method used to confirm the pathology (histological (surgery, biopsy) or by imaging in specific cases such as HCC). The method used before the treatment decision will be considered.)	Procedure.code	eucaim:hasPathologyConfirmation	SNOMED-CT, CPT4, ICD10PCS	**subclasses of "Removal"** - Biopsy (eucaim:CLIN1001712) - Excision (eucaim:CLIN1004598) **subclasses of "Imaging procedure"** - Computerized Tomography (CT Scan) of Chest, Abdomen and Pelvis (eucaim:IMG1000027) - CT of thorax with contrast (eucaim:IMG1000076) - MRI of breast for screening for malignant neoplasm (eucaim:IMG1016279) - MRI guided biopsy of prostate (eucaim:IMG1005608) - X-ray guided biopsy (eucaim:IMG1016754)
Mandatory only for the organ	**Topography** Location of the lesion, stratified in three steps: organ, region, and laterality	PrimaryCancerCondition.bodySite PrimaryCancerCondition.bodySite.location PrimaryCancerCondition.bodySite.laterality	eucaim:hasBodySite eucaim:hasLocation	ICD-O-3, SNOMED-CT, RadLex, OMOP (Cancer Modifier)	**subclasses of "Body structure"** - Breast (eucaim:CLIN1063727) - Colon (eucaim:CLIN1063722) - Cecum (eucaim:CLIN1063731) - Rectum (eucaim:CLIN1063724) - Prostate (eucaim:BP1000021)

Variable Type	Variable	CDM attribute	Hyper-ontology property	Vocabulary used (object properties) /Data type (data properties)	Possible values -Example Value
					subclasses of "Prostate" Apical anterior fibromuscular stroma of prostate (eucaim:BP1000187) - Apical peripheral zone of prostate (eucaim:BP1000017)
			eucaim:hasLaterality	RadLex, SNOMED-CT	**subclasses of "Laterality"** - Left (eucaim:IMG1016670) - Right (eucaim:IMG1016682)
Mandatory if available	**Imaging procedure protocol** (Specific protocol applied to obtain the diagnostic image)	Procedure.code	eucaim:hasImageProcedure		**subclasses of "Imaging procedure"** - Computerized Tomography (CT Scan) of Chest, Abdomen and Pelvis (eucaim:IMG1000027) - CT of thorax with contrast (eucaim:IMG1000076) - MRI of breast for screening for malignant neoplasm (eucaim:IMG1016279) - MRI guided biopsy of prostate (eucaim:IMG1005608) - X-ray guided biopsy (eucaim:IMG1016754)
Mandatory if available	**Pathology** **(**Histology and histological subtype of the lesion (in ICDO-3, if available)**)**	PrimaryCancerCondition.HistologyMorphologyBehavior	eucaim:hasHistology	ICD-O-3, SNOMED-CT	**subclasses of "ICDO3Morphology"** - Adenocarcinoma (eucaim:CLIN1047138) - Basal cell adenocarcinoma (eucaim:CLIN1049464) - Lobular carcinoma (eucaim:CLIN1052212) - Papillary carcinoma (eucaim:CLIN1063544)

Variable Type	Variable	CDM attribute	Hyper-ontology property	Vocabulary used (object properties) /Data type (data properties)	Possible values -Example Value
					subclasses of "Neoplasm" - Acinar cell carcinoma (eucaim:CLIN1049490) - Small cell carcinoma (eucaim:CLIN1049520) - Adenosquamous carcinoma (eucaim:CLIN1049457)
Mandatory if available	**Treatment** (Type of treatment received by the patient)	CancerRelatedSurgicalProcedure.code Cancer-Related MedicationAdministration.code Radiotherapy Course Summary.modality	eucaim:hasUndergone	SNOMED-CT, CPT4, ICD10PCS	**subclasses of "Administration of medication"** - Chemotherapy (eucaim:CLIN1024528) - Hormone therapy (eucaim:CLIN1016082) **subclasses of "Surgical procedure"** - Prostatectomy (eucaim:CLIN1000248) - Radical mastectomy (eucaim:CLIN1004693) - Lumpectomy of breast (eucaim:CLIN1060316) - Excision of lymph node (eucaim:CLIN1051004) **subclasses of "Radiation therapy"** - External beam radiation therapy using electrons (eucaim:CLIN1057646) - External radiation therapy using superficial radiation (eucaim:CLIN1005278)
Mandatory if available	**Date of first treatment*** (Date when first treatment occurred)	Procedure.performedDateTime	eucaim:performedDateTime	xsd:dateTime	01-01-2024

*IMPORTANT NOTE: If dates are not available in the dataset, or have been altered due to anonymization purposes, relative days to a given baseline time point should be available

according to the dataset purposes.

For Negative screening and control groups the clinical data presented in Table 21 must be available.

***Table 21**: The minimum clinical metadata for negative screening and control groups*

Variable Type	Variable	CDM attribute	Hyper-ontology property	Vocabulary used (object properties) /Data type (data properties)	Possible values -Example Value
Mandatory	**Patient ID**	Patient.identifier	eucaim:hasPatientIdentifier	Literal/String	"ECI_125HV8234BD7" X123456
Mandatory	**Population** (The categorization of the subjects in the dataset based on their status)	Group.name	eucaim:hasGroupName	SNOMED-CT, EUCAIM	**subclasses of "National Public Health Classification data"** - Negative screening (eucaim:COM1002626) - Control group (eucaim:COM1002625) Subject on Screening with a negative result; Subject on a Control group.
Mandatory	**Sex**	Patient.birthSex	eucaim:hasBirthSex	LOINC, SNOMED-CT	**subclasses of "Sex assigned at birth"** - Female (eucaim:COM1001370) - Male (eucaim:COM1001366) - Unspecified (eucaim:COM1001288)
Mandatory if available	**Date of imaging acquisition** Date when imaging study occurred for screening or control group	This is equivalent to the "Date of image acquisition" in the Imaging metadata table for the imaging scan occurred for the screening/control group.			
Mandatory	**Age (years, with one decimal)**	ImagingStudy.AgeAtScan ImagingStudy.A	eucaim:hasAgeAtScan	Decimal, UCUM	45 -Year

Variable Type	Variable	CDM attribute	Hyper-ontology property	Vocabulary used (object properties) /Data type (data properties)	Possible values -Example Value
	Age of the subject when the imaging study was acquired	geUnitConcept (Equivalent to DICOM Patient's Age (0010,1010) if available or calculated through date of birth and image acquisition date)			(eucaim:COM1000151)
In negative screening and control group cases, region and laterality are not mandatory.	**Topography (**Area exam with the imaging modality: organ**)**	ImageSeries.bodySite	eucaim:hasImageBodyPart	ICD-O-3, SNOMED-CT, RadLex, OMOP (Cancer Modifier)	**subclasses of "Body structure"** - Breast (eucaim:CLIN1063727) - Colon (eucaim:CLIN1063722) - Cecum (eucaim:CLIN1063731) - Rectum (eucaim:CLIN1063724) - Prostate (eucaim:BP1000021)

6.2 Guidelines for Federated Processing support

6.2.1 Dataset in a Local Node

Hardware requirements

The hardware requirements for Tier 3 local nodes are described in Table 22. These requirements were selected according to the expected increased workload demands of data access for federated processing. The real-world processing workload may vary depending on the set of processing tools that the local node will run locally, and additional overheads from any non-EUCAIM-specific services running in parallel on the local node.

Below, we assume the hardware requirements entailed by the most demanding processing components of the EUCAIM toolbox. However, it is important to note that, over the lifespan of the project, new software and models may emerge requiring increased computational resources.

***Table 22**: The hardware requirements for Tier 3*

Hardware	Requirement	Notes
CPU	Minimum: ● Option 1: 16 Cores >=1.8GHZ ● Option 2: 12 Cores >=3.0Ghz Recommended: 32 Cores /64 Threads 3.0Ghz	● If a GPU is not present, a server-grade, high core-count CPU is necessary for the Second Prototype. ● If not comparable by cores, the ideal thread count is 24+.
RAM	Minimum: 64GB Recommended: 128 GB ECC	● DDR5 is ideal. ● ECC memory is highly recommended for stability.
Motherboard	4+ RAM Slot	● Make sure to double check the compatibility of selected CPUs with the Chipset of the motherboard. ● In the case of DDR5, double check motherboard compatibility with DDR5.
Storage	A 1x(Dataset size) is the minimum requirement, 2x(Dataset size) is recommended. Examples include: ● 521 GB SSD Drive for Operating System (Either NVMe M.2 PCI Gen4 or SATA III) ● 1TB++ SATA III Drive (SSD or HDD) for local storage of medical data	● M.2, NVMe, Gen4 Drives are suggested for the OS ● For data storage size, Data Holders (DH) are expected to plan their purchase depending on the size of the Data they will provide. 1TB is a minimum, with some DHs already planning for 2 TB + datasets. ● For data storage, SSD are preferred for speed but are not mandatory.

GPU	>150 Tensor Cores 16GB VRAM	• Maximizing the amount of Tensor Cores is a priority, most recent GPUs will generally have higher Tensor Core counts. • Ampere and Volta architectures are preferred.
Power Supply	-	• Each DH must make calculations depending on the hardware setup that will be selected to make sure that needed Wattage is covered and ideally exceeded to prepare for any future upgrades to the machine.

We recommend starting by querying the federated nodes to specify their technical capabilities and classify them accordingly. Some initial classification ideas include: no-GPU, GPU-low, GPU-mid, or GPU-high. These categories can encompass broader specifications such as the number of CPUs, RAM memory, and hard disk space to ensure nodes are appropriately categorized and utilized.

Network requirements

The network requirements for Tier 2 nodes are considered sufficient for Tier 3 nodes as well. Each DP must make best efforts to provide the best possible connection to their Node. Network performance will directly affect node stability and can invalidate AI training or prevent successful demonstrations of the platform.

Software requirements

Tier 3 nodes must deploy and configure the Federated Execution Manager (FEM) component in order to allow federated processing of data and images, according to the goals of EUCAIM researchers. In order to initiate this process, data holders must fill in a configuration file which will document specific details related to the connection of the FEM with the local infrastructure (document in progress). Moreover, data holders must complete the configuration file for Data Materializer Tool (DMT) so FEM can allocate the right data to each experiment (EUCAIM Data Materialization definition v2[63]).

The FEM daemon supports deployment within a wide range of infrastructure types, including a highly complex Kubernetes cluster, a SLURM Workload Manager, and containers from both Singularity and Docker. It also can deal with a custom API managing an intricate job queue system like UPV's Jobman. The setup is entirely up to the local node as long as some basic formatting and return codes (e.g., standard HTTP codes and JSON responses for execution IDs and similar metadata) are respected. More detailed instructions about how to deploy and interact with the FEM component are provided.

[63] https://docs.google.com/document/d/1h5-tOBrwZKeFsoLMLvBXKZa7A6hKriCF7zq5OiobUng/edit?usp=sharing

Workflow and Data Preprocessing/Interoperability Tools

Workflow and the set of tools for Tier 3 data preparation is identical to that of Tier 2, except that all clinical data and imaging metadata will be processed by the ETL, and not just the ones for federated queries. Similarly, all clinical data and metadata may go through quality check.

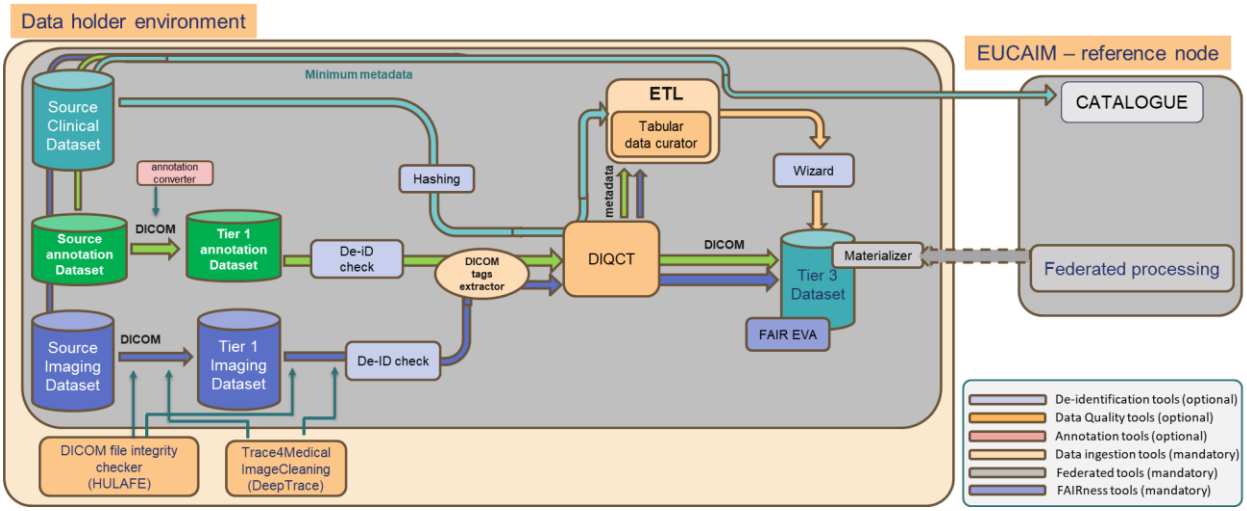


Figure 19. Workflow and the set of tools for Tier 3 data preparation

Guidelines on data structuring

Tier 3 datasets must enable EUCAIM software to operate autonomously in federated processing across any node. Therefore, it is crucial to provide imaging data in a well-structured manner with precise mapping of each dataset, patient, study, and series. Hence, DHs should consider the following aspects when compiling their Tier 3 cohorts to ensure suitability for federated processing software:

- **Data Shape and Structure**: The imaging data should be organized according to the designated hierarchical folder structure for tier-3 compliance (show *Figure 20*). Ensuring data is properly structured prevents compatibility issues, ensuring that each software/tool can work with any EUCAIM dataset as input. This can be performed using a custom procedure by the user or using the DICOM File Integrity Checker, a tool provided by EUCAIM, which includes functionality to process the input dataset and save it in the described structure. This tool is detailed in the *Supporting Material - Section 3 - DQ1 DICOM File Integrity Checker* of the deliverable *D5.4 Data Pre-processing Tools and Services*.
- **Annotations Managing**: When annotations are included on a Tier 3 dataset, i.e. a Tier 3 A+ dataset, they must be in DICOM format as EUCAIM software is prepared to use DICOM images as input[64] and must be added as an additional series in the same series hierarchy level (shown on *Figure 3*).

[64] EUCAIM software are mainly designed for DICOM format images and the ones initially designed for NIfTI format have been modified including a “DICOM to NIfTI converter” module.

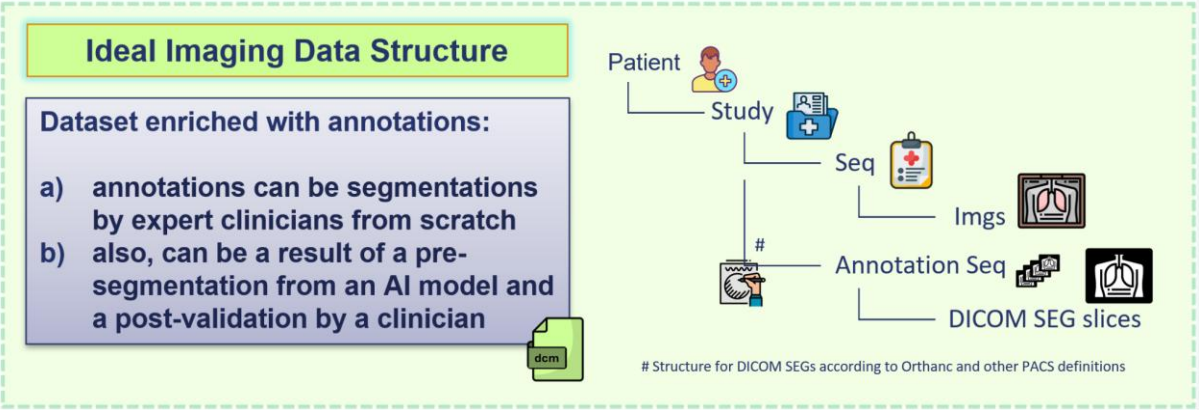


***Figure 20**. Ideal Imaging Data Structure in Tier 3 allowing federated processing capabilities enriched with annotations.*

- **Series Identification and Tagging**: The identification of relevant series is crucial for both efficient visualization and automated processing. Medical imaging studies often include a mix of series, many of which are irrelevant for secondary use or difficult to identify due to non-standardized naming conventions (e.g., the SeriesDescription DICOM tag).

  In Tier 3 datasets, normalizing these names is essential to enable tools to autonomously identify and process the required series within a federated environment. DHs should consult EUCAIM's standardized naming conventions and either update the SeriesDescription tags in their DICOM files or apply normalized names to the directory paths where Tier 3 files are stored for processing.

  The DICOM File Integrity Checker tool also can support this process by identifying and tagging relevant series based on a pre-defined list of SeriesDescription tags provided by the DH. While more automated solutions are under development, DHs may use any method that outputs series identification aligned with EUCAIM's common ontology (e.g. T2w, DWI, ADC for MRI studies). Additionally, DUs can enhance datasets by identifying relevant series using validated methods, facilitating the migration from Tier 2 to Tier 3.

  For example, a tool processing T2-weighted (T2W) MRI sequences should easily navigate each patient and study, loading only the relevant series while avoiding duplicates or unrelated data. Proper normalization and tagging during preparation ensure seamless automated processing.

Guidelines for Data Quality (MEDEX)

The same data quality framework as for Tier 1 and 2 applies to Tier 3 datasets, which must comply with the same rules (see section 4.2 and 5.2.1).

Tier 3 data holders will be provided with the same set of optional data quality tools as Tier 2 data holders (see section 5.2.1).

Guidelines for Data FAIRification (CSIC-IFCA)

For Tier 3 we expect further compliance with RDA FAIRness indicators that are listed in D4.4 Annex 6. FAIR EVA will be used through the EUCAIM Catalogue FDP to test if datasets comply with these requirements.

6.2.2 Dataset in the EUCAIM Reference Node

The federated processing service in EUCAIM performs the execution of verified software on the datasets exposed in the data holder nodes. In the case of the Reference nodes, they provide a service to securely access the data stored in their premises on processing resources. The same guidelines and specifications for tier 3 data apply to the reference nodes.

For both UPV & Euro-BioImaging reference nodes in particular, the requirements for data holders are the following. The Materializer component (the component that exposes the datasets on a read-only basis to the processing environments) is data format agnostic and will work for any dataset in the node without additional requirements from the user. Imaging data will be organized in the hierarchical model of dataset/subject/study/series/images automatically during the ingestion process. Clinical data will be provided as a JSON file extracted from the database. If the clinical data is not fully compliant to the CDM of EUCAIM, it may not be properly consumed by the processing services. The Materializer does not transform the data, although a transformation component could be run prior to the execution of a job through a Kubernetes webhook (this must be consulted with the reference node managers). It is important to state the execution environment does not allow downloading data and uses a secure proxy to access the Virtual Environment that permits the access to the data. Blind execution through a federated model will even restrict visualization of the data.

7. Limitations and Future Work

Establishing the minimum data Federation and Interoperability framework is a prominent approach to facilitate data preparation, harmonization, sharing, and integration across disparate and heterogeneous data repositories, and to simplify the on-boarding of new data sources and infrastructures. Minimum semantic and technical requirements, including mandatory clinical and imaging information and their associated semantic mappings, data preprocessing and interoperability tools, and workflows, have been provided. These requirements are given as guidelines covering the different levels of compliance with the EUCAIM data federation and interoperability framework: dataset cataloging, federated querying, and processing. Adhering to these interoperability requirements will maintain the interoperability among data holders, repositories, and infrastructures already established or expected to join the EUCAIM framework. In what follows we outline the main achievements considering each level of compliance.

- **Dataset Cataloging (Tier 1)**: At Tier 1, technical requirements have been established to support data holders to properly catalogue their datasets and make them accessible for discovery and exploration. A minimum set of interoperability requirements has been specified to standardize the definition, documentation and exchange of aggregated dataset metadata across the EUCAIM federation. This is achieved by defining EUCAIM

DCAT-AP based on DCAT-AP v3.0 and HealthDCAT-AP. Besides, a minimum set of imaging, clinical, and annotation data has been defined, supported by guidelines for dataset preparation, including guidelines on data annotation, format standardization, data de-identification, data quality, data FAIRification, and technical requirements for local node setup.

- **Federated Querying (Tier 2)**: In this Tier, a minimum set of standard, structured and interoperable, query criteria has been specified. Also, guidelines for supporting federated query have been provided, including guidelines for installing the local components and implementing a query mediator service required for federated nodes, as well as guidelines for creating a mapping component to map the requested minimum set of clinical and imaging attributes (described in section 4.2) into the EUCAIM concepts. Guidelines on format standardization, data and data annotation quality, and technical requirements for federated node setup, have also been established to support federated query.

- **Federated Processing (Tier 3)**: In this Tier, a minimum set of clinical and imaging metadata, reflecting the essential information required for basic federated processing, and their semantic mappings with the EUCAIM CDM and hyper-ontology, have been provided. Besides, guidelines supporting federated processing have been given, including, the hardware and software requirements for handling datasets in local nodes, workflows for data preprocessing and interoperability, and guidelines on data structuring, data quality, and data FAIRification.

Despite the achievements, certain limitations have been identified:

- **Interoperability scope**: in the min-FIF approach, we do not consider the interoperability between the EUCAIM data overall federation and other existing data infrastructures, such as the EHDS (European Health Data Space). The min-FIF supports and maintains data interoperability across disparate but a limited number of EUCAIM data holders, repositories, or infrastructures.

- **Domain scope**: the min-FIF covers specific cancer types (prostate, breast, colon, rectum, lung, and colorectal) considered in the EUCAIM framework based on the use cases provided by the data holders. This limitation affects the categories associated with federated query and processing (Tiers 2 and 3), such as “diagnosis”, “image modality”, “treatment” (see Tables 14 and 15, Section 5.1). For instance, specific clinical parameters, such as diagnoses (e.g., “malignant neoplasm of breast”, “malignant neoplasm of prostate”, “malignant neoplasm of colon”), and treatments (e.g., “prostatectomy”, “chemotherapy”), and imaging parameters, such as imaging modalities (e.g., “MR”, “CT”) and body parts (e.g., “breast”, “prostate”, “colon”), have been considered as query criteria for the min-FIF federated query service. Extending the framework with new cancer types, based on new data sources/repositories joining the framework, will necessitate revising the specifications to update or include new query criteria or values and to consider how these changes will affect the overall design and execution of the min-FIF federated query.

- **Complexity level**: the min-FIF approach focuses on the minimum necessary level of data integration and interoperability. It defines the semantic and technical requirements for the essential information required for basic purposes, such as browsing datasets and basic federated queries. For instance, only the mandatory clinical and imaging data has been specified, depending on the platform compliance levels (Tiers 1, 2, and 3), such as “birth sex”, “age at diagnosis”, “pathology”, and “treatment” for the clinical information, and “image modality”, “image body part”, and “manufacturer” for the imaging information. Accordingly, a minimum level of semantic data interoperability requirements has been identified to ensure consistent alignment or mapping of the provided information with the EUCAIM platform, including the CDM and hyper-ontology. Additionally, essential technical requirements, including tools, services, and workflows required for basic or minimum data preparation and integration, have been provided as guidelines. By defining these baseline standards of data federation and interoperability, the min-FIF will not support more complex or advanced use cases, which require highly specified semantic and technical requirements.

In future works, we will address these limitations using targeted strategies moving forward toward a maximized data federation and interoperability framework (max-FIF), which supports a broad insight of the semantic and technical interoperability standards or requirements. In the max-FIF, a larger scope and volume of data will be considered, where maximum interoperability requirements will be specified, supporting broader data integration. Thus, maximum clinical and imaging information will be decided covering more comprehensive and complex use cases. Accordingly, additional efforts are required to ensure semantic mappings of data, including terminology-binding between the EUCAIM CDM and hyper-ontology. Besides, a comprehensive set of technical interoperability requirements, specifying the tools and workflows required to ensure efficient data preparation and seamless integration, will be defined. The max-FIF will also support an advanced federated query service, where additional query criteria will be considered, such as tumor marker tests and results, cancer staging and grading, and imaging assessment methods and values (see Table 16, Section 5.1, for examples). Finally, a maximized interoperability scope will be considered to support the interoperability between the EUCAIM data overall federation and other existing European initiatives or infrastructures, such as the EHDS[65], XSHARE[66], and QUANTUM[67].

8. Conclusion

The EUCAIM minimum data Federation and Interoperability Framework (min-FIF) is a major achievement providing a scalable foundation for efficient data interoperability by defining a minimum set of guidelines and practices to guide data integration from heterogeneous sources and support seamless and unambiguous data exchange between disparate data infrastructures or repositories. The min-FIF helps ensure consistent access and understanding of the essential or mandatory clinical and imaging information provided by heterogeneous and disparate sources

[65] https://www.european-health-data-space.com/
[66] https://xshare-project.eu/
[67] https://quantumproject.eu/

by establishing minimum semantic and technical interoperability requirements while adhering to the European Interoperability Framework (EIF). Fulfilling these requirements will reduce the complexity of data integration and facilitate the exchange and collaboration by maintaining interoperability between heterogeneous data formats, structures, and techniques. Through the adoption of common data standards, models, and ontologies, the framework helps maintain the syntactic and semantic interoperability, ensuring consistent access and exchange of data. Besides, data preprocessing and interoperability tools and workflows, crucial to data integration, are also considered to support technical interoperability while ensuring data quality, security, and usability. Based on the min-FIF, we will continue to evolve toward the maximum data federation and interoperability framework (max-FIF), which will support a comprehensive level of semantic and technical interoperability requirements in the context of the EUCAIM federation.

Annex 1. MITK Workbench tool

The MITK (Medical Imaging Interaction Toolkit) is an open-source software library designed for the development of interactive medical image processing software. It's widely used in medical research and clinical applications for visualizing, analyzing, and segmenting medical images.

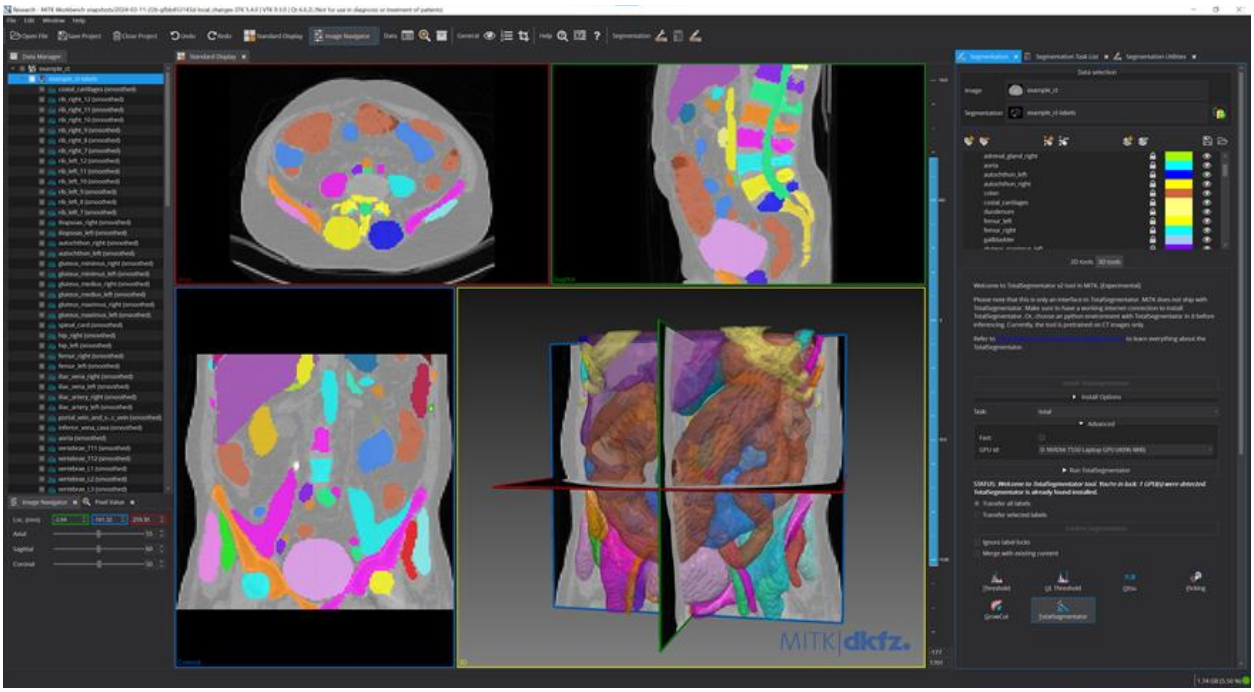
Introduction

Medical image segmentation is the process of dividing medical images, such as CT scans, MRIs, or X-rays, into meaningful regions to isolate specific structures or areas of interest. It is essential for tasks like identifying organs, detecting abnormalities, planning surgeries, and conducting research.

Segmentation can be manual (drawn or painted by the user), semi-automatic (guided by user input), or fully automatic (driven by algorithms, including AI). It often involves distinguishing tissues, lesions, or pathologies based on pixel intensity, shape, or patterns in the image. This process is fundamental in modern medical imaging workflows, enabling accurate analysis and supporting diagnostic and therapeutic decisions.

MITK Workbench: Key Features

MITK Workbench visualizing a 3D image and segmentation masks



MITK Workbench visualizing a 3D image and segmentation masks

Here are some of the key features of the MITK Workbench:

Loading Medical Data

The Workbench allows loading and viewing various medical image formats, such as DICOM, NIfTI, NRRD, HDF5, TIFF, etc. and others. Also, the DICOM Browser[68] helps in viewing and managing DICOM data, widely used in medical imaging workflows. It simplifies the process of importing, browsing, and organizing DICOM files in Picture Archiving and Communication System (PACS) servers directly from within the Workbench.

Visualization

By default, MITK provides a four-quadrant standard display[69] with axial, sagittal, coronal, and 3D render windows, enabling a comprehensive view of the dataset. Users can zoom, pan, and rotate, scroll through slices, adjust windowing settings, and manipulate 3D objects directly within the windows. The render windows support overlays for annotations, segmentation masks, and measurement tools, allowing users to tailor the display to their workflow. Any overlay can be toggled on and off in the Data Manager[70].

In addition to basic rendering, more can be done within the MITK Workbench. For example, the Level Window filter in the MITK Workbench is used to adjust the brightness and contrast of medical images. This filter enables users to map the pixel intensity values of the image to a visible grayscale range, enhancing the visual clarity of specific regions. It is especially useful for examining medical images with varying contrast, such as CT or MRI scans, where details in certain intensity ranges might otherwise be obscured.

The MxN display[71] in the MITK Workbench is a versatile visualization tool that allows users to display multiple views of medical imaging data simultaneously in a customizable grid layout. This feature is particularly useful for comparing different datasets or visualizing different slices, time points, or parameters of the same dataset side-by-side. Users can link navigation across all windows for synchronized scrolling or maintain independent controls for detailed comparisons.

In addition to viewing 3D volumes in a 3D render window, segmentation masks can also be visualized as overlapping structures in 3D.

Segmentation

Segmentation is one among the many processing options in MITK for medical images. The Segmentation Plugin allows the user to create multilabel segmentations of anatomical and pathological structures in medical images. The plugin consists of three views:

- Segmentation View: Manual and (semi-)automatic segmentation
- Segmentation Utilities View: Post-processing of segmentations

[68] https://docs.mitk.org/nightly/org_mitk_editors_dicombrowser.html
[69] https://docs.mitk.org/nightly/org_mitk_editors_stdmultiwidget.html
[70] https://docs.mitk.org/nightly/org_mitk_views_datamanager.html
[71] https://docs.mitk.org/nightly/org_mitk_editors_mxnmultiwidget.html

- Segmentation Task List View: Optimized workflow for batches of segmentation tasks based on a user-defined task list

Understanding the Segmentation View

The Segmentation View in MITK is a dedicated interface for creating, editing, and managing segmentations of medical imaging data. It hosts several tools for isolating and analyzing specific anatomical structures or regions of interest in both 2D and 3D.

Segmentation View

Labeling basics

Segmentation view consists of at least a single group called “Group 0” in which the first default label is created. More groups can be added and removed but there will always be at least a single group. Labels of the same group cannot overlap. Labels of different groups may overlap.

More information on Labels and Groups can be found in MITK’s documentation[72].

Tools

MITK offers a comprehensive set of slice-based 2D and (semi-)automated 3D segmentation tools. The manual 2D tools require user interaction and can only be applied to a single image slice, whereas the 3D tools operate on the whole image and require limited user interaction. Different toolsets can be accessed by selecting the 2D or 3D tab in the segmentation view.

[72] https://docs.mitk.org/nightly/org_mitk_views_segmentation.html#org_mitk_views_segmentationgroups

Techniques for efficient segmentation

The semi-automatic segmentation tools can produce segmentation outside the region of interest too. The Picking Tool is a feature in the MITK Segmentation plugin that allows users to extract a single region from a segmentation. Users can simply add seed points on segmentation masks of interest and discard others.

Another technique is to use the Interpolation feature for fast and efficient segmentation. The Interpolation helps mitigate the time-consuming nature of manual contouring on a slice of large image 3D volumes by suggesting likely areas of interest on neighbouring slices. While 2D interpolation works on a specific plane (e.g., axial, coronal or sagittal), 3D interpolation enables users to utilize all planes to suggest a 3D segmentation. Interpolated suggestions are displayed as outlines, until they are confirmed as part of the segmentation.

Saving and Exporting Results

Once the segmentation is completed, it can be exported in supported formats (e.g., NRRD, DICOM SEG) for further analysis or sharing. Alternatively, the current progress of the project can be saved and loaded back in MITK again in the future.

Advanced

AI assisted segmentation

MITK integrates AI tools to enhance medical image segmentation. These tools leverage advanced machine learning algorithms to automate or assist processes that traditionally require significant manual effort. These tools and their user instructions are the following: TotalSegmentator (3D tool)[73], Segment Anything Model (2D tool)[74], MedSAM (2D tool)[75], and MONAI Label (2D & 3D tool)[76].

Segmentation with Time-series Data

Dynamic 3D+t images in MITK represent volumetric datasets that vary over time, such as time-resolved MRIs, 4D CT scans, or functional imaging studies. For segmentation of 3D+t images, some tools provide the option to choose between creating dynamic and static masks. * Static masks will be defined on one time frame and will be the same for all other time frames. * Dynamic masks can be created on each time frame individually.

[73] https://docs.mitk.org/nightly/org_mitk_views_segmentation.html#org_mitk_views_segmentationTotalSegmentator
[74] https://docs.mitk.org/nightly/org_mitk_views_segmentation.html#org_mitk_views_segmentationSegmentAnything
[75] https://docs.mitk.org/nightly/org_mitk_views_segmentation.html#org_mitk_views_segmentationMedSAM
[76] https://docs.mitk.org/nightly/org_mitk_views_segmentation.html#org_mitk_views_segmentationMonaiLabel3D

Annex 2. Data Integration Quality Check Tool (DIQCT) metrics

Based on the nature and type of data (also scalar or vector), these data quality dimensions are calculated in a different way.
This document described the methodology used by the DIQCT for measuring the metrics related to the data quality dimensions that are part of the EUCAIM data quality framework (completeness, uniqueness, validity, consistency, accuracy, integrity). The tool applies to clinical data, imaging data, as well as a combination of both.

Clinical Data

- **Uniqueness**

Step 1) Analyze data to identify duplicate records in the dataset.

Step 2) Report the metric for the whole dataset

Metric: The absolute number of duplicate records identified for the dataset. In this case, a record is considered a patient entry in the dataset.

- **Validity**

Step 1) Define data format, allowable types & value ranges.

Step 2) Analyze data to identify the valid information.

Step 3) Report the metric.

Metric: The percentage of records in which all values are valid. In this case, each value inserted in the dataset is considered a record. The metric is presented per patient and for the whole dataset.

- **Accuracy**

Step 1) Define the data structure. This metric measures how well the data structure conforms to a specific template (e.g. sheets and columns in an Excel file, columns in a CSV file, etc.), mainly focusing on semi-structured data, or cases where the structure is not pre-imposed.

Step 2) Analyze the data and identify inaccuracies in the provided file.

Step 3) Report the metric for the whole dataset

Metric: The absolute number of inaccuracies in the examined file.

Clinical & Imaging Data

- **Completeness**

Step 1) Identify the critical information that needs to be present in the dataset, i.e. presence of mandatory information.

Step 2) Analyze data to identify the missing information.

Step 3) Report the metric per patient and for the whole dataset.

Metric: the percentage of records that are complete. The patients, in this case, are considered records.

- **Consistency**

Step 1) Define the integration rules, which in this case is the link between images and clinical metadata through the template.

Step 2) Analyze the data and identify if they are properly connected.

Step 3) Report the metric (per patient and for the whole data set)

Metric: The percentage of records that is properly connected. In this case, records are considered the modalities provided in the dataset.

Imaging data

- **Completeness**

It aims to quantify whether the data follow specific requirements for their analysis.

Step 1) Define the specific characteristics of image series that pose analysis requirements (slice thickness = 3mm)

Step 2) Calculate and report the metric

Metric: the percentage of patients that include at least one image series with specific characteristics. For example, if the characteristic that we aim to analyze is the slice thickness to be equal to 3mm, then the metric describes the percentage of patients that include at least one series with images of 3mm.

- **Uniqueness** (deduplication)

For imaging data, measuring uniqueness involves analyzing the data set to identify duplicate image files or series.

Step 1) Search DICOM directories and grouping files by Patient ID and SeriesInstanceUID.

Step 2) Detect duplicates within series using metadata checks and perceptual matching algorithms.

Step 3) Report the metric

Metric: The total number of identified duplicate image files or series within the dataset.

- **Validity**

The aim is to verify whether a specific de-identification protocol has been correctly applied to the imaging data.

Step 1) Check the metadata in all the DICOM files and detect any issues of compliance with the protocol.

Step 2) Generate and Report Metrics expressing the compliance with the de-identification protocol (contribute to the validity of the dataset).

Metrics:

- The first metric is the percentage of images that do not require any further de-identification and is given by the following equation:

$$ImC = 100 - \left(\frac{n_{problematic}}{N}\right) * 100$$

Where $n_{problematic}$ is the total number of images in the dataset that do have identified information and N is the total number of images in the dataset.

- The second metric is the percentage of series that do not include any image that is not de-identified correctly. The metric is given by the following equation:

$$SerC = 100 - \left(\frac{S_{problematic}}{S}\right) * 100$$

Where $S_{problematic}$ is the total number of series in the dataset that do include at least one image that was not de-identified correctly and S is the total number of series in the dataset.

- The third metric is the percentage of patients that are correctly de-identified and thus they do not include any identified image. The metric is given by the following equation:

$$PatC = 100 - \left(\frac{p_{problematic}}{P}\right) * 100$$

Where $p_{problematic}$ is the total number of patients in the dataset that do include at least one image that was not de-identified correctly and P is the total number of patients in the dataset.

Annotation data

❖ Validity

To ensure that DICOM SEG files comply with the guidelines defined by the partners in EUCAIM.

Step 1) Check the attribute types (type 1, 1C, 2, 2C and 3) and highlight issues such as missing or incorrect data. This ensures that the files meet the required standards and are correct.

Step 2) For each patient, calculate a validity percentage based on how many annotations meet the standards.

Step 3) Report the metric

Metric: An overall percentage for the dataset, indicating how correct the annotations are for all patients.

❖ Integrity

Integrity is evaluated in the annotation files that accompany the images.

Metric: the percentage of series in which the segmentation mask is relevant, is extracted. The metric is given by the following equation and contribute to the integrity of the data:

$$AnnR = \frac{\sum_{p=1}^{N} \frac{srel_p}{S_p}}{\sum_{p=1}^{N} S_p}$$

Where, $srel_p$ is the number of series of patient *p* that an annotation file is applicable. S_p is the total number of series for patient *p* and *N* is the total number of patients in the dataset.

Annex 3. Wizard tool

Software requirements

The Wizard tool will be based on the freely available ARX software[77]. Below is a general overview of the tool's software requirements directly from the official website (https://arx.deidentifier.org/). Note that these may evolve over time, so it is suggested that one should consult the official ARX documentation or releases for the most up-to-date information.

1. **Java Runtime**
    - **Java Version:** ARX is a Java-based application; it requires a Java Runtime Environment (JRE) or Java Development Kit (JDK) version 8 or higher.
    - **Platforms: platform-independent**, as long as Java is installed (Windows, Linux, macOS).
2. **Operating System**
    - **Supported OS:** Windows, Linux, macOS, or any other system capable of running Java 8+.
3. **Memory and CPU (recommended):**
    - At least 8 GB of RAM for moderate datasets (thousands of records). More memory (16 GB+) is advisable with larger or more complex data, or to use advanced privacy models or risk analysis.
4. **Additional Dependencies**
    - Integration into existing data pipelines (e.g., with other frameworks), requires to set up compatible environments or custom scripts (e.g., Java-based ETL frameworks, Docker, etc.).
5. **User Interface**
    - The tool comes in two versions, a standalone **Graphical User Interface (GUI)** version and a **Command-Line / API** version (for scripted or automated workflows, or integration in complex processing pipelines via its Java API or command-line execution (requires Java 8+).
6. **Network Connectivity (Optional)**
    - The tool itself can run offline once installed. It is obvious though, that to integrate it with other online services an active network connection is required.

The tool will be appropriately configured to meet the specific needs of the EUCAIM framework regarding the definition of sensitive metadata, quasi-identifiers and insensitive metadata. Frequently used hierarchies in the form of CSV files will be analyzed and provided by EUCAIM for each imaging modality and cancer type. Moreover, special configurations will be required regarding the specific characteristics of each dataset in order to compose a purpose-specific configuration resulting in the best generalization-suppression balance for the characteristics of the cohort, taking into account sample size, disease prevalence etc. A number of options are

[77] Fabian Prasser, Florian Kohlmayer, Ronald Lautenschlaeger, Klaus A. Kuhn. ARX – A Comprehensive Tool for Anonymizing Biomedical Data. Proceedings of the AMIA 2014 Annual Symposium, November 2014, Washington D.C., USA. (Pubmed)

provided by the tool regarding the degree of maximum record suppression, the maximum number of individuals allowed to be in equivalent classes, the number of sensitive characteristics, the higher allowed risk, etc., composing each time a scenario-specific and cohort specific Wizard configuration.

The input required is the cohort metadata in CSV or Excel format with each tuple representing one patient case. For Tier 1 data, the input will be only the DICOM header list after anonymization, while for Tier 2 and Tier 3 available clinical and imaging information will also be integrated in the input file based on the EUCAIM CDM.

After processing, the tool will provide the modified metadata according to the suggestions of the Wizard regarding optimization of security and preservation of data usability. Furthermore, a report regarding the modifications and the initial and final risk analysis will be issued to complement the data description. A schematic representation of the cohort descriptive characteristics will provide an eloquent view in the optimized version of the metadata, with respect to security as well as usability.

Installation

To install ARX using its installer, first ensure you have at least Java 8 or higher installed. Then, download the appropriate installer for your operating system from the ARX website. Double-click the downloaded file to launch the installation wizard and follow the on-screen prompts (e.g., accept the license agreement and select an installation directory). Once the setup completes, you can start ARX either via the shortcut created on your desktop (if provided) or by navigating to the installation directory and running the ARX application.

https://github.com/arx-deidentifier/arx-installer