Project title: European Federation for Cancer Images

Project acronym: EUCAIM

Grant Agreement: 101100633

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

D5.1. Early release of the Data Federation Framework

Responsible partner(s): FORTH

Author(s): Valia Kalokyri (FORTH), Haridimos Kondylakis (FORTH), Stelios Sfakianakis (FORTH), Katerina Dovrou (FORTH), Georgios Manikis (FORTH), Konstantinos Marias (FORTH), Manolis Tsiknakis (FORTH), Mirna El Ghosh (LIMICS), Melanie Sambres (LIMICS), Morgan Vaterkowski (LIMICS), Ignacio Blanquer (UPV), Jose Munuera Mora (QUIBIM), Alejandro Vergara (QUIBIM), Laure Saint-Aubert (MEDEXPRIM), Xavier Rafael-Palou (QUIBIM), Paris Laras (MAGGIOLI), Gianna Tsakou (MAGGIOLI), Tobias Kussel (DKFZ), Carlos Luis Parra Calderón (SAS), Alexandra Kosvyra (AUTH), David Rodríguez González (CSIC), Carles Hernandez-Ferrer (BSC), Leonor Cerdá (HULAFE), Matías Fernández (HULAFE), Pedro Miguel Martínez (HULAFE)

Date of delivery: [10/10/2023]

Version: final

Table of Contents

Table of Figures

List of Tables

1. Introduction

1.1 Overview of the Early Release

This deliverable marks a significant milestone in the development of the European Cancer Imaging Data Federation Framework. It introduces a preliminary proof-of-concept for the EUCAIM Data Federation Framework, demonstrating its potential to support the interoperability of federated imaging and clinical data within the realm of cancer research. This document serves as a comprehensive overview of the primary features and contributions of this early release, offering essential preliminary insights into the EUCAIM Data Federation Framework.

The early release of EUCAIM's Data Federation Framework is the result of extensive discussions and collaborative efforts among participants from various work packages within the EUCAIM consortium. Its platform encompasses several key components including a public catalogue of cancer imaging datasets, a federated search tool, foundational data models and interoperability standards, data management tools, federated node technical specifications and federating processing. Together, these elements pave the way for a more efficient and collaborative research ecosystem in the field of cancer research for researchers, data analysts and AI experts.

To begin with, EUCAIM's first platform release offers a dashboard that contains a comprehensive public catalogue of cancer imaging datasets sourced from the repositories of EU-funded AI for Health Imaging project (AI4HI) repositories. These datasets adhere to a standardized metadata schema based on the W3C DCAT model[1], ensuring consistency and ease of access and sharing for researchers and clinicians. Additionally, the platform includes a federated query tool that enables users to explore and understand the wealth of information available across federated data providers. This achievement is made possible by the development of an early proof-of-concept version of the Common Data Model and hyper-ontology, which lays the groundwork for data harmonization and integration across multiple sources. In line with our commitment to promote interoperability, a set of related common data models, protocols, formats, and terminologies specific to clinical and imaging data were explored and evaluated.

Furthermore, an initial comprehensive list of existing data preprocessing and data management and interoperability tools available to EUCAIM has been compiled. These tools and services are integral to the data integration process, ensuring data quality, security, and traceability. Lastly, early versions of the technical specifications for the setup of EUCAIM federated nodes have been defined. These specifications serve as a blueprint for establishing the infrastructure necessary to facilitate data sharing and collaboration among EUCAIM data providers.

Building on this early release, we will continue to evolve and refine the framework to better serve the needs of cancer researchers and clinicians, moving gradually from the minimum data Federation and Interoperability Framework (min-FIF) to the full Federation and Interoperability Framework (max-FIF). Data providers are given the flexibility to choose between different levels of EUCAIM interoperability compliance, according to technological and operational criteria as these are defined in deliverable D4.3.

1.2 Document Structure

This early release of the EUCAIM Data Federation Framework comprises several essential components:

[1] W3C DCAT model, https://www.w3.org/TR/vocab-dcat-3/

a) High level architecture of the EUCAIM federation: This section outlines the architecture of the Federated Cancer Image Infrastructure of EUCAIM, that defines the components and main functionality of the EUCAIM platform. The architecture will evolve along the implementation of the project.

b) Early Proof-of-Concept Common Data Model and Hyper-Ontology: One of the cornerstones of this release is the preliminary proof-of-concept version of the Common Data Model (CDM) and hyper-ontology. These foundational elements are designed to facilitate the interoperability of EUCAIM's federated imaging and clinical data, initially focused on the cancer types and data models adopted in the five AI4HI projects.

c) Protocols, Formats, and Terminologies: In alignment with the mission of EUCAIM, this release includes related protocols, formats, and terminologies for clinical and imaging data. These standards pave the way for standardized data handling across the initiative.

d) Data Preprocessing and Data/Metadata Management Tools/Services: This release introduces the diverse array of Data Preprocessing and Data/Metadata Management and Interoperability tools and services currently available to EUCAIM through the AI4HI initiatives and the EUCAIM partners. These resources are vital in ensuring data quality, compatibility, and accessibility within the federated framework.

e) Early Technical Specifications: Additionally, the early versions of the technical specifications for the setup of the EUCAIM federated nodes are outlined. These specifications provide the foundation for the distributed infrastructure that EUCAIM aims to adopt.

f) Demonstration Scenario: A description of the demonstration scenario of the early release of the EUCAIM platform is outlined in the section, showcasing how all the different components are communicating. Despite currently having a limited functionality, it will serve as a basis to understand, test, expand and build the further versions of the EUCAIM platform.

2. High-level Architecture of the EUCAIM Federation

This section describes the high-level architecture, which identifies the main components and their interactions. It serves as an overall view of the system, for better understanding the implications associated with each component.

Figure 1 depicts the architecture overview, using the following color coding: Blue for data providers (WP2 and WP7), Pink for central hub services (WP4), Orange for data modelling and matching (WP5) and Green for data processing (WP6).

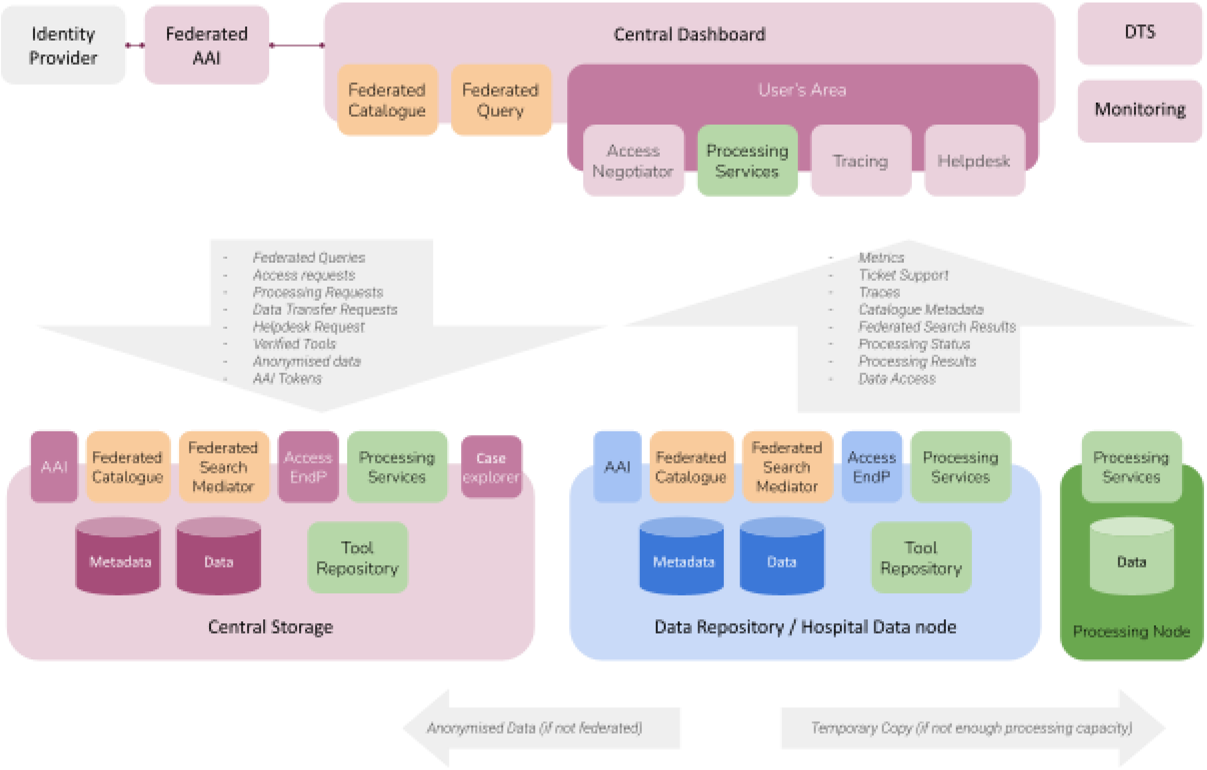


Figure 1: High-level view of the architecture.

In a nutshell, data providers can join the federation (blue) or upload anonymized data to the central storage (pink). Technically, the central storage acts as any other node of the federation. The data providers push their collections' metadata to the central services of the **federated catalogue**, which will display it to the data requesters (Tier 1 compliance). The data could be further explored through a **federated search**, which will forward the query to the providers, following a Common Data Model (CDM) schema. This applies only to data providers that comply to tier 2 or 3 of the EUCAIM Data Federation Framework as described in deliverable D4.3. Providers that do not have their data ready following the agreed CDM schema will use a mediator to match the terms in the query (Tier 2 compliance). All the services are coherently authenticated and authorized through a **Federated AAI**.

The users will request access to the dataset collections through the **Access Negotiator**, which will manage the access request process. Access request is not a trivial task and will require extensive documentation. The user will be able to explore and process the access-granted collections in the **User's Area** using the **Processing Services**. Data and Tool providers will be able to get traceability information through the **Tracing** service, which will also provide high-level provenance information to the users in the Federated Catalogue (such as the datasets used to train a tool or the popularity of a tool or a dataset). Finally, other services complement the functionality, with a **Helpdesk** for support management, and a **Data Transfer Services** to temporarily copy data on the Processing Node services.

In the following subsections each of the components is further described and elaborated.

2.1. Dashboard

The users of the EUCAIM platform are able to access the platform through the Dashboard. The Dashboard integrates all the applications of EUCAIM and provides a user space in which users manage their requests. The Dashboard includes the landing pages with the instructions and the link to the public catalogue. Through the Dashboard the user can explore the data using the federated query and request access through the negotiator system.

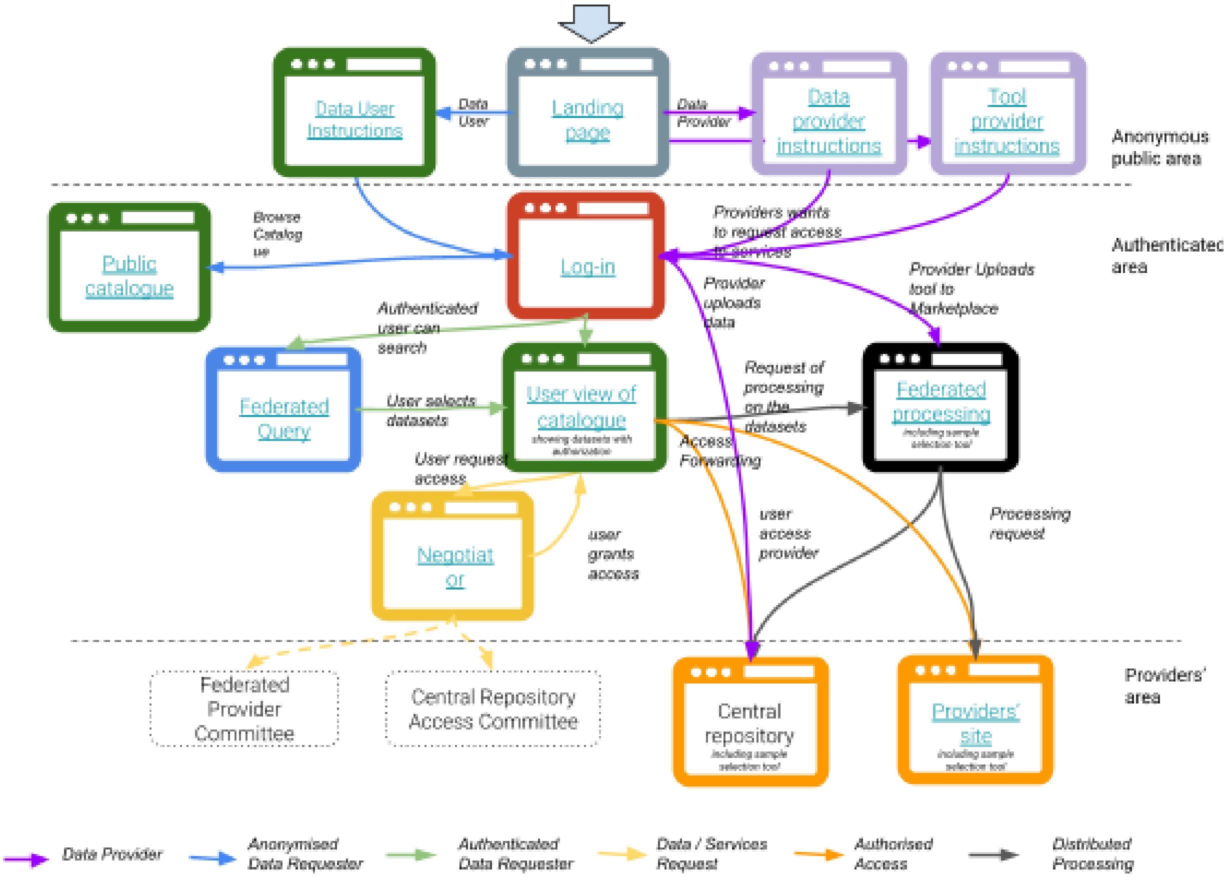


Figure 2: Dashboard workflows

Users in the Dashboard will be authenticated through LifeScience AAI, and the Dashboard will provide links to the following applications:

- Public Catalogue, implemented using Molgenis[2]. This application will feature the catalogue of datasets of the EUCAIM project. The datasets will follow the metadata specification defined in this deliverable.
- Federated Query service. This application will search on the data providers of the federation to retrieve the datasets fulfilling a specific criteria. This requires that federated providers either implement a mediator to match the queries expressed through the hyperontology to its own data and metadata schemas or transform the data to the EUCAIM's CDM. This service applies to data providers that comply with tier 2 or 3 of the EUCAIM data federation framework.

[2] https://www.molgenis.org/

- Negotiator service. This application will allow managing the access requests to the datasets of the public catalogue. It will consume the data from the public catalogue.
- My Library area. This application will provide a personalized area for each user in which they will find the access to the pending and granted requests.
- Distributed processing console, which will provide the means to run applications on the distributed environment.

Figure 3 shows a snapshot of the Dashboard landing page.

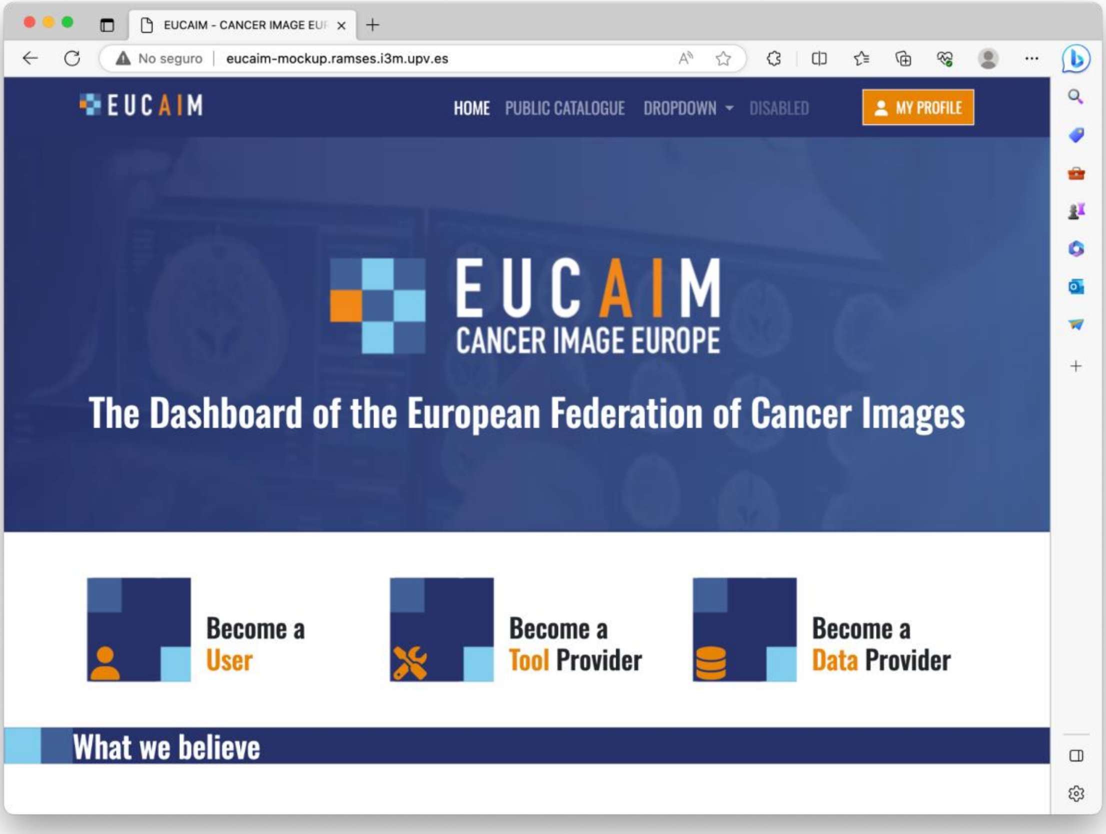


Figure 3: Dashboard Landing Page

2.2. Authentication and Authorization Infrastructure (AAI)

The services of the central hub will directly rely on the **Life Science Authentication and Authorization Infrastructure (LS AAI)**[3] for authentication and authorization, which will use external institutional IdPs for authentication. The architecture should be interoperable with external federated providers, which should trust on the LS AAI for authentication, but could manage their own AAI instances. This way, a user registered in the LS AAI could browse and explore the dataset's metadata available in the central hub, whose services will trust on the LS AAI tokens and will use the group entitlements as authorization information. The use of GA4GH Passport token sets will also be considered.

Figure 4 shows the interactions among the main components. A user will access the Dashboard (1) and get authenticated through the LS AAI (2), which will trust on the Institutional IdP (3),

[3] LS AAI, https://lifescience-ri.eu/ls-login/ls-aai-aup.html

providing a token back to the Atlas. Other services of the Atlas (4) can verify the token in the LS AAI for verifying the identity.

When a user gets the authorization to access a specific provider (I), the policy enforcement component of the Atlas could trigger the inclusion of the user in a specific group (II) (this process could be human-based). This will allow the user accessing the dashboard to access the provider services.

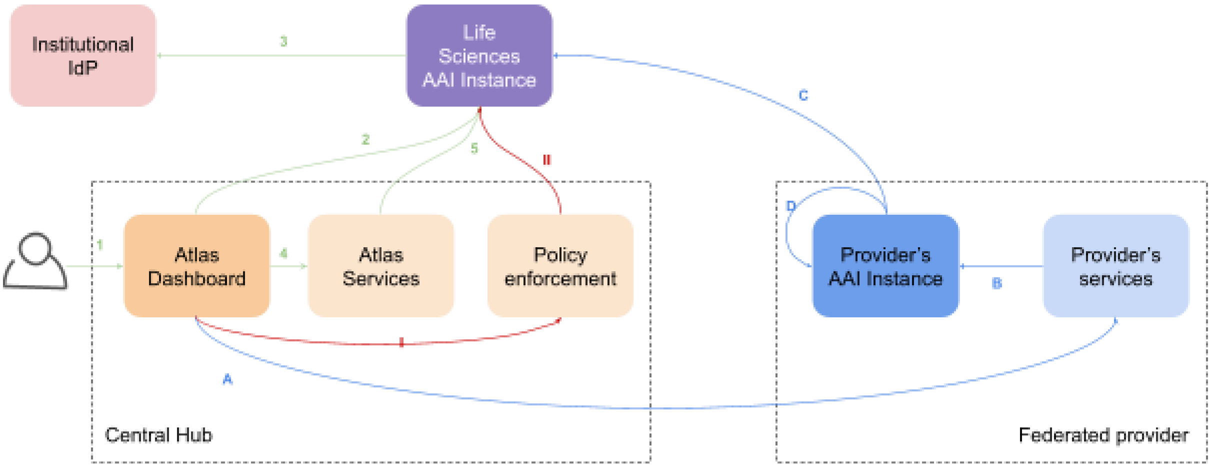


Figure 4: Architecture Diagram for the AAI services

The Atlas Dashboard will directly forward to the provider services (A), which will request the login through the provider's AAI (B). If the user is registered with LS AAI as IdP, the authentication (C) will be automatic. Then, if this is the first time a user logs in, the Provider's AAI instance can create the user automatically if the proper group in LS AAI is given (D). However, this process could be more restrictive if the provider does not accept this approach (e.g. by requesting a full registration in the system). This approach will only request that the Federated Providers trust LS AAI as IdP and the configuration of the AAI provider's instance to create users automatically (e.g. Keycloak can do it).

2.3. Public Catalogue

Data discovery is the essential first step for data re-use. The catalogue stores the metadata, offering the researchers the descriptive information about the available datasets. It will show data access conditions. At the same time, the catalogue offers a platform for data owners to display their datasets.

The Dashboard provides a link to the **public catalogue** with the metadata of the federated datasets or collections available

- Dynamically obtained through the collection of the metadata from the providers registered in EUCAIM, which will register the collections in the public catalogue.
- The specification of the metadata of the collections will be a subset of the Hyperontology, defining the fields and variables to be used.

The catalogue will include some operational information from the providers, such as the access conditions. Providers may offer different access conditions: Authorization to download the datasets, authorization to access, view and process in-situ the datasets and authorization to remotely process the datasets without the ability to access and visualize data, even remotely.

The catalogue metadata is available to anonymous users but other services such as the Federated Query are available only to authenticated users (see 2.4. Federated Query).

The catalogue is the browsing interface where relevant metadata can be exposed by data providers and found by researchers (federated catalogue explorer). The catalogue is intended to:

- Allow data providers to expose the metadata of their digital objects in a way that fulfils the FAIR (Findability, Accessibility, Interoperability and Reusability) Data Principles.
- Allow data requesters to discover information about collections that contain information that they could be interested in.
- Provide meaningful information about collections for both humans and software agents.

The **metadata elements** displayed in the catalogue are a limited list of standardized options, including access conditions; these constraints are configured by specifying a metadata schema. Metadata will be gathered from federated repositories (e.g., repositories of partner institutions, existing AI4HI projects and data warehouse catalogues) and mapped to the EUCAIM **metadata model** by the creation of data dictionaries. **Standard vocabularies** will be used to harmonize and facilitate interoperability. Figure 5 shows the architecture of the federated catalogue, following the same coloring convention of the previous figures.

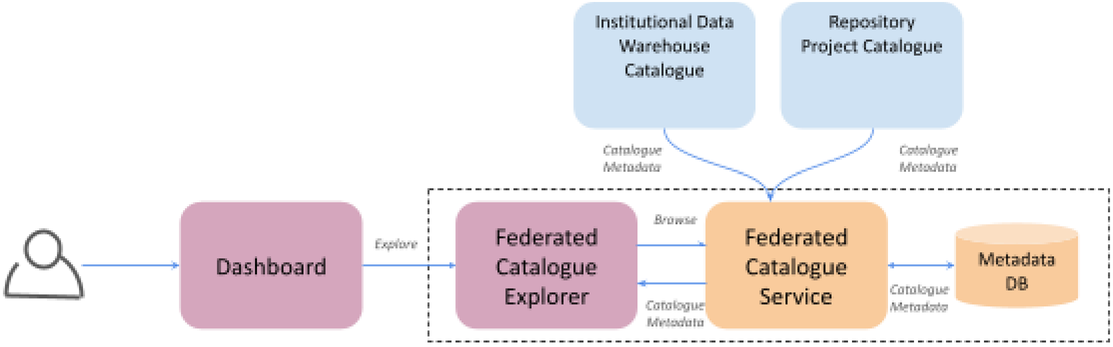

Institutional Data Warehouse Catalogue
Repository Project Catalogue
Catalogue Metadata
Catalogue Metadata
Dashboard
Explore
Federated Catalogue Explorer
Browse
Catalogue Metadata
Federated Catalogue Service
Catalogue Metadata
Metadata DB

Figure 5: Architecture of the Federated Catalogue

2.4. Federated Query

As the Public Catalogue only provides catalogue-level data exploration functionalities, the Federated Query Service extends the data exploration capabilities for datasets with granted access rights by allowing to see aggregated results for record-level search queries.

In order to enable the Federated Query and Federated Access interactions, the EUCAIM Federated Data Query Service is designed to fulfil a core functionality of sending Federated Queries within the EUCAIM data infrastructure and aggregating the results. These queries are designed to return aggregated information of the cases fulfilling the filters of the request.

The EUCAIM data infrastructure is comprised of EUCAIM Data Nodes (either Data Provider Nodes or the EUCAIM Central Node) and is extended through integration with the infrastructure of Repository Projects (i.e. AI4HI). Because these extended infrastructures hold their own Data Querying and Data access mechanisms, and their data models can be heterogeneous, each of these “Data Junctions” acts as a technical walled garden for creating a common Federated Query API to access them.

To overcome this and other resulting challenges, a Mediator component has been established. The Mediator, acting as a sort of middleware, is deployed at the site of each Data Junction, and:

- Offers the same “Federated Query Interface” /”Data Junction’s API”.
- Receives Federated Queries coming from the EUCAIM platform expressed in the federated query language, and on terms from the EUCAIM hyper-ontology.
- Transforms the query to the local data model.
- Forwards the transformed request and receives the result.
- Transforms the result to be compatible with the EUCAIM Federated Data Query Service.
- Returns the result to the EUCAIM Federated Data Query Service.

These steps describe the core functionality of the Mediator component, as it is created to support the EUCAIM Federated Data Query Service.

The result of a query will be the set of datasets that fulfil the searching criteria, including the number of cases in each dataset that match the filter. This will give an appraisal of the size of data of interest in each specific data provider. If the number of cases matching the query is below a specific threshold, the service will not provide results.

The Federated Query architecture is shown in Figure 6. This component will construct the query expressed on the Hyperontology terms and forward it to the mediator endpoints of the providers. Each provider’s mediator endpoint will adapt the query to the specific syntax and structure of the local provider and query its endpoint. The results will be transformed and sent back to the Federated Query service, which will be used by the Federated Data Explorer to display them.

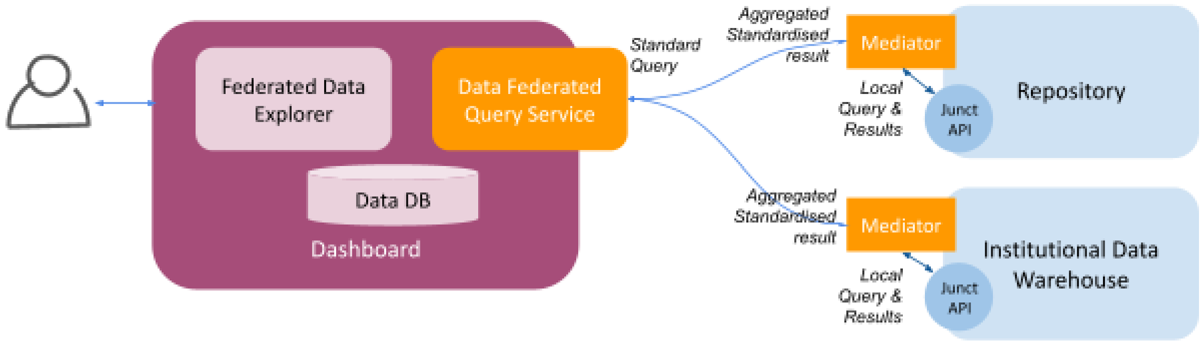


Figure 6: High Level Architecture of the Federated Query service

Query datasets fulfilling a specific criteria

An Authenticated user searches for data through the Dashboard by providing a query expression. The Query can be formulated by using a displayed search tree or by entering the search query and selecting the completion options in the search bar.

1. The Federated Query service sends a search expression to the providers registered in the platform.
   a. The Query is transformed to the specific API of each provider by the mediator.
   b. The output of the query is transformed to the EUCAIM data model and filtered according to the privacy criteria.
2. The provider returns the aggregated information
3. The results are shown in the Federated Data Explorer.

Restrictions of the Early release of the Data Federation Framework

- Aggregated patient counts, filtered based on the query, are displayed. The search criteria provided are based on the minimum common clinical and imaging data fields, described in section 3.3.
- Only FHIR and OHDSI-OMOP-based data stores can be accessed by using CQL and SQL queries, respectively.
- Only stratification and statistics are shown.

2.4.1 Technical System Design

The technical basis of the Federated Query System is the BBMRI-ERIC Sample Locator, combining multiple DKFZ developed components for federated data search and exploration.

Figure 7 gives an overview over the different interacting systems.

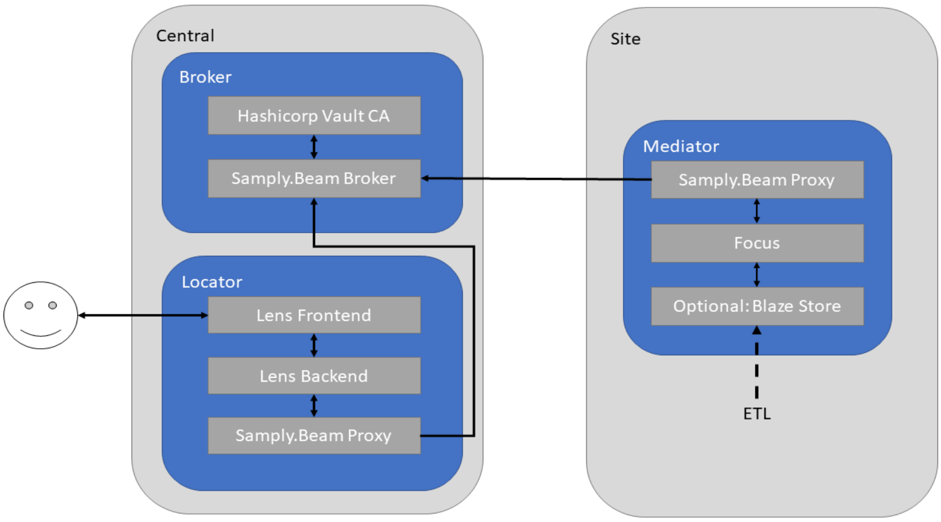


Figure 7: Components and interactions of the Federated Query System

The communication between the site's nodes and the central node is performed via Samply.Beam[4].

To use Samply.Beam, one central pod runs the Beam.Broker and the Beam certificate authority (Pod 'Broker' in Figure 7). Currently, Samply.Beam uses Hashicorp's Vault[5] as a simple and easy to maintain certificate authority. Other services, such as the Federated Processing (see 2.6. Federated Processing), could use the same Beam.Broker for their site node communication.

The central components for the Federated Query Service (Pod 'Locator') consists of:

1. The Lens-based Frontend and Backend[6]: Here, the user query is translated into an abstract syntax tree (AST), generalizing the query for easy conversion into different query languages.

[4] Samply.Beam is a message broker system, specifically designed for secure, efficient network communication across institutional barriers in secure networking environments. It communicates end-to-end encrypted via HTTP(s) using only outgoing connections, hence, it alleviates the usual firewall and proxy issues. For more information, see https://github.com/samply/beam.

[5] https://www.vaultproject.io/

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

2. A Beam.Proxy, allowing the sending of query tasks to all connected sites and the retrieval of the sites' (aggregated) results.

Finally, the site nodes (Pod 'Mediator' in Figure 7) run:

1. A Beam.Proxy to retrieve queries and publish results,
2. Focus[7], the mediator component translating the generated AST query into the target query language. Currently Focus supports CQL for querying FHIR-based data sources, AQL for OpenEHR Data Sources, BeaconV2, and SQL for OHDSI-OMOP data stores.
3. Optionally, a EUCAIM data store (e.g. Blaze for FHIR), as a target for the site's ETL processes. This component might be required by some sites for better control of accessed data and simplified ETL processes. If not used, Focus will send its queries directly to the site's (CDM compliant) data store.

The site-local Focus component is ready to introduce additional obfuscation that is differential-privacy-like statistical disclosure control, if required. As users are only able to perform federated queries on authorized data collections, the disclosure of aggregated patient counts is considered sufficiently anonymous in the piloting phase without additional obfuscation measures.

2.5. Federated Access

Several functionalities are expected for the resource negotiation (submitting, refining and approving or rejecting both access requests and tool or data provisioning) within MM2 prototype:

- Request access to a previously identified collection. The request may include a project description, a first specification of the requested items (e.g. samples and/or data, an ethical approval if it exists).
- Transfer the request to the evaluation committee.
- Follow-up on the request and its status.
- Interaction between the applicant and the desired collections.

The very first prototype of the access negotiation module will be implemented by the BBMRI-ERIC Negotiator[8] and in operational use under https://negotiator.bbmri-eric.eu. The installation will be based on the dockerized version 2 of the Negotiator software[9], which is currently in operational use at BBMRI-BBMRI-ERIC. It will be integrated with LS-AAI and the public catalogue based on Molgenis. The negotiator and the catalogue are linked so collections in the catalogue can be added to the negotiator.

For the remaining EUCAIM project and its further system versions a switch is foreseen from version 2 to version 3, which is currently under development and allowing more flexibility on the sample/data source side for defining the necessary composition of a request. Consequently, this will help improving the negotiation process between the researchers and sample/data providers.

[7] https://github.com/samply/focus
[8] https://www.liebertpub.com/doi/10.1089/bio.2020.0144
[9] https://github.com/BBMRI-ERIC/negotiator-v2/blob/master/Dockerfile

2.6. Federated Processing

The high-level architecture of the Federated Processing system involves a central service that coordinates the collaboration between the entities. As part of the Federated Processing system, we are implementing a distributed analysis engine that can deploy and support the entire life cycle of Federated Analysis or Federated Learning experiments. Using a federated learning use-case as example, we would propose the following architecture:

- Analysis Platform REST API: user interface where the user triggers an experiment and retrieves the results (The API can be called from the platform's dashboard, its command line tools, or from the EUCAIM general dashboard).
- Message broker: currently based on RabbitMQ, with which we establish/trigger the communication between the client nodes and central node.
- Management framework: In order to trigger the experiment at each node, the desired FL framework is instantiated (e.g. a docker-based local implementation of the Flower client, at the local node, and Flower server, at the master node).
- Federated Learning models: these are materialized using git and made accessible to the FL framework client docker image in a mounted folder.

In a more general case, Flower would be replaced by containers providing the desired server and client functionalities and adapted to be compatible with the Management framework.

A simplified version of the architecture is presented in Figure 8.

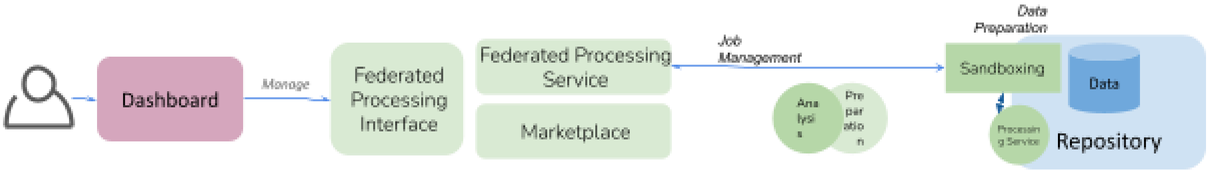


Figure 8: Architecture of the distributed processing environment.

With respect to data, applications will find the data available as a POSIX volume. This will offer the analysis tools a uniform scenario, relying on the data management infrastructure at the node for the materialization. The structure of the files in this "sandboxing area" will be clearly described, encouraging to follow a hierarchical structure (dataset / subject / study / series / image). This sandboxing area will also be used for the outputs of the processing, which will be served back to the federated processing central service.

2.7. Central Repository

EUCAIM will have a Central Repository, which is used for storing data:

- that cannot be stored inside a local organization because the data needs to be seen by others (e.g. annotated by external parties not having access to the local organization);
- the local organization is not able to store this data, either technically, financially or legally;
- where the local organization is not able to participate in a federated/distributed analysis on premise (due to any reason) and the data holder/controller wants to participate in a federated analysis;
- that is donated to the research community.

There are roughly two types of data to be stored centrally: imaging data and non-imaging data, where the non-imaging data consists of clinical (observational) variables and genetic descriptions.

The operational model is envisaged to support more than one instance of a Central Repository, i.e. replicated instances. In the envisaged deployment, we foresee two instances, which should be able to cater for the same requirements as discussed in the section above. Technically, all instances of the Central Repository should be accessible through the same kind of programmatic interfaces to be compatible with the dashboard and the other components in EUCAIM it needs interfacing with. A Central Repository instance needs to be able to take part in a Federated Processing experiment and as such needs to be able to provide interfacing that is required for that, this follows from the Federated Processing section in this document.

The choice for a specific instance of the Central Repository can be based on local regulatory or legal requirements, proximity/affinity to specialized or specific facilities or equipment, etc.

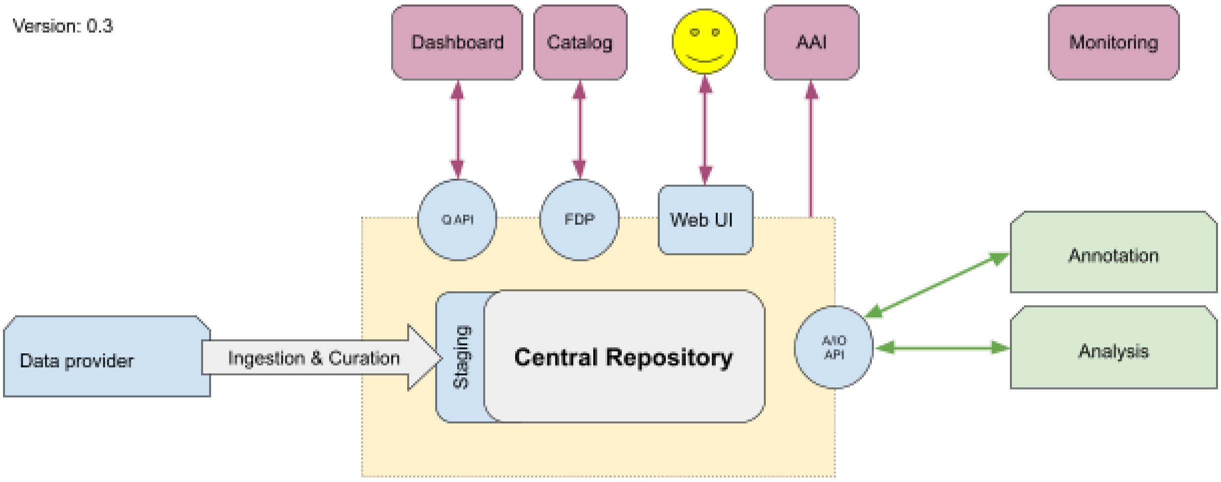


Figure 9: Architecture of the Central repository

The Central Repository will have two alternative technologies:

- **Health-RI XNAT**[10] (EMC), used in EuCanImage, euCanSHare, RadioVal, EOSC4Cancer, among others. The Data ingestion supports multiple flavours, basically anything that can talk to the XNAT DICOM Receiver:
    - CTP[11]
    - POSDA[12]
- **CHAIMELEON Repository technology**[13]. CHAIMELEON uses a central platform that combines processing and storage nodes, providing in-situ processing capabilities for the data stored. The platform uses a K8s deployment for the management of the platform services and for running the processing tasks. Data is stored in a Data Lake implemented through Ceph. The code of the deployments and the services is available in github[14].

The repository will support the proposed Data Model applicable for Imaging Data from WP5 tasks. Each central node will act as a node of the federation. The high-level architecture of the central storage is depicted in Figure 9. The components described are:

- Q API: Query API.

[10] https://xnat.bmia.nl/
[11] https://mircwiki.rsna.org/index.php?title=MIRC_CTP
[12] https://github.com/UAMS-DBMI/PosdaTools
[13] https://chaimeleon-eu.i3m.upv.es/
[14] https://github.com/chaimeleon-eu

- Catalogue: Place where to store metadata about the datasets contained in the Repository.
- AAI: Authentication & Authorization Infrastructure.
- FDP: Fair Data Point (means of exposing the metadata of a repository[15]).
- Web UI: User interface for data management, application management, data browsing and inspection, use of a case explorer, etc.
- A/IO API: Access and Data I/O API.

[15] https://www.fairdatapoint.org/

3. Early Proof-of-Concept Version of the Common Data Model and Hyper-Ontology

3.1 Introduction to the Common Data Model (CDM)

In the realm of health research and federated data analysis and learning, one of the significant challenges faced is the heterogeneity in data representations and semantics across various sources of information. Diverse data formats, structures, and interpretations hinder the seamless integration, analysis, and interpretation of data, creating complexities for researchers, data analysts and AI experts. To address this issue, the adoption of a Common Data Model (CDM) by all data sources is imperative.

A CDM is a standardized framework that defines both the structure and semantics of diverse datasets using ontologies, coding systems, and formal documentation. In a federated learning setting, such as in EUCAIM, local data nodes originate from various clinical centers, existing repositories, and research infrastructures, each potentially utilizing distinct data formats and representations for imaging and clinical oncology data. In this case, a CDM is necessary as it facilitates standardization, ensuring that all data nodes adhere to the same structure and represent their information using the same standardized terminologies. This harmonization simplifies data querying, data preparation for model training and validation, and result aggregation, enabling seamless collaboration despite the heterogeneity of data sources.

More specifically, by adhering to CDM specifications, including the use of specific terminologies and ontologies, semantic consistency across different data sources is promoted. This standardization contributes to enhanced data quality and reliability in large-scale projects, as it ensures that data have a consistent meaning and can be effectively compared and merged across various data providers[16]. In addition, the adoption of a CDM between different data sources facilitates data governance and quality assessment during the training data preparation in an AI pipeline, as it enables raw data to be subject to auditing and facilitates the effortless retrieval of data via automated processes[17]. With predefined scripts and metrics for data quality extraction, onboarding new data providers into the EUCAIM federated learning platform becomes more manageable, helping maintain overall training performance. This aids in transparency, reproducibility, and accountability in the context of federated learning, where data access by modelers is limited or restricted due to the distributed nature of data.

3.2 Challenges and Selection of the CDM for EUCAIM

One of the challenges that the EUCAIM consortium faces is the absence of standardized protocols and interoperability standards for federated learning applications in the domain of oncology, particularly when combining imaging data with clinical information. Despite an abundance of specialized data formats and models, the reuse of medical imaging data for clinical research purposes remains infrequent, especially in the context of predictive analytics and machine learning applications. Different standards cover distinct functional domains, leading to significant diversity, which in turn poses a challenge in creating a comprehensive Common Data Model (CDM) that encompasses clinical, imaging, and -omics data types.

[16] However, we shall keep in mind that the transformation induced by standardization may introduce a loss of quality (through imperfect matching) and some use cases may be efficiently executed locally and less accurately after standardization.

[17] V. Huser, M.G. Kahn, J.S. Brown, R. Gouripeddi, Methods for examining data quality in healthcare integrated data repositories, Pac. Symp. Biocomput. 23 (2018) 628–633.

The primary factor contributing to this diversity is the development of these standards within separate design frameworks, tailored to record healthcare practices and research needs while catering to diverse user communities, including clinical practitioners and researchers. Standard data representations are typically generated by healthcare providers, health insurance entities, and research institutions, drawing upon data formats from various sources such as electronic health records (EHRs), lab tests, insurance claims, and specialized electronic devices.

Within the EUCAIM project, we have examined various data models and identified two potential candidates for the Common Data Model (CDM) based on the AI4HI projects and the expertise of the consortium members. The HL7 FHIR[18] and OHDSI-OMOP[19] data models have proven their ability to support research tasks in a federated setting, encompassing oncology clinical information and imaging data. Both data models have already successfully demonstrated the ability to facilitate ML-based oncological studies[20], clinical predictive modeling[21], federated learning medical applications and other ML-based analyses[22]. Despite not being originally developed with a specific focus on oncology and imaging, both models have been extended to cater to these domains. The OMOP-CDM features both an oncology extension[23] and an imaging extension[24], while HL7 FHIR has adopted the mCODE[25] (Minimal Common Oncology Data Elements) data model for representing oncology and introduced new FHIR resources to support imaging data.

However, the question being raised is how to effectively identify and integrate healthcare interoperability standards and a data model into a federated learning platform while adhering to FAIR principles. Towards this end, some recent efforts have employed OMOP-CDM as a backend data representation and FHIR as a data transfer format among other approaches[5,26]. In addition, relevant in this regard is the FHIR Implementation Guide for the FAIRification of

[18] FHIR HL7 International. URL: https://hl7.org/fhir/ [accessed:2023-09-07]

[19] Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM). URL: https://ohdsi.github.io/CommonDataModel/ [accessed: 2013-09-07]

[20] Ahmadi N, Peng Y, Wolfien M, Zoch M, Sedlmayr M. OMOP CDM Can Facilitate Data-Driven Studies for Cancer Prediction: A Systematic Review. International Journal of Molecular Sciences. 2022; 23(19):11834. https://doi.org/10.3390/ijms23191183

[21] Khalilia, M., Choi, M., Henderson, A., Iyengar, S., Braunstein, M., & Sun, J. (2015). Clinical predictive modeling development and deployment through FHIR web services. In AMIA Annual Symposium Proceedings (Vol. 2015, p. 717). American Medical Informatics Association.

[22] Choudhury, A., van Soest, J., Nayak, S., & Dekker, A. (2020, June). Personal health train on fhir: A privacy preserving federated approach for analyzing fair data in healthcare. In International Conference on Machine Learning, Image Processing, Network Security and Data Sciences (pp. 85-95). Singapore: Springer Singapore.

[23] Belenkaya R, Gurley MJ, Golozar A, et al: Extending the OMOP common data model and standardized vocabularies to support observational cancer research. JCO Clin Cancer Inform 5:12-20, 2021

[24] Park, C. *et al.* Development and Validation of the Radiology Common Data Model (R-CDM) for the International Standardization of Medical Imaging Data. *Yonsei Med J* **63**, S74 (2022).

[25] Osterman, Travis J., May Terry, and Robert S. Miller. "Improving cancer data interoperability: the promise of the Minimal Common Oncology Data Elements (mCODE) initiative." JCO Clinical Cancer Informatics 4 (2020): 993-1001.

[26] Lo, S. K., Lu, Q., Wang, C., Paik, H. Y., & Zhu, L. (2021). A systematic literature review on federated machine learning: From a software engineering perspective. ACM Computing Surveys (CSUR), 54(5), 1-39.

health data collections, currently in STU status (Standard for Trial Use[27]), developed and co-led by HL7 and SAS in the context of the FAIR4Health project (SAS was the FAIR4Health project coordinator and is also an EUCAIM partner). Regarding this Implementation Guide, it should be noted that this version is based on a not-yet-published FHIR version (HL7 FHIR 4b), with all the limitations this brings. This choice has been made to take advantage of the new capabilities offered by the Evidence-Based Medicine resources, including Citation.

In addition, the HL7 Vulcan initiative[28] should also be considered. This initiative aims at facilitating the integration of care and research activities to improve patient lives, reduce costs and improve efficiency through the use of HL7 FHIR interoperability standards. This FHIR Acceleration Program develops FHIR resources needed to execute prioritized use cases of secondary use of « real-world » data and especially EHR data. The main goal of the HL7 Vulcan FHIR implementation 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.

3.2.1 CDM business requirements

Prior to selecting a Common Data Model (CDM), we conducted an analysis of the main requirements, expectations, and constraints from various stakeholders. Our approach involved engaging with representatives from the AI4HI projects and requesting specific information, as follows:

- The specific cancer types that each project focused on.
- The clinical questions/use cases addressed by each project.
- The clinical and imaging data used to answer these questions, including mandatory and optional information.
- The format of the raw data available and whether standardized terminologies were used for different data types, along with the versions of these terminologies.
- The anonymization techniques/profiles employed by each project to ensure compliance with GDPR and national data privacy laws.
- Details about the modalities of radiological images collected and the imaging metadata associated with them, or extracted, if applicable.
- Information regarding the format of segmentation masks, if they exist.
- The chosen common data model and whether it covers all data types, with a straightforward mapping from the raw data.

The result of this process is shown in Table 1.

Table 1: Summary of AI4HI clinical and imaging information

	EuCanImage	ProCAncer-I	INCISIVE	CHAIMELEON	PRIMAGE
Cancer covered	Rectum, Breast, CRLM, Liver	Prostate	Lung, Prostate, Breast, Colorectal	Lung, Prostate, Breast, Colon, Rectum	Neuroblastoma, Diffuse Intrinsic Pontine Glioma (DIPG)

[27] http://hl7.org/fhir/uv/fhir-for-fair/2022Jan/
[28] http://hl7.org/fhir/uv/vulcan-rwd/#overview

CDM	FHIR	OMOP (oncology+radiology extension)	FHIR	OMOP	DICOM MIABIS
Terminology	SNOMED CT; ICD-O3; RxNorm; LOINC	SNOMED CT; ICD-O3;NICt; PI-RADS, RadLex; LOINC; NAACCR; CPT4; cancer modifiers	SNOMED CT; ICD10;ACD/RADS; ATC; LOINC	SNOMED CT; ICD;LOINC; RxNorm, ATC; CPT4 NAACCR; Cancer modifiers	-
Radiological images collected	CT; MRI; MG	mpMRI	CT MRI, X-Ray; US; MG; PET	CT; MRI; MG	CT; MRI; PET; MIGB
Metadata collected from DICOM?	Yes	Yes	-	Yes	Yes
Segmentation Mask?	Yes	Yes (MR; DICOM SEG)	-	No	Yes
Histopathological images collected?	No	No	Only for colorectal cancer	No	No
Mutation status collected in eCRF?	Only for breast cancer	No	Yes	Yes	Yes
Biological results collected in eCRF?	Yes	Yes	Yes	Yes	Yes

Following the collection of information from the AI4HI projects, several group meetings were conducted with different domain experts within the consortium, including AI experts, data providers, software engineers, and legal teams, to define the data model business requirements for the project. The most critical requirements are presented below:

- EUCAIM should support as many input formats as possible for raw clinical data, which may or may not comply with interoperability standards and could be in proprietary formats.
- The data model should be customizable and extensible, allowing adaptation to the specific data format specifications relevant to the project.
- The data model should be terminology-agnostic, accommodating different terminologies seamlessly.
- Minimization of the effort required from clinical data managers to prepare data for federated training through the platform.

- Data pertaining to the same patient across multiple data providers should be identifiable as belonging to the same subject.
- The data model must fully comply with GDPR and national privacy laws.
- The data model should comprehensively represent all target data types at their intended level of detail, including clinical, demographic, radiomic, and laboratory data.
- The data model should support additional information extracted from raw data, potentially through automated AI algorithms, such as radiomics features.
- The data model must provide a uniform interface for accessing and querying data for the purpose of training federated AI models.
- Data transformations from the raw source to the AI training dataset should be as straightforward as possible.
- The quality of data exposed by the CDM for training predictive models must be at least as high as the quality of the raw input data.
- The data model should be structured in a way that simplifies the retrieval of records in the training dataset, regardless of the training plan for an AI algorithm.

3.2.2 CDM selection

Based on our requirements analysis, both FHIR and OMOP-CDM were deemed generally capable of supporting all the information foreseen for the EUCAIM project, as well as all the necessary processes for querying and transforming information required by the AI algorithms. In the sequel, we provide an explanation of the process followed for selecting the appropriate CDM.

OHDSI stands as a firmly established observational health analytics platform, with active participation from researchers, developers, and clinicians. Initially, its primary focus was on digitizing and analyzing health insurance claims data within the United States. However, over the recent years, the platform has grown to encompass patient electronic health records (EHR) and even clinical trial data. OHDSI is primarily tailored to serve scientists, offering them a comprehensive tool set for observational research. The tooling and foundational data model, OMOP-CDM, are purposefully crafted to produce high-quality and reproducible real-world evidence.

FHIR on the other hand, known as Fast Healthcare Interoperability Resources, serves as a universal standard for exchanging patient health records among different EHR-systems. It operates by utilizing resources bundled in profiles, which define the representation of specific information elements. There are approximately 100-150 resources available, covering various aspects such as diagnostics, medication, physical measurements, and more. This way, it can also be considered as a simplified model represented by JSON or XML objects (the FHIR Resources), where each resource consists of a number of logically related data elements. What sets FHIR apart is its inclusion of an API description that exposes the underlying EHR system's resources. This feature enables other applications to securely access and retrieve data from EHR systems while also being able to push relevant data back into them. In general, FHIR’s main purpose is to act as a facilitator for data transfer to other applications using its resource profiles and API definition. EHR-systems, such as openEHR, can use FHIR to make their data accessible to other applications. These applications may include consumer-oriented health apps that provide patients with valuable insights into their health data or research platforms like OHDSI.

Furthermore, already from our previous work[29], we have identified several overlaps in the capabilities offered by FHIR and OMOP-CDM. Table 2 presents the mapping between those two models, identifying that in principle they can both serve well as a layer of standardization for clinical research data within one's own research network.

Table 2: Mappings between FHIR and OMOP-CDM. (adapted from [11])

	FHIR	OMOP-CDM
Demographic	Patient, Observation	Person, Observation, Measurement
Treatments	Procedure, ServiceRequest	Drug_Exposure, Procedure_Occurrence
Diagnosis	Observation, DiagnosticReport	Condition_Occurrence
Conditions/Clinical Manifestations	Condition, Observation	Condition_Occurrence, Observation
Laboratory	Observation, Measure, MeasureReport	Measurement
Longitudinal data (History)	Most FHIR resources of interest have a date field	Condition_Occurrence, Observation with dates
Terminologies/ Ontologies	CodeSystem ValueSet, ConceptMap	Concept, Vocabulary, Domain, Concept_class, Concept_relationship, Concept_synonym, Concept_ancestor

As such, we can make the following observations:

- FHIR is designed as a flexible and extensible standard, which allows for the representation of various healthcare concepts. However, this flexibility can make it challenging to enforce strict data modeling and data integrity constraints, leading to potential data quality issues. In addition, this flexibility can lead to complex data structures and relationships which makes it more challenging to implement and maintain FHIR-based systems/servers which can also be based on different storages and accessed/queried with no standardized query languages (e.g. mongoDB for storing JSON resources, with its MQL (mongo query specific language) or XML-based databases). Furthermore, achieving semantic interoperability can also be challenging, since each data provider may use different standardized terminologies (or custom vocabularies) to represent their data, which makes the interpretation of data to vary between different organizations and implementations, leading to potential misinterpretation of data. Finally, FHIR is designed to evolve over time, with new versions and updates regularly released. Managing versioning and ensuring backward

[29] Cremonesi, Francesco, et al. "The need for multimodal health data modeling: A practical approach for a federated-learning healthcare platform." Journal of Biomedical Informatics 141 (2023): 104338.

compatibility between different FHIR implementations can be complex and may require significant effort.

- OMOP-CDM, on the other hand, provides a well-defined, standardized data model with a relational database schema. This structure enforces strong data modeling and integrity constraints, ensuring consistency and data quality. Its tabular structure reduces storage overhead as well, compared to hierarchical models like FHIR, and makes it easier to query and retrieve patient records without the resource fragmentation, which can be seen in FHIR. Furthermore, OMOP-CDM focuses on standardized coding systems for healthcare concepts, enhancing semantic interoperability through its OMOP standardized vocabularies, which ensures that data elements have consistent meanings across different implementations. Finally, OMOP-CDM maintains a relatively stable schema over time, reducing the challenges associated with versioning and backward compatibility. This stability is particularly important in the context of longitudinal research studies.

Based on the above, we can conclude that OMOP-CDM is more appropriate to be used as a common data model, persisting all information that will be used within EUCAIM. However, we also opt to facilitate FHIR resources as well as FHIR messages, which can enable transfer of EHR data into the common data model.

Nevertheless, since for this current early release of the data federation framework, the EUCAIM data infrastructure is comprised of the AI4HI project repositories, each serving as a separate EUCAIM Data Node that holds their own data models with no plan of transforming the data, we have also introduced a mediator component to allow flexibility, which however forces the local nodes to be responsible for the translations/transformations.

3.3 Hyper-Ontology for Interoperability

With the goal to support data interoperability and enable data integration among the various providers in the EUCAIM federation, a hyper-ontology is required. Moreover, one major challenge of this project is to facilitate interoperability among data that have been stored and modelled using diverse Clinical Data Models. An ontology in healthcare is a structured and formal representation of medical knowledge, concepts, and relationships within the healthcare domain. It provides a standardized framework for arranging and classifying medical data, enabling accurate interpretation, data integration, and semantic understanding across various healthcare systems and applications. Over the last few decades, ontologies have been developed in a number of medical domains, including those related to cancer and imaging fields[30,31]. Medical ontologies provide the conceptual basis for information exchange while standards ensure consistency in the information exchange across various systems[32]. National Cancer Institute thesaurus (NCIt) is the most commonly used ontology in cancer research.

[30] Silva, M.C., Eugénio, P., Faria, D., Pesquita, C., 2022. Ontologies and Knowledge Graphs in Oncology Research. Cancers (Basel) 14, 1906. https://doi.org/10.3390/cancers14081906

[31] Mayo, C. et al. Operational Ontology for Oncology (O3) – A Professional Society Based, Multi-Stakeholder, Consensus Driven Informatics Standard Supporting Clinical and Research use of "Real - World" Data from Patients Treated for Cancer: Operational Ontology for Radiation Oncology. International Journal of Radiation Oncology*Biology*Physics (2023) doi:10.1016/j.ijrobp.2023.05.033.

[32] Karami M, Rahimi A. Semantic Web Technologies for Sharing Clinical Information in Health Care Systems. Acta Inform Med. 2019 Mar;27(1):4-7. doi:10.5455/aim.2019.27.4-7. PMID: 31213735; PMCID: PMC6511266

SNOMED CT or the UMLS Metathesaurus are also popular and include a large number of mappings enabling interoperability.

3.3.1 Objective

The hyper-ontology aims to maintain and support semantic interoperability to query heterogeneous data sets of interest while addressing semantic search purposes (Figure 10). It defines the ontology-based standard and structured vocabulary with which federated queries are exchanged among applications and semantic annotations are assigned to images. It performs semantic alignment, mapping, and merging processes, supporting the ontology's model interoperability and reusability. Besides, the hyper-ontology's goal is to ensure an integration with the local common data models (CDMs), permitting a consistent mapping with local nodes.

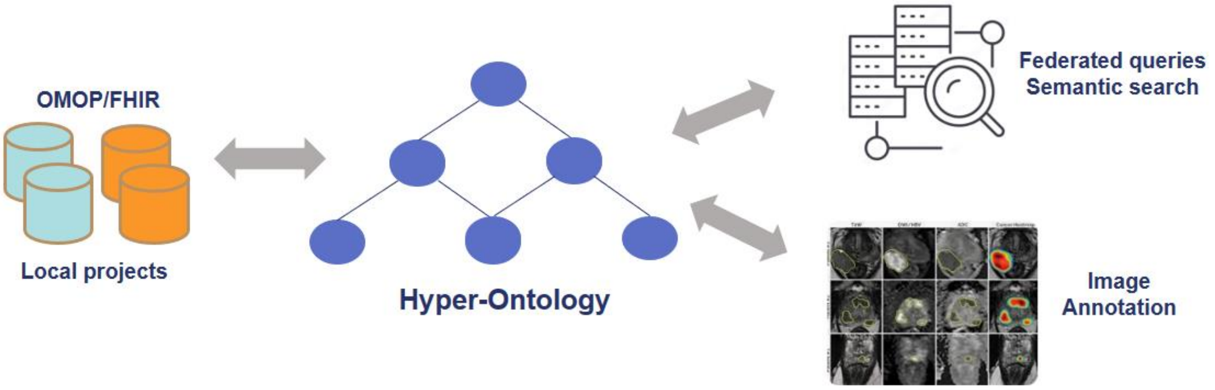


Figure 10: Hyper-Ontology's objective

3.3.2 Methodology

For building the hyper-ontology, we propose a collaborative-iterative ontology development process (Figure 11). In this process, six main phases are defined:

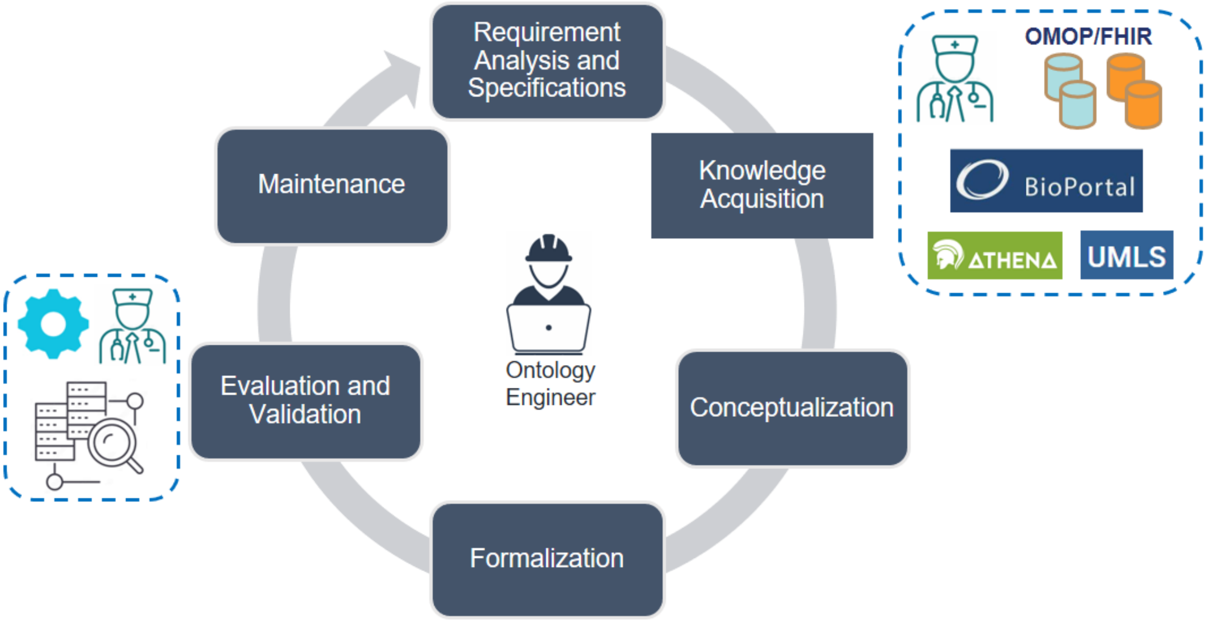


Figure 11: The Hyper-Ontology development process

1) *Requirement analysis and specifications*: specifies what should be represented in the hyper-ontology by considering the main user requirements. Based on the user

requirements of the semantic search and federated queries presented at the Consortium on June 22, 2023 (see Figure 12 for an example), we identified four main specifications: i) mandatory clinical knowledge; ii) mandatory imaging knowledge; iii) query criteria (e.g., age group, gender, image modality, diagnosis, body part, etc.); iv) common knowledge of all cancer types (e.g., date of birth, age at diagnosis, date of incident, topology, morphology, etc.). Furthermore, discussions with the application and partner experts are required to revise or rectify the requirements.

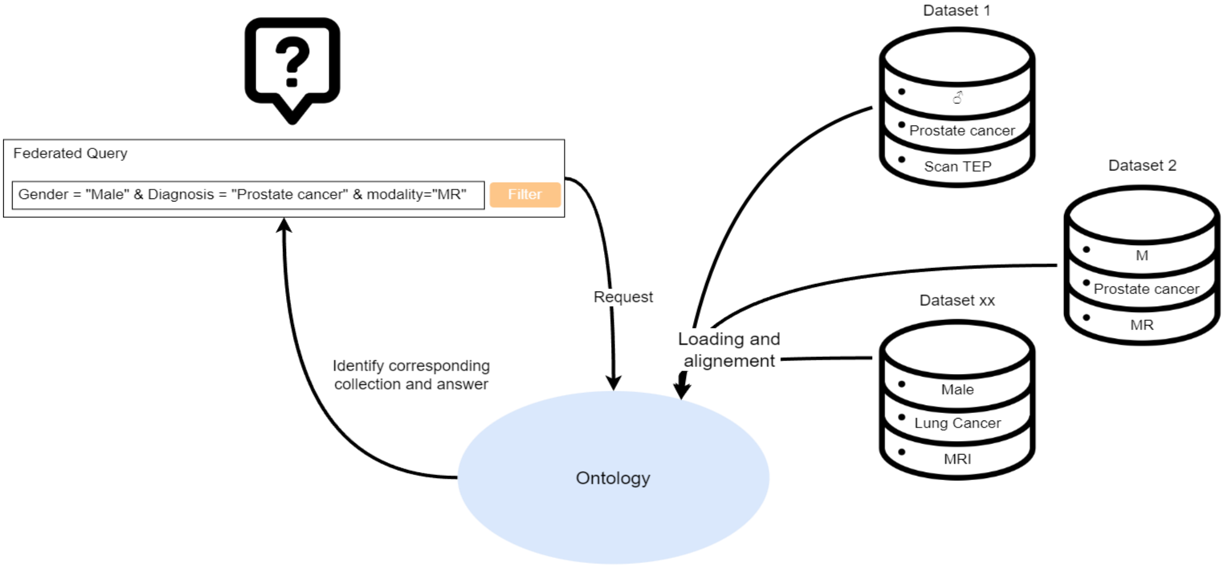


Figure 12: An example of federated query using terms from the hyper-ontology (e.g., Male, Prostate cancer, MRI)

2) *Knowledge acquisition*: defines the main sources of knowledge (e.g., experts, AI4HI projects, existing validated biomedical ontologies, reference terminologies, standards, etc.) to be considered in the hyper-ontology. A strategy is required to handle the diversity of the sources and the decision of the most convenient knowledge to be considered in the ontology.

   For the moment, mandatory clinical knowledge is collected from OMOP and FHIR based AI4HI projects by requesting the standardized concepts defined in those.

   The status of the clinical knowledge extracted from these projects is analysed. For the OMOP-based projects, the extracted concepts are considered taking into account the OHDSI standardization rules, especially those addressed for the definition of the cancer extension of the OMOP model[33].

   For the FHIR-based projects, the extracted concepts are considered taking into account the HL7 FHIR implementation guides relevant in the context of the EUCAIM project such as HL7 Vulcan Real-World Data[34], HL7 mCODE[35]).

   Figure 13 depicts an example of standardized concepts collected from the OMOP local projects (*concept_name_1*) and the (hierarchical) mappings performed to these

[33] Belenkaya R, Gurley MJ, Golozar A, et al: Extending the OMOP common data model and standardized vocabularies to support observational cancer research. JCO Clin Cancer Inform 5:12-20, 2021

[34] https://build.fhir.org/ig/HL7/vulcan-rwd/index.html

[35] HL7.FHIR.US.MCODE\Home - FHIR v4.0.1

concepts (*concept_name_2*) using Athena. Athena[36] is a searchable database maintained by the Observational Health Data Sciences and Informatics (OHDSI) that is available to researchers to help identify codes and match them to their standard Observational Medical Outcomes Partnership (OMOP) equivalent. The following columns are harvested: table, vocabulary, concept id, concept name, concept class, and relationship type (e.g., is-a, subsumes, maps to, mapped from, etc.). The blue-shaded concepts and mapped concepts are available in the AI4HI OMOP-based projects and will be used to build the ontology content. However, the white-shaded concepts are not available in the AI4HI projects, but they will be considered, in specific cases depending on the relationship type, to enrich the ontology content. For instance, if the linked concept, which is not available in local nodes (e.g., Central zone of prostate, concept_id='4089569'), is a superclass of a concept available in local nodes (e.g., Central zone of left half-prostate, concept_id='37396912'), it will be considered to build the hierarchy and classify the concepts (see Figure 16). However, to prevent additional granularity level in the hierarchy, which is not required, the concept (not available in the AI4HI nodes) mapped as a subclass (of a concept available in local nodes) will not be considered in the hyper-ontology.

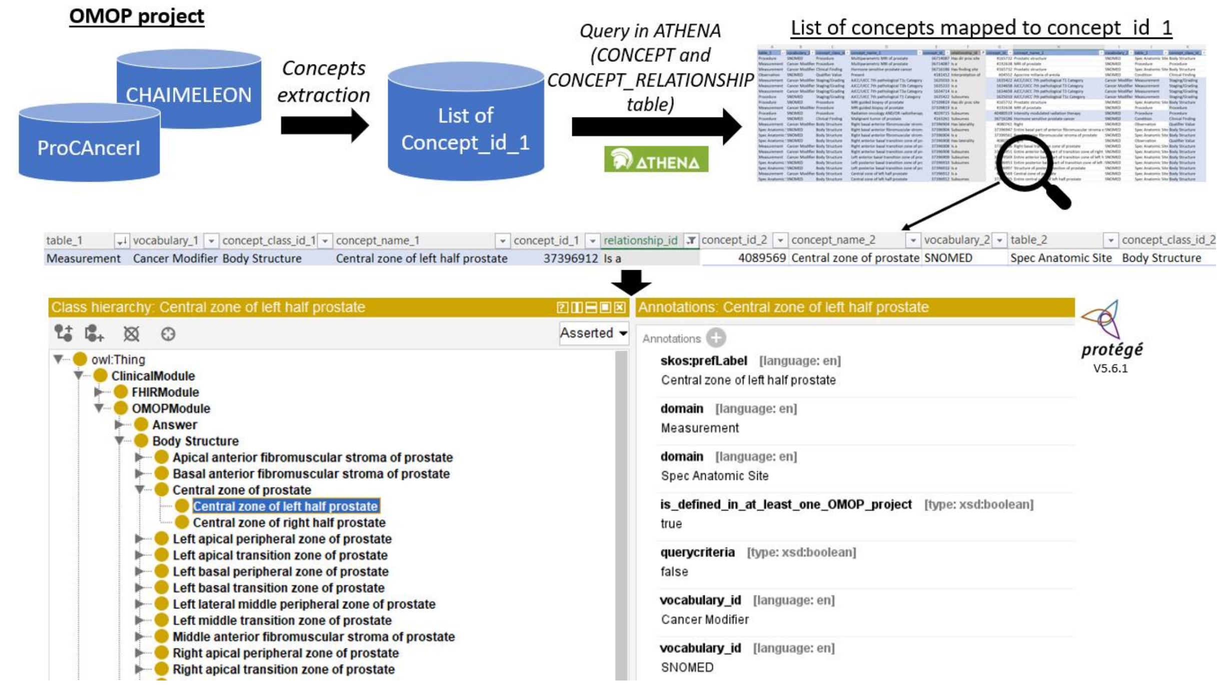

OMOP project
CHAIMELEON
ProCAncerI
Concepts extraction
List of Concept_id_1
Query in ATHENA (CONCEPT and CONCEPT_RELATIONSHIP table)
ATHENA
List of concepts mapped to concept_id_1
protégé
V5.6.1

table_1	vocabulary_1	concept_class_id_1	concept_name_1	concept_id_1	relationship_id	concept_id_2	concept_name_2	vocabulary_2	table_2	concept_class_id_2
Measurement	Cancer Modifier	Body Structure	Central zone of left half prostate	37396912	Is a	4089569	Central zone of prostate	SNOMED	Spec Anatomic Site	Body Structure

Class hierarchy: Central zone of left half prostate

Asserted

- owl:Thing
  - ClinicalModule
    - FHIRModule
    - OMOPModule
      - Answer
      - Body Structure
        - Apical anterior fibromuscular stroma of prostate
        - Basal anterior fibromuscular stroma of prostate
        - Central zone of prostate
          - Central zone of left half prostate
          - Central zone of right half prostate
        - Left apical peripheral zone of prostate
        - Left apical transition zone of prostate
        - Left basal peripheral zone of prostate
        - Left basal transition zone of prostate
        - Left lateral middle peripheral zone of prostate
        - Left middle transition zone of prostate
        - Middle anterior fibromuscular stroma of prostate
        - Right apical peripheral zone of prostate
        - Right apical transition zone of prostate

Annotations: Central zone of left half prostate

Annotations

skos:prefLabel [language: en]
Central zone of left half prostate

domain [language: en]
Measurement

domain [language: en]
Spec Anatomic Site

is_defined_in_at_least_one_OMOP_project [type: xsd:boolean]
true

querycriteria [type: xsd:boolean]
false

vocabulary_id [language: en]
Cancer Modifier

vocabulary_id [language: en]
SNOMED

Figure 13: Methodology and example of mappings applied to standardized concepts collected from OMOP projects

For the medical imaging knowledge, the main challenge is that it is defined within the AI4HI projects using the DICOM standard, which is not considered as a standardized terminology (apart from the ProCAncer-I AI4HI project that has used Radlex[37] to standardize a set of important DICOM imaging metadata). Meanwhile, ongoing research contributions aiming to align the imaging data to OMOP by extending the OMOP tables will guide the selection of reference terminologies to be considered within

[36] https://athena.ohdsi.org/

[37] Radiological Society of North America (RSNA). RadLex: Radiology Lexicon. Available at: https://www.rsna.org/radlex. Accessed Sep 1, 2023.

the hyper-ontology (Park et al., 2022[38]; Kim et al., 2022[39]; Peng et al., 2023[40]). Although, pertinent results are not yet achieved on the standardization level of the imaging data. RadLex, is considered as a major terminology for defining standard imaging concepts, such as *Imaging modality*, *Laterality*, etc.

3) *Conceptualization*: we propose a multi-layered-modular architecture of the hyper-ontology and a hybrid approach, composed of bottom-up and top-down strategies, to develop the content of the layers and modules (Figure 14).

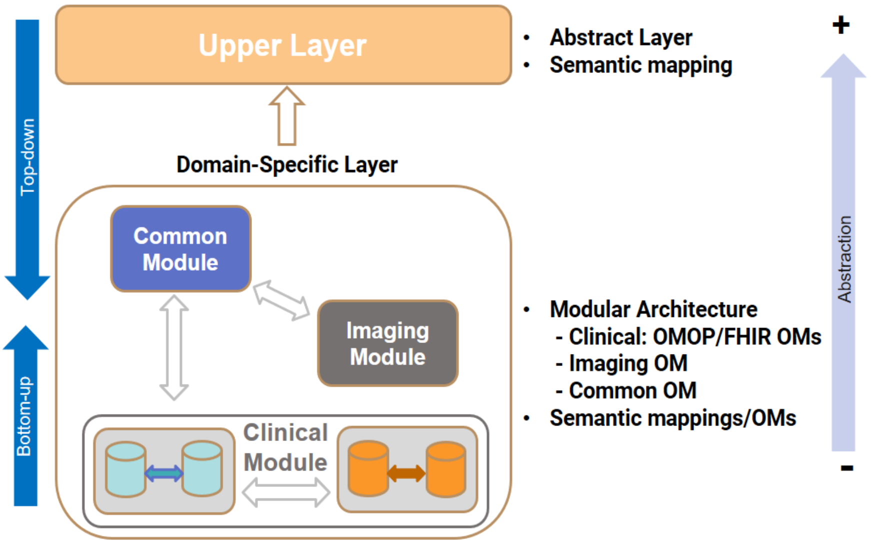


Figure 14: Two-layered-modular architecture of the Hyper-Ontology. OM: Ontology-module.

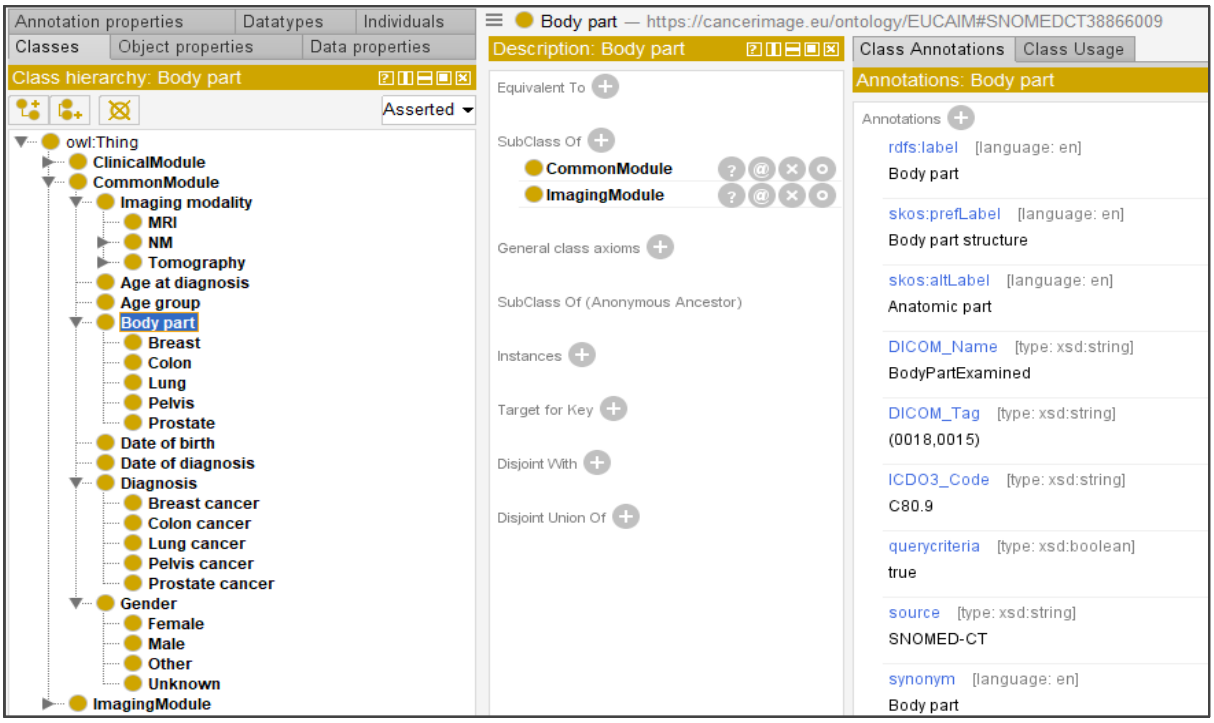
Two main layers are identified for the moment: *the upper layer,* that defines concepts at the abstract level, and *domain-specific layer* that represents the main concepts at the domain-specific level. Three main modules are defined at the domain-specific level:

i) *Common Module*: defines the common knowledge between all cancer types (e.g., *date of birth*), including the query criteria (e.g., *body part*, *diagnosis,* and *gender*) (see Figure 15 for an example). This module is explicitly defined as a separate module in version 0.4 alpha of the hyper-ontology to fulfill the requirements of the federated queries for the MM2;

ii) *Clinical Module*: represents the mandatory clinical knowledge collected from the AI4HI projects. In this module, two separate sub-modules are provisionally defined: OMOP and FHIR. While the OMOP module defines the standardized concepts defined in the OMOP projects, the FHIR module represents the standardized concepts defined in the FHIR projects. The concepts are organized in hierarchies based on the hierarchical mappings harvested from Athena and UMLS. Excerpts of these hierarchies (Prostate cancer for the OMOP module and Breast cancer for the FHIR module) are depicted in Figure 16. Moreover, non-hierarchical mappings will be performed to

[38] Park, C. *et al.* Development and Validation of the Radiology Common Data Model (R-CDM) for the International Standardization of Medical Imaging Data. *Yonsei Med J* 63, S74 (2022)
[39] Kim TH. *et al*. Development and validation of a management system and dataset quality assessment tool for the Radiology Common Data Model (R_CDM): A case study in liver disease. Int J Med Inform. 2022 Apr 1;162:104759. doi: 10.1016/j.ijmedinf.2022.104759. Epub ahead of print. PMID: 35390589
[40] Peng Y, *et al* . An ETL-process design for data harmonization to participate in international research with German real-world data based on FHIR and OMOP CDM. Int J Med Inform. 2023 Jan;169:104925. doi: 10.1016/j.ijmedinf.2022.104925. Epub 2022 Nov 10. PMID: 36395615.

maintain the semantic context of the hyper-ontology. Commonly defined concepts are expected between the OMOP and FHIR modules (e.g., Chemotherapy). They will be tracked using the concept IDs retrieved from OMOP and FHIR CDMs. For instance, the concept IDs will be defined as semantic annotations to follow the CDM source of concepts. Therefore, we suppose that an explicit definition of OMOP and FHIR modules is not required in further work.



Annotation properties | Datatypes | Individuals
Classes | Object properties | Data properties

Class hierarchy: Body part

Asserted

- owl:Thing
  - ClinicalModule
  - CommonModule
    - Imaging modality
      - MRI
      - NM
      - Tomography
    - Age at diagnosis
    - Age group
    - Body part
      - Breast
      - Colon
      - Lung
      - Pelvis
      - Prostate
    - Date of birth
    - Date of diagnosis
    - Diagnosis
      - Breast cancer
      - Colon cancer
      - Lung cancer
      - Pelvis cancer
      - Prostate cancer
    - Gender
      - Female
      - Male
      - Other
      - Unknown
  - ImagingModule

Body part — https://cancerimage.eu/ontology/EUCAIM#SNOMEDCT38866009

Description: Body part

Equivalent To

SubClass Of
- CommonModule
- ImagingModule

General class axioms

SubClass Of (Anonymous Ancestor)

Instances

Target for Key

Disjoint With

Disjoint Union Of

Class Annotations | Class Usage

Annotations: Body part

Annotations

rdfs:label [language: en]
Body part

skos:prefLabel [language: en]
Body part structure

skos:altLabel [language: en]
Anatomic part

DICOM_Name [type: xsd:string]
BodyPartExamined

DICOM_Tag [type: xsd:string]
(0018,0015)

ICDO3_Code [type: xsd:string]
C80.9

querycriteria [type: xsd:boolean]
true

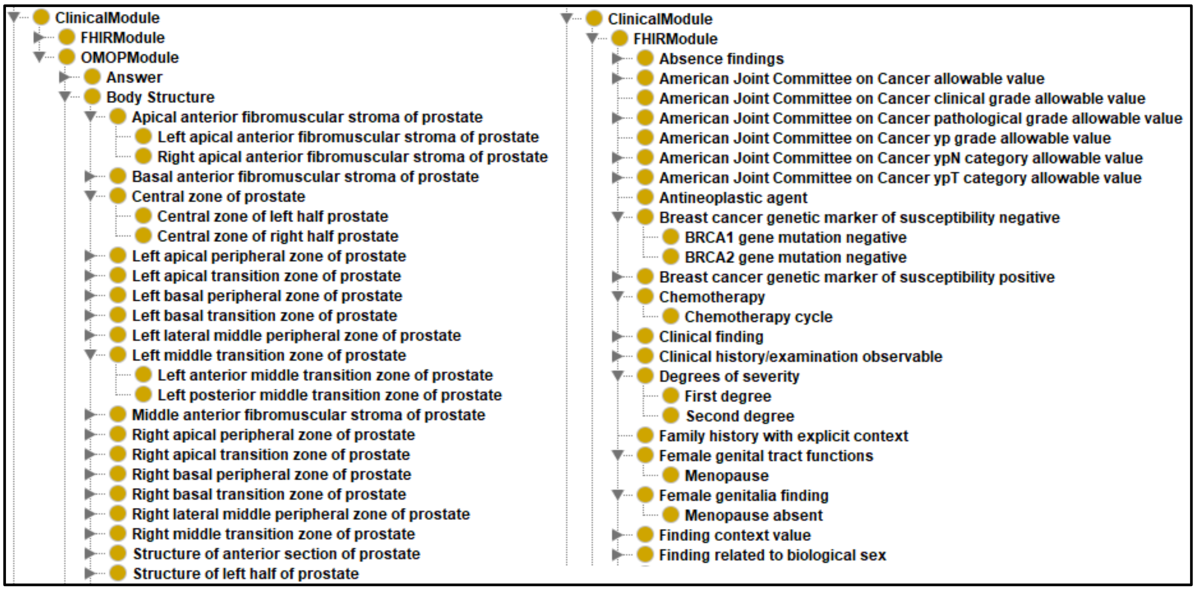
source [type: xsd:string]
SNOMED-CT

synonym [language: en]
Body part

Figure 15: An example of query criteria (Body part) represented in Protege.

iii) *Imaging Module*: two main perspectives are proposed for building this module. The first is to define the mandatory imaging knowledge considered in the DICOM metadata of the local projects, such as, among others, *body part*, *modality*, and *laterality*. Standardized terminologies, such as RadLex, are envisaged to represent the imaging knowledge in the ontology by applying semantic similarity techniques between the mandatory DICOM attributes and the labels of standard concepts. For instance, Figure 17 depicts the RadLex's concept *Imaging modality (*RID10311), defined in the hyper-ontology, and its alignment to the DICOM attribute *Modality* (0008,0060). Interestingly, adding *DICOM_name* and *DICOM_tag* annotations permits the integration between the hyper-ontology and the local CDMs on the imaging metadata level. Besides, for the imaging knowledge, which is considered in the DICOM standard as attribute values (e.g., *MR*, *CT*, *PT* for *Modality*) and is required for executing the federated queries (e.g., *MRI* (see Figure 12)), we rely on the standard terminologies (e.g., RadLex) to define them by considering the specified user requirements (e.g., synonym, acronym, etc.) (see Figure 18). In this case, the integration is performed on the imaging dala level of the local nodes.

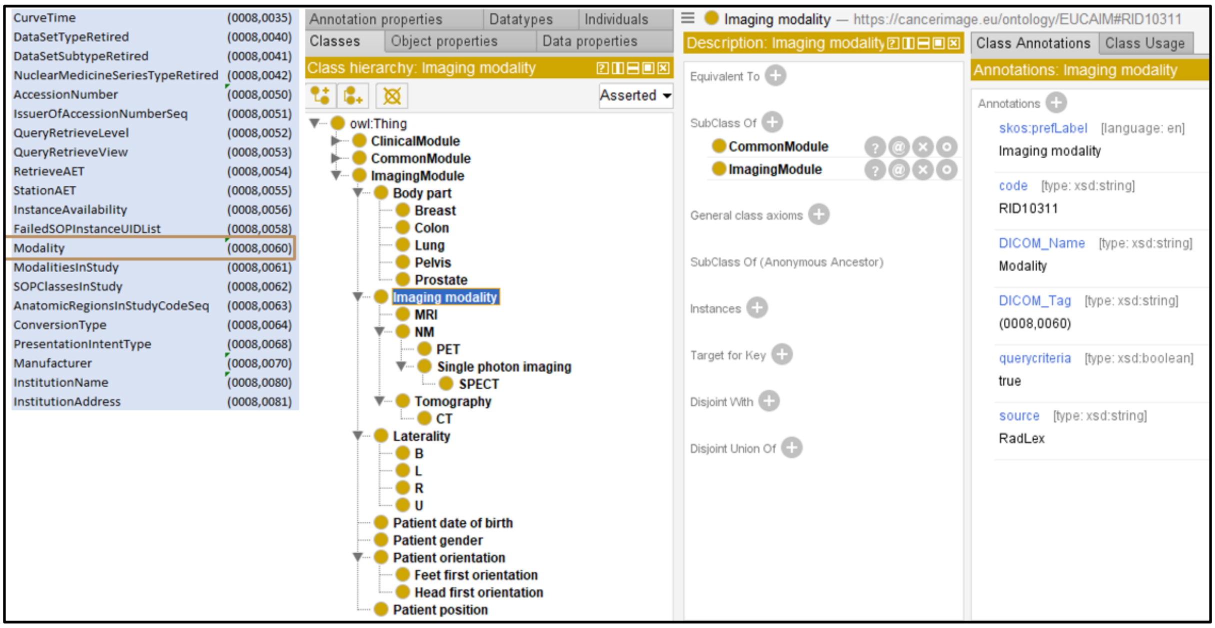
Preliminary work of semantic mapping has been performed in the context of the ontology modules aiming to align the standardized concepts to validated ontological resources and reference terminologies (e.g., SNOMED CT[41], RadLex[42], Gender[43], ICD-O3[44]). For instance, the synonym, acronym, preferred and alternative labels are defined in the current version of the ontology (examples are shown in Figure 15 and Figure 17).



- ClinicalModule
  - FHIRModule
  - OMOPModule
    - Answer
    - Body Structure
      - Apical anterior fibromuscular stroma of prostate
        - Left apical anterior fibromuscular stroma of prostate
        - Right apical anterior fibromuscular stroma of prostate
      - Basal anterior fibromuscular stroma of prostate
      - Central zone of prostate
        - Central zone of left half prostate
        - Central zone of right half prostate
      - Left apical peripheral zone of prostate
      - Left apical transition zone of prostate
      - Left basal peripheral zone of prostate
      - Left basal transition zone of prostate
      - Left lateral middle peripheral zone of prostate
      - Left middle transition zone of prostate
        - Left anterior middle transition zone of prostate
        - Left posterior middle transition zone of prostate
      - Middle anterior fibromuscular stroma of prostate
      - Right apical peripheral zone of prostate
      - Right apical transition zone of prostate
      - Right basal peripheral zone of prostate
      - Right basal transition zone of prostate
      - Right lateral middle peripheral zone of prostate
      - Right middle transition zone of prostate
      - Structure of anterior section of prostate
      - Structure of left half of prostate

- ClinicalModule
  - FHIRModule
    - Absence findings
    - American Joint Committee on Cancer allowable value
    - American Joint Committee on Cancer clinical grade allowable value
    - American Joint Committee on Cancer pathological grade allowable value
    - American Joint Committee on Cancer yp grade allowable value
    - American Joint Committee on Cancer ypN category allowable value
    - American Joint Committee on Cancer ypT category allowable value
    - Antineoplastic agent
    - Breast cancer genetic marker of susceptibility negative
      - BRCA1 gene mutation negative
      - BRCA2 gene mutation negative
    - Breast cancer genetic marker of susceptibility positive
    - Chemotherapy
      - Chemotherapy cycle
    - Clinical finding
    - Clinical history/examination observable
    - Degrees of severity
      - First degree
      - Second degree
    - Family history with explicit context
    - Female genital tract functions
      - Menopause
    - Female genitalia finding
      - Menopause absent
    - Finding context value
    - Finding related to biological sex

Figure 16: An excerpt of OMOP and FHIR ontology modules represented in Protege



CurveTime	(0008,0035)
DataSetTypeRetired	(0008,0040)
DataSetSubtypeRetired	(0008,0041)
NuclearMedicineSeriesTypeRetired	(0008,0042)
AccessionNumber	(0008,0050)
IssuerOfAccessionNumberSeq	(0008,0051)
QueryRetrieveLevel	(0008,0052)
QueryRetrieveView	(0008,0053)
RetrieveAET	(0008,0054)
StationAET	(0008,0055)
InstanceAvailability	(0008,0056)
FailedSOPInstanceUIDList	(0008,0058)
Modality	(0008,0060)
ModalitiesInStudy	(0008,0061)
SOPClassesInStudy	(0008,0062)
AnatomicRegionsInStudyCodeSeq	(0008,0063)
ConversionType	(0008,0064)
PresentationIntentType	(0008,0068)
Manufacturer	(0008,0070)
InstitutionName	(0008,0080)
InstitutionAddress	(0008,0081)

Class hierarchy: Imaging modality

Asserted

- owl:Thing
  - ClinicalModule
  - CommonModule
  - ImagingModule
    - Body part
      - Breast
      - Colon
      - Lung
      - Pelvis
      - Prostate
    - Imaging modality
      - MRI
      - NM
        - PET
        - Single photon imaging
          - SPECT
      - Tomography
        - CT
    - Laterality
      - B
      - L
      - R
      - U
    - Patient date of birth
    - Patient gender
    - Patient orientation
      - Feet first orientation
      - Head first orientation
    - Patient position

Imaging modality — https://cancerimage.eu/ontology/EUCAIM#RID10311

Description: Imaging modality

Equivalent To

SubClass Of
- CommonModule
- ImagingModule

General class axioms

SubClass Of (Anonymous Ancestor)

Instances

Target for Key

Disjoint With

Disjoint Union Of

Annotations: Imaging modality

Annotations

skos:prefLabel [language: en]
Imaging modality

code [type: xsd:string]
RID10311

DICOM_Name [type: xsd:string]
Modality

DICOM_Tag [type: xsd:string]
(0008,0060)

querycriteria [type: xsd:boolean]
true

source [type: xsd:string]
RadLex

Figure 17: An excerpt of Imaging Module represented in Protege (examples of DICOM attributes to the left)

Concerning the strategies, the bottom-up approach is applied mainly to integrate the mandatory clinical and imaging knowledge collected from the local AI4HI projects into the hyper-ontology. A hierarchy of standard concepts is built based on the mandatory knowledge and semantically enriched by applying semantic mappings. The top-down

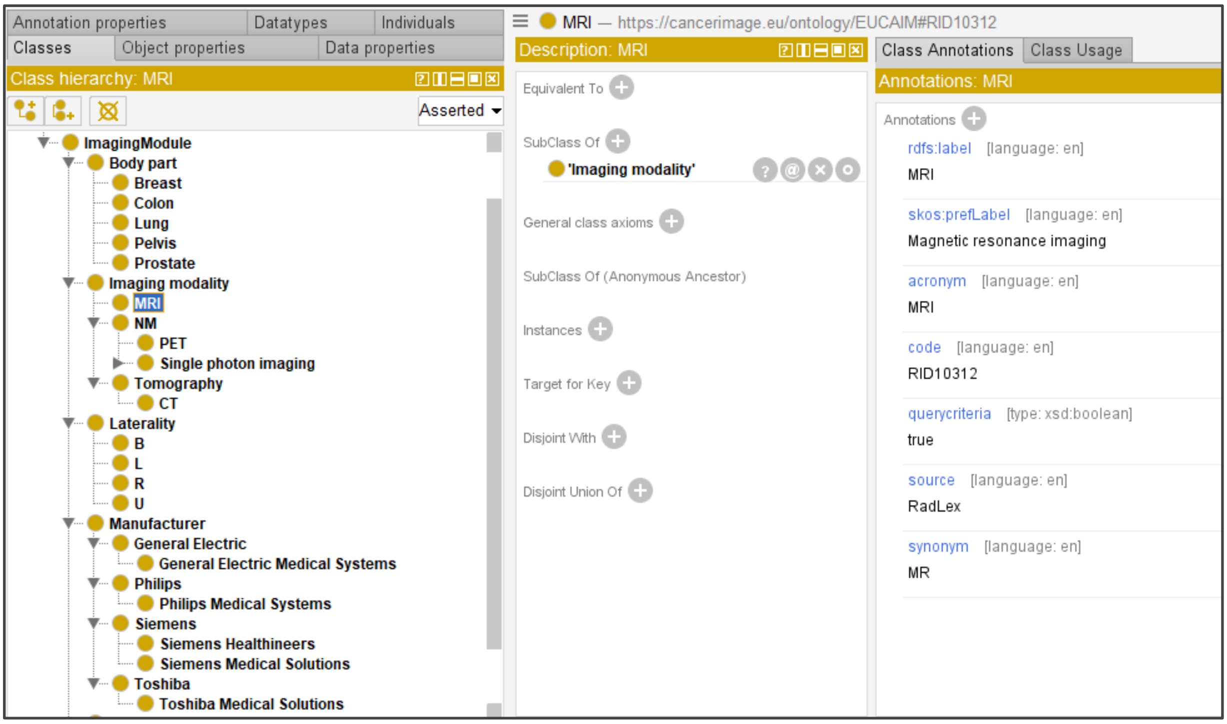
[41] SNOMED Clinical Terms. Available from URL: https://www.snomed.org/

[42] Radiological Society of North America (RSNA). RadLex: Radiology Lexicon. Available at: https://www.rsna.org/radlex. Accessed Sept 1, 2023.

[43] https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:gender

[44] https://seer.cancer.gov/icd-o-3/

strategy focuses on coherent modeling and semantic representation of the acquired knowledge in the ontology by relying on validated foundational ontologies. Also, approaches, such as, among others, the Oncology extension of the OMOP CDM and R-CDM (Radiology-CDM)[45,46], will be analyzed for the conceptualization of the OMOP and Imaging modules.



rdfs:label [language: en]
MRI

skos:prefLabel [language: en]
Magnetic resonance imaging

acronym [language: en]
MRI

code [language: en]
RID10312

querycriteria [type: xsd:boolean]
true

source [language: en]
RadLex

synonym [language: en]
MR

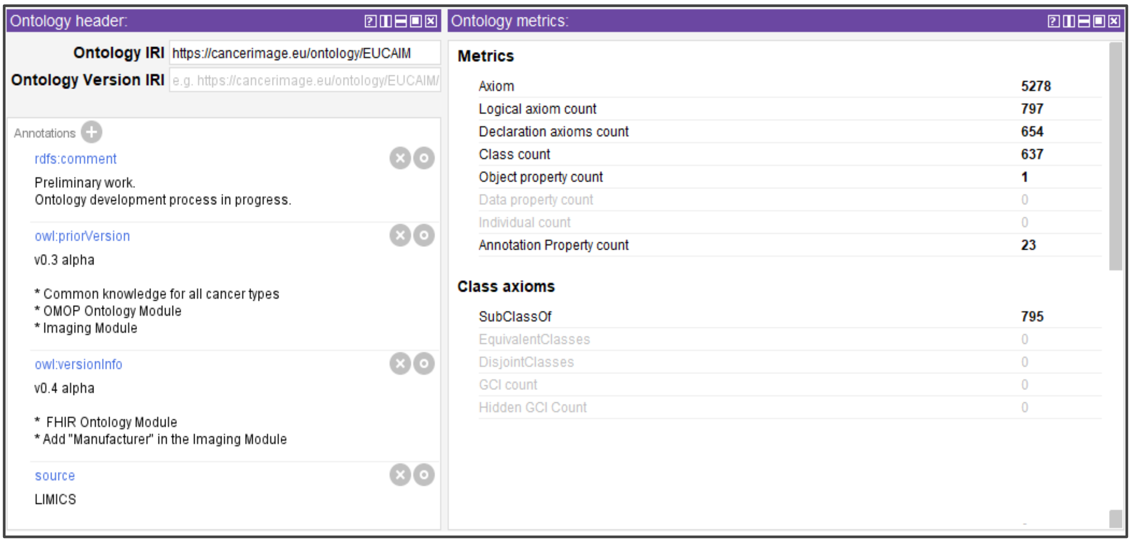
Figure 18: An excerpt of Imaging Module represented in Protege

4) *Formalization*: the hyper-ontology is formalized in OWL and shared for querying or semantic search applications (Figure 19). For the metrics, 637 classes and 795 hierarchical relations are defined in the version 0.4 alpha.

5) *Evaluation and Validation*: with the help of domain experts, the content of the hyper-ontology, such as the semantic alignments, synonyms, preferred labels, is evaluated. Also, using automatic strategies, the ontology consistency and structure are assessed. To test the consistency of the ontology, semantic reasoners, such as HermiT and Pellet, are applied, ensuring that the ontology is free of contradictions. For the structure-based evaluation, ontology criteria, such as size and complexity, are quantified. We are interested in quantifying structural measures since they are positively correlated with the semantic accuracy of the knowledge modeled in the ontology[47].

45. Belenkaya, R. et al. Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research. JCO Clinical Cancer Informatics 12–20 (2021) doi:10.1200/CCI.20.00079.
46. Park, C. et al. Development and Validation of the Radiology Common Data Model (R-CDM) for the International Standardization of Medical Imaging Data. Yonsei Med J **63**, S74 (2022).
47 Sanchez, D., Batet, M., Martinez, S. and Ferrer, J.D. Semantic variance: An intuitive measure for ontology accuracy evaluation, Engineering Applications of Artificial Intelligence 39 (2015), 89–99. doi:10.1016/j.engappai.2014.11.012.

Besides, the ontology validation is required to test the performance of the ontology model. Therefore, we need to use the ontology in a semantic search or querying tasks and verify the accuracy of the results and if the specified requirements are fulfilled.



Ontology header:

Ontology IRI https://cancerimage.eu/ontology/EUCAIM
Ontology Version IRI e.g. https://cancerimage.eu/ontology/EUCAIM/

Annotations

rdfs:comment
Preliminary work.
Ontology development process in progress.

owl:priorVersion
v0.3 alpha

* Common knowledge for all cancer types
* OMOP Ontology Module
* Imaging Module

owl:versionInfo
v0.4 alpha

* FHIR Ontology Module
* Add "Manufacturer" in the Imaging Module

source
LIMICS

Ontology metrics:

Metrics

Axiom	5278
Logical axiom count	797
Declaration axioms count	654
Class count	637
Object property count	1
Data property count	0
Individual count	0
Annotation Property count	23

Class axioms

SubClassOf	795
EquivalentClasses	0
DisjointClasses	0
GCI count	0
Hidden GCI Count	0

Figure 19: Version 0.4 alpha of the Hyper-Ontology

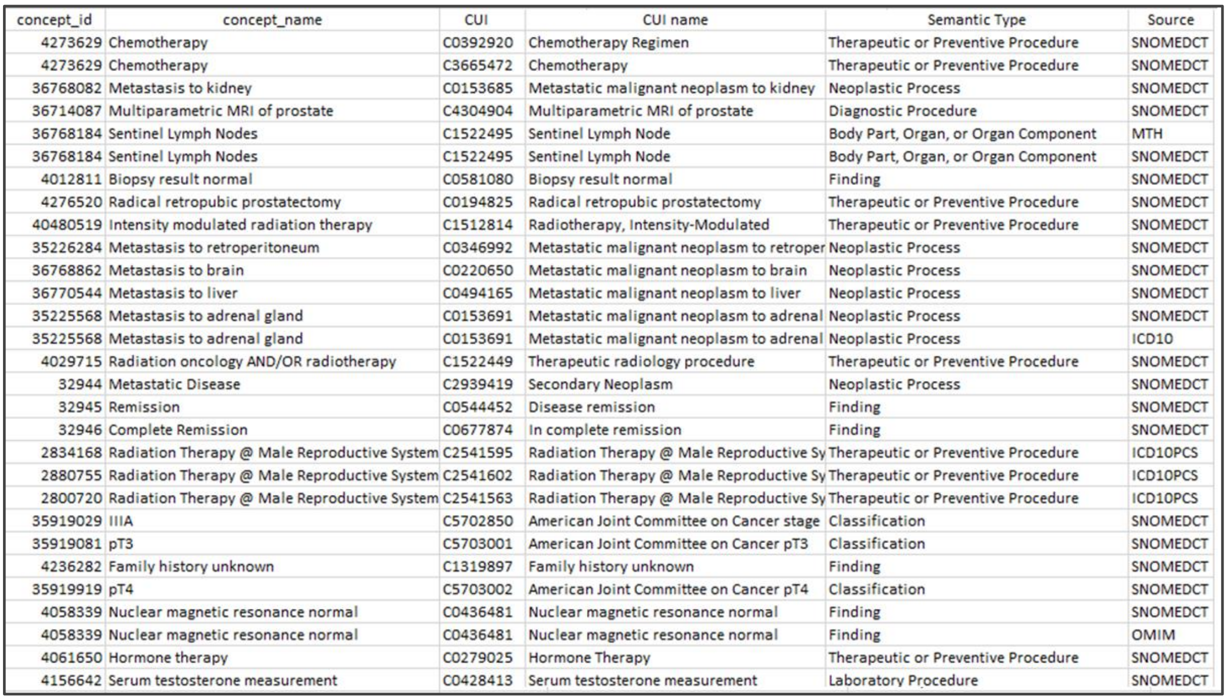
6) *Maintenance*: based on the feedback of the domain experts and the results of the ontology validation, a revision and correction phase is required to enhance the semantic content of the ontology.

For this MM2 Milestone, we provide the WIP (Work In Progress) OWL file (version 0.4 alpha) of the **EUCAIM's Hyper-Ontology** (link).

In further works, we will revise the preliminary results of the hyper-ontology based on the results of the evaluation and validation phase. Moreover, we will revise the ontology specifications and requirements with the experts (application, partners, and medical) to produce the ontology requirements specification document (ORSD)[48] required to support the hyper-ontology development process. This activity will be carried out at the beginning of the ontology project and in parallel with the knowledge acquisition phase. Therefore, the knowledge acquisition phase will be considered to acquire the mandatory clinical knowledge by requesting the OMOP and FHIR based AI4HI projects. Also, the imaging knowledge will be revised to specify the mandatory knowledge required in the hyper-ontology. Semantic mappings will also be performed by applying semantic/syntactic similarity techniques to align the hyper-ontology content with validated ontological resources and reference terminologies, and enrich the ontology content semantically. The UMLS (Unified Medical Language System) is envisaged to support the ontology development process specifically for building the upper layer and maintain the interoperability in the hyper-ontology. Figure 20 shows an example of aligning clinical concepts collected from OMOP based projects (*concept_id* and *concept_name*) to UMLS by identifying their CUIs (Concept Unique Identifiers), CUI names (a concept can have many different names defined in different source vocabularies), semantic types (e.g., Finding, Laboratory procedure, etc.), and sources (e.g., SNOMED, ICD10, etc.). As an example, the semantic types will be considered in the hyper-ontology. For instance, the OMOP and/or FHIR

[48] Suárez-Figueroa, M.C., Gómez-Pérez, A., Villazón-Terrazas, B. (2009). How to Write and Use the Ontology Requirements Specification Document. In: Meersman, R., Dillon, T., Herrero, P. (eds) On the Move to Meaningful Internet Systems: OTM 2009. OTM 2009. Lecture Notes in Computer Science, vol 5871. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-05151-7_16

concepts having a ‘Finding’ as semantic type will be identified (or categorized) as findings in the hyper-ontology.

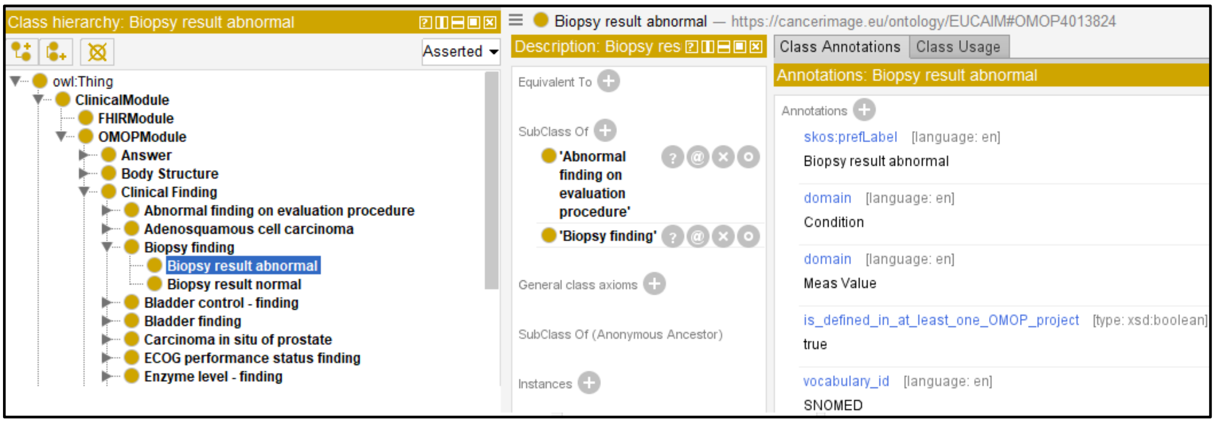


concept_id	concept_name	CUI	CUI name	Semantic Type	Source
4273629	Chemotherapy	C0392920	Chemotherapy Regimen	Therapeutic or Preventive Procedure	SNOMEDCT
4273629	Chemotherapy	C3665472	Chemotherapy	Therapeutic or Preventive Procedure	SNOMEDCT
36768082	Metastasis to kidney	C0153685	Metastatic malignant neoplasm to kidney	Neoplastic Process	SNOMEDCT
36714087	Multiparametric MRI of prostate	C4304904	Multiparametric MRI of prostate	Diagnostic Procedure	SNOMEDCT
36768184	Sentinel Lymph Nodes	C1522495	Sentinel Lymph Node	Body Part, Organ, or Organ Component	MTH
36768184	Sentinel Lymph Nodes	C1522495	Sentinel Lymph Node	Body Part, Organ, or Organ Component	SNOMEDCT
4012811	Biopsy result normal	C0581080	Biopsy result normal	Finding	SNOMEDCT
4276520	Radical retropubic prostatectomy	C0194825	Radical retropubic prostatectomy	Therapeutic or Preventive Procedure	SNOMEDCT
40480519	Intensity modulated radiation therapy	C1512814	Radiotherapy, Intensity-Modulated	Therapeutic or Preventive Procedure	SNOMEDCT
35226284	Metastasis to retroperitoneum	C0346992	Metastatic malignant neoplasm to retroper	Neoplastic Process	SNOMEDCT
36768862	Metastasis to brain	C0220650	Metastatic malignant neoplasm to brain	Neoplastic Process	SNOMEDCT
36770544	Metastasis to liver	C0494165	Metastatic malignant neoplasm to liver	Neoplastic Process	SNOMEDCT
35225568	Metastasis to adrenal gland	C0153691	Metastatic malignant neoplasm to adrenal	Neoplastic Process	SNOMEDCT
35225568	Metastasis to adrenal gland	C0153691	Metastatic malignant neoplasm to adrenal	Neoplastic Process	ICD10
4029715	Radiation oncology AND/OR radiotherapy	C1522449	Therapeutic radiology procedure	Therapeutic or Preventive Procedure	SNOMEDCT
32944	Metastatic Disease	C2939419	Secondary Neoplasm	Neoplastic Process	SNOMEDCT
32945	Remission	C0544452	Disease remission	Finding	SNOMEDCT
32946	Complete Remission	C0677874	In complete remission	Finding	SNOMEDCT
2834168	Radiation Therapy @ Male Reproductive System	C2541595	Radiation Therapy @ Male Reproductive Sy	Therapeutic or Preventive Procedure	ICD10PCS
2880755	Radiation Therapy @ Male Reproductive System	C2541602	Radiation Therapy @ Male Reproductive Sy	Therapeutic or Preventive Procedure	ICD10PCS
2800720	Radiation Therapy @ Male Reproductive System	C2541563	Radiation Therapy @ Male Reproductive Sy	Therapeutic or Preventive Procedure	ICD10PCS
35919029	IIIA	C5702850	American Joint Committee on Cancer stage	Classification	SNOMEDCT
35919081	pT3	C5703001	American Joint Committee on Cancer pT3	Classification	SNOMEDCT
4236282	Family history unknown	C1319897	Family history unknown	Finding	SNOMEDCT
35919919	pT4	C5703002	American Joint Committee on Cancer pT4	Classification	SNOMEDCT
4058339	Nuclear magnetic resonance normal	C0436481	Nuclear magnetic resonance normal	Finding	SNOMEDCT
4058339	Nuclear magnetic resonance normal	C0436481	Nuclear magnetic resonance normal	Finding	OMIM
4061650	Hormone therapy	C0279025	Hormone Therapy	Therapeutic or Preventive Procedure	SNOMEDCT
4156642	Serum testosterone measurement	C0428413	Serum testosterone measurement	Laboratory Procedure	SNOMEDCT

Figure 20: Example of aligning standardized OMOP concepts to UMLS

3.4 Integration of CDM and Hyper-Ontology

The hyper-ontology should ensure consistent integration with the local CDMs permitting the mapping between local nodes and the ontology model on the clinical and imaging levels. In the hyper-ontology (see section 3.3), the clinical and imaging knowledge is represented using standardized terminologies and internationally accepted biomedical ontologies, such as SNOMED CT, RadLex, ICD-O-3, etc. Besides, mappings among terminologies are also considered in the hyper-ontology permitting the interoperability and the semantic enrichment of the ontology model.



Class hierarchy: Biopsy result abnormal
Asserted
Biopsy result abnormal — https://cancerimage.eu/ontology/EUCAIM#OMOP4013824
Description: Biopsy res
Class Annotations
Class Usage
Annotations: Biopsy result abnormal
Equivalent To
SubClass Of
General class axioms
SubClass Of (Anonymous Ancestor)
Instances
Annotations

- owl:Thing
  - ClinicalModule
    - FHIRModule
    - OMOPModule
      - Answer
      - Body Structure
      - Clinical Finding
        - Abnormal finding on evaluation procedure
        - Adenosquamous cell carcinoma
        - Biopsy finding
          - Biopsy result abnormal
          - Biopsy result normal
        - Bladder control - finding
        - Bladder finding
        - Carcinoma in situ of prostate
        - ECOG performance status finding
        - Enzyme level - finding

SubClass Of: 'Abnormal finding on evaluation procedure'; 'Biopsy finding'

skos:prefLabel [language: en]
Biopsy result abnormal

domain [language: en]
Condition

domain [language: en]
Meas Value

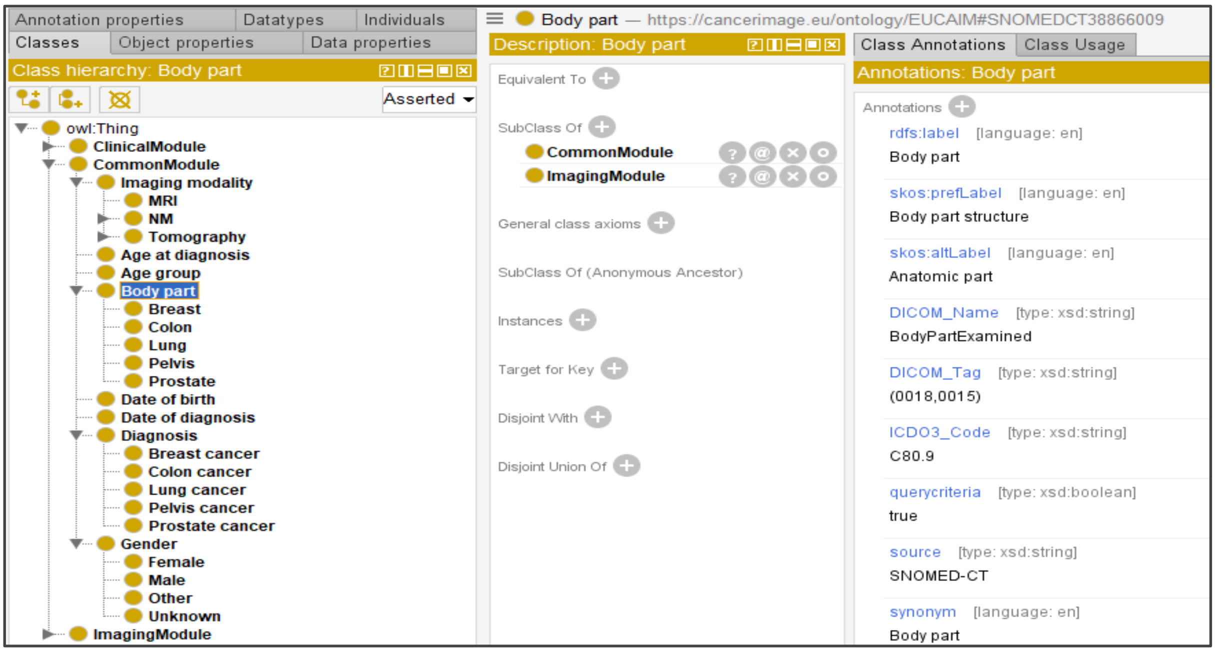
is_defined_in_at_least_one_OMOP_project [type: xsd:boolean]
true

vocabulary_id [language: en]
SNOMED

Figure 21: Example of a standard clinical concept (Biopsy result abnormal)

To provide consistent integration, we differentiate between clinical and imaging levels and metadata and data levels. For the clinical level, the data defined in local nodes are interoperable and aligned to standardized concepts by applying the ETL process. Therefore, the mappings between the local CDMs and the hyper ontology are performed through the standard concepts defined on both sides. For instance, the concept Biopsy result abnormal (concept_id = '4013824' in OMOP) is defined in the hyper-ontology as a SNOMED concept and in the local nodes aligned to OMOP, assuring the mapping between both sides.

Concerning the metadata level, the hyper-ontology aligns with the public metadata catalogue (section 3.5 Public Catalogue - Metadata Model by considering the main categories and sources of concepts. For instance, Body part (see Figure 22) is specified into five types in the hyper-ontology: Breast, Colon, Lung, Pelvis, and Prostate.



Body part — https://cancerimage.eu/ontology/EUCAIM#SNOMEDCT38866009

```
owl:Thing
  ClinicalModule
  CommonModule
    Imaging modality
      MRI
      NM
      Tomography
    Age at diagnosis
    Age group
    Body part
      Breast
      Colon
      Lung
      Pelvis
      Prostate
    Date of birth
    Date of diagnosis
    Diagnosis
      Breast cancer
      Colon cancer
      Lung cancer
      Pelvis cancer
      Prostate cancer
    Gender
      Female
      Male
      Other
      Unknown
  ImagingModule
```

Equivalent To

SubClass Of
CommonModule
ImagingModule

General class axioms

SubClass Of (Anonymous Ancestor)

Instances

Target for Key

Disjoint With

Disjoint Union Of

Annotations

rdfs:label [language: en]
Body part

skos:prefLabel [language: en]
Body part structure

skos:altLabel [language: en]
Anatomic part

DICOM_Name [type: xsd:string]
BodyPartExamined

DICOM_Tag [type: xsd:string]
(0018,0015)

ICDO3_Code [type: xsd:string]
C80.9

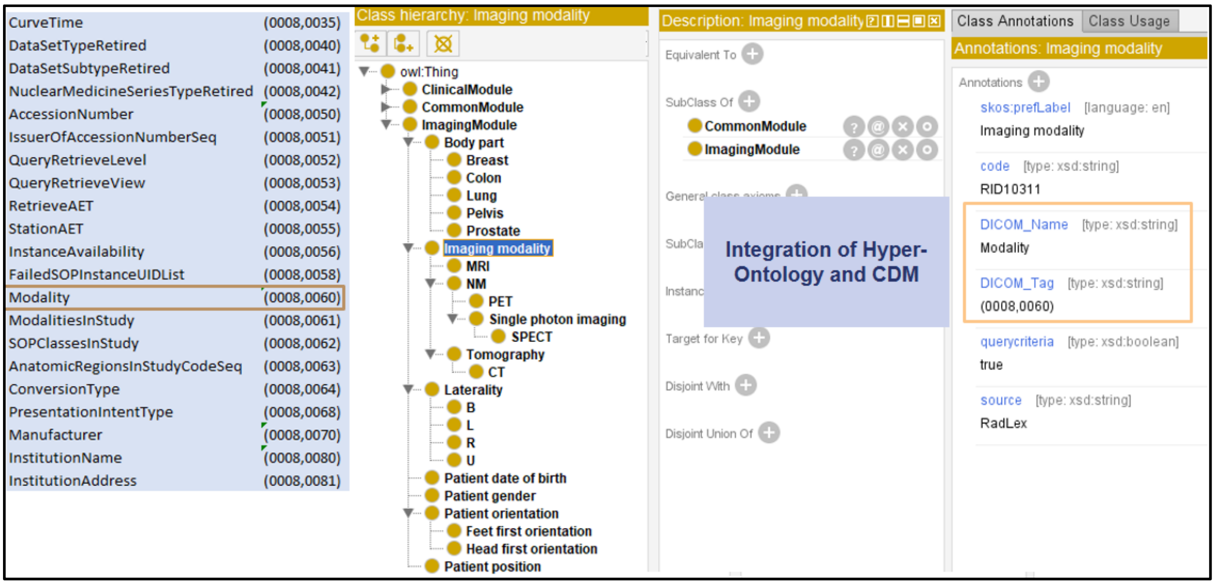
querycriteria [type: xsd:boolean]
true

source [type: xsd:string]
SNOMED-CT

synonym [language: en]
Body part

Figure 22: Excerpt of the hyper-ontology around the concept Body part

These types is a subset of the body parts considered in the different AI4HI projects. Additional types, such as Abdomen, Hip, Head, etc. will be considered to maintain the integration.


Class hierarchy: Imaging modality
Description: Imaging modality
Class Annotations
Class Usage
Annotations: Imaging modality
Integration of Hyper-Ontology and CDM

CurveTime	(0008,0035)
DataSetTypeRetired	(0008,0040)
DataSetSubtypeRetired	(0008,0041)
NuclearMedicineSeriesTypeRetired	(0008,0042)
AccessionNumber	(0008,0050)
IssuerOfAccessionNumberSeq	(0008,0051)
QueryRetrieveLevel	(0008,0052)
QueryRetrieveView	(0008,0053)
RetrieveAET	(0008,0054)
StationAET	(0008,0055)
InstanceAvailability	(0008,0056)
FailedSOPInstanceUIDList	(0008,0058)
Modality	(0008,0060)
ModalitiesInStudy	(0008,0061)
SOPClassesInStudy	(0008,0062)
AnatomicRegionsInStudyCodeSeq	(0008,0063)
ConversionType	(0008,0064)
PresentationIntentType	(0008,0068)
Manufacturer	(0008,0070)
InstitutionName	(0008,0080)
InstitutionAddress	(0008,0081)

```
owl:Thing
  ClinicalModule
  CommonModule
  ImagingModule
    Body part
      Breast
      Colon
      Lung
      Pelvis
      Prostate
    Imaging modality
      MRI
      NM
        PET
        Single photon imaging
          SPECT
      Tomography
        CT
    Laterality
      B
      L
      R
      U
    Patient date of birth
    Patient gender
    Patient orientation
      Feet first orientation
      Head first orientation
    Patient position
```

Equivalent To

SubClass Of
CommonModule
ImagingModule

General class axioms

SubCla

Instanc

Target for Key

Disjoint With

Disjoint Union Of

Annotations

skos:prefLabel [language: en]
Imaging modality

code [type: xsd:string]
RID10311

DICOM_Name [type: xsd:string]
Modality

DICOM_Tag [type: xsd:string]
(0008,0060)

querycriteria [type: xsd:boolean]
true

source [type: xsd:string]
RadLex

Figure 23: Example of mapping a standard imaging concept (Imaging modality) to DICOM attribute (Modality)

For the imaging level, the local nodes commonly define the imaging data using the DICOM standard, which is not a standardized vocabulary. Meanwhile, ongoing research contributions are in progress aiming to align the imaging data to OMOP CDM by extending the OMOP tables. Although, pertinent results are not yet achieved on the standardization level of the imaging data. In this regard, the mapping between the local nodes and the hyper-ontology is performed using the semantic annotations defined in the ontology model (e.g., DICOM_name, DICOM_tag). These annotations provide the mapping of standard imaging concepts (e.g., Imaging modality, MRI, etc.) defined in standardized terminologies (e.g., RadLex) to DICOM attributes (e.g., Modality) and attribute values (e.g., MR, CT, etc.) in the local nodes, permitting the integration on the imaging metadata and data levels.

Figure 23 depicts an example of a metadata mapping of a RadLex's concept (Imaging modality (RID10311)) defined in the hyper-ontology to a DICOM attribute (Modality (0008,0060)) in local nodes.

3.5 Public Catalogue - Metadata Model

In the context of the EUCAIM project, the necessity of a comprehensive metadata catalogue and a tailored metadata model is of paramount importance. This significance is also underscored in the proposal for a regulation on the European Health Data Space (EHDS) (Article 55), where a metadata catalogue "shall inform the data users about the available datasets and their characteristics. Each dataset shall include information concerning the source, the scope, the main characteristics, nature of electronic health data and conditions for making electronic health data available".

EUCAIM aims to advance cancer research and patient care by facilitating the sharing, integration, and analysis of medical imaging data across European institutions through the EUCAIM metadata catalogue.

3.5.1 Background

While established models like DCAT (Data Catalog Vocabulary)[49] and its Application Profile, DCAT-AP[50], serve as solid foundations for describing datasets and data catalogues, the unique complexities of cancer-related data necessitate a more specialized approach. In this section, we explore the development of a first version of a customized metadata catalogue and model for the EUCAIM project, addressing the limitations of existing frameworks and capturing essential attributes that drive seamless data sharing, interoperability, and valuable insights.

DCAT and DCAT-AP offer standardized approaches for describing datasets and catalogues, emphasizing key metadata elements. They are both part of the Semantic Web framework and are used to facilitate the discovery and sharing of datasets across different platforms. However, they serve slightly different purposes and have different levels of specificity.

DCAT is a general specification developed by the W3C (World Wide Web Consortium) for describing datasets and data catalogues on the web. It provides a basic framework for describing metadata about datasets, data distributions (different versions or formats of the same dataset), and data catalogues. DCAT focuses on the core metadata elements required to describe datasets and their relationships within a catalogue.

DCAT-AP is a more specific and specialized version of DCAT. It is designed to meet the requirements of European public administrations for sharing and publishing data on national

[49] https://www.w3.org/TR/vocab-dcat-3

[50] https://joinup.ec.europa.eu/collection/semic-support-centre/solution/dcat-application-profile-data-portals-europe/release/300

and European data portals. DCAT-AP builds on top of DCAT but adds additional constraints and requirements specific to the needs of public administrations, ensuring interoperability and consistency across different data portals within the European Union. DCAT-AP defines a set of mandatory and optional metadata elements, controlled vocabularies, and guidelines for data catalogue descriptions.

However, cancer imaging data introduces intricacies that surpass the general scope of these models. The EUCAIM project requires granularity in describing cancer-specific attributes, including tumor location, histological type, treatment modalities, and patient demographics. Additionally, ethical considerations, data anonymization techniques, and compliance with health data privacy regulations become crucial elements. To comprehensively address these requirements, a tailored metadata model that extends beyond the conventional DCAT-AP framework is required.

To accomplish this, our initial step was to outline the essential information elements that data providers must provide to comprehensively describe their datasets and associated characteristics. This requirement is also specified in Article 58 of the EHDS proposal and should pertain to both clinical and imaging metadata.

3.5.2 Methodology

Our approach towards establishing the minimum information that should accompany the medical images and describing the datasets to be registered in the EUCAIM catalogue was multifaceted:

- We initially adopted a bottom-up approach, gathering the obligatory information mandated by the AI4HI projects for various cancer types considered within these projects. This procedure began by collecting the mandatory information per cancer type, starting with prostate and breast cancer – common denominators across three respective projects. By collating this information, we concluded at a unified set of minimum data elements demanded by the projects for these two cancer types. Moving forward, our approach will extend to encompass additional cancer variants.

- Additionally, we explored the initiatives undertaken by the European Network of Cancer Registries (ENCR), with a focused examination of the ENCR Recommendations document outlining the Standard Dataset specifications[51] as well as the proposal addressing cancer data quality checks[52].

- Concurrently, we sought to leverage and build upon the work of other European initiatives, such as the BBMRI-ERIC biobank metadata catalogue[53], the EIBIR public metadata catalog[54] adopting the DICOM-MIABIS model[55], as well as the IHE Radiology

[51] https://encr.eu/sites/default/files/Recommendations/ENCR-Recommendation-standard-dataset_Mar2023.pdf

[52] https://encr.eu/sites/default/files/A_proposal_on_cancer_data_quality_checks-one_common_procedure_for_European_cancer_registries_1_1.pdf

[53] https://www.bbmri.nl/services/samples-images-data/catalogue

[54] https://molgenis.eibir-edc.org/menu/main/app-molgenis-app-biobank-explorer

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

White Paper that pertains to AI Interoperability in Imaging and includes a set of required data elements useful for AI model development[56].

3.5.2.1 AI4HI mandatory clinical data

Starting with the bottom-up strategy, the mandatory clinical elements for the prostate and breast cancer related datasets are shown in Table 3 and Table 4 respectively.

Table 3: Mandatory clinical information for prostate cancer related projects - with dark blue color are denoted the common attributes among all three projects, and with light blue color the common ones between two projects.

ProCAncer-I	CHAIMELEON	INCISIVE
Patient Identification Number	Patient Identification Number	Patient Identification Number
Sex	Sex	Sex
Age(Year Of Birth)	Age(Year Of Birth)	Age
Histological Type	Histological Type	Histological Type
Gleason1	Gleason1	
Gleason2	Gleason2	
PI-RADS		PI-RADS
Tumor Location (More Granular)		Tumor Location (side)
Index Lesion		Tumor Grade
Biopsy or Prostatectomy performed		Current State
MRI positive		Prostate Volume

Table 4: Mandatory clinical information for breast cancer related projects.

EUCANIMAGE	INCISIVE	CHAIMELEON
Patient Identification Number	Patient Identification Number	Patient Identification Number
Sex	Sex	Sex
Age	Age	Age (Year Of Birth)
Histological Type	Histological Type	Histological Type
biopsy cT	cT	
biopsy cN	cN	
Tumor Grade	Tumor Grade	
	cM	
Menopausal Status	Current State	

[56] https://www.ihe.net/uploadedFiles/Documents/Radiology/IHE_RAD_White_Paper_AI_Interoperability_in_Imaging.pdf

ER Percent Positive	BIRADS	
PR Percent Positive	Max Tumor Diameter	
HER2 IHC Status	Breast Location & Laterality	
Ki67 Percent Positive	Pathological Lymph Nodes, Other Breast Lesions	
Breast Cancer Molecular Subtype		

It's evident that across all projects involving prostate and breast cancer types, there exists a shared set of information, irrespective of the specific clinical questions or use cases that each project addresses. This shared information encompasses essential demographic data, including the sex and the age at which the patient received a diagnosis of a malignant tumor. Additionally, crucial tumor-related details, such as morphology and topography, provide the basis for deducing the diagnosis itself, as a diagnosis is formed by the combination of the morphology and topography of tumors.

3.5.2.2 ENCR Recommendations: Standard Dataset Specifications and Cancer Data Quality Checks

The European Network of Cancer Registries (ENCR)[57] offers invaluable guidance for determining the essential information related to cancer clinical information. Through its Recommendations document, which outlines standardized dataset specifications, it facilitates setting up the ground for consistent and harmonized data collection across various cancer registries. These recommendations ensure that critical clinical information, such as patient identification, sex, age, histological type, tumor location, and tumor grade, are uniformly captured by all cancer registries. Furthermore, the ENCR's proposal on cancer data quality checks underscores the significance of maintaining accurate, complete, and reliable data within cancer registries.

Incorporating the ENCR Recommendations into the EUCAIM project's metadata model not only ensures alignment with established best practices but also enhances the credibility and usefulness of the project's efforts within the broader landscape of cancer research and patient care.

Therefore, based on the information extracted from the AI4HI projects as well as the ENCR recommendations, we concluded to the following minimum clinical information that should accompany the cancer images in order to join the EUCAIM federation and for comprehensively describe the clinical information of their datasets.

Table 5: Minimum clinical information for joining the EUCAIM federation

Variable description	Format	Missing/Unknown Values	Allowed Values
Patient Identification Number	String	Not Allowed	Not allowed to have duplicates in the same dataset.
Sex	String	“Unknown” concept	Gender Vocabulary

[57] https://www.encr.eu/

			Male, Female, Other, Unknown
Date of birth (year)	Integer (YYYY)	Estimate if not known/Empty if age at diagnosis is available	>1873 and <=current year
Date of tumor incidence (year)	Integer (YYYY)	Not Allowed/Empty if age at diagnosis is available	>1873 and <=current year
Age at diagnosis in years*	Integer	Not allowed if date of birth and date of tumor incident are missing	>=0 and <150
Morphology (Histological type)	String	Not allowed	ICD-O3, SNOMED-CT
Topography (Anatomic Location)	String	Not allowed	ICD-O3, SNOMED-CT
Diagnosis (Histology+Anatomic Location)	String	Not allowed	ICD-O3, SNOMED-CT

* If date of birth (year) and date of tumor incidence (year) are missing or unknown, the age at diagnosis should be present.

Note, that in case of "image-only" datasets only morphology and topography information is required.

3.5.2.3 AI4HI mandatory imaging metadata

A similar approach was employed to evaluate the essential imaging metadata that should accompany the cancer images. Initially, we analyzed the imaging metadata requirements of the AI4HI projects for defining the cohorts specific to the distinct use cases each project addresses. A list of 73 imaging attributes (DICOM tags) kept in every AI4HI project (with the exception of EuCanImage that do not have the imaging metadata available) have been identified, and are described in Annex C: List of imaging attributes kept for all AI4HI projects . From there, a set of variables have been selected as the minimum set of imaging metadata required for describing the images to be part of EUCAIM based on the metadata used for cohort discovery within the AI4HI projects. The result is summarized in Table 6: Minimum imaging metadata for joining the EUCAIM federation.

Table 6: Minimum imaging metadata for joining the EUCAIM federation

Variable description	Corresponding attributes' name and tag code	Format	Missing/Unknown Values	Allowed Values
Patient Identification Number	Patient ID (0010,0020)	String	Not Allowed	
Image Modality	Modality (0008,0060)	String	Not Allowed	DICOM, Radlex
Image Body Part	BodyPartExamined (0018,0015)	String	Not Allowed	DICOM, SNOMED-CT, ICD-O3

Image Vendor	Manufacturer (0008,0070)	String	Not Allowed	Medical device manufacturers
Date of image creation (year)		Integer (YYYY)	Empty if not known	<=current year
Number of dates from the date of tumor incidence		Integer	Not allowed if only "Age at diagnosis" is present.	Integer

This list is prone to change along with the project progression. Indeed, other variables may appear of great importance and be added in this minimum set, depending on the modality of the images (e.g. for MR images "SliceThickness" is one essential metadata attribute that has proved to be important for AI model development etc.)

3.5.2.4 EUCAIM Dataset Metadata Elements

Therefore, based on the mandatory clinical and imaging information, as well as the mandatory information as this is specified by the DCAT and DCAT-AP models for describing datasets, we concluded to the following set of metadata elements required to describe the datasets to be registered into the EUCAIM metadata catalogue and comply to tier 1 of the Data Federation Framework. Note that this corresponds to the minimal information required in order to join the EUCAIM federation, and if the users would like to assess further the dataset compliance to their needs, more fine-grained queries will be possible, based on the hyper-ontology concepts, such as TNM staging, tumor grade etc. (tier 2 or 3). Furthermore, this minimal information depends on whether the dataset to be registered in the EUCAIM platform is an "image-only" dataset. In this case, a subset of the metadata is obligatory as described below in Table 7.

Table 7: EUCAIM Public Metadata Catalogue - Dataset Attributes

Attribute Name	Type	Required	Cardinality	Terminologies/Description
Dataset Identifier	String	YES	1	A unique identifier for the dataset.
Dataset Name	String	YES	1	A clear and concise name for the dataset.
Dataset Description	String	YES	1	A detailed description of the dataset's content, purpose, and scope.
Dataset Type	Categorical	YES	1-many	The categorization of the dataset. Possible values include: Original Dataset, Annotated Dataset, Processed Dataset
Dataset Access	Categorical	YES	1	The accessibility level of the dataset, indicating how users can obtain and interact with the data. The following values clarify the access methods available: ● **By request**: Access to dataset that requires users to submit a formal request, specifying the purpose or justification for accessing the dataset, within the framework of a research project. ● **Restricted Access**: Access limited to authorized individuals or organizations based on specific permissions or roles.

				● **Commercial Access**: Access to datasets available through a paid subscription model (Access granted through commercial licenses, where users pay a fee to access and use the data).
Dataset Collection Method	Categorical	YES	1-many	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. Possible values: ● **Patient-based**: Data records are organized individually based on patients. Each data entry corresponds to a single patient's information, not necessarily specific to a clinical use case. ● **Cohort**: Data records are grouped according to specific medical studies or research projects. This grouping includes all relevant data elements such as imaging scans, clinical assessments, lab results, etc., related to a particular study or clinical use case. ● **Only-Image**: Data elements in this category exclusively consist of imaging data and associated metadata. Clinical information is not included; only metadata present in the DICOM headers is provided. ● **Longitudinal**: Data elements are structured to cover multiple time points for either a particular patient or study. This structure enables the analysis of changes over time, making it suitable for longitudinal studies. ● **Case-control**: Data records are divided into two distinct groups: cases and controls. Cases encompass subjects with the disease or condition under study, while controls include subjects who do not have the disease or condition. ● **Disease-specific**: Data records are gathered from subjects who have already developed a particular disease. This category is particularly focused on subjects with the specified condition.
Dataset Terms of Use	Categorical	YES	1-many	The terms and conditions that govern dataset usage. Possible values are based on the Data Use Ontology.[58]
Dataset Intended Purpose	String	YES	1	The primary objective for which the dataset was created.

[58] https://www.ebi.ac.uk/ols4/ontologies/duo?tab=classes

Dataset Contact Point	Vcard	YES	1	Contact information using VCard format for dataset-related inquiries.
Dataset Metadata Issued	Datetime	YES	1	The date when the dataset's metadata was generated.
Dataset Last Modified	Datetime	YES	1	Most recent date on which the dataset was changed, updated or modified.
Dataset Version	String	YES	1	The version number or identifier for the dataset.
Dataset Provider	String	YES	1	The entity responsible for providing the dataset.
Age Low	Integer	YES*	1	The minimum age of subjects in the dataset.
Age High	Integer	YES*	1	The maximum age of subjects in the dataset.
Age Median	Integer	YES*	1	The median age of subjects in the dataset.
Sex	Categorical	YES*	1-many	Sex distribution of subjects in the dataset by using the OMOP Gender Vocabulary
Topography	Categorical	YES	1-many	Anatomical sites specified using ICD-O3, SNOMED-CT.
Diagnosis	Categorical	YES	1	Diagnostic information using ICD-O3, SNOMED CT.
Image Modality	Categorical	YES	1-many	The imaging modality used (e.g., DICOM, Radlex).
Image Body Part	Categorical	YES	1-many	Anatomical areas captured in the images using DICOM.
Image Vendor	String	YES	1-many	Manufacturer of the imaging device (DICOM tag (0008,0070)).
Image Creation Year(s)	Integer	NO	1-many	A year range that the actual (DICOM) images were created/acquired (if this has not been changed in the anonymization process). If this is not available, an estimation should be added.
Number of Subjects	Integer	YES	1	Total count of unique individuals in the dataset.
Number of Studies	Integer	YES	1	Total count of DICOM studies.
Number of Series	Integer	YES	1	Total count of DICOM series within the dataset.
Image size (GB)	Integer	NO	0-1	The size of the dataset in gigabytes.

* Not mandatory in case the dataset is an “Only-image” dataset.

4. Protocols, Formats, and Terminologies for Clinical/Imaging Data

4.1 Protocols & Formats for Clinical and Imaging Data

In the context of EUCAIM, various protocols, formats and terminologies were explored and evaluated. In the following sections, we will provide descriptions of each of these elements, some of which were already introduced in section 3.

HL7 Reference Information Model (RIM)

HL7 version 3 standards are built on messaging that uses the Reference Information Model (RIM)[59]. The HL7 RIM specifies the grammar of HL7 v3 messages. HL7 RIM is an object-oriented data model based on UML (Unified Modeling Language) specialized by HL7, which became an ISO standard in 2006. HL7 RIM structures information and maps a wide range of clinical concepts intended to cover the entire healthcare domain. For example, HL7 RIM can be used to document the actions taken to treat a patient. This model is used along with a library of data types and with a set of vocabularies. Some of the properties of HL7 v3 models are coded and combined with a set of possible values. These values, used for the instantiation of messages or HL7 v3 documents, are coded concepts defined with reference to a given code system in a specific context. In practice, a coded concept is unique in a specific context, and when possible, it reuses the concepts from a reference terminology (e.g. SNOMED-CT, LOINC) instead of creating an instance in the HL7 vocabulary.

HL7 Clinical Document Architecture (CDA)

One of many models provided by HL7 v3 RIM, the CDA (Clinical Document Architecture) [60]model is a standard that enables structuring the clinical information and its semantics within an exchanged clinical document. It is made up of a header and a body structure. The header provides data that are relatively neutral from a medical point of view, making it possible to link documents with the context of medical care, to classify it in appropriate categories and facilitate long term accessibility. Headers are always structured. On the other hand, the body is structured by sections that contain a narrative block followed by entries, which can be coded using standard vocabularies

HL7 Fast Healthcare Interoperability Resources (FHIR)

Fast Healthcare Interoperability Resources (FHIR) 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 health.

Several HL7 FHIR initiatives focus on enabling the secondary use of EHR data for research and public health through implementation guides such as the HL7 Vulcan **Retrieval of real-world data for clinical research IG.**

The **HL7 Vulcan initiative** aims to facilitate the integration of care and research activities to improve patient lives, reduce costs and improve efficiency with HL7 FHIR interoperability standards. This FHIR Acceleration Program develops FHIR resources needed to execute prioritized use cases of secondary use of « real-world » data and especially EHR data. The main goal of the HL7 Vulcan **FHIR implementation guide (IG) for Retrieval of real-world data for clinical research** is to help define a minimal set of clinical research FHIR resources

[59] http://www.hl7.org/implement/standards/rim.cfm
[60] K. W. Boone, The CDA, TM book, 2011.

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 that are needed for conveying 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 that are needed to achieve different outcomes (e.g. FHIR to CDISC, FHIR to OMOP, etc.).

The HL7 **Vulcan Real World Data IG** project identified two phases to requesting data from an EHR: i) Determine patients based on inclusion and exclusion criteria and ii) Retrieve healthcare data for the specific patients of the targeted population.

- Step 1: Cohort Building[61]:This first phase is building a cohort by querying for patients based on a set of inclusion and exclusion criteria. A set of patient identifiers is retrieved and becomes input into the next phase of requesting data. Through the analysis of first use cases of cohort building, the HL7 Vulcan initiative identified a set of data elements that were needed to identify patients that meet a set of inclusion and exclusion criteria for cohort feasibility determination
- Step 2: Retrieve Healthcare Data: After specific patients have been determined, the next phase is to access and/or retrieve their healthcare data. This guide inherits from the International Patient Access (IPA) and uses the IPA as the means to transfer healthcare data. This guide lists specific search parameters needed to find healthcare data for a specific patient and it lists the specific resources and data elements that have been deemed important to return for the purposes of research.

Complementary initiatives provide additional specifications in particular domains of interest e.g. **mCODE** in the oncology domain. In addition, the CodeX initiative[62] develops and evaluates mCODE implementation in the context of care and also clinical research use cases in the oncology domain.

OpenEHR

OpenEHR[63] is a set of open specifications for an Electronic Health Record (EHR) architecture designed to enable semantic interoperability of health information between EHR systems – without proprietary formats and vendor lock-in of data. There is ongoing alignment and implementation with a number of key standards, such as with IHTSDO, the international DCM (Detailed Clinical Models) group, ISO 13606, HL7, and ASTM CCR.

The openEHR Foundation is a not-for-profit company whose founding shareholders are University College London, UK and Ocean Informatics pty, Australia. The OpenEHR Foundation aims to enable ICT to effectively support healthcare and medical research with a knowledge-oriented computing framework that includes ontologies, terminology and a semantically enabled health computing platform. According to the openEHR Foundation, the principal challenge for health ICT is to represent the semantics of the sector. The openEHR Foundation's aims of 'future-proof and flexible' are exemplified by the stated goal of supporting the economically viable construction of maintainable and adaptable health computing systems and patient-centric EHRs.

[61] http://hl7.org/fhir/uv/vulcan-rwd/patients.html
[62] https://confluence.hl7.org/display/COD
[63] http://www.openehr.org/home.html

The openEHR information architecture consists of four levels of information organisation:a) User Interface; b) Templates - data capture sets for each business process event, c) Archetype Model - the standardised semantics of the data points in data capture sets, defined on the basis of topic. These are core application requirements formalized in an Archetype Definition Language (ADL); d) Reference Model - standardised data representation, enabling interoperability. The information architecture enables a standardised querying capability based on archetype paths and terminology as well as a standardised interface to terminology for inferencing. A computational view of the architecture includes a Service Model, where major services are defined, largely derived from existing work in OMG Corbamed, CEN HISA.

CDISC

The CDISC[64] was initiated in 1997 by the FDA (Food and Drug Administration) in order to develop standards for the acquisition, exchange, submission and storage of clinical data in clinical research. The primary goal of CDISC is to develop rules in order to submit standardized records to regulatory authorities. The other CDISC objectives are acquisition, exchange and storage of data. The records follow CDISC reporting rules and protocols, simplifying their interpretation and auditing by regulatory agencies, as well as decreasing the burden on trial physicians and speeding the entire clinical development cycle. This homogenisation also improves internal training and simplifies the set-up of new tests, both accelerating the adoption of systems of data capture and improving data exchange with partners and long term storage of clinical data. To ensure full neutrality with respect to the market economy in the world of research, the development of these standards is independent of computer platforms and solution vendors. CDISC has identified eXtensible Markup Language (XML) as the cornerstone of its plan because this language has a good reputation in industry.

Integrating the Health Enteprise

Integrating the Health Enteprise[65] (IHE) promotes the coordinated use of standards such as DICOM and HL7 to address specific clinical needs in support of optimal patient care. The objective of IHE is not to define new standards but to support the use of existing ones. The IHE process is defined in ISO28380, and according to that one starts with an existing interoperability problem e.g. scheduled workflow in radiology. It defines a technical framework for the implementation of established messaging standards to achieve specific clinical goals. By defining the desired behavior of systems and the content of transactions, IHE diminishes a large part of the ambiguity of single standards. These definitions are made available to the general public in so-called technical frameworks. In each integration profile, actors and transactions are defined. Then, it develops using existing standards a domain specific profile; IHE profiles address critical interoperability issues related to information access for care providers & patients, clinical workflow, security, administration, information infrastructure.

DICOM

The Digital Imaging and Communications in Medicine (DICOM)[66] standard is a crucial framework for the seamless exchange and management of medical image data and associated information across different healthcare systems and devices. At its core, DICOM standardizes

[64] http://www.cdisc.org/
[65] http://www.ihe.net
[66] https://dicom.innolitics.com/ciods

the structure and format of various elements within medical imaging, such as X-ray, MRI, CT scans, ultrasounds, and more. This standardization enables medical professionals to view, share, and analyze images reliably, regardless of the manufacturer, modality, or software used to acquire or process the images.

DICOM achieves this interoperability through a combination of defined data structures, encoding rules, and network communication protocols. Central to the standard is the concept of a DICOM Information Object, which stores pertinent information like patient demographics, acquisition parameters, and image pixel data. These attributes are structured in a consistent way, making the data comprehensible 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 time, DICOM has evolved by adding new modules and supplements to incorporate advancements in technology and changes in medical practices. These updates accommodate emerging imaging modalities, enhance security, and support the integration of medical imaging into broader electronic health record (EHR) systems.

DICOM SEG

DICOM-SEG[67] is a crucial aspect of the DICOM standard, designed 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 different imaging devices and healthcare institutions, facilitating multi-center studies. Its utility lies in its ability to improve collaboration, research, and patient care by standardizing how segmented regions are represented within DICOM images.

NIfTI

The Neuroimaging Informatics Technology Initiative (NIfTI) standard[68] 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 originally geared towards neuroimaging, NIfTI's adaptability extends to oncology research, particularly in scenarios involving brain tumors where it is leveraged for the storage of 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.

[67] https://dicom.innolitics.com/ciods/segmentation
[68] https://nifti.nimh.nih.gov/

4.2 Terminologies

The need for standardized terminologies is crucial, especially in medical informatics. It ensures communication and exchange of information between healthcare professionals, researchers and computer systems. Moreover, it is essential for data consistency, research integrity and improvement in healthcare quality and outcome.

Depending on the type of medical data, different terminologies have been chosen.

A specificity of our project is that it is based on existing, previous work, namely the different AI4HI projects. Our objective is to adapt to these projects, despite the disparity in the representations and vocabularies used.

For example, the vocabularies used for procedures were SNOMED CT, CPT4 or ICD10PCS, depending on the project. In order to facilitate interoperability between projects, the EUCAIM hyper-ontology needs to support different terminologies.

For the first version of the hyper-ontology (v0.4 alpha), we decided to focus on the OHDSI recommendation and the vocabulary used by the AI4HI projects (which uses the OMOP CDM or HL7/FHIR).

- **SNOMED-CT (Systematized Nomenclature of Medicine Clinical Terms)[69]** is an internationally recognized clinical terminology and coding system. It organizes medical concepts hierarchically, providing precise and standardized descriptions for healthcare information.
- **ICD10PCS (International Classification of Diseases, Tenth Revision, Procedure Coding System)** is a standardized medical coding system that classified procedures and surgeries.
- **ICD-O-3, the WHO international Classification of Diseases for Oncology, 3rd edition** is an international standard classification of tumor and related diseases. It provides specialized representation of cancer histology, topography, and behavior. (Belenkaya 2020[70])
- **LOINC (Logical Observation Identifiers Names and Codes) [71]** is a standardized system used for identifying and naming laboratory tests, observations and clinical measurements. Because LOINC is one of the standard vocabulary in OMOP, it was used, as Radlex, for the R-CDM (Park 2022[72]).
- **DICOM (Digital Imaging and Communications in Medicine) [73]** serves as a standard protocol within the medical field for transmitting, storing, and exchanging medical images and related information such as patient data, imaging parameters etc.[74]. Additionally, DICOM can be considered a terminology that standardizes and defines terms for describing medical images and associated data, such as the image type, modality, patient position, and orientation, among others.

[69] SNOMED Clinical Terms. Available from URL: https://www.snomed.org/

[70] Belenkaya, R. et al. Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research. JCO Clinical Cancer Informatics 12–20 (2021) doi:10.1200/CCI.20.00079.

[71] LOINC, the international standard for identifying health measurements, observations, and documents. Available from URL: https://loinc.org/.

[72] Park, C. et al. Development and Validation of the Radiology Common Data Model (R-CDM) for the International Standardization of Medical Imaging Data. Yonsei Med J 63, S74 (2022).

[73] https://dicom.innolitics.com/ciods

[74] https://dicom.nema.org/medical/dicom/current/output/chtml/part16/chapter_L.html#table_L-5)

- **RadLex** is a (glossary) standardized radiology lexicon and ontology used to describe and unifying medical imaging concepts. Its hierarchical organization and relationships between terms improves the accuracy and consistency of radiology information exchange. Recently, the Radlex playbook has been adopted by Park et al. (Park 2022) to standardize the terminology system of the radiology common data model (R-CDM) compatible with the HL7/FHIR and linked with the OMOP CDM (for the international standardization of medical imaging data.)
- **NAACCR, the North American Association of Central Cancer Registries,** is a standardized terminology for cancer data reporting in North America. The terminology covers a wide range of precise codes and definitions to categorize diverse elements of cancer cases, including details about tumor attributes like size, stage, and grade, patient demographic information such as age, gender, and race, as well as the methods of treatment, such as surgery, chemotherapy, and radiation therapy.
- **Cancer modifiers** [75,76] is a vocabulary developed based on the content of LOINC, NCIt, NAACCR, and CAP. It refers to concepts and definitions used within the field of cancer data to provide additional details or modifiers to describe specific characteristics of cancer cases. These modifiers can include factors like staging and grading, laterality (whether the cancer is on the left or right side of the body), and histology (the type of tissue or cell involved in the cancer). However, cancer modifiers did not include ontological relationships.
- **The "Episode" vocabulary in OMOP CDM** is used to represent periods of continuous healthcare activity related to a specific medical condition or event for a patient. An episode typically captures a sequence of related healthcare interactions and interventions related to a particular condition, such as a disease, an injury, or a medical procedure.
- **UCUM (Unified Code for Units of Measure)** terminology is a standardized system of codes and definitions used to represent units of measurement in various fields, including healthcare, science, and engineering.

Table 8: List of vocabularies utilized/supported by the hyper-ontology version 0.4 alpha

DOMAIN	TERMINOLOGY
Patient (date of birth, date at diagnosis, age)	RadLex ; DICOM; SNOMED CT;
Gender, sex	Gender
Body part	SNOMED CT; ICDO3; DICOM
Diagnosis	SNOMED CT
Clinical findings	SNOMED; ICD O 3
Family History	SNOMED CT
Procedure	SNOMED CT;

[75] Dymshyts, D. Ontology of Cancer Diagnosis in the OMOP Vocabulary.
[76] Belenkaya, R. et al. Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research. JCO Clinical Cancer Informatics 12–20 (2021) doi:10.1200/CCI.20.00079.

Measurement (Lab Result, Duration, Timepoint, Grade, Category)	SNOMED CT; NAACCR; LOINC; cancer modifier
Observation	SNOMED CT; LOINC
Episode	Episode;
Image modality	RadLex
Image laterality	RadLex
Unit	UCUM

5. Data Preprocessing and (Meta)Data Management & Interoperability Tools/Services

5.1 Overview of Data Preprocessing Tools/Services

The EUCAIM project is dedicated to offering a comprehensive suite of tools and services designed to streamline the data preprocessing process. This ensures that data is quality-checked, harmonized, made GDPR-compliant, and annotated. These crucial steps enable the data to be seamlessly shared within the EUCAIM repository, facilitating its subsequent utilization in various data-driven applications such as development, validation, and other related scenarios.

In the scope of the project, an extensive analysis has been performed to identify the pipelines and tools used in each of the AI4HI projects for the different preprocessing steps such as data quality and data cleaning, harmonization, annotation, de-identification and FAIRification, always taking into account the specific needs of medical imaging data.

Table 9, adapted from[77], summarizes the approaches of the five AI4HI projects, the EuCanImage, INCISIVE, ProCAncer-I, PRIMAGE and CHAIMELEON.

Table 9: Summary of the approaches of the AI4HI projects

	EuCanImage	INCISIVE	ProCAncer-I	PRIMAGE	CHAIMELEON
Architecture	Accommodating both decentralized and centralized storage	Hybrid (federated and centralized storage)	Centralized	Centralized and replicated	Centralized
Data models and types of data	DICOM-MIABIS FHIR (and terminologies supported by FHIR + extensions)	FHIR SNOMED-CT LOINC DICOM	OMOP CDM with extensions	DICOM-MIABIS OMOP CDM	DICOM-MIABIS OMOP CDM (terminology IDs mainly) Structure of the eCRF
Deidentification process	Pseudonymized data	Pseudonymized data	Fully anonymized data	Pseudonymized data	Pseudonymized initially for curation and then fully anonymized data at the central repository
Curation tools	Image anonymization/pseudonymization, quality control and annotation, non-imaging data anonymization and homogenization	Image de-identification tools, quality control, and annotation tools	Image quality control, anonymization, motion-correction, co-registration & annotation.	Image labelling, quality checking, annotation, denoising, motion correction, registration	Data completeness and consistency tools, image quality checking, Image anonymization, annotation,

[77] Kondylakis, H., Kalokyri, V., Sfakianakis, S. et al. Data infrastructures for AI in medical imaging: a report on the experiences of five EU projects. Eur Radiol Exp 7, 20 (2023). https://doi.org/10.1186/s41747-023-00336-x

					segmentation and harmonization

In this section, we provide descriptions of both the tools and a summary of the approaches. A comprehensive explanation of each approach adopted by the AI4HI projects, can be found in Annex B: AI4HI approaches on data pre-processing. Please note that additional tools may be incorporated into the final catalogue of tools as the project develops.

Before being part of the final catalogue of tools, each of the tools will go under approval performed by the EUCAIM Technical Committee, which will ensure their compliance to the quality and security requirements. Once the tool is finally approved to be part of the EUCAIM platform, it will be registered in ELIXIR bio.tools[78] under the EUCAIM identifier. A summary of the main characteristics of each of tools is shown in Table 10. The full description of the characteristics can be found in Annex D.

Table 10: Summarized catalogue of tools.

Name	Partner - project	CPU/GPU	Modality type	Current status
Breast dense tissue segmentation - ITI BREAST Calculate	ITI	CPU	Imaging - FFDM	Validated
MR-based neuroblastoma tumour detection and segmentation	HULAFE - PRIMAGE	Both	Imaging - T2w or T2w fat sat MRI	Containerized
MR-based DIPG tumour detection and segmentation	HULAFE - PRIMAGE	Both	Imaging - T1w and/or T2w/FLAIRI	Containerized
MR-based glioblastoma tumour detection and segmentation	HULAFE	Both	Imaging - T1Wce + T2W + FLAIR	Containerized
CT-based neuroblastoma tumour detection and segmentation	HULAFE - PRIMAGE	Both	CT CE Sequence	Developed
nnUnet	DKFZ	Both	Imaging	Developed
nnDetect	DKFZ	Both	Imaging	Developed
MITK	DKFZ	CPU	Imaging	Containerized
Multi-regional prostate segmentation	Quibim	CPU	Imaging - T2w MRI	Validated
Data harmonization				
MRI_pixel_intensity_harmonization	Quibim - CHAIMELEON	Both	MRI (T2W, T2Flair)	Developed

[78] https://bio.tools/

CT_pixel_intensity_harmonization	Quibim - CHAIMELEON	Both	CT	Developed
Trace4Harmonization	DeepTrace Technologies	CPU	Numeric variables (e.g., radiomic features)	Developed
Biologically motivated normalization techniques	FORTH- ProCAncer-I	CPU	MRI T2W (only prostate images)	Containerized
Feature based Harmonization (ComBat)	FORTH - ProCAncer-I	CPU	Numeric variables (e.g., radiomic features) in pkl format and MRI data in dicom format	Containerized
Neuroblastoma_T1W/T2W/DCE_harmonization	HULAFE - PRIMAGE	CPU	MRI T2W,T1W, DCE (abdomen only)	Containerized
Neuroblastoma_DWI_harmonization	HULAFE - PRIMAGE	CPU	MRI DWI (abdomen only)	Containerized
DIPG_T1W_harmonization	HULAFE - PRIMAGE	CPU	MRI T1W (brain only)	Containerized
DIPG_T2W/DCE_harmonization	HULAFE - PRIMAGE	CPU	MRI T2W and DCE (brain only)	Containerized
DIPG_DWI_harmonization	HULAFE - PRIMAGE	CPU	MRI DWIW (brain only)	Containerized

In the following subsections, we will describe the approaches to be followed by EUCAIM for each of the data preprocessing categories.

5.2 Data quality and data cleaning

Real-world medical imaging data are not always of highest quality, as they may contain incomplete information, or have poor resolution, noise or artefacts when it comes to imaging exams. In addition, given the diversity of the information that may appear in the data coming from different and heterogeneous sources, tools dedication to quality control and improvement is warranted to ensure data quality, overcome dissimilar formats, availability with respect to adequate and uniform sampling, and to assess whether the data falls within expected value ranges. This applies both at the dataset level as well as the unitary data level. To address these challenges, the EUCAIM consortium has identified a range of tools developed by its partners, several of which are used in the AI4HI projects. These tools are dedicated to 1) assessing the quality level of the data, and/or 2) improving the quality of the data, focusing on aspects such as completeness and coherence of clinical data, noise reduction, artefact removal, and bias correction on imaging data.

To identify the EUCAIM approach related to data quality and cleaning, we initially examined the AI4HI approaches in the domain. Specifically, in the AI4HI projects, robust data quality assessment and cleaning methodologies have been implemented to ensure reliable and reproducible medical data analysis.

- **PRIMAGE** employs comprehensive checks for data accuracy, completeness, consistency, and integrity, along with specialized image preprocessing and quality assessment tools.
- **ProCAncer-I** focuses on enhancing MR data quality through preprocessing tools, bias field correction, motion correction, image enhancement, and noise reduction techniques.
- **CHAIMELEON** emphasizes quality control for both imaging and clinical data, with on-site image quality assessments and manual clinical data quality checks.
- **EuCanImage** employs radiologist and automated image quality assessment phases, using a reference-free metric (PIQUE[79]) for image quality evaluation.
- **INCISIVE** follows a structured approach, including data quality metrics, rule definition, and data quality assessment, to ensure data integrity and quality throughout the project.

More detailed information about each AI4HI approach on the data quality and cleaning aspects is included in Sub-Annex B.1: Data quality and data cleaning.

Based on the tools developed in the context of the AI4HI projects, two approaches will be considered in **EUCAIM**:

1. Data quality assessment - this approach should provide as output a global index of quality.
2. Data cleaning processing - dedicated tools will create a new version of the data or dataset (a “cleaned” copy) that will first be evaluated by a dedicated internal committee, without being accessible to the EUCAIM community.

Currently, five tools have been identified for quality assessment, and six for data cleaning in the catalogue of tools. All tools deemed relevant and compatible with the EUCAIM initiative will be made accessible, reusable and interoperable within the EUCAIM platform, in order to be integrated to a federated data analysis pipeline. These tools will be applied either on the on-boarding of the data provider into the EUCAIM federation, in order to “certify” the quality of the data and its compliance to the federation rules, or during the pre-processing stages of an AI model pipeline. Below is a description of each tool made available for EUCAIM by consortium members (see Annex D: Catalogue of tools for more details for each of the tools).

5.2.1 Data quality assessment tools

Since EUCAIM recognizes the importance of assessing data quality at multiple levels, we present below the tools, categorized into three groups: those focusing on data quality at the individual data level, those targeting quality at the dataset level, and those addressing data quality at both individual and dataset levels.

Dataset level

- **Extended A Priori Probability (EAPP).** ITI has worked in dataset bias discovery to develop machine-learning solutions. As a result of this work, they have defined a semi-

[79] Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, S. S. Channappayya and S. S. Medasani, "Blind image quality evaluation using perception based features," 2015 Twenty First National Conference on Communications (NCC), Mumbai, India, 2015, pp. 1-6, doi: 10.1109/NCC.2015.7084843.

supervised baseline metric: the Extended A Priori Probability (EAPP[80]). This metric is intended for binary classification tasks that consider not only the a priori probability but also some possible bias present in the dataset and other features that could provide a relatively trivial separability of the target classes. The procedure involves (1) multiobjective feature extraction and (2) a clustering stage in the input space with autoencoders, and (3) a subsequent combinatory weighted assignment from clusters to classes depending on the distance to the nearest cluster for each class.

- **Data Integration Quality Check Tool (DIQCT).** AUTH has been developing a data integration quality check tool[81] to use in the INCISIVE project, which aims to identify whether data follows the corresponding data harmonisation requirements. It is a rule-based, extensible and dockerized tool that checks various parameters: the clinical metadata integrity and validity, the dataset completeness, the integrity between images and clinical metadata provided, the de-identification protocol applied, the DICOM images validity, the imaging analysis requirements and the existence of annotation. The tool produces reports that inform the user on corrective actions prior to data upload. The tool can be used on the user-side including also a user interface and on server side running in cli applied in various datasets.

Individual data level

- **Image Qure.** In the context of the PRIMAGE project, Medexprim worked on developing a tool called Image Qure, whose main goal is to automatically analyze medical images and provide detailed quality metrics plus a global probability for this image to be classified as bad quality. The first version of this tool takes MR T2WI as input and returns as output 3 metrics regarding signal and noise (Signal to noise ratio, Signal variance and Contrast to noise ratio), 3 metrics that aim to detect most frequent MR artefacts (Foreground-background energy ratio, Entropy focus criterion, Coefficient of Joint variation) and a bad quality global probability given the combined analysis of these metrics. Next versions of this tool aims to include other MR sequences.
- **DICOM File Integrity Checker.** The DICOM File Integrity Checker from HULAFE is a specialised tool to ensure the quality and integrity of DICOM files. Its primary function is to conduct thorough quality checks on DICOM datasets, with a specific focus on aspects such as the correct number of files, the detection of corrupted files, and the identification of missing files. This tool plays a crucial role in maintaining the reliability and accuracy of medical imaging data, a paramount requirement in the field of healthcare. The tool operates by analyzing DICOM files located in a designated input path. Whether arranged hierarchically following the “Patient/Study/Series/Image DICOM hierarchy” or not, the output is consistently organized according to the cited hierarchy. It scans and evaluates each DICOM file, verifying its integrity and completeness. The results of this analysis are then compiled into a comprehensive report. This report not only highlights any corrupted or missing files but also identifies noteworthy series within the dataset. The DICOM File Integrity Checker is powered by Python and Docker, enabling efficient processing of DICOM files on the CPU. With a processing time of under a second per image, the tool ensures timely results. It is

[80] V. O. castelló, F. J. Pérez-benito, O. D. T. Catalá, I. S. Igual, R. Llobet and J. -C. Perez-Cortes, "Extended a Priori Probability (EAPP): A Data-Driven Approach for Machine Learning Binary Classification Tasks," in IEEE Access, vol. 10, pp. 120074-120085, 2022, doi: 10.1109/ACCESS.2022.3221936.

[81] Kosvyra, A., Filos, D., Fotopoulos, D., Tsave, O. and Chouvarda, I., 2022, July. Data Quality Check in Cancer Imaging Research: Deploying and Evaluating the DIQCT Tool. In 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 1053-1057). IEEE.

versatile, accommodating various DICOM file modalities. Dockerization simplifies deployment, and the tool can access data via specified filesystem paths. Security measures are in place to protect data integrity, and the tool operates independently, free from external internet dependencies.

- **Time coherence tool.** HULAFE also developed a time coherence tool in the scope of the PRIMAGE project. It consists of a Python program designed to validate the chronological order and logical consistency of dates associated with a patient's medical history. It takes in data like diagnosis dates, doctor visit dates, image scan dates, and treatment dates for a patient. The program uses a set of predefined rules to check if the sequence of dates makes sense. For example, one rule may be that a treatment date cannot occur before the diagnosis date. If any rule is violated, the program will highlight that specific data line in red and show a warning banner. This helps catch situations where dates may have been entered incorrectly or in the wrong order. Having accurate, chronological data is important for understanding the progression of a patient's health over time and providing proper care. The time coherence tool aims to identify potential timeline inconsistencies or errors so they can be validated and corrected if needed. By automating these coherence checks, it makes it easier to keep date logs complete and coherent across multiple patient records.

Both individual and dataset levels

- **Quality Image assessment metrics for XNAT platform.** CNR-IBB is developing a tool that can assess different metrics about image quality, including SNR, CNR and perceptual image quality indexes for clinical image datasets based on DICOM data. These tools have been developed to be compatible within the XNAT platform, exploiting Docker containers to scale up the computational power and can be used both at subject level and at project (whole dataset) level.

It is important to note that some tools share some features with other tools from other tool categories, such as de-identification tools (see as an example DIQCT that – among other tasks – checks the de-identification protocol applied). Similarly, some tools from the harmonisation category quite logically share some features of our data cleaning tools.

5.2.2 Data cleaning tools

Below we describe the data cleaning tools available to the EUCAIM community. All apply at the individual data level.

- **N4 Bias Filter.** In the context of ProCAncer-I, FORTH developed a tool based on the state-of-the-art N4ITK[82] method that is used for bias field correction of the MR images. Two main options are offered: (1) Apply N4 filter to MR prostate images (either with the default optimal parameters values or with parameters values defined by the user); (2) Find the optimal configuration of the N4 filter for specific T2W pelvic image/images by experimenting on various N4 configurations and automatically measuring the Full Width at Half Maximum (FWHM) of the periprostatic fat distribution. For the second option, the K-means algorithm is used to identify the periprostatic fat distribution with the high intensity signal in the filtered image. The filter configuration that results in the minimum

[82] Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging. 2010 Jun;29(6):1310-20. doi: 10.1109/TMI.2010.2046908. Epub 2010 Apr 8. PMID: 20378467; PMCID: PMC3071855.

FWHM for each patient is considered as optimum, indicating homogeneous tissue representation[83].

- **Deep Learning Noise Reduction (DLNR).** Another tool developed by FORTH in the context of ProCAncer-I is the Deep Learning Noise Reduction (DLNR) tool, which is meant to reduce the noise on already noisy MR prostate images. Note: Quality control is required prior to applying DLNR, and the module will compromise the image quality if input examinations are of high-quality.
- **ML model for MR series categorization.** HULAFE is developing an ML-based tool using artificial intelligence that can categorize MRI series using standardized non-free-text DICOM tags. The categorization includes identifying the type of sequence (e.g. spin echo, gradient echo), the weighting (e.g. T1W, T2W, DCE), the presence of fat suppression, and detecting non-relevant or junk series (e.g. localizers, calibrations, screenshots...). This development originated in the PRIMAGE project using neuroblastoma and DIPG imaging data and is currently being adapted for the CHAIMELEON project to be applied in MR images of prostate and breast cancer.
- **Curation of MR dynamic sequences.** There exists a large variability in storing MR dynamic sequences, such as diffusion-weighted imaging (DWI) and dynamic contrast-enhanced (DCE) MR images, in image repositories. These sequences could be stored as a combined series or as individual series, making it extremely difficult to determine which series should be grouped together as part of the same dynamic sequence. To address this, HULAFE developed a custom feature based on Euclidean similarities between individual series. This feature uses diverse non-free-text MR DICOM tags corresponding to each MR series. Its purpose was to identify individual series that belonged to the same dynamic sequence, allowing for their proper combination. With this tool, the PRIMAGE image repository was harmonised by consistently storing different individual MR series in their combined forms whenever necessary.
- **RACLAHE filter.** The RACLAHE filter is another tool developed by FORTH in ProCAncer-I to locate the prostate's whole gland in T2 MR axial images and enhance that area by applying CLAHE algorithm. The filter proved to be effective on segmentation tasks as it improves the segmentation performance on 5 DL models. The RACLAHE filter was developed using a ProstateX[84] patient cohort (204 patients). A U-Net bounding box model is utilized to isolate the whole gland from the outer parts of the examination and the CLAHE algorithm is applied in the isolated region whole gland region to further enhance that area.
- **NLmCED denoising filter.** CNR-IBB partner developed a denoising filter tool called NLmCED, which is a combination between two powerful denoising methods aimed to reduce Rician Noise in MR images. The filter can provide increased quality images for T1w, T1 + Gadolinium (contrast enhanced), T2w and FLAIR MR images. It has proven its efficiency also in reducing noise in CEST-MRI brain images.
- **Trace4MedicalImageCleaning.** Partners at DeepTrace Technologies are developing a tool whose aim is to detect and possibly remove text in medical images, specifically in ultrasound and mammographic 2D studies. Studies must be passed as input to the tool as DICOM files. The tool uses both information extracted from the image and from the DICOM metadata, and by applying machine-learning techniques it is aimed at detecting the presence of text in the images. In such a way, it will be able to signal if the quality

[83] Dovrou A, Nikiforaki K, Zaridis D, Manikis GC, Mylona E, Tachos N, Tsiknakis M, Fotiadis DI, Marias K. A segmentation-based method improving the performance of N4 bias field correction on T2weighted MR imaging data of the prostate. Magn Reson Imaging. 2023 Sep;101:1-12. doi: 10.1016/j.mri.2023.03.012. Epub 2023 Mar 31. PMID: 37004467

[84] Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman. "ProstateX Challenge data", The Cancer Imaging Archive (2017). DOI: 10.7937/K9TCIA.2017.MURS5CL

of the study is compromised by the presence of text superimposed on the image. If possible, in specific cases, the tool will also remove the text and clean the interested region.

5.3 Data harmonization

Data harmonization techniques are usually applied either before or after the extraction of radiomics features. These techniques aim to reduce the variability in image appearance across different sites, vendors and acquisition protocols.

In the context of data harmonization in AI4HI projects, various approaches have been adopted to standardize and enhance the quality of medical data.

- In the **PRIMAGE** project, an image harmonization pipeline was created to reduce variability between vendors, magnetic fields, acquisition protocols, and sites, incorporating denoising, bias field correction, spatial resampling, and intensity normalization.
- **ProCAncer-I** employed image-based normalization techniques and ComBat method for harmonizing MR T2W prostate images and radiomics features. Feature-based harmonization in ProCAncer-I aimed to mitigate "center-effect" variability in radiomics features using ComBat.
- **CHAIMELEON** introduced a self-supervised learning approach to improve image texture, contrast, and noise in MRI and CT images.
- In **INCISIVE**, a structured procedure involving template creation, review, consensus, standardization, and refinement was followed to define clinical metadata rules, while workshops between medical experts and technical partners established imaging data requirements, including de-identification protocols and annotation procedures.

More detailed information about each AI4HI approach regarding data harmonization is included in Sub-Annex B.2: Data harmonization.

In the **EUCAIM project,** we will consolidate and extend aforementioned methods with the intention to offer a rich and useful collection of data harmonization techniques for end users and for multiple clinical targets. To do this, we will provide two different types of data harmonization methods, one for medical images (e.g. MRI, CT) and others for numeric (i.e., radiomic) features. Accepted input data formats will be DICOM and NIFTI for raw images and CSV files for radiomic features. The data harmonization methods will be parameterized and containerized (e.g. docker images) in order to be accessible, reusable and interoperable within the EUCAIM architecture/platform. The overall idea is to be able to compose the provided harmonization tools within a federated data analysis EUCAIM pipeline for addressing different clinical targets. These containerized tools will be configured to read and write to certain data folders which will be mapped from the physical machine, so that previous methods can provide the data to harmonize and posterior methods collect the outcomes from these methods. Additional technical specifications (such as cpu, ram, disk space) will be identified in order to be able to run properly these methods in the local/central node.

Below is a description of each tool made available for EUCAIM by consortium members (see Annex D: Catalogue of tools for more details). Note that, as mentioned earlier in the data cleaning tools section, some harmonisation tools logically present some features dedicated to data cleaning (see as an example the harmonisation tools from HULAFE’s pipeline).

- **MRI_pixel_intensity_harmonization**. Partners at Quibim, in the context of CHAIMELEON project, are developing a tool to generate synthetic and harmonized images from a given MRI image set.
- **CT_pixel_intensity_harmonization**. Quibim, from the CHAIMELEON project, will provide another tool to generate synthetic and harmonized images from a given CT image set.
- **Trace4Harmonization**. DeepTrace Technologies provides a novel tool that aims at harmonizing numerical values extracted from medical images acquired with different models of image-acquisition systems.
- **Biologically motivated normalization techniques**. FORTH partner, in the context of ProCAncer-I, will be in charge of delivering normalization methods aiming to reduce the variability in the intensity values of the MR prostate images due to different scanners, acquisition protocols and conditions, based on the intensity values of specific tissues. Three biologically-motivated normalization techniques are provided: (1) The fat-based normalization method (2) The muscle-based normalization method (3) The single tissue (fat or muscle) piece-wise normalization method
- **Feature based Harmonisation (ComBat)**. In addition, FORTH partner will lead a harmonization method that aims to reduce the variability in the radiomics features of the MR prostate images due to different scanners, acquisition protocols and conditions by using empirical Bayesian methods to estimate differences in radiomics values and then expressing them in a common space (location/scale adjustment). There are two methods: 1. ComBat method, which shifts radiomics features to the overall mean and pooled variance of all centers and 2. M-ComBat method, which shifts radiomics, features to the mean and variance of the chosen reference center.
- **Neuroblastoma_T1W/T2W/DCE_harmonisation**. HULAFE partner will hand over a tool to harmonise abdominal images and increase image quality of T1W, T2W and DCE by removing image noise and field inhomogeneities using the anisotropic diffusion filter and N4 bias field correction filter.
- **Neuroblastoma_DWI_harmonisation**. Also HULAFE will contribute with another tool to harmonise abdominal images and increase image quality of DWI by removing image noise, using the SUSAN filter using intensity Otsu threshold as brightness threshold
- **DIPG_T1W_harmonisation**. The aim of this tool developed by HULAFE is to harmonise brain images and increase image quality of T1W by removing image noise and field inhomogeneities using the anisotropic diffusion filter and N4 bias field correction filter.
- **DIPG_T2W/DCE_harmonisation**. The aim of this tool provided by HULAFE is to harmonise brain images and increase image quality of T2W and DCE by removing image noise and field inhomogeneities using the bilateral filter and N4 bias field correction filter.
- **DIPG_DWI_harmonisation**. This tool is proposed by HULAFE and aims to harmonise brain images and increase image quality of DWI by removing image noise using the SUSAN filter using intensity Otsu threshold as brightness threshold.

5.4 Data de-identification

Data de-identification tools are essential for erasing personal information from both image and clinical data, ensuring GDPR compliance. In the realm of data de-identification and privacy protection across various AI4HI projects, distinct approaches have been adopted to safeguard sensitive clinical and imaging data.

- **CHAIMELEON** employs a two-step anonymization process for patient identifiers and DICOM images, utilizing random generation and shifting of dates to ensure compliance and ethics committee approval.
- **EuCanImage** relies on pseudonymization with the EuCanImage-ID and strict DICOM tag removal for anonymizing both patient and clinical data, using GDPR-compliant REDcap eCRF software.
- In **INCISIVE**, the focus is on de-identifying DICOM data, where extensive study led to the selection of the CTP Anonymizer for GDPR-compliant de-identification, with an emphasis on patient privacy while maintaining usability for AI developers.
- **ProCAncer-I** employs a double-level anonymization approach, with blacklisting and whitelisting, adhering to consortium-defined rules and DICOM standard modifications for image anonymization.
- **PRIMAGE** relies on pseudonymization via EUPID, phonetic hash-based pseudonyms, and DICOM tag removal, ensuring patient privacy and confidentiality while accessing Real World Data for tumor behavior prediction models.

While all projects have adopted strict de-identification measures, based on the confidentiality profile defined by the DICOM Standard, and have taken both technical and organizational measures so that data users cannot reidentify patients, there are significant differences in approaches. Some projects have opted for pseudonymization, while others perform anonymization. In pseudonymization, the data provider keeps the key to the patients' original identity. This has several advantages:

- It facilitates the curation process where authorized users can access the EHR data to verify the data completeness and address any doubts when dealing with outliers or incoherent data.
- Additional time points can be added to the data and are linked with previously collected information.
- Patient withdrawals can be handled.
- Patients can be reached by caregivers in case of incidental findings in a research project.

In anonymization, de-identification is irreversible. No table of correspondence is kept, even at the hospital level, so it is no longer possible to identify a person. Anonymous data is no longer considered personal data. Therefore, anonymous data falls outside of the scope of GDPR. For this reason, anonymization may be preferred by some institutions, as otherwise, it would be difficult for them to assume any responsibility over personal data that are stored in an environment that is not managed by them. Anonymization, however, requires specific precautions:

- As it is not possible to add timepoints, only full cases with longitudinal information and evaluation during full follow-up period can be sent.
- Specific measures need to be taken to ensure no duplicates of patients are sent.
- The European Data Protection Board (EDPB) has warned that anonymization is difficult to achieve and maintain over time "taking into consideration the available technology at the time of the processing and technological developments".
- Additional safeguards should be taken to ensure reidentification would be impossible to achieve, such as ensuring that downloading data is not possible and that any further processing of the data is monitored and justified.

Standardized guidelines would avoid individual projects having to make a choice and would reassure data holders.

The approach followed in **EUCAIM** will be the anonymization for the Central Repository, as a general rule. However, only for the Federated Nodes, pseudonymization can be considered, after ensuring GDPR compliance. The data included in the project will be de-identified, following a set of guidelines defined during the project. There will also be a de-identification profile of the DICOM tags, defining which should be removed, modified (and how) and which can be kept as they are. The first version of this de-identification profile is already defined (Annex C: List of imaging attributes kept for all AI4HI projects and is under approval by the EUCAIM legal team to ensure GDPR compliance.

As stated earlier there is a need to device ways ensuring that data pertaining to the same patient across multiple data providers should be identifiable as belonging to the same subject. Our current understanding and analysis points towards the fact that duplication detection will be guaranteed at the central storage, where appropriate EUCAIM services will seek to identify identical studies and/or patients.

Within the AI4HI projects, a variety of tools have been employed to accomplish this task. Furthermore, open-source tools may also be a valuable consideration for achieving this objective. A description of the initial tools that are being explored to be part of the EUCAIM catalogue is provided below:

- **Quibim Precision anonymization tool:** It is a pseudonymization tool. When a new patient is incorporated in the associated database, a new and unique pseudonym is given for its pseudonymization using European Unified Patient Identity Management (EUPID)[85]. When uploading an associated imaging study, this pseudonym is used to substitute personal data in the DICOM files such as the Patient Name or the Patient ID, and all the DICOM tags with sensitive information as stated in the DICOM standards PS3.15 are removed or emptied from the uploaded files. Additionally, the platform includes a tool to remove any sensitive burned data within the image. By drawing a rectangle on a specific region, the tool erases this area of the image before uploading by assigning background pixel values to the whole delineated region.
- **Radiomics Enabler®:** Radiomics Enabler® is a commercial web app used to extract large amounts of specific studies/series, but also allows for applying project-custom de-identification pipelines. It is combined with the Clinical Trials Processor (CTP), an open-source solution edited by the RSNA, which automates the de-identification process, filtering and specific routing according to sets of rules defined in editable scripts. The CTP DicomAnonymizer script allows defining how to handle each DICOM tag from the images (keep, remove, modify). The CTP DicomFilter enables isolating junk series in quarantine. Finally, the CTP DicomPixelAnonymizer adds a covering rectangle mask to frames based on the manufacturer and modality to hide potential identifying text information. Radiomics Enabler® also has a dateshift function, that shifts the dates information in the DICOM tags, to further de-identifiy the data. This tool has been used in the CHAIMELEON project to extract, de-identify and export data from sites to the CHAIMELEON repository.
- **RSNA DICOM Anonymizer Tool (CTP Anonymizer):** The DicomAnonymizerTool is a command-line program that processes a single file or a tree of directories using pipeline stages like those in Clinical Trial Processor (CTP). It uses the "DICOM Anonymizer"

[85] https://pubmed.ncbi.nlm.nih.gov/27139382/

functionality with a project specific configuration file. This tool has been the one used for the de-identification process in ProCAncer-I.

- **Mainzelliste:** It is an open-Source web-based pseudonymization and record linkage solution in use and co-developed at many institutions[86].
- **MainSEL:** Short for "Mainzelliste Secure EpiLinker"; an extension to Mainzelliste to perform Record Linkage using Secure Multi-Party Computation, thus without revealing input data.

5.5 Data annotation

The annotation tools focus mainly on tasks related to segmentation and detection. The aim is to incorporate these tools into the annotation environment located at the central node. This allows the annotation processes to be automated, reducing the need for time-consuming tasks and avoiding the need to create annotations from scratch.

Similar to the rest of the approaches in the previous sections, we first focused on the AI4HI approaches, and more specifically on the ProCAncer-I and PRIMAGE projects, as the rest of the AI4HI projects either did not include a specific annotation task or it was carried out by a partner outside the project.

- In **ProCAncer-I**, an integrated annotation tool environment adhering to the DICOM standard has been established, featuring a user-friendly interface for image manipulation and annotation, a secure back-end/proxy handling authentication and authorization, and an automatic prostate segmentation algorithm utilizing Convolutional Neural Networks (CNNs) for DICOM image segmentation.
- **PRIMAGE**, on the other hand, addressed the challenge of pediatric cancer tumor segmentation, involving manual delineation led by expert radiologists and the development of CNN-based automatic segmentation methods for neuroblastoma and Diffuse Intrinsic Pontine Glioma (DIPG). The approach in PRIMAGE combined manual and semi-automatic segmentations, significantly reducing the time needed for manual segmentation while achieving high accuracy. These segmentation models have been integrated into the PRIMAGE platform, facilitating their utilization in medical imaging analysis.

More information on the approaches followed by the AI4HI projects are outlined in the Sub-Annex B.4: Data annotation.

In **EUCAIM**, two general annotation pathways are considered, depending on where the annotation is performed and stored: either in the local node or in the central repository.

On the one hand, the annotation process in the central node follows a similar approach to the one presented in ProCAncer-I and PRIMAGE. Quibim DICOM Web Viewer will be integrated into the EUCAIM platform as the annotation environment, maintaining a DICOM-in - DICOM-out approach. As described in ProCAncer-I, this environment 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, various annotation tools provided by EUCAIM partners will be integrated into the viewer, allowing for automatic annotation of the images (similar to the prostate segmentation algorithm in ProCAncer-I and the neuroblastoma and DIPG segmentation algorithms in PRIMAGE). The primary focus of these models is the segmentation task, as it is the most time-consuming activity. These

[86] https://bitbucket.org/medicalinformatics/mainzelliste/src/master/

algorithms are executed in Docker containers, invoked by the backend, which will also manage the annotation results and associated information.

On the other hand, local annotation is conducted when a site is capable of performing manual data annotation using in-house annotation software. The resulting annotations must adhere to the standard format (DICOM). Subsequently, a quality check must be conducted to assess whether the annotations comply with the standard. After confirming these are correct, they can be stored in the local node for further federated processing, or they can be ingested by the central node for storage there. A specific case of local annotation arises when annotations are collected from the AI4HI projects. In such cases, the annotations are already performed and will be transferred to the central repository, with a conversion to DICOM SEG if needed.

As previously mentioned, the proposed standard format for EUCAIM is DICOM; thus, the annotation standard format will be DICOM SEG. Another accepted annotation format is NIfTI, as long as it includes the original images in DICOM format, enabling conversion from NIfTI to DICOM SEG.

The tools considered for the data annotation step are the following:

- **Breast dense tissue segmentation ITI BREAST Calculate**[87]. The tool provides automatic segmentation of dense tissue in digital mammographies. The segmentation is intended to be modified for any specialist if it does not match its criterion. This tool provides a fully automated method based on deep learning to estimate breast density, including breast detection, pectoral muscle exclusion, and dense tissue segmentation. A novel confusion matrix (CM)—YNet model for the segmentation step is employed. This architecture includes networks to model each radiologist's noisy label and gives the estimated ground-truth segmentation as well as two parameters that allow interaction with a threshold-based labelling tool.
- **MR-based neuroblastoma tumour detection and segmentation**. The tool performs an automatic segmentation of neuroblastoma tumours on T2w MR images.
- **MR-based DIPG tumour detection and segmentation**. The tool performs an automatic segmentation of DIPG tumours on T1w and/or T2w/FLAIR MR images.
- **MR-based glioblastoma tumour detection and segmentation**. The tool performs an automatic segmentation of glioblastoma tumours and its subregions (enhanced tumour, peritumoral edema, non-enhanced/necrotic tumour, and total tumour) across four MR sequences (T1w, T1w-contrast enhanced, T2w, and FLAIR).
- **CT-based neuroblastoma tumour detection and segmentation**. The tool performs an automatic segmentation of the neuroblastoma tumours on CT images.
- **nnUnet**. nnU-Net is a semantic segmentation method for training and inferencing that automatically adapts to a given dataset. It will analyse the provided training cases and automatically configure a matching U-Net-based segmentation pipeline. It includes many pretrained models (e.g. MRI cardiac, MRI abdominal organ, CT thorax, or CT Liver).
- **nnDetection**. Tool designed for the simultaneous localization and categorization of objects in medical images. This technique is intended to automate and systematise the process of configuring object detection methods in the medical field. It allows training and inferencing.

[87] Larroza, A.; Pérez-Benito, F.J.; Perez-Cortes, J.-C.; Román, M.; Pollán, M.; Pérez-Gómez, B.; Salas-Trejo, D.; Casals, M.; Llobet, R. Breast Dense Tissue Segmentation with Noisy Labels: A Hybrid Threshold-Based and Mask-Based Approach. Diagnostics 2022, 12, 1822. https://doi.org/10.3390/diagnostics12081822

- **MITK**. Comprises (1) Interactive UI based Application for image analysis (including registration and segmentation, considering classical manual tools, but also sophisticated options like GrowCut, TotalSegmentator, SegmentAnything, and nnUnet), and (2) collection of command line tools for basic image processing automatization (e.g. file conversion, registration, stitching, resampling...).
- **Multi-regional prostate segmentation**. The tool performs an automatic segmentation of the different regions of the prostate (CZ+TZ, PZ, SV) using MRI images.

5.6 Data Fairness

Compliance with the FAIR principles implies considering multiple dimensions. The EUCAIM approach is based on the RDA recommendations[88], but during the project, we will also define further FAIR attributes related specifically to Cancer Imaging data. On top of this, the sensitive nature of the data will have to be taken into account when adopting the FAIR principles, setting some limitations to the fulfilment of some of them.

While all AI4HI projects implemented comprehensive approaches to data management, when it comes to the adoption of the FAIR principles their implementations differ both in principles and scope, with CHAIMELEON and ProCAncer-I being currently the closest to the RDA recommendations.

- **CHAIMELEON** ensures findability through structured identifiers and Zenodo deposition, enhancing accessibility via open description landing pages, but with authorization and Terms of Usage requirements. Interoperability is achieved through MIABIS and DICOM-MIABIS, while reusability benefits from DICOM retention and clear ethical approval processes.
- **ProCAncer**-I employs Biotronics3D, OMOP-CDM, Radiology-CDM, and AI-Passport not only for achieving data FAIRification but also and AI model FAIRification, alongside infrastructure standardization.
- In contrast, **INCISIVE** employs unique identifiers, an in-house querying mechanism, FHIR-HL7 CDM, and detailed dataset metadata lists for FAIRification, lacking specific FAIR tools.

More information on the AI4HI approaches are in Sub-Annex B.5: Data FAIRification.

In the **EOSC-Synergy** H2020 project, **CSIC** developed a tool called **FAIR EVA** (evaluator, validator & advisor) that has been selected for its deployment in the EUCAIM infrastructure.

FAIR EVA has been developed to check the FAIRness level of digital objects from different repositories or data portals. It requires the object identifier (preferably persistent and unique identifier) and the repository to check. It also provides a generic and agnostic way to check digital objects. FAIR evaluator is a service that runs over the web. It can be deployed as a stand-alone application or in a Docker container. It implements different web services: the API that manages the evaluation and the web interface to facilitate accessing and user-friendliness.

FAIR evaluator implements a modular architecture to allow data services and repositories to develop new plugins to access its services. In addition, some parameters can be configured like the metadata terms to check, controlled vocabularies, etc. For the initial iteration, the vanilla version of the tool will be deployed, but during the project, a plugin will be developed to include the agreed new FAIR attributes to be checked.

[88] RDA FAIR Data Maturity Model. https://www.rd-alliance.org/system/files/FAIR%20Data%20Maturity%20Model_%20specification%20and%20guidelines_v0.90.pdf

In addition, the use of templates in CEDAR[89] is another powerful option for effective implementation of the FAIR principles, which will also be explored.

In **EUCAIM**, we will follow RDA recommendations, revising the indicators specified by them; but during the project, we will also define further FAIR attributes related specifically to Cancer Imaging data. Also, we will adopt and adapt the EOSC-Synergy FAIR EVA tool for checking the datasets FAIRness. We summarize some of the main points in the following:

- **Findability**: We will ensure that digital objects are assigned unique permanent IDs. A way of making such IDs "official" is using an approved service like identifier.org and Zenodo (which provides DOI's for the datasets), for publishing the datasets metadata (as it is already accessible in the public catalogue of EUCAIM). Due to the potential growth of the data collections, we will consider if a EUCAIM controlled alternative would be more suitable in later stages of the project.
- **Accessibility**: We will provide access to the public metadata, to access the actual data will require to follow the EUCAIM established procedures.
- **Interoperability**: Imaging data will be stored in DICOM and/or Nifti formats that are widely used in clinical environments (the former) and research (both). Full documentation of the data models and formats used will be publicly available.
- **Reusability**: The use of DICOM and/or Nifti for the imaging data, the adoption of a well-documented CDM and appropriately licensed data, and data sharing agreements, so that they can be replicated and/or exploited in different settings, will ensure the reusability of the data. Additionally, full documentation about the conditions of dataset access and re-usage will be made available.

5.7 Metadata Management and Interoperability Tools/Services

5.7.1 Imaging metadata interoperability

Extracting and organizing imaging metadata from DICOM files effectively, is the first key step for acquiring the most important information for querying and assessing the suitability of the available EUCAIM datasets in terms of imaging related information. In this section, we will provide an overview of the key aspects related to imaging metadata management and interoperability, as these are currently defined.

DICOM, as already mentioned in section 3, is the standard to be used in the context of EUCAIM for the exchange, storage, and management of medical images and associated data. DICOM files consist of the so-called DICOM tags, which are a structured set of metadata attributes that contain essential information about the patient, imaging equipment, acquisition parameters, and image characteristics, among others.

The first step in utilizing the DICOM metadata is the extraction of the most important tags from the available DICOM files. This process involves parsing the DICOM header to identify and capture specific attributes of interest. The starting point to define the minimum set of imaging metadata for EUCAIM, was to gather information from each AI4HI project on imaging metadata requirements. A list of imaging attributes *common* to all AI4HI projects with available imaging metadata has been collected (Annex C: List of imaging attributes kept for all AI4HI projects), and from there a set of mandatory variables have been selected as the *minimum* set for EUCAIM for data queries and dataset registration. This has already been thoroughly described in Section 3.5 Public Catalogue - Metadata Model (see Table 6: Minimum imaging metadata for

[89] Musen MA, O'Connor MJ, Schultes E, Martínez-Romero M, Hardi J, Graybeal J. Modeling community standards for metadata as templates makes data FAIR. Sci Data. 2022;9(1):696. Published 2022 Nov 12. doi:10.1038/s41597-022-01815-3

joining the EUCAIM federation). Note that the set of DICOM tags that will be exploited in EUCAIM depends on the EUCAIM anonymization profile to be employed.

Once the DICOM tags are extracted, it is imperative to store and organize them effectively to support the objectives of our project. By analyzing the imaging metadata management approaches in the AI4HI projects, we observed that they handle imaging metadata differently. Some have set up a dedicated storage entity for imaging metadata: this is the case for **ProCAncer-I** that has a repository storing the DICOM header information at the *image level.* This repository contains a columnar database installation, "ClickHouse"[90], which is able to answer imaging metadata related queries in sub-second-latency despite searching over 5 million images. Furthermore, in ProCAncer-I, a set of the most important imaging metadata used for defining cohorts, as these were defined by the clinical and AI experts, are extracted and standardized through the Radlex terminology by using the radiology extension of the project's OMOP database. In **EuCanImage**, imaging data was exported to and processed by an external partner and the related metadata is not available for querying. The other projects (**PRIMAGE, CHAIMELEON**) have the imaging metadata stored individually in the header of the images, on their central repository. In these projects, only a subset of the complete imaging metadata (e.g., modality, sequence type, echo time, repetition time) is queryable. In **INCISIVE** imaging metadata are stored along with the DICOM images as the header of the image, in the PACs server. They are not included in the common data model, thus they cannot be queried upon.

In **EUCAIM**, we plan the OMOP radiology extension to serve as a specialized framework for standardizing imaging metadata that will allow queries by using a combination of clinical and imaging metadata. This allows us to map DICOM tags to a structured, standardized data model, facilitating data integration and analysis. In addition, the FHIR resources for representing DICOM studies and series are also to be explored and utilized.

Key features of the imaging extension data model to be used include:

- Standardized Vocabulary: It uses standardized medical vocabularies (e.g., SNOMED CT, ICDO-3, Radlex) to represent concepts, ensuring semantic interoperability as described in section 3.3.
- Hierarchical Structure: The extension provides a hierarchical structure to represent various levels of information, from patients and studies to series.
- Mapping to DICOM Tags: It includes mappings that relate DICOM tags to standardized concepts, allowing for the integration of DICOM metadata into the common data model.
- Query and Analysis Support: The extension facilitates efficient querying and analysis of radiological data, in combination with the clinical information that pertains to patients.

5.7.2 Clinical data interoperability

For achieving clinical data interoperability, our primary objective is to provide the means and the tools that will transform the clinical data from the organizations willing to participate in the EUCAIM federation into interoperable (meta)data conforming to the EUCAIM CDM. The metadata will be used to make federated nodes searchable using the tools for data search present in the central node. This transformation will enable the dissemination of clinical data from different centers to all EUCAIM platform users. To accomplish this, ETL (Extract, Transform, Load) techniques will be employed. ETL are processes to collect data from various sources, transform it into a structured format, and load it into a data warehouse or another database system for analysis.

[90] https://clickhouse.com/

EUCAIM will adapt to current approaches employed by other federated data research infrastructures. Additionally, EUCAIM will offer the possibility of anonymizing and loading the data and metadata into the central node if a federated query is not possible.

There are three main considerations in order to prepare an ETL process to translate all the relevant information into EUCAIM:

- The Common Data Model (CDM) must be defined to transform inbound data and metadata into EUCAIM's Data Model.
- The Data Sources must be correctly identified. Initially, the project will extract its information from AI4HI Horizon 2020 Projects (CHAIMELEON, ProCAncer-I, PRIMAGE, etc.) but in the future new sources (both established repositories and databases from clinical centers) may be added.
- Finally, the workflow for the integration of information must be established. Depending on the data sources, it is possible that several data and metadata ingestion strategies will be performed to correctly collect all data.

The ETL process presents some challenges for the ingestion of data and metadata:

- There may be non-structured data present (such as free text reports) that may require pre-processing before the data extraction. This issue is solved in the sections of this chapter. There may also be non-standard information (such as custom vocabularies) that may require a complex transformation.
- EUCAIM will deal with several heterogeneous systems from multiple data providers, that is, the system used by one partner to store the information (infrastructure, hardware, software) may be completely different from the one employed by another partner. This may make the ETL process more complex, as it will have to be designed as generic as possible to fit all the different possibilities.
- Lastly, the ETL design should be both scalable and easily maintainable. Due to the considerable complexity of the task at hand, as well as the aforementioned challenges, achieving this is a substantial endeavour.

5.7.2.1 Proposed workflow

To perform the ETL processes for clinical data, the following workflow is proposed:

1. As part of T2.1, data providers will attend an onboarding process, where the requisites in terms of data, hardware and software will be specified. Data providers will conform to the requirements of EUCAIM.

2. As part of T5.3, data providers will submit their clinical data and metadata to a pre-processing stage. This stage will contain the following phases, described in the previous sections of this chapter:

    a. Quality control
    b. Data cleaning
    c. Data harmonisation
    d. Data de-identification
    e. Data FAIRification

3. Finally, once the clinical data is pre-processed and conformed to EUCAIM's CDM, the data will be extracted, transformed, and loaded into the federated network. To do so, EUCAIM will provide a toolset and training of open-source tools to the data providers, who will be the ones responsible for performing this process. EUCAIM will remain

available to provide support to prospective data partners in order to complete these steps.

The ETL toolset depends on the CDM selected by EUCAIM and the data pre-processing performed. We envision two main approaches: Systems employing HL7 messaging standards and systems employing the OMOP Common Data Model.

Please note that we aim to allow flexibility on the way local nodes structure their data within the federation. In certain instances, data providers may already employ a well-established ("standardized") Data Model like OMOP-CDM, FHIR (e.g. as in the case of AI4HI projects), while in others, they may opt for entirely ad-hoc/bespoke models. The primary objective of the project is to establish a unified "Common" Data Model to be used/imposed for any new Data Provider that joins the federation. In some cases (and especially for new providers) an ETL process will be defined to transform local data models to the EUCAIM CDM (OMOP/FHIR). To shield the platform from the "particularities" of each provider within the federation, a "Mediator" component has been also defined and described (section 2.4. Federated Query and section 6.3.4 ETL/Mediator described later in the document). It is worth mentioning that there may be specific implementations of the Mediator tailored to the local data model in use (for example, a specific implementation for OMOP and another for FHIR). Regardless of the specific implementation, all mediators are expected to offer the same "interface" which is exposed to the central infrastructure, in addition to any local extensions required for data access during federated processing, data visualization, permitted data processing, and similar functionalities.

5.7.2.1.1 HL7/FHIR Integration

HL7 (Health Level Seven International) is a set of standards for the exchange, integration, sharing, and retrieval of electronic health information. HL7 facilitates the communication of data between different healthcare applications, and is widely adopted around the globe for healthcare information system interoperability. HL7 covers a wide range of healthcare operations, from patient admissions to billing to laboratory results. This standard enables consistent data representation and communication across different healthcare applications.

There are multiple versions of HL7, with HL7 v2.x and HL7 v3 being the most prominent. HL7 v2.x is more commonly used, known for its pipe-delimited format, while HL7 v3 is XML-based. Additionally, HL7 FHIR is a newer standard under the HL7 umbrella, providing a more modern approach to healthcare data exchange using RESTful APIs and JSON or XML formats.

HL7 provides a framework for the safe and efficient transfer of clinical, administrative, and financial information between healthcare system components. HL7 is open-source, under the Mozilla Public License, V1.1[91]

HL7 FHIR (Fast Healthcare Interoperability Resources) introduces a new approach to structuring and exchanging healthcare data compared to previous HL7 versions. Instead of the traditional message-based model, FHIR is built around the concept of resources. A resource is a discrete chunk of healthcare data that has a known structure and a common set of behaviors. Examples of resources include Patient, Encounter, Observation, Medication, and many others. Each resource type has a standardized definition, including its attributes and their data types, relationships to other resources, and a set of constraints. This ensures consistency across different implementations. Resources can reference each other, enabling complex relationships between different types of data. For example, an Encounter resource might reference a Patient

[91] https://www.mozilla.org/en-US/MPL/1.1/

resource to indicate who the encounter was with. FHIR resources can be versioned, allowing for the tracking of changes over time.

FHIR resources can be represented as JSON or XML, allowing for easy interoperability with modern web technologies and systems. FHIR is designed with the principles of REST (Representational State Transfer). It uses standard HTTP-based CRUD (Create, Read, Update, Delete) operations for interacting with resources. This allows for straightforward integration with web applications and mobile apps.

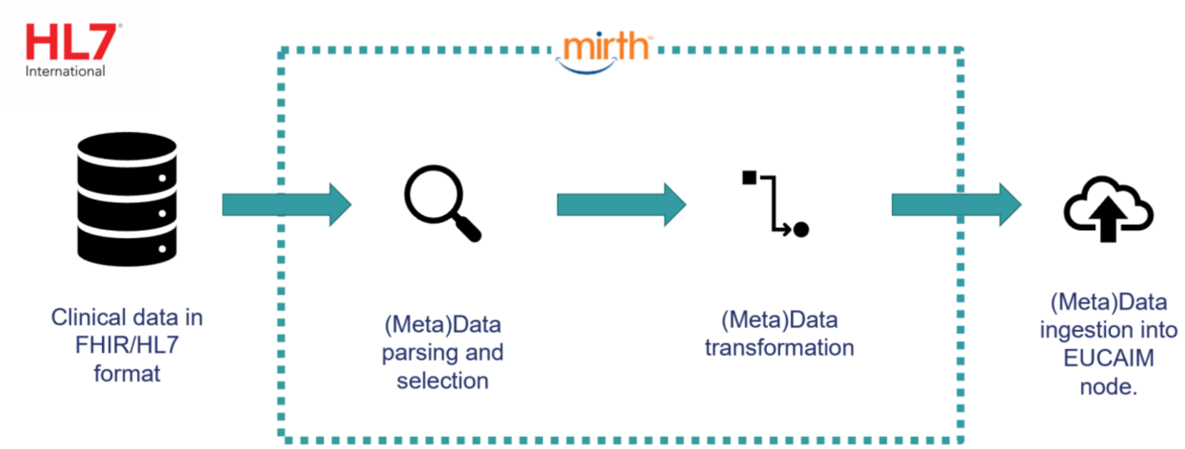

HL7 International
mirth
Clinical data in FHIR/HL7 format
(Meta)Data parsing and selection
(Meta)Data transformation
(Meta)Data ingestion into EUCAIM node.

Figure 24: ETL process for FHIR repositories.

It is important to note that while FHIR defines the structure and exchange format of health data, the actual storage mechanism (e.g., which database system or storage solution) is not prescribed by FHIR. Organizations can choose their storage solutions based on their requirements, as long as they can produce and consume data in the FHIR format when interfacing with external systems.

To ingest resources into EUCAIM a tool called MIRTH Connect[92] will be used to query, parse, transform and load the relevant information (either data or metadata) into EUCAIM, as can be seen in the diagram in Figure 24.

MIRTH connect is an open-source healthcare integration engine designed to facilitate the interoperability of health information systems. It provides tools to develop, test, and deploy data exchange interfaces using a variety of communication standards, especially HL7. With a user-friendly interface, it allows users to transform, filter, and route incoming and outgoing messages (that is, clinical data), making it easier to connect disparate systems in healthcare settings.

5.7.2.1.2 OMOP Integration

The Observational Medical Outcomes Partnership[93] (OMOP) Common Data Model (CDM) is a standardized data model and methodology aimed at transforming disparate observational databases into a common format. This harmonization allows for the consistent and efficient analysis of healthcare data across multiple sources. By providing a consistent way to represent healthcare data, OMOP CDM facilitates collaborative research, large-scale analytics, and real-world evidence generation in medicine.

[92] https://www.nextgen.com/solutions/interoperability/mirth-integration-engine
[93] https://www.ohdsi.org/

OMOP uses a relational database structure to store data, providing a consistent and organized way to store diverse healthcare data, which, once transformed and loaded into the model, can be easily queried and analyzed across different datasets and institutions.

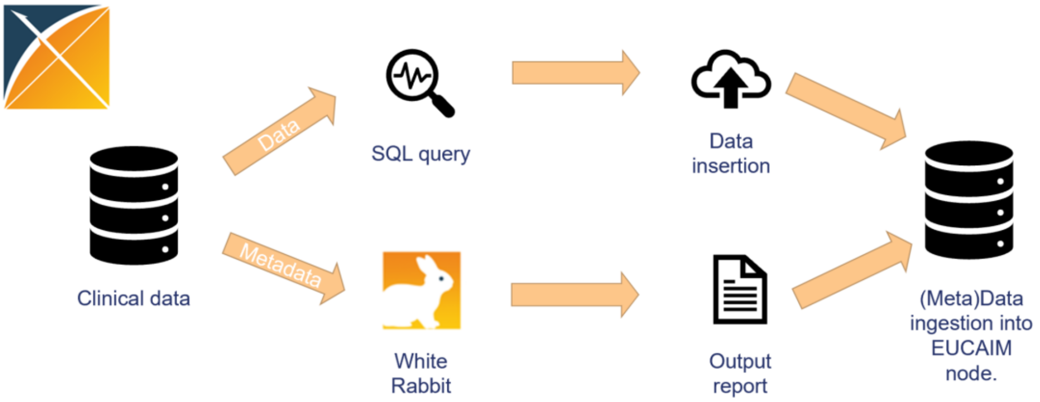
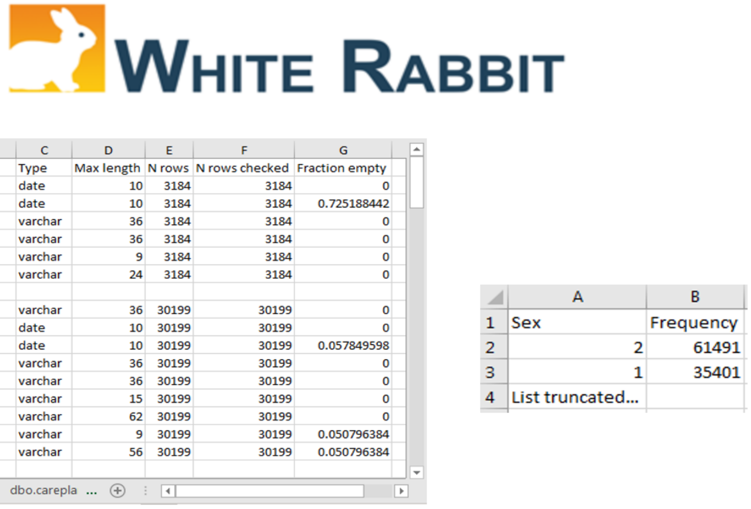


Figure 25: ETL process for OMOP-based repositories.

The OMOP CDM is organized into tables, each designed to capture specific types of healthcare data. Some tables are related to patient care (like visits, conditions, procedures), while others store reference information (like vocabularies). The vocabularies are standardized vocabularies, ensuring different terms from various sources can be mapped to a common set of concepts. Data tables are organized by domain, such as “Condition”, “Drug”, “Procedure”, etc. Each of these domains captures specific aspects of information. All of this helps to capture a longitudinal health record of the patients, making it possible to track patient health events over time.

The process to integrate data or metadata in EUCAIM from OMOP-based systems will be different, as can be seen in the diagram of Figure 25.



	A	B	C	D	E	F	G
1	Table	Field	Type	Max length	N rows	N rows checked	Fraction empty
2	dbo.allergies	start	date	10	3184	3184	0
3	dbo.allergies	stop	date	10	3184	3184	0.725188442
4	dbo.allergies	patient	varchar	36	3184	3184	0
5	dbo.allergies	encounter	varchar	36	3184	3184	0
6	dbo.allergies	code	varchar	9	3184	3184	0
7	dbo.allergies	description	varchar	24	3184	3184	0
8							
9	dbo.careplans	id	varchar	36	30199	30199	0
10	dbo.careplans	start	date	10	30199	30199	0
11	dbo.careplans	stop	date	10	30199	30199	0.057849598
12	dbo.careplans	patient	varchar	36	30199	30199	0
13	dbo.careplans	encounter	varchar	36	30199	30199	0
14	dbo.careplans	code	varchar	15	30199	30199	0
15	dbo.careplans	description	varchar	62	30199	30199	0
16	dbo.careplans	reasoncode	varchar	9	30199	30199	0.050796384
17	dbo.careplans	reasondescription	varchar	56	30199	30199	0.050796384
18							

	A	B
1	Sex	Frequency
2	2	61491
3	1	35401
4	List truncated...	

Figure 26: Metadata extraction on OMOP databases through the White Rabbit tool

On the one hand, since OMOP is a relational database, it can be queried using SQL commands. This will allow the extraction of the relevant data. This can be done with different software solutions, depending on the flavor of SQL employed at each site. Then, the data (that will be already structured and clean) may be uploaded into an EUCAIM node.

Another tool available in the EUCAIM consortium by its partner IQVIA, is the OMOP Converter, which will also be explored and compared to the solutions given by the OHDSI community. More details on the OMOP Converter can be found in Annex A: OMOP Converter.

On the other hand, metadata will follow a somewhat different process. To extract metadata from the OMOP database, a tool called WhiteRabbit[94] will be used. WhiteRabbit is a software tool that will perform a scan of the source data, providing detailed information on the tables, fields, and values that appear in a field. This scan will generate a report with all the relevant metadata of the database. Figure 26 shows different examples of the output report from WhiteRabbit, where information regarding the different tables from the database can be seen:

5.7.2.1.3 Other cases

There may be other possibilities in which additional work must be performed in order to extract the necessary clinical information from a database. For example, the transformation of an OMOP database to a FHIR one for EUCAIM (or vice versa), or cases where a custom solution must be performed.

In the case of the transformation of a FHIR database into an OMOP one, it is only necessary to create a process to collect information from FHIR messages and insert them into the OMOP database. FHIR-based databases usually have an API built in order to exploit its information.

The case of transforming an OMOP database into a FHIR one is a little more complex. To transform OMOP to FHIR an interface is needed to collect data from the database. Then, the information would be sent using FHIR messages. Fortunately, there is an existing tool that can fulfill this role, the OHDSI Web API[95].

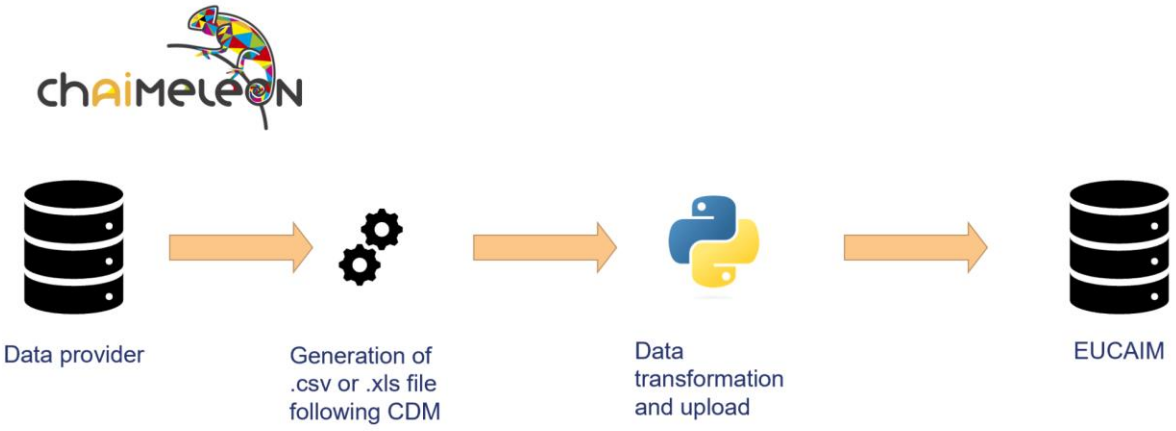

chAImeleoN
Data provider
Generation of .csv or .xls file following CDM
Data transformation and upload
EUCAIM

Figure 27: Alternate workflow of transforming local data into the EUCAIM CDM.

The OHDSI WebAPI is a Java-based application that is designed to provide a set of RESTful web services for interacting with one or more databases converted to the OHDSI Common Data Model (CDM) v5. This API can be used as a first step to connect into a OMOP CDM and then an additional layer could be created to transform the OMOP CDM concepts into FHIR resources.

Finally, there may be the need for exploring more customized solutions. For this, we propose the workflow employed for the CHAIMELEON project with some of the data partners. In this workflow the data provider was contacted and asked to create and populate a .csv or .xls file with all their clinical cases. Then, a Python script was written to transform the incoming data

[94] https://www.ohdsi.org/analytic-tools/whiterabbit-for-etl-design/
[95] https://github.com/OHDSI/WebAPI/wiki

from the data centers into CHAIMELEON CDM compliant data. A system of tables was then created to fit them into the OMOP CDM and sent into the local node for processing. In case a data partner has data that cannot be extracted using one of the previous methods, this will allow them to successfully share their data with EUCAIM. Figure 27 shows the proposed workflow.

6. Technical Specifications for the Set-Up of the EUCAIM Federated Nodes

6.1 Introduction to Federated Nodes

The Federated Nodes of EUCAIM refer to the pool of individual *software deployments, which* have been configured to connect and integrate with the EUCAIM infrastructure in order to enable the federation of data.

Federated Nodes are defined by this, basic, functionality of data federation, which is enabled through the containerized deployment of EUCAIM components, which have been configured individually for each node to allow it to:

- Provide a structured and queryable data storage service.
- Provide Data Federation functionalities through the deployment of the Federated Query and other services.
- Perform the loading of contributed data to the data storage service through additional components, which have been described in Section 5.

As stated, this is the bare-minimum descriptive type of Federated Node, essentially a Federated Node that provides data storage and federation; but no data processing capabilities. For the Federated Nodes, which will offer data processing capabilities, additional components will be deployed which additionally enable the Federated Node to:

- Provide a persistent volume as a staging area for data processing, with access configurations according to the business logic and defined privacy processes of EUCAIM.
- Load data from the local data storage service into the staging area, independent of the type of implementation of the Federated Query & ETL workflow.
- Deploy data processing tools and give them access to the staged data.
- Monitor the execution of data processing, collect, and report the results back to the corresponding services.

An EUCAIM Federated Node can and will be deployed in various types of infrastructures; as stated, some nodes will offer data processing while others will only offer data storage and federation. This multiplicity is increased when including the integration processes necessary to deploy an operational Federated Node in the infrastructure of an AI4HI project or an ERIC. As such, the components, which will be later presented in Section 6.3 Federated Node Componentsare grouped in modules such as “EUCAIM Processing Module”, the “EUCAIM Query Module” and the “Data Module”. The thought process being that while adaptations might be needed at a software level to adjust to the various intricacies and existing configurations of complicated infrastructures, EUCAIM partners will be able to change the minimum required parts of a module’s software in order to enable the integration.

6.2 Architecture Diagram

A draft of a draft architecture of the Federated Node is presented in this section in Figure 28, providing a preview of the software components that are expected to be deployed and outlining a general flow of requests from the EUCAIM Central Repository towards the Federated Node, as well as the internal data flows between Federated Node components.

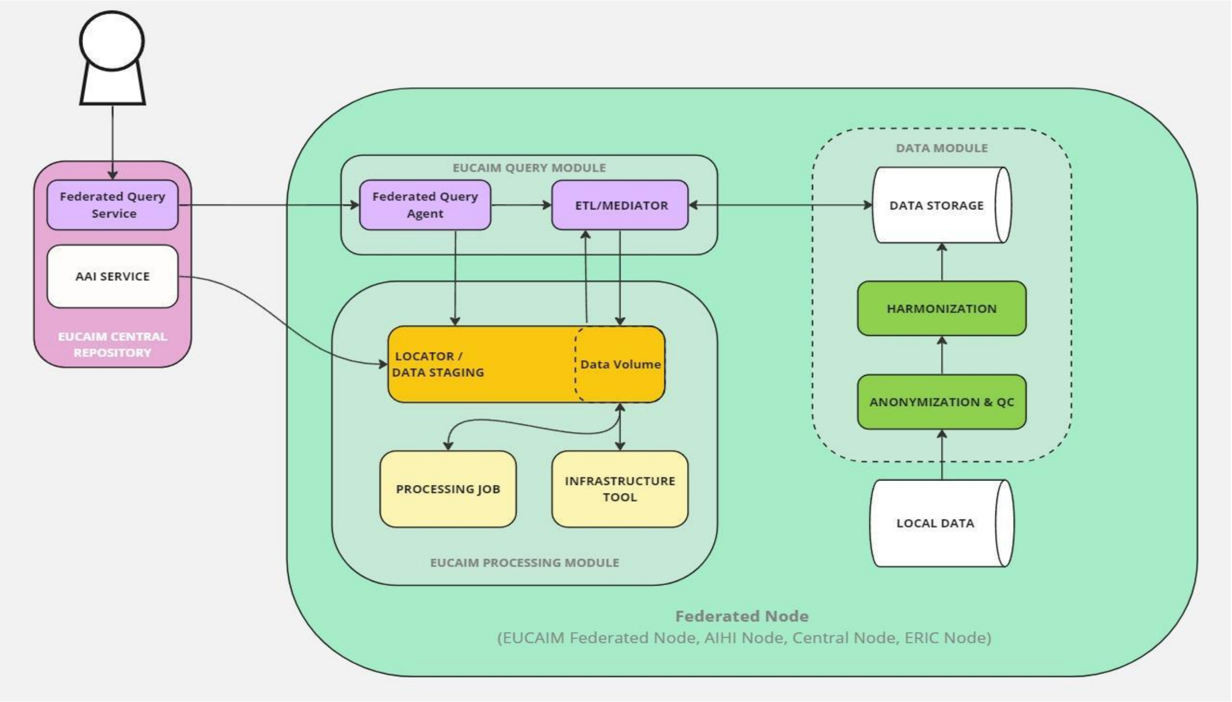


Figure 28: Draft Architecture Diagram /w data flows

6.3 Federated Node Components

6.3.1 Local Data

This component refers to a native storage volume attached to the federated node; this component stores any local, file data that is to be ingested into the EUCAIM system through the data preparation process. This component can have multiple iterations, ranging from a simple dedicated hard drive volume, to a network-attached drive, or even omitted completely if data can be ingested through other means, such as an accessible database for example.

6.3.2 Data Storage

The Data Storage component is a locally-deployed database that is selected to align with the standards used in the local data model (OMOP/FHIR/other). Its purpose is to store the data that will be available for query and managed-access to the EUCAIM federation. During the data preparation stage, this component will store the data resulting from the data preparation tools used, and during runtime this component will serve analytics or data as requested by the local ETL/Mediator component. The method of data storage (database/data lake/data warehouse) is being evaluated as of writing, as different methods present different benefits and, of course, different challenges for adjusting the deployment to the varied types of nodes integrated in EUCAIM's federated infrastructure.

6.3.3 Federated Query Agent

The Federated Query Agent refers to a federation-framework dependent component whose role is to maintain a connection with the central EUCAIM infrastructure, provide secure channels of communication for software, and to forward requests through these channels to deployed components, enabling those component's services.

6.3.4 ETL/Mediator

The ETL/Mediator component is temporarily named like this due to its role in the infrastructure, which is to perform queries for the data located in the Data Storage component of the Federated Node upon request and to provide these responses to the Federated Query service. Furthermore, this component is called to extract this data upon request and load it on a storage volume provided for a processing job through the Locator. The final implementation of this component is still to be determined as depending on the nature of the host infrastructure and the locally available data model, it might be called to perform a variety of data operations. More information on this component's functionality can be found in section 2.4,

6.3.5 Locator

The Locator component provides its services when a request is made for a subset of a data collection to be loaded and provided for a data processing project. The component carries out the following:

- Reserves a logical data volume for the purposes of the data processing project.
- Provides the data volume via reference to the ETL/Mediator service, which is responsible for extracting the data from the local Data Storage mechanisms.
- Configures access control for the provided volume and contained data, based on the requesting user's level of access contained in the AAI authorization tokens and according to EUCAIM's business logic.
- Once data loading is complete, it provides the data volume to the corresponding project and the data processing tools, which have been deployed for the purposes of the project.
- Preserves the data volume reserved for each project, following the specified business logic and provided parameters for data volume de-allocation (project cancellation, time-constraints, data removal requests etc.).
- Gracefully de-allocates the data volume after the required monitoring metrics and data processing results have been collected.

The locator component should follow simple and low-level implementations for managing data volumes in order to minimize overhead and allow for as many managed data volumes running in parallel as possible. Finally, the Locator component has the capacity to provide common, data staging and preparation functionalities for data processing such as feature extraction, analytics, reporting, layered de-identification etc.

6.3.6 Harmonization, Anonymization & QC

Harmonization, Anonymization & QC refer to the data pre-processing components of the EUCAIM federated node Data Module. The specific components are deployed as part of the Data Module "stack", which might contain additional components in the future. The Harmonization component refers to a collection of tools aimed at harmonizing the data available locally before storage, and similarly, the Anonymization & QC component will provide a collection of tools to be used to de-identify and validate the quality and integrity of data before they are stored in the local Data Storage mechanism.

6.3.7 Data discoverability

Data discoverability refers to the data components that aim to provide interoperability of data, metadata and access mechanisms of EUCAIM with other European initiatives.

Beacon aims to facilitate the responsible and secure discoverability of sensitive data regardless of where the data is located. Beacon is a query protocol established by the Global Alliance for Genomics and Health (GA4GH)[96] that defines a standard for the federated discovery of genomics and clinical data. Beacon v2 allows retrieval of aggregated information with different levels of granularity: boolean, counts or records. Access to Beacons information is securable through institutional systems for authentication and authorization, allowing access to only

[96] https://beacon-project.io/

trusted individuals and/or for specific purposes. Individual Beacon instances can be lit by both big or small organizations/projects and can be assembled into a single searchable network.

Technically, Beacon is composed of two basic components: the framework and the model. The Beacon framework describes the structure of the API requests, responses, parameters and common components, while the Beacon model describes the data model itself, meaning domain specific attributes and the relationships between them. To achieve interoperability between different Beacon instances the protocol includes a recommended data model for genomics and clinical data. However, since the framework and the model are independent from each other, the Beacon model can be adapted to other disciplines than genomics keeping its value. The development of a common Beacon model for images will contribute towards the interoperability of different European infrastructures and initiatives related to clinical imaging data, radiology, imaging repositories or digital pathology imaging.

Beacon is being developed and widely adopted by diverse EU and national projects: Federated EGA, Elixir, GDI and BIGPICTURE among others. Interestingly, the national level project IMPACT is working towards the development of a pilot Beacon for images, which can set the bases for further development and adoption.

6.3.8 Additional components

As of writing, further additions to this collection of components are being made through the project's research efforts as well as the validation activities aimed at the MM2 prototype. This list will be clearly defined and updated further in future technical deliverables outlining the architecture of EUCAIM.

6.4 Requirements for Node Setup

6.4.1 Hardware Requirements

6.4.1.1 Data Storage Requirements

Organizations must procure and set up the storage infrastructure ensuring that every hosted federated node aligns with EUCAIM's data storage specifications. These needs, based on participation tiers, begin with storing the organization's contributed dataset and can be expanded to provide supplementary storage for localized data processing projects.

An **assumed** dataset size of **under 4 TB** is used as an example in the table below.

Hardware	Minimum	Recommended
Storage	2x4TB Raid	2x8TB Raid
Power Supply	While power supply of drives is minimal, it should be included in calculations if many drives are present, especially for server unit form factors.	

In this case, the provision and installation of a 2x 4TB drives enables:

I. The redundancy of the data through a RAID configuration (see section 6.2.3.2 for more information)

II. The storage of the dataset, which is assumed to be under 4 TB of the purposes of this example. This allows a minimum data compatibility with Tier 2 of EUCAIM's node integration plan.

In the case that the storage is expanded to 2x8TB drives

I. RAID redundancy of the data is preserved, with 8TB of storage available

II. The organization is able to duplicate parts of, or, the entirety of the dataset locally in order to perform tests, to prepare for anonymization, to restructure folders or any other data preparatory action.

III. The federated node can allocate parts of its storage for the execution of data processing jobs, the download and execution of EUCAIM tools through the tool catalogue, and to preserve data processing projects for a continuous period.

As such, it is suggested that data providers procure and provide redundancy for 2 or 3 times the storage required to store the federated dataset locally, as this will create flexibility for future activities and reinforce the capabilities of the EUCAIM federated infrastructure.

6.4.1.2 Processing Requirements - Data Federation - Tier 2 (See D4.3 for details)

Hardware	Minimum	Recommended
CPU	8C/16T 2.5Ghz+	16C/32T 3.0Ghz+
RAM	64 GB ECC	128 GB ECC
Power Supply	Depends on the selected CPU and form factor of the FN	

6.4.1.3 Processing Requirements - Data Processing - Tier 3

Hardware	Minimum	Recommended
CPU	16C/32T 2.5Ghz+	32C/64T 3.0Ghz+
GPU	>150 Tensor Cores, >16GB VRAM	>300 Tensor Cores, >32GB VRAM
RAM	128GB ECC	256GB ECC
Power Supply	Depends on the added power consumption of one or more GPUs	

6.4.2 Software Requirements

For the operation and integration of the federated node, organizations will have to:

I. Install a compatible operating system, it is recommended to install stable Linux distributions such as Ubuntu, CentOS, or Debian. These distributions have proven reliability and are compatible with most software stacks required for EUCAIM.

II. Install the necessary software to enable the download, deployment and management of the EUCAIM federated node components. Guidelines for installation of the required tools and the EUCAIM software stack are available in the online EUCAIM GitHub repository and are accessible to EUCAIM developers and integrators.

6.4.3 Infrastructure Configuration

6.4.3.1 Installation of Physical Infrastructure

All procured physical infrastructure (processing, electrical, or network units), essential for the federated node, should be securely positioned, safeguarded from external hazards and detrimental environments. Adequate precautions must be taken against potential risks like food, liquids, or cleaning agents. EUCAIM is producing relevant specifications with WP3 to address these issues.

In addition, physical infrastructure related to the operation of the federated node should reside in a restricted access zone, allowing entry only to approved individuals.

6.4.3.2 Network Configurations

Each federated node must be connected to the public internet via a wired connection. Relevant network infrastructural adjustments, such as firewall configurations should be made to enable specific network port inbound or outbound access to the public internet.

Organizations which use added security protocols (e.g., VPN, virtual, reverse proxy networks, packet monitoring), must notify and collaborate with EUCAIM's technical support team. This ensures that the federated node maintains robust internet connectivity while upholding EUCAIM's security and privacy guidelines.

6.4.3.3 Storage Configurations

Attached storage volumes must be configured by the organization in order to be used by the operating system for storage of data, installation of software etc. It is recommended to identify the purpose of each attached storage volume and create guidelines for naming the folder structure in order to facilitate the administration of the node's data and locally deployed software.

To ensure data integrity and accessibility, multi-layered redundancy at both the infrastructure and organisational levels is advised. An initial recommendation includes the deployment of a RAID configuration. Additionally, a comprehensive backup plan should be established to consistently safeguard the data's latest version.

6.4.3.4 Access Configurations

Physical infrastructure connected or associated with the federated node installation should reside in a restricted access area, permitting entry exclusively to authorized individuals. Any access events must be continually monitored and recorded.

To enhance security, individual user accounts should be created on the Linux machine for every authorized individual. This approach not only promotes accountability but also minimizes risks associated with shared access. For remote management, SSH access can be granted for authorized technical staff which should be provided with user credentials and SSH keys, ensuring encrypted and secure access. Regular audits are suggested in order to review access logs and ensure no unauthorized access attempts.

6.4.3.5 Operation and Monitoring of federated node

It is suggested that each organization monitors the health and performance of their hosted federated node using automation tools, which can be implemented to track metrics and set up alerts for anomalies in each metric. Organizations should prioritize regular data and configuration backups, timely system updates, and patch applications. For optimal long-term health of the federated node, organizations should schedule periodic maintenance and document any configurations and changes made to the federated node for future reference.

However, we should also state that some services of the federation will be monitored from the central node. The Federated Query endpoint for example will be monitored to ensure the availability of the site, notifying the provider in case of failure.

6.4.4 Technical Support & Administration

The technical support for Federated Nodes will be a joint effort from the EUCAIM technical support team, the EUCAIM tool support teams and the available and authorized technical staff at each data site. This technical support will also include guidelines provided, through defined channels, for the procurement, installation, configuration and operation of the Federated Nodes. Furthermore, technical administration will be provided by EUCAIM's leading partner for integration, as necessary installation & integrations of infrastructure software (Kubernetes, Docker, monitoring etc.) will require configurations to be applied as they have been defined through the project's activities and through iteration of the integration team's deployment strategy. Reference implementations of the different required tools and infrastructures will be made available.

It is worth stating also, that the project intends to provide appropriate support, through its helpdesk (help.cancerimage.eu). More information on the processes of support and inter-organizational communication are presented in the following section.

6.5 Data Exchange and Communication Protocols

6.5.1 General Channels of Communication

During the early stages of the deployment of Federated Nodes, it is expected that communication with data sites/organizations will be initially performed through email (and the helpdesk when available), as partners get to know each other and build a consensus for the required activities.

For any information that is not directly related to the Federated Nodes and for general administration and coordination efforts, emails will remain a good way to maintain a history of conversations and share information openly with partners.

Additionally, relevant official documentation will be created for the purposes of creating a common understanding between partners, data sites and their managing organizations. This documentation will take multiple forms, such as:

- The EUCAIM glossary, which aims to align partner's understandings of specific terms. This will be a vital resource to streamline coordination and communication between partners of various types, especially considering the large and disparate nature of EUCAIM's consortium.
- The EUCAIM Federated Node Guidelines will be assembled at a later stage as a specific document to be shared with potential Data sites interested in joining the EUCAIM infrastructure. The initial contents of these guidelines are defined within this document, and, as the various technical strategies mature and evolve over the project's duration, they will be refined, validated, and then compiled into the referenced guidelines document.

6.5.2 Secure channels of communication

Following these early steps, the EUCAIM Helpdesk will be the official channel of communication for any topic that directly regards the Federated Nodes, their status and configurations. Any information that is shared regarding the Federated Nodes should be considered confidential by default, and only shared over trusted channels.

6.5.3 Data Exchange for Credentials

The effective and secure management of credentials is crucial in managing the security of federated nodes. Partners should always exchange credentials over secure communication methods and avoid the use of emails or standard messaging platforms. Instead, it is recommended to at the very least make use of messaging channels or services, which guarantee the encryption of messages sent.

A project-wide strategy for exchange of credentials is being researched, since solutions such as a public key infrastructure (PKI), key management solutions, or enforced multi-factor authentication, while hard to implement, enforce and maintain, could provide the necessary layers of security.

6.5.4 Data Exchange for Files

In the cases that the transfer of data is required either to or from the federated node, strict measures should be adopted when planning and executing the transfer. It is necessary to make use of secure transfer protocols such as SFTP (Secure File Transfer Protocol) or SCP (Secure Copy Protocol), which encrypt files during transfer. Furthermore, it is recommended for the data to be catalogued before transfer and to perform data-validation techniques such as checksum or hashing in order to verify their integrity post-transfer. In the cases that data must be transferred outside the host organization’s network it is highly recommended to make use of VPN (Virtual Private Network) tunnel technologies between the origin and the destination infrastructures in order to preserve the security and privacy of data during transit.

7. Demonstration Scenario

EUCAIM has committed to deliver in PM9 a first version of the platform. This platform will be a first proof of concept of the services of the platform and despite having limited functionality, it will serve as a basis to understand, test, expand and build the further versions of the platform.

The objective of this first release is:

- Deploy the main core services that have been implemented and customized for EUCAIM in a cloud-based infrastructure.
- Create a registry of the data available at the moment of the release of the prototype, from the AI4HI projects and other collaborative activities.
- Expose the tangible assets of EUCAIM's platform so the consortium understands the functionality better.
- Create awareness and raise interest from external providers and requesters.
- Identify the main difficulties raised when deploying complex tools across the constellation of data nodes.

The platform incorporates:

- A Dashboard with a personal area supporting LS AAI.
- A Catalogue with 35 dataset from AI4HI registered (>200K series of images).
- A Federated Query engine with limited functionality.
- An Access Negotiation service.
- A Helpdesk Service.

The current limitations are:

- Federated query functionality is limited in the number of providers and in its functionality.
- Dashboard integrates a basic functionality for the "My library" area.
- Negotiator system is not automatically integrated.
- The AAI supports LS login, but as separate components.
- The Distributed processing is a non-integrated proof of concept.
- Access requests depend on the providers' side (agreements and on their way).
- At the moment, the federated processing orchestration is not linked to the main platform. However, an initial orchestration and Federated learning experiment has been performed for demonstration purposes.

The components of the platform are described in the following subsections.

7.1. Core services infrastructure

The core services of EUCAIM run in a Kubernetes cluster deployed on an on-premise cloud at the UPV. This cluster comprises a front-end and three nodes with a total of 28 cores and 64 GB RAM. The services in the infrastructure have a network disk for the persistent layer and each core service is deployed in a separate namespace, so operation can be distributed. All the applications are described as IaC recipes to guarantee reproducibility. Services run as Kubernetes deployments for higher high-availability.

Access to the services is provided through ingress. Any domain in the form *.eucaim.cancerimage.eu is redirected to a nginx proxy in the front end of the cluster. The services are uniquely identified in the following addresses:

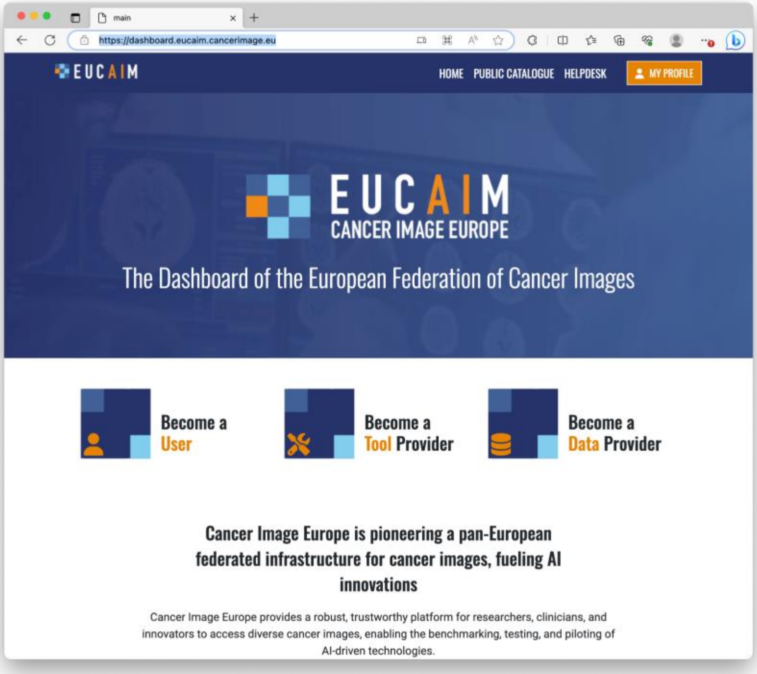
- dashboard.eucaim.cancerimage.eu. The Dashboard of the platform, main entry point for the platform.

- catalogue.eucaim.cancerimage.eu. The public catalogue of the platform, containing the collections registered in EUCAIM.
- explorer.eucaim.cancerimage.eu. The federated search application.
- negotiator.eucaim.cancerimage.eu. The access negotiation application.

7.2. Dashboard

The Dashboard is the entry page for the platform. It links the catalogue and the general pages for the providers, as well as the link to the LS AAI and the information pages.

The Dashboard is available in https://dashboard.eucaim.cancerimage.eu/. It contains links to the HelpDesk and the Catalogue, as well as the login service. Figure 29 shows a snapshot of the Dashboard.



The Dashboard of the European Federation of Cancer Images

Cancer Image Europe is pioneering a pan-European federated infrastructure for cancer images, fueling AI innovations

Cancer Image Europe provides a robust, trustworthy platform for researchers, clinicians, and innovators to access diverse cancer images, enabling the benchmarking, testing, and piloting of AI-driven technologies.

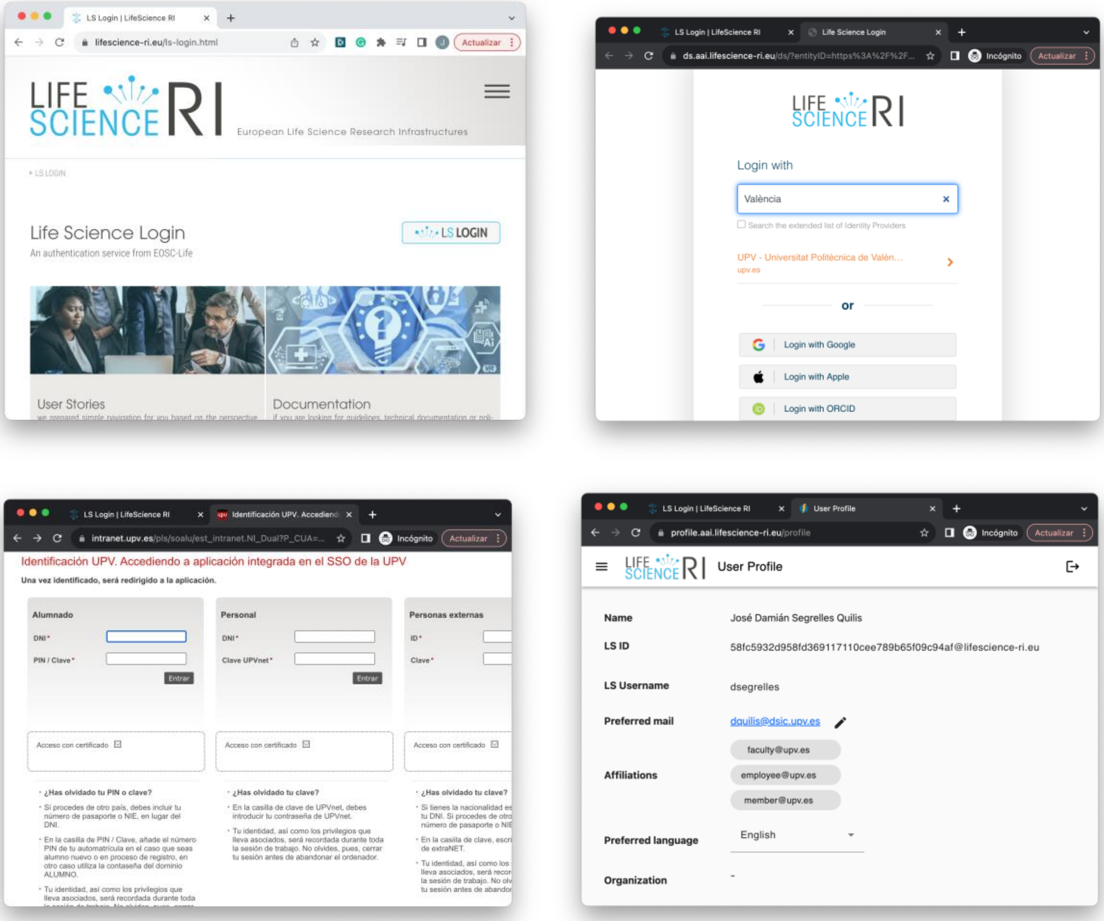
Figure 29: Snapshot of EUCAIM's Dashboard. Links to the Information pages, helpdesk and catalogue are included, as well as to the login page

In order to access the internal pages, the users have to register in the Life Science Login AAI. For this purpose, a user has to go through the following steps:

- Create an account in the LS Login[97], linked to their institutional Identity Provider (Figure 30).
- Request the membership to the group of EUCAIM through the Help Desk system. The requests will be managed by EUCAIM's Group manager, who will add the user to the proper group (Figure 31).
- Once a user has a valid account, she can access the login pages from the Dashboard. The user has to access the dashboard login button and accept the acceptable users' policy. The dashboard will reflect the users' name (Figure 32)

[97] https://lifescience-ri.eu/ls-login.html

The membership to EUCAIM will only entitle the user to perform federated searching, request access to data and have an internal area.


LS Login | LifeScience RI
lifescience-ri.eu/ls-login.html
European Life Science Research Infrastructures
LS LOGIN
User Stories
Documentation
Life Science Login
ds.aai.lifescience-ri.eu/ds/?entityID=https%3A%2F%2F...
Login with Google
Login with Apple
Login with ORCID
Identificación UPV. Accediend
intranet.upv.es/pls/soalu/est_intranet.NI_Dual?P_CUA=...
Entrar
Entrar
User Profile
profile.aai.lifescience-ri.eu/profile
Incógnito
Actualizar

Life Science Login

An authentication service from EOSC-Life

Login with

València

□ Search the extended list of Identity Providers

UPV - Universitat Politècnica de Valèn...
upv.es

or

Identificación UPV. Accediendo a aplicación integrada en el SSO de la UPV

Una vez identificado, será redirigido a la aplicación.

Alumnado
DNI*
PIN / Clave*

Personal
DNI*
Clave UPVnet*

Personas externas
ID*
Clave*

Acceso con certificado

- ¿Has olvidado tu PIN o clave?
- Si procedes de otro país, debes incluir tu número de pasaporte o NIE, en lugar del DNI.
- En la casilla de PIN / Clave, añade el número PIN de tu automatrícula en el caso que seas alumno nuevo o en proceso de registro, en otro caso utiliza la contaseña del dominio ALUMNO.
- Tu identidad, así como los privilegios que lleva asociados, será recordada durante toda

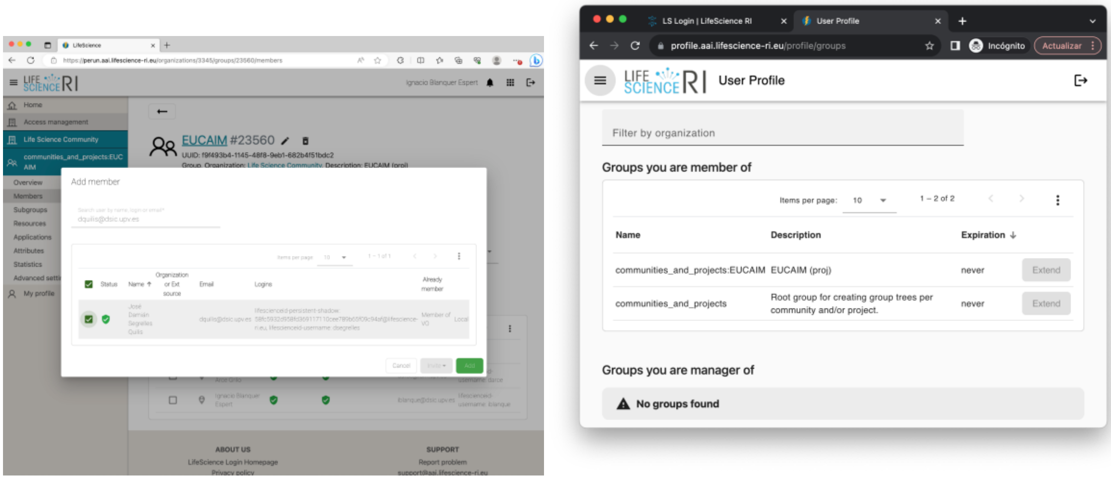
- ¿Has olvidado tu clave?
- En la casilla de clave de UPVnet, debes introducir tu contraseña de UPVnet.
- Tu identidad, así como los privilegios que lleva asociados, será recordada durante toda la sesión de trabajo. No olvides, pues, cerrar tu sesión antes de abandonar el ordenador.

- ¿Has olvidado tu clave?
- Si tienes la nacionalidad es tu DNI. Si procedes de otro número de pasaporte o NIE
- En la casilla de clave, escri de extraNET.
- Tu identidad, así como los lleva asociados, será recor la sesión de trabajo. No olv tu sesión antes de abandor

User Profile

Name	José Damián Segrelles Quilis
LS ID	58fc5932d958fd369117110cee789b65f09c94af@lifescience-ri.eu
LS Username	dsegrelles
Preferred mail	dquilis@dsic.upv.es
Affiliations	faculty@upv.es employee@upv.es member@upv.es
Preferred language	English
Organization	-

Figure 30: Process of creating an account. Access Life Sciences Login page (top left); login the RI and select the user's Identity Provider from the list of trusted entities (top right); login with the institutional IdP service (bottom left); and check the account details (bottom right)



EUCAIM #23560

UUID: f9f493b4-1145-48f8-9eb1-682b4f51bdc2

Group. Organization: Life Science Community. Description: EUCAIM (proj)

Add member

Search user by name, login or email*: dquilis@dsic.upv.es

Items per page: 10 1 – 1 of 1

☑	Status	Name	Organization or Ext source	Email	Logins	Already member	
☑	✓	José Damián Segrelles Quilis		dquilis@dsic.upv.es	lifescienceid-persistent-shadow: 58fc5932d958fd369117110cee789b65f09c94af@lifescience-ri.eu, lifescienceid-username: dsegrelles	Member of VO	Local

User Profile

Filter by organization

Groups you are member of

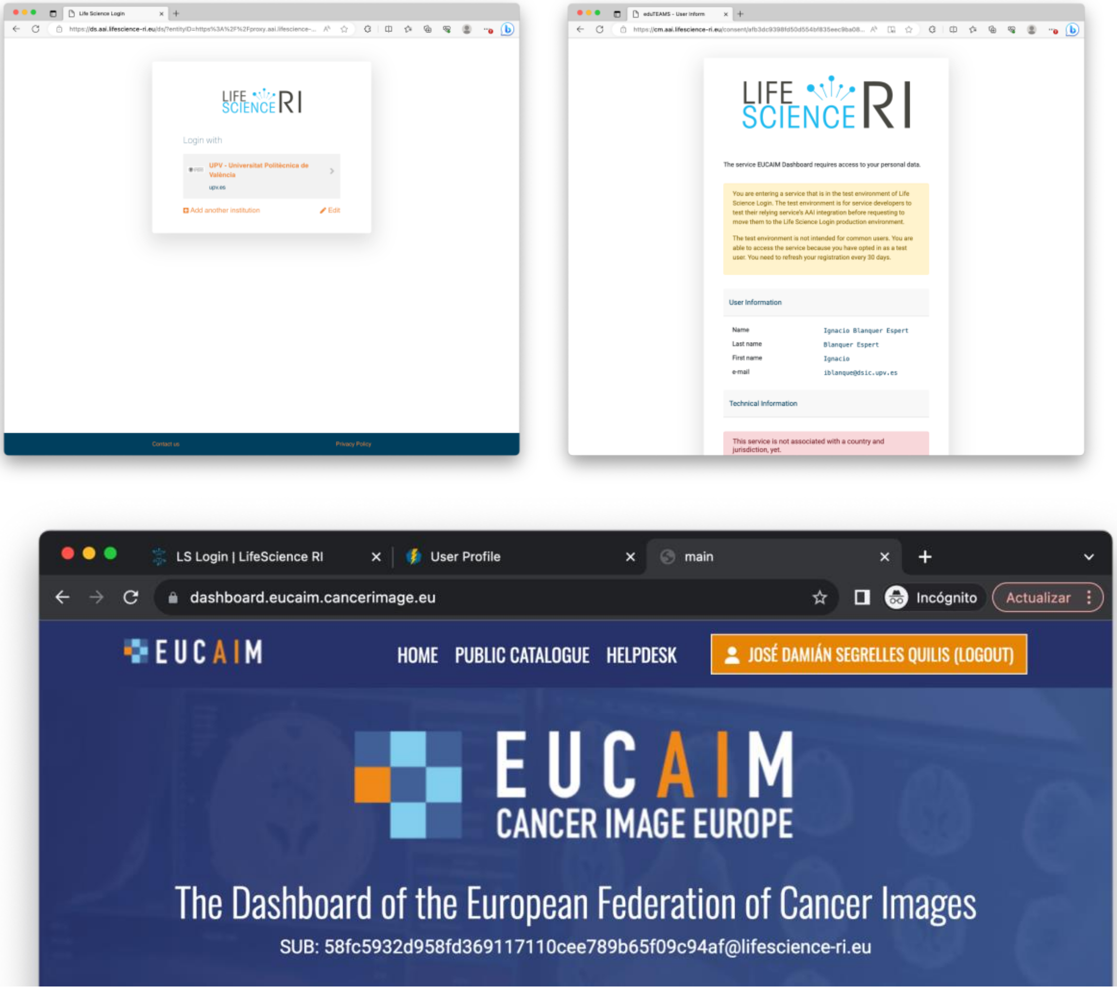
Items per page: 10 1 – 2 of 2

Name	Description	Expiration	
communities_and_projects:EUCAIM	EUCAIM (proj)	never	Extend
communities_and_projects	Root group for creating group trees per community and/or project.	never	Extend

Groups you are manager of

⚠ No groups found

Figure 31: Adding the user to EUCAIM's group (left) from the administrative console and checking the group permissions from the users' account panel (right).



Life Science Login
https://ds.aai.lifescience-ri.eu/ds/?entityID=https%3A%2F%2Fproxy.aai.lifescience-...
LIFE SCIENCE RI
Login with
UPV - Universitat Politècnica de València
upv.es
Add another institution
Edit
Contact us
Privacy Policy
eduTEAMS - User Inform
https://cm.aai.lifescience-ri.eu/consent/afb3dc9398fd50d554bf835eec9ba08...
LIFE SCIENCE RI
LS Login | LifeScience RI
User Profile
main
dashboard.eucaim.cancerimage.eu
Incógnito
Actualizar
EUCAIM
HOME
PUBLIC CATALOGUE
HELPDESK
JOSÉ DAMIÁN SEGRELLES QUILIS (LOGOUT)
EUCAIM
CANCER IMAGE EUROPE

The service EUCAIM Dashboard requires access to your personal data.

You are entering a service that is in the test environment of Life Science Login. The test environment is for service developers to test their relying service's AAI integration before requesting to move them to the Life Science Login production environment.

The test environment is not intended for common users. You are able to access the service because you have opted in as a test user. You need to refresh your registration every 30 days.

User Information

Name	Ignacio Blanquer Espert
Last name	Blanquer Espert
First name	Ignacio
e-mail	iblanque@dsic.upv.es

Technical Information

This service is not associated with a country and jurisdiction, yet.

The Dashboard of the European Federation of Cancer Images

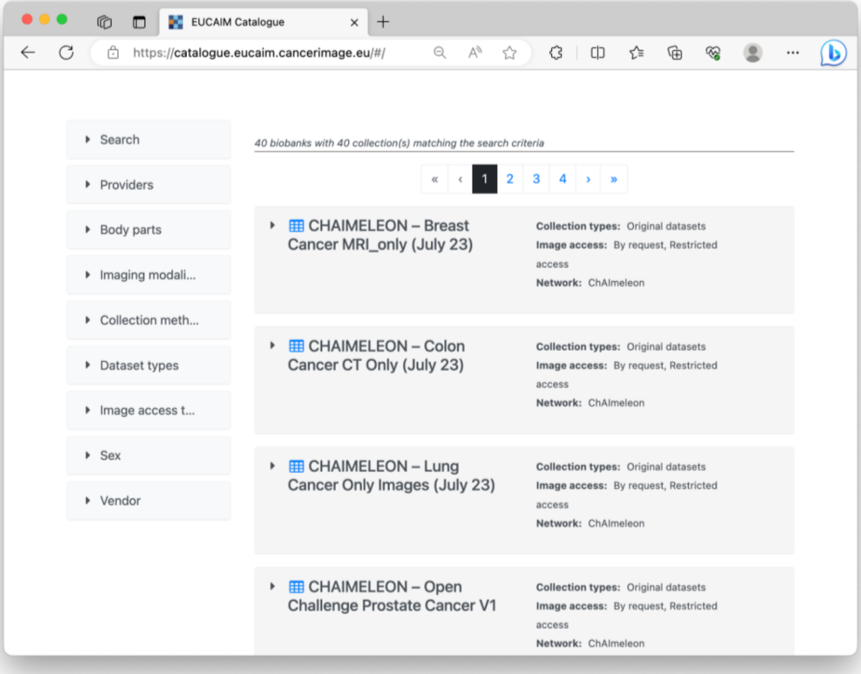
SUB: 58fc5932d958fd369117110cee789b65f09c94af@lifescience-ri.eu

Figure 32: Once a user has clicked on the login button, the user will has to select the preferred IdP and login (top left). After successfully login in the system, the user has to accept the access policy (top right) and the Dashboard will show her name (bottom)

7.3. Catalogue

The catalogue of EUCAIM is based on Molgenis. A customized version has been deployed and populated with the ontology data and the datasets from the AI4HI. In total, the 40 datasets registered include 20.000 subjects and more than 200.000 image series from 9 cancer types (breast, colon, lung, prostate, rectum, liver, glioma, neuroblastoma, glioblastoma). A snapshot of the catalogue application is shown in Figure 33.

The catalogue has a set of options to filter the collections that fit specific criteria. The information is obtained from the dataset metadata. The datasets are organized by biobanks, and biobanks by providers. The items of the metadata for both the collections and the biobanks are described in Figure 34.



EUCAIM Catalogue

https://catalogue.eucaim.cancerimage.eu/#/

- Search
- Providers
- Body parts
- Imaging modali...
- Collection meth...
- Dataset types
- Image access t...
- Sex
- Vendor

40 biobanks with 40 collection(s) matching the search criteria

« ‹ 1 2 3 4 › »

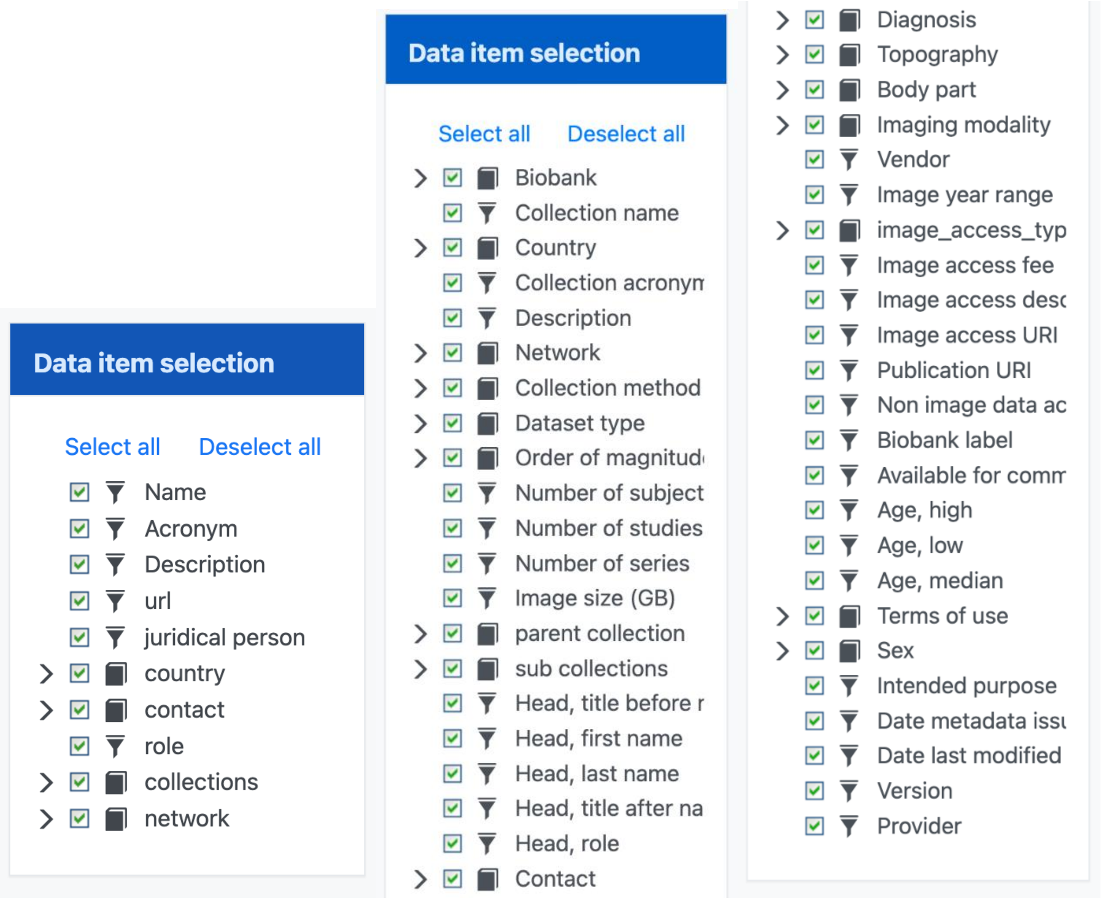
CHAIMELEON – Breast Cancer MRI_only (July 23)
Collection types: Original datasets
Image access: By request, Restricted access
Network: ChAImeleon

CHAIMELEON – Colon Cancer CT Only (July 23)
Collection types: Original datasets
Image access: By request, Restricted access
Network: ChAImeleon

CHAIMELEON – Lung Cancer Only Images (July 23)
Collection types: Original datasets
Image access: By request, Restricted access
Network: ChAImeleon

CHAIMELEON – Open Challenge Prostate Cancer V1
Collection types: Original datasets
Image access: By request, Restricted access
Network: ChAImeleon

Figure 33: Public Catalogue of EUCAIM



Data item selection

Select all Deselect all

- ☑ Name
- ☑ Acronym
- ☑ Description
- ☑ url
- ☑ juridical person
- › ☑ country
- › ☑ contact
- ☑ role
- › ☑ collections
- › ☑ network

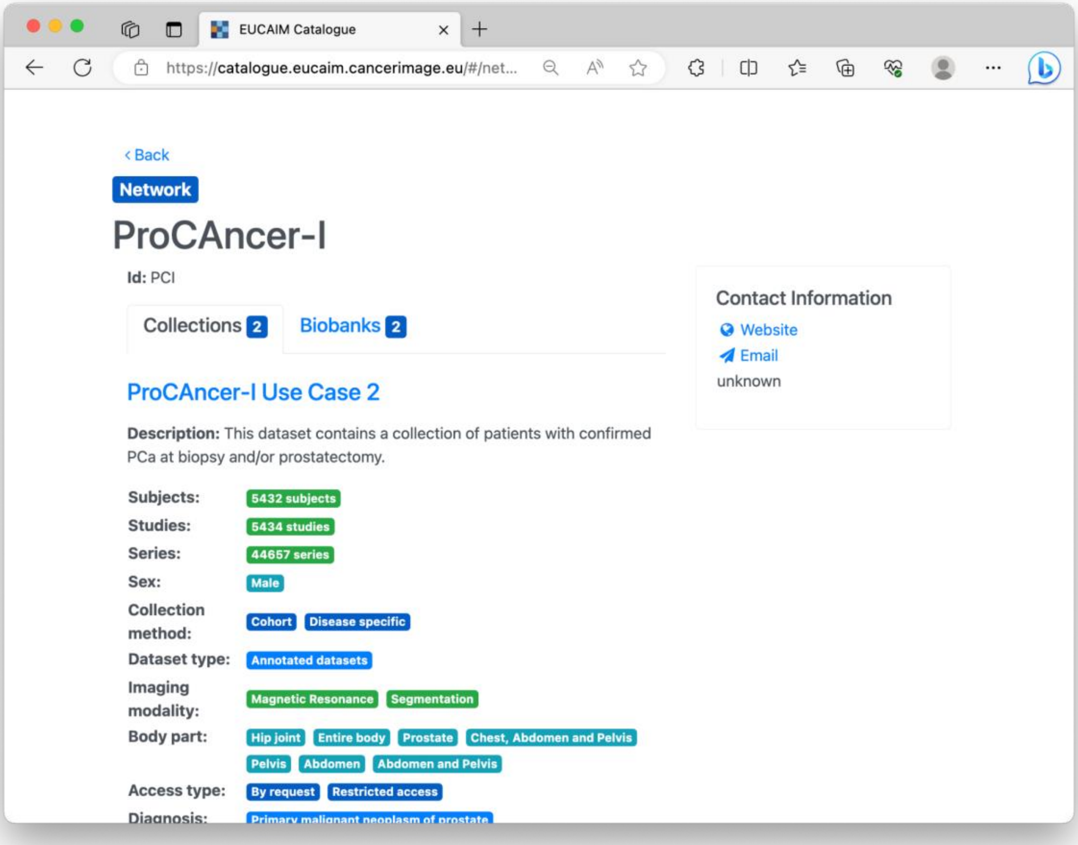
Data item selection

Select all Deselect all

- › ☑ Biobank
- ☑ Collection name
- › ☑ Country
- ☑ Collection acronym
- ☑ Description
- › ☑ Network
- › ☑ Collection method
- › ☑ Dataset type
- › ☑ Order of magnitud
- ☑ Number of subject
- ☑ Number of studies
- ☑ Number of series
- ☑ Image size (GB)
- › ☑ parent collection
- › ☑ sub collections
- ☑ Head, title before r
- ☑ Head, first name
- ☑ Head, last name
- ☑ Head, title after na
- ☑ Head, role
- › ☑ Contact
- › ☑ Diagnosis
- › ☑ Topography
- › ☑ Body part
- › ☑ Imaging modality
- ☑ Vendor
- ☑ Image year range
- › ☑ image_access_typ
- ☑ Image access fee
- ☑ Image access desc
- ☑ Image access URI
- ☑ Publication URI
- ☑ Non image data ac
- ☑ Biobank label
- ☑ Available for comm
- ☑ Age, high
- ☑ Age, low
- ☑ Age, median
- › ☑ Terms of use
- › ☑ Sex
- ☑ Intended purpose
- ☑ Date metadata issu
- ☑ Date last modified
- ☑ Version
- ☑ Provider

Figure 34: Metadata for the Biobanks (left) and the collections (right)

Information about a collection can be obtained by selecting the collection. By clicking on the "collection network", all the biobanks and collections of an institution are listed. Figure 35 shows the network of the provider ProCAncer-I.



‹ Back

Network

ProCAncer-I

Id: PCI

Collections 2 | Biobanks 2

Contact Information

Website

Email

unknown

ProCAncer-I Use Case 2

Description: This dataset contains a collection of patients with confirmed PCa at biopsy and/or prostatectomy.

Subjects:	5432 subjects
Studies:	5434 studies
Series:	44657 series
Sex:	Male
Collection method:	Cohort, Disease specific
Dataset type:	Annotated datasets
Imaging modality:	Magnetic Resonance, Segmentation
Body part:	Hip joint, Entire body, Prostate, Chest, Abdomen and Pelvis, Pelvis, Abdomen, Abdomen and Pelvis
Access type:	By request, Restricted access
Diagnosis:	Primary malignant neoplasm of prostate

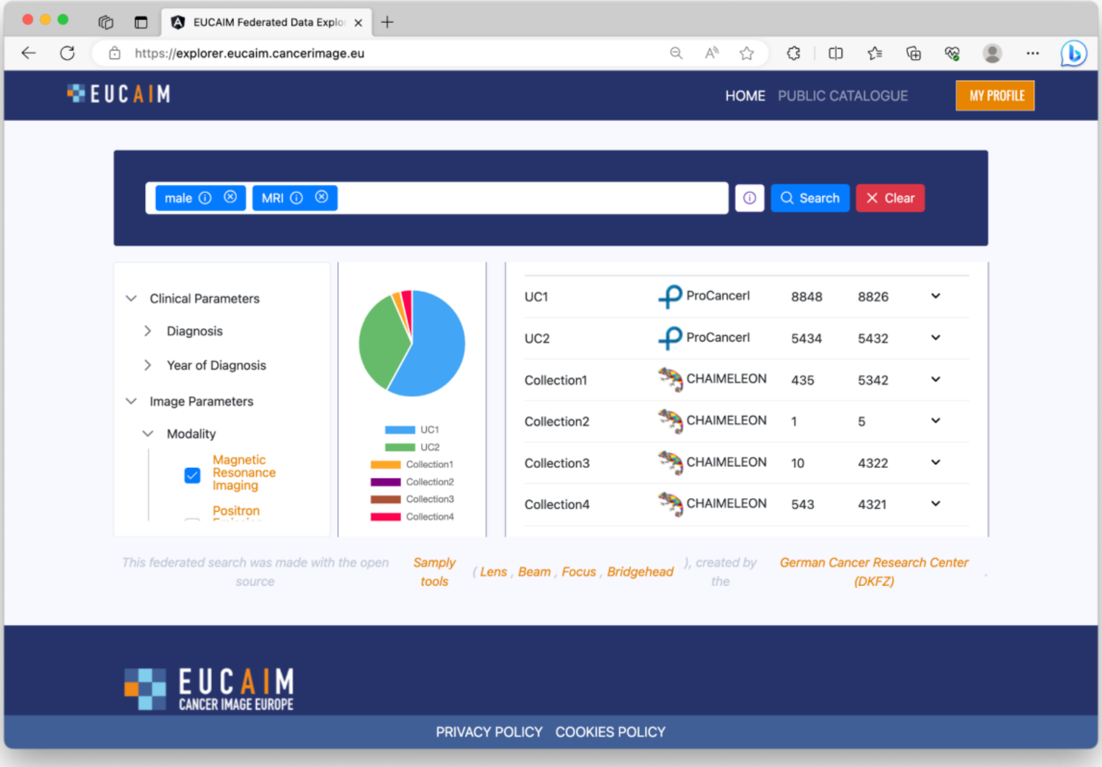
Figure 35: Provider's network view

7.4. Federated Query

The federated query is a service that enables users to retrieve the number of cases that fulfil a specific searching criteria. The information of the catalogue can only show the datasets that have cases matching the filtering criteria, but it does not guarantee that the dataset will have a case that will match all of them. For example, the metadata of a dataset could indicate that it comprises lung cancer CT and MRI images for both male and female, but it may be possible that males have only CT images, and this could not be discovered unless actual data is explicitly queried. The federated query provides this functionality.

The federated query service interacts with an agent at the provider side (the mediator) that transforms the queries expressed in the hyperontology terms to the actual format of the provider. This has been implemented for both CHAIMELEON and ProCAncer-I OMOP-based repositories. However, CHAIMELEON uses static data for the time being in comparison to the ProCAncer-I node that returns answers based on real data of its registered datasets.

The federated query has a similar appearance as the filtering component of the catalogue, providing more detailed information. A snapshot of the application is shown in Figure 36.



UC1	ProCancerI	8848	8826	⌄
UC2	ProCancerI	5434	5432	⌄
Collection1	CHAIMELEON	435	5342	⌄
Collection2	CHAIMELEON	1	5	⌄
Collection3	CHAIMELEON	10	4322	⌄
Collection4	CHAIMELEON	543	4321	⌄

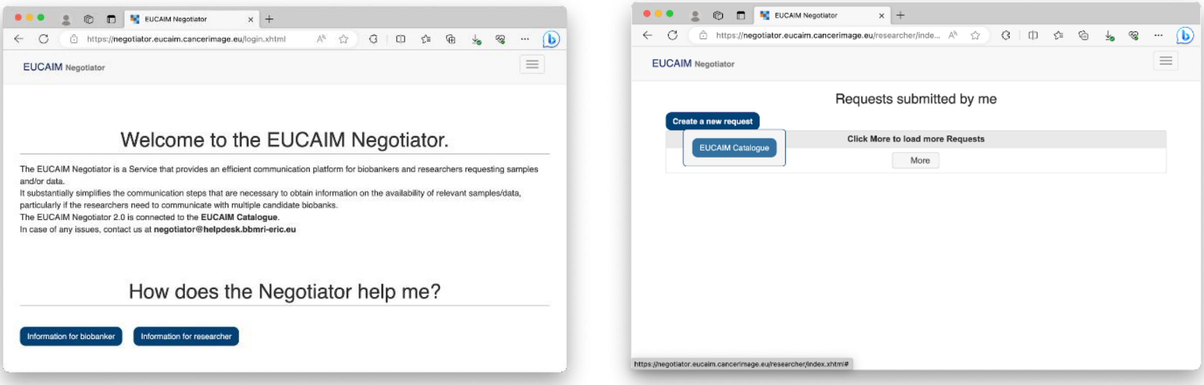
This federated search was made with the open source Samply tools (Lens , Beam , Focus , Bridgehead), created by the German Cancer Research Center (DKFZ) .

Figure 36: Federated Query snapshot after searching for studies of males with MRI

The access to the Federated query is enabled in the Dashboard for authenticated users.

7.5. Negotiator

Although access to a dataset will require direct contact with the provider at this point in time of the project, a negotiator service has been deployed. This service will manage the life cycle of requests and it will be triggered directly from the public catalogue, only for authenticated users. The service is linked to the internal section of the Dashboard, and available under the URL https://negotiator.eucaim.cancerimage.eu. This service requires a valid LS AAI user, and provides the user with a link back to the catalogue. Figure 37 shows the snapshots of the service.


EUCAIM Negotiator
https://negotiator.eucaim.cancerimage.eu/login.xhtml
EUCAIM Negotiator
Information for biobanker
Information for researcher
https://negotiator.eucaim.cancerimage.eu/researcher/inde...
Create a new request
EUCAIM Catalogue
More
https://negotiator.eucaim.cancerimage.eu/researcher/index.xhtml#

Welcome to the EUCAIM Negotiator.

The EUCAIM Negotiator is a Service that provides an efficient communication platform for biobankers and researchers requesting samples and/or data.
It substantially simplifies the communication steps that are necessary to obtain information on the availability of relevant samples/data, particularly if the researchers need to communicate with multiple candidate biobanks.
The EUCAIM Negotiator 2.0 is connected to the **EUCAIM Catalogue**.
In case of any issues, contact us at **negotiator@helpdesk.bbmri-eric.eu**

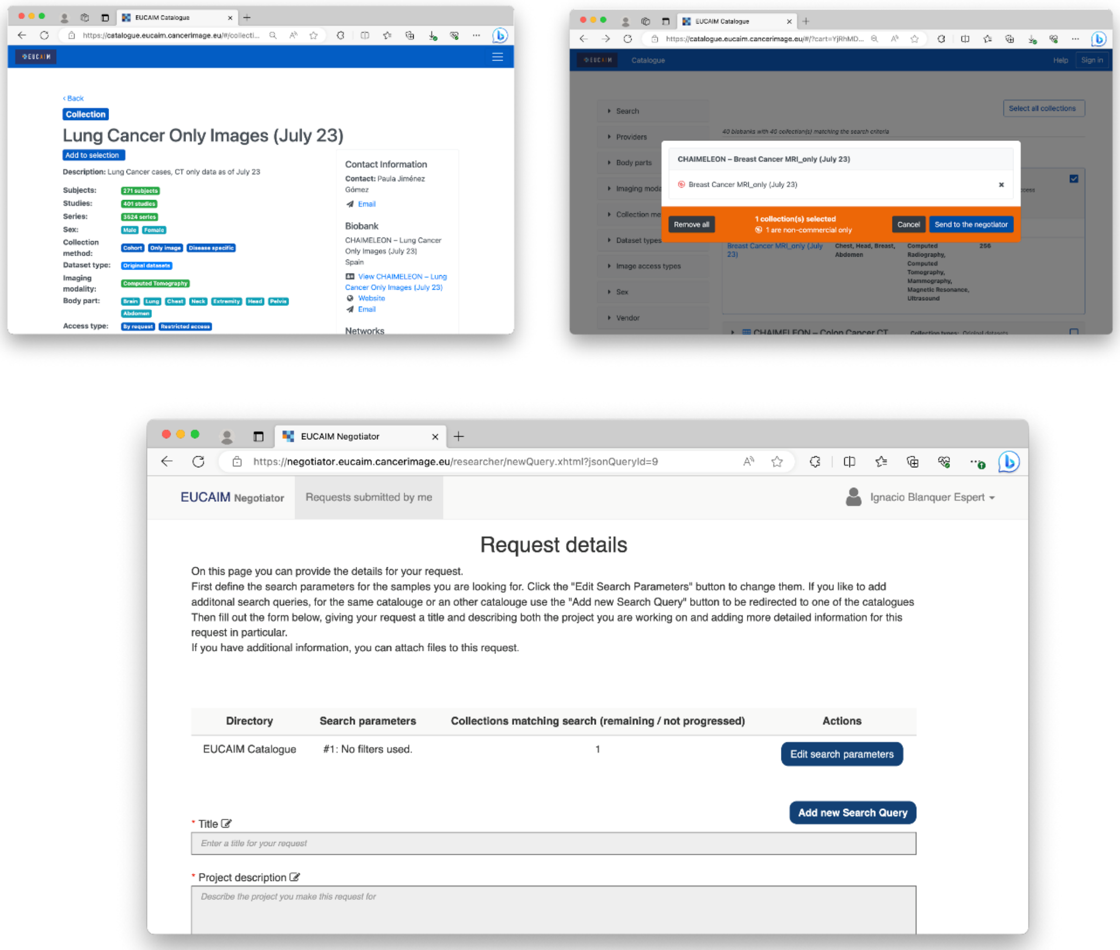
How does the Negotiator help me?

Requests submitted by me

Click More to load more Requests

Figure 37: Snapshots of the Negotiator, before and after logging into the platform

By accessing the catalogue from the negotiator, a user can select a dataset and add it to the negotiator request area. The process of requesting access will be available by the end of October 2023. Figure 38 shows a snapshot of the catalogue with the “Add to negotiator” button activated and a pop-up window to enable sending the requests to the negotiator.



Lung Cancer Only Images (July 23)

Description: Lung Cancer cases, CT only data as of July 23

Subjects:	271 subjects
Studies:	401 studies
Series:	3524 series
Sex:	Male Female
Collection method:	Cohort Only image Disease specific
Dataset type:	Original datasets
Imaging modality:	Computed Tomography
Body part:	Brain Lung Chest Neck Extremity Head Pelvis Abdomen
Access type:	By request Restricted access

Contact Information

Contact: Paula Jiménez Gómez

Email

Biobank

CHAIMELEON – Lung Cancer Only Images (July 23)
Spain

View CHAIMELEON – Lung Cancer Only Images (July 23)
Website
Email

Networks

40 biobanks with 40 collection(s) matching the search criteria

CHAIMELEON – Breast Cancer MRI_only (July 23)

Breast Cancer MRI_only (July 23)

1 collection(s) selected
1 are non-commercial only

Breast Cancer MRI_only (July 23) | Chest, Head, Breast, Abdomen | Computed Radiography, Computed Tomography, Mammography, Magnetic Resonance, Ultrasound | 256

CHAIMELEON – Colon Cancer CT | Collection types: Original datasets

Request details

On this page you can provide the details for your request.
First define the search parameters for the samples you are looking for. Click the "Edit Search Parameters" button to change them. If you like to add additonal search queries, for the same catalogue or an other catalogue use the "Add new Search Query" button to be redirected to one of the catalogues
Then fill out the form below, giving your request a title and describing both the project you are working on and adding more detailed information for this request in particular.
If you have additional information, you can attach files to this request.

Directory	Search parameters	Collections matching search (remaining / not progressed)	Actions
EUCAIM Catalogue	#1: No filters used.	1	Edit search parameters

* Title
Enter a title for your request

* Project description
Describe the project you make this request for

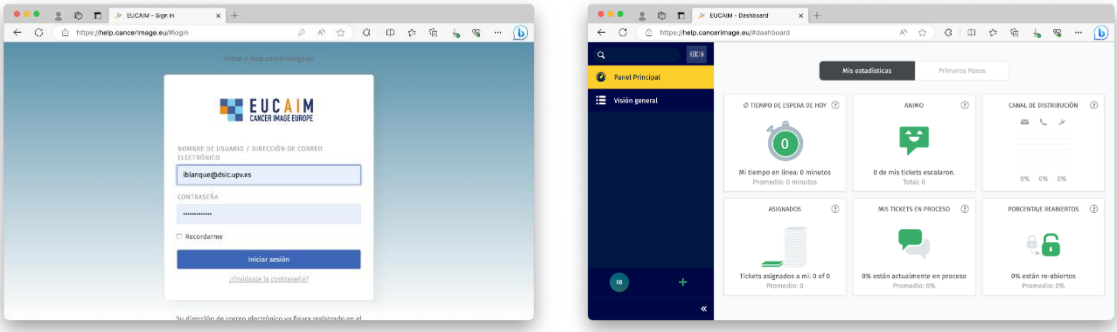
Figure 38: Snapshots of a Dataset page in the catalogue (top left) with the “Add to negotiator” button enabled and the list of selected datasets (top right). Submission form when sending the request to the negotiator (bottom).

7.6 Helpdesk

The helpdesk is the main communication channel of the users and providers with the support team of the platform. It will be used to manage:

- Incidences of the services.
- Bugs or new requirements.
- Requests of memberships.
- Any other formal communication that is subject to auditing.

The Helpdesk platform is deployed externally to the core services and it is accessible in the URL https://help.cancerimage.eu/#login, although it is linked in the main page of the Dashboard. The authentication system is currently being migrated to LS AAI, so users authenticated in the dashboard will seamlessly access the help desk. The users will then be able to create and follow their requests (tickets) through the interface. A snapshot is provided in Figure 39.



iblanque@dsic.upv.es

□ Recordarme

Su dirección de correo electrónico ya figura registrado en el

0

Mi tiempo en línea: 0 minutos
Promedio: 0 minutos

0 de mis tickets escalaron.
Total: 0

0% 0% 0%

Tickets asignados a mi: 0 of 0
Promedio: 0

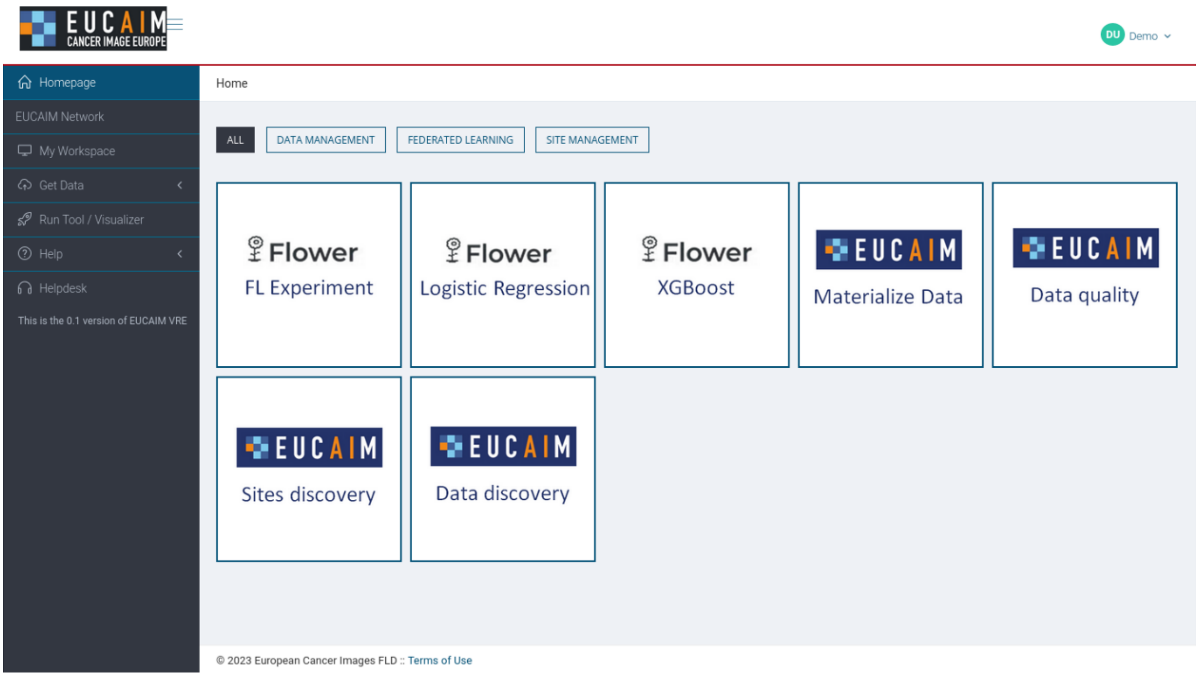
0% están actualmente en proceso
Promedio: 0%

0% están re-abiertos
Promedio: 0%

Figure 39: Snapshot of the help desk system. Login page (left) and user's panel (right)

7.7 Federated processing

The federated processing orchestrator is a service providing an API to start, pause, stop, and retrieve the results of analytical experiments. The tool, providing an API, will be integrated in a dashboard based on BSC's openVRE platform (see figure 40 for a screenshot of the initial prototype).



This is the 0.1 version of EUCAIM VRE

© 2023 European Cancer Images FLD :: Terms of Use

Figure 40: Federated processing's orchestrator dashboard

The orchestrator can raise multiple tools and software pieces packaged as docker containers and trigger their execution at selected EUCAIM sites. Tools would include analysis but also federated learning platforms. At the current moment, and still detached from the main platform, we deployed three federated learning platforms (**UB's Flower stack, Substra, and FedBio-Med**) on three data nodes (**UB, BSC, FORTH**). For demonstration, we have prepared two models for diagnosis using clinical and image datasets to be trained on a federated environment[98]. The goal was to run the same two experiments using the three platforms and gather the resulting models for future exploration of similarities and differences, as well as to collect any issues on deploying the platforms and running the experiments. The datasets used were from the public domain and divided in third to be distributed into each data node.

[98] https://github.com/EUCAIM/fl_demonstrator

Table 11: Deployment of federated learning platforms according to data nodes

Data node / Platform	**UB's**	**Substra**	**FedBio-Med**
UB	Yes	Yes	Yes
BSC	Yes	Yes	Yes
FORTH	Yes	No	Yes

8. Conclusion

This deliverable presents the first release of the Data Federation Framework of the EUCAIM project. It represents a major achievement, presenting an initial proof-of-concept for the EUCAIM high-level Architecture and Data Federation Framework, showcasing how it can potentially facilitate the seamless exchange of federated imaging and clinical data for cancer research. This document provides a thorough overview of the key functionalities and contributions of this early version, providing crucial initial insights into the EUCAIM Data Federation Framework.

It begins with a high-level description of the architecture of the EUCAIM federation presenting the various components, i.e. the dashboard, the AAI, the public catalogue, the federated query, access and processing as well as the central repository. Then it focuses on selecting a common data model adopted for the EUCAIM documenting that in essence both FHIR and OMOP-CDM are supported, however with a focus on the second one. Further, the first version of the hyper-ontology is presented showing also how to integrate CDM with the hyper-ontology and how this is linked with the metadata model of the federated catalogue. Then, an overview is presented on protocols, formats and terminologies in the domain of clinical and imaging data that was explored.

The data preprocessing and (meta)data management and interoperability tools/services are presented in the sequel followed by an initial technical specification for the set-up of the EUCAIM federated nodes. More specifically the architecture of the federated nodes is presented, as well as the hardware and software requirements for node setup and deployment. Data exchange and communication protocols of the federated nodes is outlined explaining how the federated nodes will be interoperable with the whole EUCAIM infrastructure. Besides interoperability of the federated nodes, interoperability with other European initiatives is also explored by the use of Beacon for images. Finally, a demonstration scenario is outlined and presented showing how all components are currently implemented and working. This platform will be a first proof of concept of the services of the platform and despite it will have limited functionality, it will serve as a basis to understand, test, expand and build the subsequent versions of the platform.

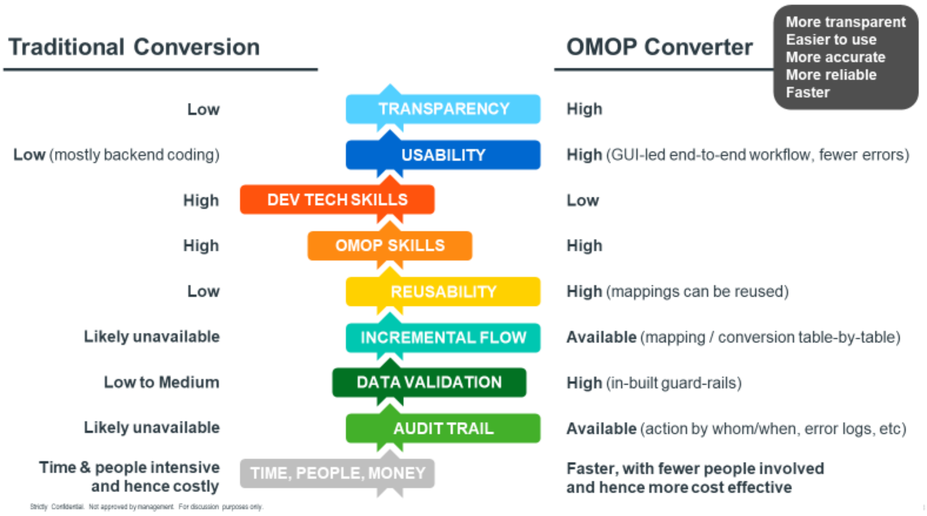
Concluding, we have to note that the current document depicts the current status of the EUCAIM data federation framework, which is evolving, continuously updated and improved. The final version will be subsequently reported at dedicated future deliverables.

Annexes

Annex A: OMOP Converter

OMOP Converter is a software solution that automates the conversion to OMOP format, only requiring users to import their data, complete the mapping fields, check the results and automatically export to OMOP format, all in the same graphical solution, which requires little technical knowledge.

The solution can be deployed on-premises or in the cloud. Compared to traditional conversions, OMOP Converter makes the process more transparent, reusable, reliable and cost-effective, reducing the cost of maintaining high quality OMOP datasets. Advantages of using:



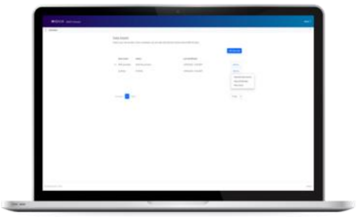
Traditional Conversion		OMOP Converter
Low	TRANSPARENCY	High
Low (mostly backend coding)	USABILITY	High (GUI-led end-to-end workflow, fewer errors)
High	DEV TECH SKILLS	Low
High	OMOP SKILLS	High
Low	REUSABILITY	High (mappings can be reused)
Likely unavailable	INCREMENTAL FLOW	Available (mapping / conversion table-by-table)
Low to Medium	DATA VALIDATION	High (in-built guard-rails)
Likely unavailable	AUDIT TRAIL	Available (action by whom/when, error logs, etc)
Time & people intensive and hence costly	TIME, PEOPLE, MONEY	Faster, with fewer people involved and hence more cost effective

More transparent
Easier to use
More accurate
More reliable
Faster

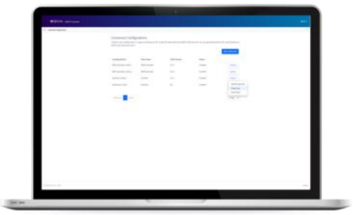
Strictly Confidential. Not approved by management. For discussion purposes only.

OMOP Converter emulates the workflow of a traditional conversion project, the main difference being based on the fact that the ETL specification document, in this case, will no longer represent a work input for the developer, but an iterative output with the mappings that are completed interactively in the application.

The solution distinguishes four distinct steps in the conversion workflow:

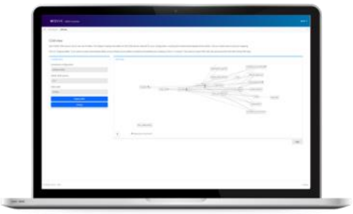


1. Once a dataset is imported and its schema is confirmed, the software starts to generate profiles of that source data

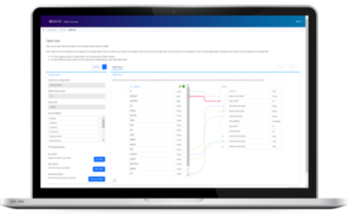


3. Assignments can be created from new or previously copied configurations, saving time for data updates.

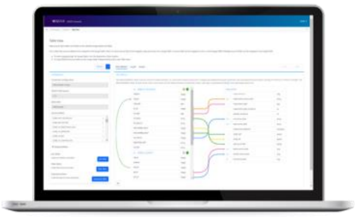
2. Source data profiling reveals important information, such as value distributions



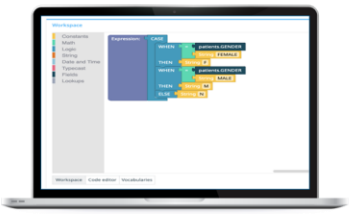
4. It is always clear to the user which CDM tables need to be mapped visually, facilitating the process.



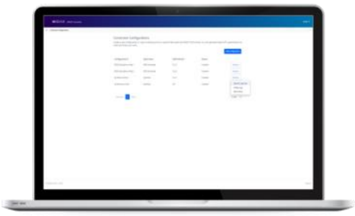
5. Mapping between source and target fields with a simple click.



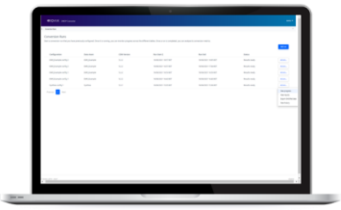
7. Several source tables can be assigned to a target table.



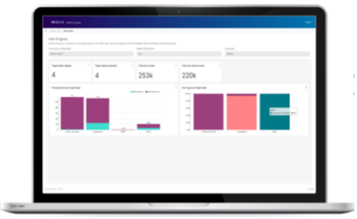
6. Simple or complex field logic is always on.



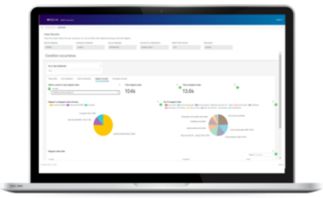
8. The ETL specification document is available as tables are assigned.



9. Users can perform on-the-fly conversion on a table-by-table basis.



10. Users can monitor and observe a conversion in real time and check for any errors.



11. Conversion results are available for each run and per target table, they can be shared or any analysis can be specified.

OMOP Converter is type-agnostic, facilitates OMOP conversions and automates the execution of these conversions. It supports OMOP v5.2, v5.3.1 and v6 versions and supports oncology extensions.

The solution includes several types of data validation as well as internal rules:

- Data structure rules define the structure for each supported CDM conversion, including table structure and key relationships.
- Other data conformance rules, for example, verify date fields that are within a reasonable range.
- Other rules, for example relationships between vocabulary related fields (source value, source concept, target concept, as well as relationships between key identifiers in source and CDM data.

For this project, within the services included in the IQVIA hours for these WP5 for the OMOP conversion, this SW is also included to help make it more agile.

Annex B: AI4HI approaches on data pre-processing

Sub-Annex B.1: Data quality and data cleaning

We will describe below the different approaches for data quality assessment and data cleaning in the AI4HI projects.

PRIMAGE

The PRIMAGE project implemented a range of data quality and curation methodologies to enhance the reliability of medical data analysis and the reproducibility of AI outcomes.

Within the e-form data entry process, there were checks for data quality dimensions, including validity, accuracy, consistency, integrity, and completeness. This involved ensuring that data values fell within specified domains and adhered to defined formats, ranges, and types. Numeric variables, for instance, had to conform to specified value ranges, and certain fields necessitated integer values without decimal points.

- Data accuracy was upheld through the use of data sources from different hospitals and official registries, ensuring the real-world alignment of the data with patients' medical histories.
- Consistency checks were conducted to identify and prevent duplicate records in the database. The project also verified that the sequence of dates, such as diagnosis dates and treatment dates, was logically coherent within each patient's medical history.
- To maintain data integrity, rules were established for certain fields. For instance, it was required that the date of the first treatment initiation must precede the date of image acquisition.
- Completeness of data was ensured by mandating that all fields were reviewed and filled. Any variable not collected could not be left empty and had to be explicitly marked as such.

Additionally, a time coherence tool was developed in Python to validate the chronological order and logical consistency of dates associated with patients' medical histories. This tool used predefined rules to check if the sequence of dates aligned logically. Any discrepancies were flagged, providing an opportunity for review and correction.

For categorizing MR series, an automated MR series classifier utilized machine learning applied to DICOM metadata. This approach facilitated efficient labeling of numerous MR series, even without the need for a graphics processing unit (GPU). This classifier had been exposed to diverse MR images across different pediatric cancer types and anatomical locations.

A custom feature based on Euclidean similarities was used to group individual MR series into the same dynamic sequence, addressing the challenge of variable storage formats for MR dynamic sequences.

Furthermore, an image preprocessing pipeline had been developed to enhance the reproducibility of radiomics features and image biomarkers. This pipeline included denoising, bias field correction, spatial resampling, and intensity normalization, tailored to specific tumor types (see Data Harmonization section for more details).

Finally, an Image Qure tool assessed the quality of MR T2WI images, providing detailed quality metrics on artifacts, signal, noise, and an overall probability of image quality.

Collectively, these methodologies contributed to the robustness and quality assurance of medical data analysis within the PRIMAGE project, ensuring reliable and reproducible AI outcomes.

ProCAncer-I

In ProCAncer-I project, several pre-processing functionalities were implemented in order to enhance data quality. Magnetic Resonance (MR) data from the project's clinical sites are collected capturing the variability across the different sites. While it is impossible to control or predict all sources of variation, currently many research efforts focus on image quality enhancement and data cleaning, such as removing bias field effects and noise, in order to improve both the performance and the trustworthiness of the machine and deep learning models. To this end, inhomogeneity correction, noise filtering, and image enhancement tools were implemented in order to minimize intensity related variations that could potentially affect and degrade the performance of the models. These preprocessing tools are independently offered in containerized packages and the users can select the tools they want to exploit in their pipelines.

Regarding bias field signal correction, a popular image filter was used for correcting the bias field, which is a low-frequency variation within their acquired MR signals, resulting in intensity inhomogeneities across the image. Such inhomogeneities can alter the textural descriptors in the radiomics extraction process hampering the planned modeling efforts. The N4ITK method is the state-of-the-art method for bias field correction and thus it is selected for bias field correction. However, The N4ITK has several parameters that their values should be defined. The optimal values of these parameters have been overlooked and most studies use the default values. To this end, a tool for applying the N4 bias field correction method and identifying its optimal configuration was developed for MR T2W prostate images, improving the performance of the bias field correction method.

A well-known methodology to reduce image value variability emanating from different imaging systems and protocols is image enhancement. ProCAncer-I proposes an original image enhancement technique based on the CLAHE method, named Region Adaptive CLAHE (RACLAHE), to serve as a model-invariant preprocessing method for improving prostate and prostatic zone segmentation. The aim was to provide a universal pipeline that will enhance models performance regardless of the choice of model for the segmentation task.

When collecting a large number of images from diverse clinical sites there is always the need to address the problem of noise which is also present when high-resolution examinations are obtained with a fast acquisition protocol. In ProCAncer-I, a novel Deep Learning noise reduction tool (deep learning-based model) was developed. Deep learning (DL) techniques for noise reduction have gained popularity because of their texture reconstruction and edge-preserving properties. A key advantage of using DL-based models over traditional image processing denoising methods, such as average or median filtering, is that the parameters of the deep model are optimized during training, while with traditional denoising, a predefined method is used and cannot be tuned for the examined set of data. Furthermore, unlike traditional methods, deep learning denoising preserves the edges and granular details in texture. This is not the case with traditional methods, which corrupt the original signal and make the image blurry.

CHAIMELEON

Quality Control of Imaging data:

Deliverable *5.1*

Image quality assessment in CHAIMELEON is dedicated to making sure that only the relevant imaging series were exported from sites to the central platform. This control is performed on site, before export, using the extraction and de-identification tool Radiomics Enabler (see above) : at extraction, any series that is not relevant to the project (e.g.: irrelevant modality) may be excluded using a DicomFilter script. Additionally, the site clinicians have the possibility to visually review the extracted imaging exams, and further exclude any series or patients based on their visual assessment.

Quality Control of Clinical data:

Clinical data quality assessment is fully manual in CHAIMELEON. Clinical data are collected using the ezCRF tool provided by Medexprim, a secure web application providing a custom eCRF that is linked to the imaging extraction and de-identification tool (Radiomics Enabler), making sure the patient pseudonym is consistent for clinical and imaging data. The user interface allows to issue queries when a mandatory field is incomplete, or when inconsistencies between fields are detected.

EuCanImage

EuCanImage defined a set of tools and procedures to ensure quality control of both imaging data and its associated clinical data.

Quality Control of Imaging data:

Image quality assessment is conducted in two phases. The first phase of quality assessment is performed by radiologists during the annotation process. If images in a given study are of too poor quality to permit annotation, the radiologist will reject the entire study. These data will be removed from the annotation and data storage pipeline and the submitting data controller will be requested to provide a replacement study. This is the ONLY point at which studies will be rejected and replaced. The second phase, once annotation is performed, an automated process collects the data acquisition protocol for each image series and calculates a perceptual quality metric, PIQUE, on a representative sample of data from each image series.

It is common practice to use general quality metrics that require the selection of a reference range, such as root means square error (RMSE), peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). However, in EuCanImage we decided to use a reference free metric to permit automation and simplify the quality assessment process. PIQUE (also referred to as PIQE) estimates block-wise distortion and measures the local variance of perceptibly distorted blocks to compute a quality score. PIQUE scores fall in the range [0,100]. The PIQUE score is inversely correlated to the perceptual quality of an image. A low score value indicates high perceptual quality and high score value indicates low perceptual quality. PIQE is a MATLAB^{TM} (Mathworks^{TM} Natick, MA) function that EuCanImage implemented in Python based on the algorithm proposed by Venkatanath ,et al[99]. Our version was validated against the MATLAB^{TM} function using data from The Cancer Imaging Archive.

The PIQUE algorithm computes the Mean Subtracted Contrast Normalized (MSCN) coefficient for each pixel in the input image. The image is then divided into non overlapping blocks and high spatially active blocks are identified based on the variance of the MSCN coefficients. In each active block distortion is evaluated using the MSCN coefficients and threshold criteria used to score the blocks as distorted with blocking artifacts, with Gaussian noise, or

[99] Venkatanath N, Praneeth D, Bh MC, Channappayya SS, Medasani SS, editors. Blind image quality evaluation using perception based features. 2015 twenty first national conference on communications (NCC); 2015: IEEE.

undistorted. The PIQUE score is computed as the mean of scores in the distorted blocks. Based on the literature a relative quality scale has been defined by Mathworks^{TM} and illustrated in the following Table 12:

Table 12: Relative Quality Scales

Relative Quality	PIQUE Score Range
Excellent	[0, 20]
Good	[21, 35]
Fair	[36, 50]
Poor	[51, 80]
Bad	[81, 100]

Quality Control of Clinical data:

In EuCanImage, clinical data is collected using the REDCap software, a secure web application for building custom eCRFs and managing online data capture mainly for clinical research studies. It has a user-friendly interface that allows field validation, custom logic patterns, calculated fields and a customizable data quality control module. Quality control rules include two levels: a) pre-defined data rules, standard rules pre-established by the app aimed to to avoid blank values, incorrect data types, invalid values or outliers among others b) custom rules, designed to fulfil the specific needs of the project. Additionally, EuCanImage has developed a quality assessment and quality scoring tool to evaluate three critical data dimensions: completeness, conformance and plausibility.

INCISIVE

In order to ensure the data quality, INCISIVE project followed a procedure with specific steps:

- *Step 1.* Data Quality Metrics

Data quality refers to the suitability of data to serve its intended purpose. The evaluation of data quality focuses on identifying whether the data supports the project needs in 4 dimensions namely:

Completeness: the comprehensiveness or wholeness of the data. There should be no gaps or missing information for data to be truly usable. Completeness was assessed by : 1/ identifying all critical information that needs to be present in the dataset and making them mandatory; 2/ Analyzing the data and identifying the missing information; 3/ Reporting the metric, i.e the percentage of records that are complete. The patients, in this case, are considered as records.

Validity: How well data conforms to required value attributes. Data should follow specific rules that were set from the beginning of the study. Validity was assessed by: 1/ defining the data format, allowable types & value ranges; 2/ analyzing the data and identifying the valid information (implemented only for the mandatory fields); 3/ Reporting metric, i.e the percentage of records in which all values are valid.

Consistency: process of keeping information uniform & homogeneous across different sites. The first step of the methodology for measuring this dimension is the definition of standards that will be followed by all sites. The second step is to analyze the data and identify the inconsistencies between sites. The third step is to report the metric, which is the percentage of values that match across different records. Consistency was assessed by : 1/ defining standards that will be followed by all sites; 2/ analyzing the data and identifying the inconsistencies between sites; 3/ reporting the metric, i.e the percentage of values that match across different records.

Integrity: Data Integrity refers to the extent to which all data references have been joined accurately. The first step of the methodology for measuring this dimension is the definition of the integration rules, which in this case is the link between images and clinical metadata through the template. The second step is to analyze the data and identify if they are properly integrated. The third step is to report the metric, which is the percentage of data that are properly integrated.

- Step 2. Definition of Rules and Requirements

To define the requirements for the clinical metadata collection and overcome the homogenization challenges in functional, semantics and privacy levels, an iterative procedure took place following steps described in the Data harmonization processes section (see below 5.1.2.2).

- *Step 3*. Data Quality Assessment

The Data Quality Assessment was performed in two levels with the use of the Data Quality Integration Check Tool **(reference to where the tool is reported in the document)**. In the first level, the tool is used by the user and the user has to correct all reported errors prior to the data upload. In the second level, the tool was modified to run on the server side as a data quality assessment workflow and assess the quality of the uploaded data by reporting the values of the metrics specified in step 1.

Sub-Annex B.2: Data harmonization

Following, we describe the data harmonization approaches adopted in different AI4HI projects:

In the **PRIMAGE** project, an image harmonization pipeline was created with the main objective to increase the reproducibility of radiomics features and image biomarkers reducing the variability between vendors, magnetic fields, acquisitions protocols and sites.

Specifically, the pipeline was composed of four different stages: denoising, bias field correction, spatial resampling and intensity normalization. For the denoising filter stage, a study similar to the one performed in [100] was carried out to choose the most optimal filter for each sequence based on the different quality metrics. The best filter parameters were selected using one database and validated with a second one. Each tumor (Neuroblastoma and DIPG) has its own pipeline configuration with the most optimal parameters for each region (abdomen and brain, respectively).

Regarding the bias field correction, the N4 filter from ANTS was chosen for correcting field inhomogeneities. The process to select the most optimal parameters was like the one mentioned above with the exception that artificial field inhomogeneities were added instead of noise. A brief summary of the denoising and bias field correction filters and parameters is shown in the following Table.

Table 13: Summary of denoising and bias field correction in PRIMAGE project

		T1W	T2W	DCE	DWI
NB	**Denoising**	ADF Iterations=2 Conductance=1	ADF Iterations=2 Conductance=1	ADF Iterations=2 Conductance=1	SUSAN FWHM = 2 Threshold= 1.2 Otsu
	Bias Field Correction	N4 (ANTS) Iterations =[50,30] BSpline =50 Shrink Factor=2	N4 (ANTS) Iterations =[50,30] BSpline =50 Shrink Factor=2	N4 (ANTS) Iterations =[50,30] BSpline =50 Shrink Factor=2	-
DIPG	**Denoising**	ADF Iterations=1 Conductance=0.5	Bilateral Domain sigma=0.5 Range sigma=60	Bilateral Domain sigma=0.5 Range sigma=60	SUSAN FWHM = 2 Threshold= 1.5 Otsu
	Bias Field Correction	N4 (ANTS) Iterations =50 BSpline =50 Shrink Factor=2	N4 (ANTS) Iterations =50 BSpline =50 Shrink Factor=2	N4 (ANTS) Iterations =50 BSpline =50 Shrink Factor=2	-

In the case of DIPG, spatial resampling was applied to a 1x1x1 mm isotropic voxel to extract radiomic features On the other hand, for neuroblastoma and due to the variability in longitudinal

[100] Fernández Patón, Matías, et al. "MR denoising increases radiomic biomarker precision and reproducibility in oncologic imaging." *Journal of Digital Imaging* 34.5 (2021): 1134-1145.

voxel size (between 3 to 10 mm), spatial resampling was only applied in transversal plane, maintaining voxel size in the longitudinal plane. Neuroblastoma radiomics features were calculated in 2D. A z-score normalisation was applied for both T1W and T2W as well as for DIPG and Neuroblastoma.

This harmonisation process is integrated into the radiomics or biomarker imaging calculation flow in a docker image using a json file with the DICOM path as input. For this purpose it makes use of quiblib, a library developed by QUIBIM to facilitate the interaction with the PRIMAGE platform, SimpleITK library for Anisotropic Diffusion Filter (ADF) and bilateral filter, FSL software for SUSAN filter and ANTS for N4 bias field correction filter.

Apart from image harmonization, feature harmonization techniques have been applied before predictive models training. Radiomics features are sensitive to variations caused by different scanners, reconstruction methods, or acquisition protocols among others, leading to a challenge known as batch effect. To mitigate this issue and ensure reliable results, data harmonization is a crucial step in multi-site studies. ComBat is a data harmonization method that seeks to eliminate the non-biological variability in multi-site studies as well as reduce the number of features with a considerably different distribution that has shown satisfactory results in different radiomics harmonization studies[101,102]. ComBat standardizes the means and variances of radiomics features across different batches allowing for more accurate comparisons and integration of data from different sources to develop radiomic models. The Nested ComBat methodology was applied, which provides a sequential workflow for radiomics feature harmonization to compensate for multicenter heterogeneity caused by multiple batch effects[103].

In the **ProCAncer-I** project, three biologically motivated image-based normalization techniques were developed to harmonize MR T2W prostate images and the well-known ComBat method was also incorporated for radiomics-based harmonization.

Image-based normalization: The challenge of varying intensities in MRI scans with arbitrary units poses a central obstacle to quantitative analysis. A uniform representation of the intensities among identical tissue types within and across patients is required. To this end, three biologically motivated normalization techniques were developed to harmonize MR T2W prostate data: (a) the fat-based (b) the muscle-based, and (c) the single tissue (fat or muscle) piece-wise normalization method. The tool is packaged in a docker image. The module requires as input data the folders of patients with the images in NifTI format (*.nii, *.nii.gz). The output is the normalized images, which are exported in the same file format. Initially, the N4 bias field correction method is applied and an approximation of the fat and muscle tissue is calculated by using the K-means algorithm.

In the fat- and muscle-based normalization method, the image is normalized according to statistics derived from the tissue's (fat or muscle) distribution, similarly to the White Stripe

[101] R. Da-ano et al., «Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies», Sci. Rep., vol. 10, n.o 1, p. 10248, jun. 2020, doi: 10.1038/s41598-020-66110-w.

[102] F. Orlhac et al., «How can we combat multicenter variability in MR radiomics? Validation of a correction procedure», Eur. Radiol., vol. 31, n.o 4, pp. 2272-2280, abr. 2021, doi: 10.1007/s00330-020-07284-9.

[103] H. Horng et al., «Generalized ComBat harmonization methods for radiomic features with multi-modal distributions and multiple batch effects», Sci. Rep., vol. 12, n.o 1, p. 4493, mar. 2022, doi: 10.1038/s41598-022-08412-9.

normalization technique[104]. More precisely, the mean value of the voxels of the fat/ muscle tissue is subtracted by the original image intensity values and then divided by the standard deviation of the fat/muscle tissue. In the single tissue piece-wise normalization, the intensity values of either fat or muscle tissue are used as input in Nyul and Udupa's piece-wise histogram normalization algorithm[105] to extract landmarks and generate a standard scale. The landmarks are determined through the training phase (60% of the patients' images of the dataset is used as a training set), leveraging the distribution of a specific (either fat or muscle) reference tissue.

Feature-based harmonization. Radiomics, an evolving field within medical image analysis, seeks to enable the conversion of images into quantitative features. These features play a crucial role in facilitating earlier and more precise clinical decision-making [106,107]. Typically, these features encompass aspects like volume, shape, intensity, and texture, and they are computed using various mathematical equations from the domain of image analysis. However, these calculations can be susceptible to bias and variation due to multiple factors, such as differences between vendors, variations in acquisition protocols, and settings for image reconstruction [108]. In the scientific literature, this variability is commonly referred to as the "center-effect" [109]. This variability can have a substantial impact on the absolute values and statistical distribution of radiomics features, consequently affecting the robustness and applicability of any subsequent analyses based on them [110].

The feature-based harmonization tool proposed in ProCAncer-I is based on the 'Combine Batched' method (ComBat), originally introduced by Jean-Philippe Fortin [111], and its alternatives (e.g., M-ComBat). The primary goal is to mitigate inherent variability in radiomics features attributed to the "center-effect" by expressing them in a common space. This tool has been packaged in a Docker image and requires two input arguments: the images in DICOM format (.dcm) and the corresponding radiomics features in pickle format (.pkl). The outputs include harmonized radiomics features, estimated parameters from the harmonization process, and additional method details, all exported in pickle format. Within the Docker file, a 'hypes.json' file is available for declaring the 'center_effect' parameter (e.g., Manufacturer or ManufacturerModelName), as well as the 'reference_batch' (e.g., 'true' for M-ComBat or 'false' for ComBat). In the case of the M-ComBat method, the scanner providing the most data in absolute numbers is automatically selected as a reference batch. It is essential to ensure that

[104] Shinohara RT, Sweeney EM, Goldsmith J, Shiee N, Mateen FJ, Calabresi PA, Jarso S, Pham DL, Reich DS, Crainiceanu CM; Australian Imaging Biomarkers Lifestyle Flagship Study of Ageing; Alzheimer's Disease Neuroimaging Initiative. Statistical normalization techniques for magnetic resonance imaging. Neuroimage Clin. 2014 Aug 15;6:9-19. doi: 10.1016/j.nicl.2014.08.008. Erratum in: Neuroimage Clin. 2015;7:848. PMID: 25379412; PMCID: PMC4215426.

[105] Nyúl LG, Udupa JK. On standardizing the MR image intensity scale. Magn Reson Med. 1999 Dec;42(6):1072-81. doi: 10.1002/(sici)1522-2594(199912)42:6<1072::aid-mrm11>3.0.co;2-m. PMID: 10571928.

[106] Gillies, Robert J., Paul E. Kinahan, and Hedvig Hricak. "Radiomics: images are more than pictures, they are data." Radiology 278.2 (2016): 563-577.

[107] Van Timmeren, Janita E., et al. "Radiomics in medical imaging "How-to" guide and critical reflection." Insights into Imaging 11.1 (2020): 1-16.

[108] Da-Ano, R.; Visvikis, D.; Hatt, M. Harmonization Strategies for Multicenter Radiomics Investigations. *Phys. Med. Biol.* 2020, *65*, 24TR02.

[109] Da-Ano, R., et al. "Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies." Scientific Reports 10.1 (2020): 1-12.

[110] Stamoulou, Elisavet, et al. "Harmonization strategies in multicenter MRI-Based radiomics." *Journal of Imaging* 8.11 (2022): 303.

[111] Fortin, Jean-Philippe, et al. "Harmonization of cortical thickness measurements across scanners and sites." Neuroimage 167 (2018): 104-120.

the number of patients per scanner manufacturer or per scanner manufacturer model name exceeds five for the tool to function correctly.

In the **CHAIMELEON** project, a novel methodology for image harmonization was developed. One of the most common approaches consists of adopting GAN-type[112] networks to conveniently separate content from style, that is, being able to maintain the image content of a sample, but based on contrast, textures, and subtle image features of the second sample. In contrast, in this project the proposed method relied on a self-supervised learning approach, and the idea consisted of learning to transform the original images into others with better texture, contrast, luminance and less noise.

This methodology was adopted for two different image modalities MRI and CT. For the first image type, the goal was to improve contrast and brightness while for the second type to improve texture and noise reduction.

The approach can be decomposed into two different stages. First, we need to generate smart alterations of the images, for this the original image is transformed in a frequency space to get two-embedded (i.e. module and phase) features that allows synthesising a large number of images by just changing the value of these features. For this, we needed to identify those samples with the same or similar acquisition parameters (e.g. echo time, inversion time, repetition time) inside the same protocol, to harmonize around them. Next, and once we have the images on which we are going to harmonize, we proceed to make the contrast alterations. Second, we trained a UNet-like deep neural network architecture to learn the mapping between altered and ground truth images. To validate the solution, the original and synthetic sets were compared in different tasks related to prostate cancer segmentation and survival analysis. In general, a systematic increment of 3-5% in segmentation was achieved as well as a reduction of variance and outliers in most of the features extracted per image.

In the **INCISIVE** project, to define the requirements for the clinical metadata collection and overcome the homogenization challenges in functional, semantics and privacy levels, an iterative procedure took place following several steps:

1. Identification: Proposition of a template per cancer type based on bibliography and medical experience.
2. Review: The templates were circulated through the Data providers, reviewed, and discussed.
3. Merge: A consensus of each template was extracted and discussed in a meeting to resolve homogenization issues.
4. Redefine: The data providers were asked to provide a sample case. The cases were reviewed for integrity and privacy issues.
5. Standardize: Standardization of the fields content and adopted terminologies.
6. Review and Refine: The templates were circulated again for verification.

This procedure is described in Figure 41: Data collection procedureFigure 41 below:

This procedure defined the rules for the clinical metadata:

1. The structure of the clinical metadata records, in the form of a template in .xls format.

[112] Bashyam, Vishnu M., et al. "Deep generative medical image harmonization for improving cross-site generalization in deep learning predictors." *Journal of Magnetic Resonance Imaging* 55.3 (2022): 908-916.

2. The patient encoding.
3. The value ranges and allowable types of each field, based on the terminologies.
4. The fields that were considered as mandatory and should be present in every record
5. The definition of the timepoints in which data should be collected and the time boundaries of the timepoints.
6. The definition of a complete case
7. The dataset structure and patient folders and studies naming conventions

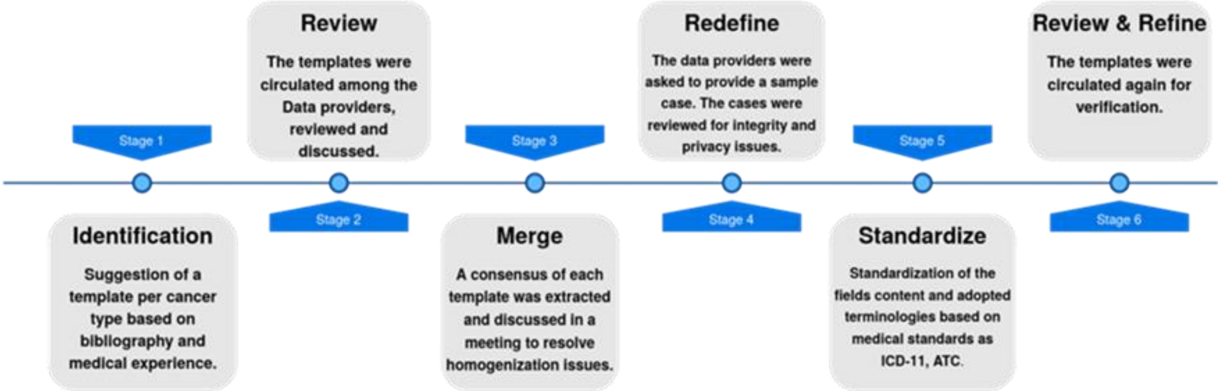

Stage 1
Identification
Suggestion of a template per cancer type based on bibliography and medical experience.
Stage 2
Review
The templates were circulated among the Data providers, reviewed and discussed.
Stage 3
Merge
A consensus of each template was extracted and discussed in a meeting to resolve homogenization issues.
Stage 4
Redefine
The data providers were asked to provide a sample case. The cases were reviewed for integrity and privacy issues.
Stage 5
Standardize
Standardization of the fields content and adopted terminologies based on medical standards as ICD-11, ATC.
Stage 6
Review & Refine
The templates were circulated again for verification.

Figure 41: Data collection procedure

With regards to the imaging data, the requirements were the result of workshops between medical experts and technical partners. The rules that came up from these workshops are:

1. The data de-identification protocol that all imaging data must follow
2. The image analysis requirements referring to image properties such as pulse sequence and slice thickness
3. The annotation procedure that should be followed for each image modality and each cancer type as well as the correspondence of the annotation file with the series.
4. The imaging modalities that should be provided for each cancer type

Sub-Annex B.3: Data de-identification

The approach of the various projects is summarized in the sequel:

CHAIMELEON: The project consortium opted to store anonymized data, including clinical and imaging data, on the central repository. Anonymization involved two steps for both patient identifiers and DICOM images. In the first step, patient identifiers were randomly generated and linked to their original patient identifier, while DICOM images were pseudonymized, either with or without accompanying eCRFs, using the CTP Anonymizer in the Radiomics Enabler tool (MEDEXPRIM) installed on premise. Direct identifiers were removed or replaced with pseudonyms, and a table of correspondence was kept within the hospital. In the second step, a new patient identifier was generated, and all dates were shifted to maintain longitudinal information. The key for the correspondence to the new patient identifier is deleted at the export of data. This final anonymization process was necessary before sending the data to the central repository, as it facilitated ethics committee approval at clinical sites and ensured compliance with regulations across institutions.

EuCanImage: EuCanImage is a centralized project built on existing services and repositories, incorporating three key data processors: Collective Minds Radiology (CMRAD), Euro-BioImaging XNAT, and the European Genome Archive (EGA). Pseudonymization is a crucial step to ensure patient privacy, and it involves encrypting the patient's medical record ID using the SHA512/256 hash algorithm. This pseudonymized ID, called EuCanImage-ID, is used for data correlation across different data types in the platform. DICOM images are anonymized by removing specific tags containing personal information while preserving clinically relevant tags. Clinical data is collected using the GDPR-compliant REDcap eCRF software, with direct identifiers replaced by the EuCanImage ID and indirect identifiers modified to reduce identifiability. Some clinical sites use a novel data collection tool to extract and anonymize data from EMR systems before uploading it to REDcap. Throughout the process, stringent data minimization and anonymization practices are followed, ensuring compliance with privacy regulations.

INCISIVE: The INCISIVE project aims to establish a hybrid repository of clinical data and DICOM images, which will be utilized for the development, training, and validation of AI algorithms in cancer management and follow-up. Ensuring GDPR compliance, the de-identification process focuses on DICOM data. The project team conducted an extensive study of DICOM protocols, including NEMA and TCIA, and identified the CTP Anonymizer as an open-source and user-friendly de-identification tool. Collaborating with data providers, AI developers, and legal partners, they formulated the INCISIVE de-identification protocol. The protocol prioritizes patient privacy while enabling usability for AI developers. It involves de-identifying patient names, identifiers, unique IDs, and dates in the DICOM images. A naming convention was proposed, combining data provider-specific codes and sequential patient numbers. Hash functions were used for de-identifying other identifiers, while dates were modified to maintain the original offset between patient examinations. Additionally, irrelevant DICOM fields were either removed or replaced. The project also developed its own de-identification tool based on the NEMA protocol, empowering data providers to choose the level of privacy for their data by customizing DICOM field removal or de-identification options.

ProCAncer-I: The ProCAncer-I project's primary objective is to develop a large imaging repository containing approximately 17,000 anonymized prostate multi-parametric (mp) MRI or bi-parametric MRI examinations from 9 clinical centers across Europe. The data is divided based on 9 specific clinical use cases, and additional clinical information is included to establish ground truth for each case. To ensure patient privacy, a double-level anonymization approach is employed. The first step, called blacklisting, allows each data provider to apply their established anonymization workflows. The second step, called whitelisting, involves applying common rules defined by the consortium to all data during upload to the repository. The anonymization profile focuses on limiting the amount of information to attributes relevant to the use cases without compromising the data value for possible future uses. Image anonymization follows the DICOM standard and involves modifying tags based on agreed-upon rules, preserving data quality and vendor information. Clinical data is integrated through an eCRF form, and no automatic extraction is performed to minimize the risk of sensitive information being exposed. Dates are relatively counted, not given in absolute terms, and certain clinical aspects are restricted to ensure privacy.

PRIMAGE: PRIMAGE focuses on Real World Data, gathered from collaborating hospitals, registries, and clinical trials, to develop in-silico tumor behavior prediction models. Access is granted to registries and clinical trial databases like SIOPEN-r-net and GPOH, with secondary use of available clinical data following Ethics Committee approval and data protection rules. Data from the SIOPEN-r-net database was pseudonymized using EUPID, allowing linkage of datasets without using directly identifying data. EUPID is also employed in the PRIMAGE project for pseudonymization, creating unique pseudonyms based on phonetic hashes for patient data. When a new patient is added to the PRIMAGE database, a unique pseudonym is generated for their pseudonymization. Imaging studies associated with the patient are uploaded using the pseudonym, replacing personal data in DICOM files and removing sensitive DICOM tags to ensure data privacy and confidentiality.

Sub-Annex B.4: Data annotation

The focus has been on the ProCAncer-I and PRIMAGE projects, as the rest of the AI4HI projects either did not include a specific annotation task or it was carried out by a partner outside the project.

ProCAncer-I. In ProCAncer-I, an annotation tool environment integrated with the rest of the ProCAncer-I system is implemented. The platform follows a DICOM in - DICOM out approach, utilizing DICOM files for both inputs and outputs, according to the DICOM standard (NEMA PS3 / ISO-12052). The annotation environment comprises three main components:

1. User Interface: This interface facilitates interactions and functionalities for image annotation. It allows loading and viewing images, along with existing annotations. Tools for image manipulation such as scrolling, zooming, panning, and contrast adjustment are provided. Unique annotation tools include a brush for manual voxel-wise segmentation and an automatic prostate segmentation algorithm.

2. Back-end/Proxy: The backend is implemented as a Single-Page Application (SPA) using JavaScript and resides within the user's web browser. An API Gateway acts as a proxy to provide secure access to the ProCAncer-I image and metadata repositories. It handles authentication and authorization, ensuring secure user interactions. The proxy operates a static asset web server, manages user sessions, and performs credentials translation for secure architecture. It also invokes the automatic prostate segmentation tool, and updates the DICOM Image Repository and Metadata Repository with segmentation results and associated information.

3. Automatic Prostate Segmentation Algorithm: A Python-based algorithm using Convolutional Neural Networks segments the prostate into central/transitional zone, peripheral zone, and seminal vesicles. This algorithm is integrated within the annotation environment and operates on T2w MR series. This algorithm is run as a Docker image, and generates DICOM Seg files.

PRIMAGE. The PRIMAGE project tackled the challenge of automatically segmenting primary tumors in pediatric cancers, specifically neuroblastoma and Diffuse Intrinsic Pontine Glioma (DIPG). Expert tumor delineation was needed as no dedicated open-access annotated data or tool existed for this task. A methodology was developed for manual segmentation, involving experienced pediatric radiologists defining modality, strategy, and tools for accurate segmentation.

Manual tumor delineation was carried out using the ITK-SNAP tool by four radiologists. A lead radiologist per tumor type led the segmentation, and additional radiologists performed inter-observer variability studies. The resulting segmentation masks underwent quality checks and were curated for accuracy and alignment for the actual tumor boundaries. These masks formed a ground truth sub-repository.

Neuroblastoma's automatic segmentation employed CNN algorithms, such as U-Net, on T2-weighted MR images. Data was divided into 80% training and 20% validation sets, stratified by features. The nnU-Net framework created the final model for PRIMAGE. An interactive learning approach enhanced accuracy by combining manual and semi-automatic segmentations, reducing manual segmentation time by over 70%. Evaluation used DICE coefficient and false positive/negative rates. The model was validated on an international dataset, achieving a high median DSC of 0.997. It had a 6% failure rate in identifying/segmenting tumors. Performance remained consistent regardless of MR factors or tumor location. Visual inspection took around 7.9 seconds, while manual editing averaged 124 seconds.

The DIPG dataset was relatively small, comprising 70 MR exams from 52 patients. These 70 images were divided into training, validation, and test sets (consisting of 49, 7, and 14 images, respectively). To augment the dataset, they segmented the images into patches, each containing the tumor with a 50% overlap between patches. Two distinct models were trained: one using only T1-weighted images and another using a combination of T1-weighted and T2-weighted/FLAIR images. Both models were based on the 3D U-Net architecture. Despite the limited number of cases, the results demonstrated a DICE coefficient exceeding 0.88 for the combined T1-weighted and T2-weighted/FLAIR model, which is a remarkable outcome considering the challenge of defining DIPG tumor edges. Notably, the combined model outperformed the model trained solely with T1-weighted images.

Both segmentation models were integrated into the PRIMAGE platform and are executed in Docker containers. The generated output is a segmentation mask in JSON format, that can be loaded into the PRIMAGE DICOM viewer.

Sub-Annex B.5: Data FAIRification

We present the information about data FAIRification from CHAIMELEON, ProCAncer-I and INCISIVE, focusing on the first as it is the one with the most comprehensive approach.

CHAIMELEON revised the 41 FAIR principles indicators specified from the Research Data Alliance (RDA) and verified them with respect to specific restrictions and requirements of the project.

- **Findability**: Dataset Service uses a structured incremental and hashed identifier model to guarantee uniqueness of the PID. Then uses the Zenodo API to automatically create a deposition of dataset metadata (only metadata is published). This is done for published datasets and optionally for released datasets. The publication at Zenodo gives visibility to the dataset and allows the project to obtain a DOI.
- **Accessibility**: The PID forward to an open description landing page where basic aggregated metadata is shown, along with the instructions of how to access the data. Actual access to the data requires authorisation and the signature of the Terms of Usage. No anonymous access is provided to the actual data. PIDs are registered in external repositories (Zenodo) along with the metadata record to guarantee that it is available even if the CHAIMELEON platform is not operative or if the dataset has been invalidated. Data identifiers are also persisted on Tracer. This is a CHAIMELEON core service to trace the use of datasets.
- **Interoperability**: The metadata is coded following the MIABIS model (Minimum Information About BIobank data Sharing), extended to consider DICOM data (DICOM-MIABIS) and provided into JSON. Medical data is stored in DICOM files and clinical data into OMOP-like records.
- **Reusability**: The use of DICOM and MIABIS facilitate the reusability. The medical images keep most of the DICOM information included in the DICOM tags, except those that have been removed in the anonymisation process performed in the provider centres. Data access is granted only to users that have signed the Terms of Usage and who have obtained an Ethical approval from a recognized Ethical committee. This is clearly stated in the metadata landing page, although the exact details of the approval of a request depend on the criteria of the scientific committee of the platform.

ProCAncer-I uses Biotronics3D platform for the management of imaging data, and MOLGENIS as a metadata catalogue.

- **Findability:** Persistent and globally unique ids are assigned to each dataset coupled with standardized metadata. In addition, the publication of datasets at Zenodo gives visibility to the data and allow us to obtain a DOI (Digital Object Identifier), which serves as a globally unique reference (Permanent ID Url) for each dataset.
- **Accessibility:** Each Dataset is accessible with dedicated pages that serve as an entry point to its metadata through Molgenis. Furthermore, through the implementation of a Fair Data Point (FDP), which offers metadata in an RDF (Resource Description Framework) format, machine-readability is exemplified. This dual approach ensures accessibility both for humans and machines, facilitating a diverse set of data consumers.
- **Interoperability**: Image data comply to the DICOM standard, clinical data follow an OMOP-CDM format along with an oncology and radiology extension, and the dataset metadata adhere to the W3C DCAT (Data Catalog Vocabulary) model. This standardization of data facilitates seamless data exchange and collaboration across the healthcare and AI research domains.

- **Reusability**: ProCAncer-I data are also reusable with all the steps of the data creation processes to be meticulously recorded. Moreover, we prioritize privacy protection through advanced anonymization strategies and implementation of data access licenses, for enhancing data reuse in the future.

Apart from applying FAIR principles to datasets, ProCAncer-I applies these principles to AI models **for building trust and fostering responsible AI development**. The concept of an "AI Model Passport" has been introduced as a comprehensive solution to track the entire development process of AI models while ensuring adherence to the FAIR principles. Through the AI model passport, the data and the models are **findable**, by having a unique identifier, ensuring they are identifiable and citable, they are **accessible**, permitting anyone to download or utilize them based on specific access condition/limitations, they are interoperable since the AI Model Passport adheres to standardized formats and tools, with well-documented interfaces that outline input and output specifications, along with any preprocessing steps applied, and finally they are **reusable**, featuring comprehensive instructions and guidelines for effective implementation, training, and utilization.

INCISIVE: not developed or adopted specific data FAIRification tools, but implements the following:

- **Findability**: unique identifiers (country level).
- **Accessibility**: querying mechanism developed in-house.
- **Interoperability**: FHIR-HL7 CDM.
- **Reusability**: dataset metadata list that describes the dataset in detail.

Annex C: List of imaging attributes kept for all AI4HI projects

Attributes Keywords	CTP RSNA / NEMA 2023 TAGS	Module Classification	Tag Description
SpecificCharacterSet	(0008,0005)	Image	Character Set that expands or replaces the Basic Graphic Set.
ImageType	(0008,0008)	Image	Image identification characteristics.
SOPClassUID	(0008,0016)	Image	Uniquely identifies the SOP Class.
Modality	(0008,0060)	Series	Type of equipment that originally acquired the data used to create the images in this Series.
Manufacturer	(0008,0070)	Equipment	Manufacturer of the equipment that produced the Composite Instances.
ManufacturerModelName	(0008,1090)	Equipment	Manufacturer's model name of the equipment that produced the Composite Instances.
BodyPartExamined	(0018,0015)	Series	Text description of the part of the body examined.
ScanningSeq	(0018,0020)	Image	Description of the type of data taken.
SeqVariant	(0018,0021)	Image	Variant of the Scanning Sequence.
ScanOptions	(0018,0022)	Image	Parameters of scanning sequence.
MRAcquisitionType	(0018,0023)	Image	Identification of data encoding scheme.
AngioFlag	(0018,0025)	Image	Angio Image Indicator. Primary image for Angio processing.
SliceThickness	(0018,0050)	Image	Nominal slice thickness, in mm.
RepetitionTime	(0018,0080)	Image	The period of time in msec between the beginning of a pulse sequence and the beginning of the succeeding (essentially identical) pulse sequence.
EchoTime	(0018,0081)	Image	Time in ms between the middle of the excitation pulse and the peak of the echo produced (kx=0). In the case of segmented k-space, the TE(eff) is the time between the middle of the excitation pulse to the peak of the echo that is used to cover the center of k-space (i.e., -kx=0, ky=0).
InversionTime	(0018,0082)	Image	Time in msec after the middle of inverting RF pulse to middle of excitation pulse to detect the amount of longitudinal magnetization. Required if Scanning Sequence (0018,0020) has values of IR.
NumberOfAverages	(0018,0083)	Image	Number of times a given pulse sequence is repeated before any parameter is changed
ImagingFrequency	(0018,0084)	Image	Precession frequency in MHz of the nucleus being addressed

ImagedNucleus	(0018,0085)	Image	Nucleus that is resonant at the imaging frequency.
EchoNumber	(0018,0086)	Image	The echo number used in generating this image. In the case of segmented k-space, it is the effective Echo Number.
MagneticFieldStrength	(0018,0087)	Image	Nominal field strength of MR magnet, in Tesla
SpacingBetweenSlices	(0018,0088)	Image	Spacing between slices, in mm. The spacing is measured from the center-to-center of each slice.
NumberOfPhaseEncodingSteps	(0018,0089)	Image	Total number of lines in k-space in the 'y' direction collected during acquisition.
EchoTrainLength	(0018,0091)	Image	Number of lines in k-space acquired per excitation per image.
PercentSampling	(0018,0093)	Image	Fraction of acquisition matrix lines acquired, expressed as a percent.
PercentPhaseFieldOfView	(0018,0094)	Image	Ratio of field of view dimension in phase direction to field of view dimension in frequency direction, expressed as a percent.
PixelBandwidth	(0018,0095)	Image	Reciprocal of the total sampling period, in hertz per pixel.
SoftwareVersion	(0018,1020)	Equipment	Manufacturer's designation of software version of the equipment that produced the Composite Instances.
SpatialResolution	(0018,1050)	Equipment	The inherent limiting resolution in mm of the acquisition equipment for high contrast objects for the data gathering and reconstruction technique chosen. If variable across the images of the Series, the value at the image center.
TriggerTime	(0018,1060)	Image	Time, in msec, between peak of the R wave and the peak of the echo produced. In the case of segmented k-space, the TE(eff) is the time between the peak of the echo that is used to cover the center of k-space. Required for Scan Options (0018,0022) that include heart gating (e.g.: CG, PPG)
CardiacNumberOfImages	(0018,1090)	Image	Number of images per cardiac cycle.
TriggerWindow	(0018,1094)	Image	Percent of R-R interval, based on Heart Rate (0018,1088), prescribed as a window for a valid/usable trigger.
ReconstructionDiameter	(0018,1100)	Image	Diameter in mm. of the region from within which data were used in creating the reconstruction of the image. Data may exist outside this region and portions of the patient may exist outside this region.
ReceiveCoilName	(0018,1250)	Image	Receive coil used.
TransmitCoilName	(0018,1251)	Image	Transmit coil used.
AcquisitionMatrix	(0018,1310)	Image	Dimensions of the acquired frequency /phase data before reconstruction.
InPlanePhaseEncodingDirection	(0018,1312)	Image	The axis of phase encoding with respect to the image.

FlipAngle	(0018,1314)	Image	Steady state angle in degrees to which the magnetic vector is flipped from the magnetic vector of the primary field.
VariableFlipAngleFlag	(0018,1315)	Image	Flip angle variation applied during image acquisition.
SAR	(0018,1316)	Image	Calculated whole body Specific Absorption Rate in watts/kilogram.
dBDt	(0018,1318)	Image	The rate of change of the gradient coil magnetic flux density with time (T/s).
PatientPosition	(0018,5100)	Series	Patient position descriptor relative to the equipment.
DiffusionBValue	(0018,9087)	Image	Diffusion sensitization factor in sec/mm2. This is the actual b-value for original frames and those derived from frames with the same b-value, or the most representative b-value when derived from images with different b-values.
DiffusionGradientOrientation	(0018,9089)	MR diffusion macro attributes	The direction cosines of the diffusion gradient vector with respect to the patient
SeriesNumber	(0020,0011)	Series	A number that identifies this Series.
AcquisitionNumber	(0020,0012)	Image	A number identifying the single continuous gathering of data over a period of time that resulted in this image.
InstanceNumber	(0020,0013)	Image	A number that identifies this image.
ImagePosition	(0020,0032)	Image	The x, y, and z coordinates of the upper left hand corner (center of the first voxel transmitted) of the image, in mm.
ImageOrientation	(0020,0037)	Image	The direction cosines of the first row and the first column with respect to the patient.
Laterality	(0020,0060)	Series	Laterality of (paired) body part examined. Required if the body part examined is a paired structure and Image Laterality (0020,0062) or Frame Laterality (0020,9072) or Measurement Laterality (0024,0113) are not present.
TemporalPositionIdentifier	(0020,0100)	Image	Temporal order of a dynamic or functional set of Images.
NumberOfTemporalPositions	(0020,0105)	Image	Total number of temporal positions prescribed.
TemporalResolution	(0020,0110)	Image	Time delta between Images in a dynamic or functional set of Images.
ImagesInAcquisition	(0020,1002)	Image	Number of images that resulted from this acquisition of data
PositionReferenceIndicator	(0020,1040)	Frame of Reference	Part of the imaging target used as a reference.

SliceLocation	(0020,1041)	Image	Relative position of the image plane expressed in mm.
StackID	(0020,9056)	Frame content macro attributes	Identification of a group of frames, with different positions and/or orientations that belong together, within a dimension organisation.
InStackPositionNumber	(0020,9057)	Frame content macro attributes	The ordinal number of a frame in a group of frames, with the same Stack ID (0020,9056).
SamplesPerPixel	(0028,0002)	Image	Number of samples (planes) in this image.
PhotometricInterpretation	(0028,0004)	Image	Specifies the intended interpretation of the pixel data.
Rows	(0028,0010)	Image	Number of rows in the image.
Columns	(0028,0011)	Image	Number of columns in the image.
PixelSpacing	(0028,0030)	Image	Physical distance in the patient between the center of each pixel, specified by a numeric pair - adjacent row spacing (delimiter) adjacent column spacing in mm.
BitsAllocated	(0028,0100)	Image	Number of bits allocated for each pixel sample. Each sample shall have the same number of bits allocated. Bits Allocated (0028,0100) shall be either 1, or a multiple of 8.
BitsStored	(0028,0101)	Image	Number of bits stored for each pixel sample. Each sample shall have the same number of bits stored.
HighBit	(0028,0102)	Image	Most significant bit for pixel sample data. Each sample shall have the same high bit. High Bit (0028,0102) shall be one less than Bits Stored (0028,0101).
PixelRepresentation	(0028,0103)	Image	Data representation of the pixel samples. Each sample shall have the same pixel representation.
SmallestImagePixelValue	(0028,0106)	Image	The minimum actual pixel value encountered in this image.
LargestImagePixelValue	(0028,0107)	Image	The maximum actual pixel value encountered in this image.
WindowCenter	(0028,1050)	Image	Window Center for display.
WindowWidth	(0028,1051)	Image	Window Width for display.
RescaleIntercept	(0028,1052)	DX image module attributes	The value b in relationship between stored values (SV) and the output units specified in Rescale Type (0028,1054).
RescaleSlope	(0028,1053)	CT image module attributes	m in the equation specified by Rescale Intercept (0028,1052).

Annex D: Catalogue of tools

Name	Project and partner	License	URL	GPU/ CPU	Dockerized	Works with data located in an specific path	Compliant with the minimal security measures	Modality type	Inputs	Outputs	Current status
Breast dense tissue segmentation ITI BREAST Calculate	ITI	LGPL	https://www.mdpi.com/2075-4418/12/8/1822	CPU	Yes	Yes	No	Imaging	FFDM in DICOM format	Parametric segmentation in JSON format and segmentation in a numpy object	Validated
MR-based neuroblastoma tumour detection and segmentation	HULAFE - PRIMAGE	TBD	https://pubmed.ncbi.nlm.nih.gov/35954314/	Both	Yes	Yes	Yes	Imaging - MRI	T2w in DICOM format	Segmentation in DICOM SEG format or NIFTI	Containerized
MR-based DIPG tumour detection and segmentation	HULAFE - PRIMAGE	TBD	-	Both	Yes	Yes	Yes	Imaging - MRI	T1w and/or T2w/FLAIR in DICOM format	JSON file or NIFTI	Containerized
MR-based glioblastoma tumour detection and segmentation	HULAFE	TBD	-	Both	Yes	Yes	Yes	Imaging - MRI	T1Wce + T2W + FLAIR in NIFTI format	Segmentation in NIFTI format	Containerized
CT-based neuroblastoma tumour detection and segmentation	HULAFE - PRIMAGE	TBD		Both	Yes	Yes	Yes	CT (CE)	CT CE sequence	Segmentation in DICOM SEG format or NIFTI	Developed
nnUnet	DKFZ	Apache-2.0	https://github.com/MIC-DKFZ/nnUNet	Both	Yes	Yes	Yes	Depends on the model	depends on the model	Segmentation in DICOM SEG format or NIFTI	Developed
nnDetection	DKFZ	Apache-2.0	https://github.com/MIC-DKFZ/nnDetection	Both	Yes	Yes	yes	Depends on the model	depends on the model	Object detection	Developed

MITK	DKFZ	BSD 3 Clause	https://github.com/MITK/MITK	CPU	Yes	Yes	yes	Any	DICOM, NIFTI, NRRD and other formats supported by ITK	DICOM, NIFTI, NRRD and other formats supported by ITK	Containerized
Multi-regional prostate segmentation	Quibim		https://pubmed.ncbi.nlm.nih.gov/36690774/	CPU	Yes	Yes	Yes	T2w MRI	T2w in DICOM format	DICOM SEG	Validated