**Project title:** European Federation for Cancer Images

**Project acronym:** EUCAIM

**Grant Agreement:** 101100633

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

# D2.1: Onboarding invitation package

| | |
|---|---|
| **Partner(s):** | HULAFE |
| **Author(s):** | Miguel Ángel Herrero Ramiro (HULAFE), Luis Marti-Bonmati (HULAFE), Ana Miguel Blanco (HULAFE), Hanna Leisz (DKFZ), Heinz-Peter Schlemmer (DKFZ), Laure Saint-Aubert (MEDEX) |
| **Date of delivery:** | 29/06/2023 |
| **Version:** | 2 |

| | |
|---|---|
| **Author(s):** | Miguel Ángel Herrero Ramiro (HULAFE), Luis Marti-Bonmati (HULAFE), Ana Miguel Blanco (HULAFE), Hanna Leisz (DKFZ), Heinz-Peter Schlemmer (DKFZ), Laure Saint-Aubert (MEDEX) |
| **Reviewers (WP)** | |
| **WP1** | Peter Gordebeke (EIBIR), Monika Hierath (EIBIR) |
| **WP2** | Javier Blázquez (SERMAS), Javier Soto (SERMAS) |
| **WP3** | Ricard Martínez Martínez (UV), Janos Meszaros (KUL) |
| **WP4** | Ignacio Blanquer (UPV), Esther Bron (EMC), Ignacio Gómez-Rico (HULAFE) |
| **WP5** | Manolis Tsiknakis (FORTH), Valia Kalokyri (FORTH), Gianna Tsakou (MAG), Stefanie Charalambous (MAG), Olga Tsave (AUTH), Octavio Herranz (QUIBIM) |
| **WP6** | Josep Gelpí (BSC), Alfonso Valencia (BSC), Salvador Capella (UB) |
| **WP7** | Sara Zullino (EATRIS), Katrine Riklund (UMU) |
| **WP8** | Linda Chaabane (EUBI & CNR), Mario Aznar (MAT), Amelia Suarez (MAT) |
| **WP9** | |
| **COO** | Peter Gordebeke (EIBIR), Monika Hierath (EIBIR) |
| **SCO** | Luis Marti-Bonmati (HULAFE) |
| **Date of delivery:** | 30/06/2023 |
| **Version:** | 2 |
| **Due date:** | Month 6 |
| **Actual delivery date:** | 29/06/2023 |
| **Type:** | R - Document, report |
| **Dissemination level:** | PU - Public |

# Table of contents

# List of abbreviations and acronyms

AAI: Authentication and Authorisation Infrastructure.

AI: Artificial Intelligence.

AI4HI: Artificial Intelligence for Health Imaging.

API: Application Programming Interface.

BBMRI-ERIC: Biobanking and Biomolecular Resources Research Infrastructure – European Research Infrastructure Consortium.

CDM: Common Data Model.

CSO: Chief Security Officer.

DPO: Data Protection Officer.

DGA: European Data Governance Act.

DICOM: Digital Imaging and Communication In Medicine.

DICOM-SEG: Digital Imaging and Communication In Medicine Segmentation object.

EATRIS: European Advanced Translational Research Infrastructure in Medicine.

EHDS: European Health Data Space.

ELIXIR: Distributed Infrastructure for Life Science Data.

ELSI: Ethical, Legal and Social Issues.

ETL: Extract Transform Load.

EUCAIM: EUropean Federation for CAncer IMages.

Euro-BioImaging: European *landmark* research infrastructure for biological and biomedical imaging as recognised by the European Strategy Forum on Research Infrastructures (ESFRI).

FAIR: Findable, Accessible, Interoperable and Reusable.

GDPR: General Data Protection Regulation.

GPU: Graphics Processing Unit.

HTTPS: Hypertext Transfer Protocol Secure.

NGS: Next-Generation Sequencing.

OMOP: Observational Medical Outcomes Partnership.

PREM: Patient Reported Experience Measure.

PROM: Patient Reported Outcomes Measure.

QoL: Quality of Life.

RAID: Redundant Array of Independent Disks.

REST: Representational State Transfer.

RI: Research Infrastructure.

SAML: Security Assertion Markup Language.

SLA: Service Level Agreement.

TNM: Tumour, Node, Metastasis.

WSI: Whole-Slide Imaging.

# Overview of the EUCAIM project and research infrastructure

## Our Mission

The goal of EUropean Federation for CAncer IMages (EUCAIM) is to build a pan-European digital federated infrastructure of cancer-related radiological and nuclear medicine images and other related digital information, which will be used to develop Artificial Intelligence (AI) tools for Precision Medicine. Our mission is to facilitate seamless access to de-identified, high-quality real-world data, and to foster collaboration among clinicians, researchers and innovators.

We aim to accelerate the development and benchmarking of AI-driven cancer management tools, empowering healthcare professionals to make data-driven decisions in diagnosis and treatment, and ultimately enhancing patient outcomes.

## Our Vision

We envision EUCAIM as the leading European hub for cancer research communities, interoperable oncological data and AI innovation. By unifying fragmented datasets into an extensive Atlas of Cancer Images and by ensuring data sovereignty and adherence to ethical standards, we will unlock the untapped potential of imaging and big data in oncology, driving forward a future where every patient receives personalised, and effective care.

## EUCAIM in detail

EUCAIM is a pan-European digital federated infrastructure of de-identified cancer medical images compliant with FAIR principles (Findable, Accessible, Interoperable and Reusable) [3] coming from Real-World data. The infrastructure is being designed as an experimentation platform to facilitate and foster the development and benchmarking of AI-based cancer management tools towards precision medicine in cancer diagnosis and treatment. To allow this, EUCAIM will provide a comprehensive Dashboard for data discovery, federated search, metadata harvesting, annotation and distributed processing, including federated and privacy-preserving learning techniques. In addition to and as part of its federated data infrastructure, EUCAIM will also build a Central Hub governing the Atlas of Cancer Images, which will be interoperable with the other European Health Data Space (EHDS) components while preserving the data sovereignty of providers.

EUCAIM will target clinicians, researchers, and innovators, providing the means to build reproducible clinical decision-making systems supporting diagnosis, treatment, and predictive medicine. This Infrastructure will benefit citizens through improved healthcare procedures and will stimulate the European market’s innovation through new tools and services.

EUCAIM will also contribute to shaping the legal grounds for such operation on a pan-European scale, adapting to the particularities of different countries in the management of clinical data. To do so, EUCAIM will implement a federation of data providers compliant with commonly agreed legal grounds, defining common data models, ontologies, quality standards, FAIR principles, and de-identification procedures. EUCAIM will align with the EHDS initiative toward a sustainable flagship repository of high-quality data and tools.

The EUCAIM infrastructure is being created through the EUCAIM project, a 4-year initiative which started on January 1st, 2023 that is co-funded by the European Union under the Digital Europe program, call DIGITAL-2022-CLOUDAI-02-CANCER-IMAGE “Federated European infrastructure for cancer images data".

Specifically, the EUCAIM project pursues 11 Specific Objectives (SO):

> SO1. Set up the Ethical, Legal and Security framework of EUCAIM, which will define the data access and transfer agreements, the de-identification and anonymisation procedures, and the legal bounds of the project.
>
> SO2. Set up a Coordinating Entity that will host the services of the Central Hub and will define the legal model, the rules for participation (for data and service providers and consumers), the recognition models and the operative procedures.
>
> SO3. Integrate and implement the Central Services that will provide a platform for data discovery, data querying and access to de-identified high-quality data on the federated nodes.
>
> SO4. Follow a data protection and privacy-by-design and by-default approach (as established in article 25 of the General Data Protection Regulation (GDPR) [1] to define an Authentication and Authorisation Infrastructure (AAI) and to implement the privacy-preserving technologies needed to fulfil the security agreements.
>
> SO5. Define the common data models, interoperability guidelines, best practices, FAIR metrics, tools and standards for the integration of federated data and metadata.
>
> SO6. Integrate a set of key data providers of cancer images coming from existing repositories, hospital coalitions, Research Infrastructures, networks and other data providers in the consortium.
>
> SO7. Integrate a distributed processing environment including appropriate processing tools, federated learning and computing intensive frameworks with seamless access to the data resources to implement the on-demand processing by the research users.
>
> SO8. Monitor data provisioning, data access, data processing, users, data accesses and other key metrics of the repository for reporting, evaluating and assessing the functionality of the platform.
>
> SO9. Define and implement the operating bodies of EUCAIM, which will oversee the access, scientific guidance, technical support, training and monitoring of the EUCAIM infrastructure.
>
> SO10. Create an environment for supporting a collaborative network across already existing Research Infrastructures such as the European Advanced Translational Research Infrastructure in Medicine (EATRIS), the Distributed Infrastructure for Life Science Data (ELIXIR), the Biobanking and Biomolecular Resources Research Infrastructure – European Research Infrastructure Consortium (BBMRI–ERIC) and the European *landmark* research infrastructure for biological and biomedical imaging (Euro-BioImaging).
>
> SO11. Define a sustainability plan and implement the necessary structures to be able to operate the repository as a Research Infrastructure beyond the end of the project.

## What needs does EUCAIM address?

The use of AI on health data is generating promising decision-support tools to assist clinicians in cancer management, as an increasing number of imaging-based AI approaches are proving to have vast potential to become useful clinical tools in different areas of application, such as recurrence and survival prediction, prediction of tumour molecular features and association with tumour spread, amongst many others. Despite these major advancements, the development of imaging-based AI tools relies on the availability of large, quality-controlled datasets, which currently remains a major challenge. This is because health data, including medical images, are highly distributed and fragmented across Europe. As a result, the generation of imaging biobanks is a resource-intensive endeavour, facing multiple technical and operational difficulties such as image and data harmonisation, data curation, image preprocessing, and image annotation as well as various legal and ethical restrictions.

EUCAIM aims to address the fragmentation of the existing cancer image repositories by building a distributed, federated Atlas of Cancer Images, which will include data related to both common and rare types of cancer from pre-existing initiatives (related research infrastructures and five successful Horizon 2020 projects), as well as clinical images, pathology, molecular and laboratory data from multiple hospitals across Europe.

## The EUCAIM project consortium

The EUCAIM project consortium brings together clinical data providers, researchers, research infrastructures, and industry with mature solutions addressing the challenges of implementing such a cancer imaging infrastructure. The key players in making EUCAIM a reality include experts in the following fields:

- Healthcare providers (primary use clinical repositories and hospitals, data warehouse architectures, clinical data holders and controllers, medical imaging).
- IT infrastructure developers and service providers.
- Common data models and data analysis (automated annotation, curation, de-identification, Observational Medical Outcomes Partnership (OMOP) data mapping, harmonisation).
- Biostatistics and epidemiology, developers of AI tools for cancer management (i.e. data users).
- Ethical, legal and societal issues, including authorities and policy makers.
- Project governance and sustainability.
- Dissemination and communication.
- Project management and financing.

## High-level architecture of the EUCAIM infrastructure

EUCAIM will follow a hybrid federated-centralised model, preserving the independence of Research Infrastructures and existing thematic, national or institutional repositories and providing centralised governance corresponding to a higher coordination layer that provides a cohesive and coherent structure in the access to the data. EUCAIM will also support observational studies and will provide long-term preservation and sustainability for the data collected in these studies. This architecture is implemented through the following main types of services:

- A public catalogue that gathers the collections from the different providers with the collections' metadata and enables the user to browse and explore the existing data.

- A federation layer comprising services that make the data compliant with FAIR principles, connecting the different repositories to the EUCAIM core services' layer using a Federated Query service and a Hyperontology that will facilitate querying the data stored in the federated providers platforms regardless of the data models that they have adopted.

- A set of core services providing a coherent and seamlessly connected Authentication and Authorisation Infrastructure (based on Life Sciences AAI), a monitoring service to improve operation, a traceability service to log all the data-related interactions of users and services to support the recognition models and to monitor the fulfilment of the Terms of Usage, a federated processing service, and a third-party data transfer service for temporary copies of datasets on High-Performance Computing sites.

- A set of organisational services such as a helpdesk system to support users and manage incidences and a security incident response service.

- A Dashboard that integrates all the functionality in a coherent environment, enabling users to browse and search datasets, request access to them, access data in the federation according to each repository's access conditions, and browse tools and pipelines to run them on a containerised environment at the provider's side.

*Figure 1* shows the architecture of the EUCAIM repository. Users will register on the EUCAIM platform, which will entitle them to browse and search the datasets available in the federation. The user will be provided with aggregated information and metadata from the dataset matching their criteria, indicating the provenance, the access conditions and the processing services available at the provider side. Users will be able to request access to the actual data, which will be selectively granted. Subject to the Access Conditions of the providers, granted users will be able to explore, process the data on-site and/or through federated processing services, using containerised applications from the federation marketplace. In the case of intensive data processing, data could be temporarily transferred to a High-Performance Computing service. The central storage of EUCAIM will be integrated as any other node in the federation.

Providers will commit to a given service level which will define the expected availability of services, the quality and quantity of the data and the access conditions. EUCAIM will provide them with tools, interoperability plugins and services to facilitate the integration into the federation, and will provide recognition by publishing aggregated data on the usage of the data of each provider. Providers will also benefit from networking opportunities, the creation of communities and the participation in projects by being members of EUCAIM.

The federated central hub will create the governance of the EUCAIM infrastructure to regulate the onboarding of data and tools providers and data users. The onboarding procedure will define the Service Level Agreement, the technical interoperability, data quality metrics, compliance to FAIR principles and data standards so datasets from providers can be findable in the central hub Dashboard. The Dashboard will provide the first step for the data access process, collecting the necessary information and forwarding the access request to the provider for the final decision.

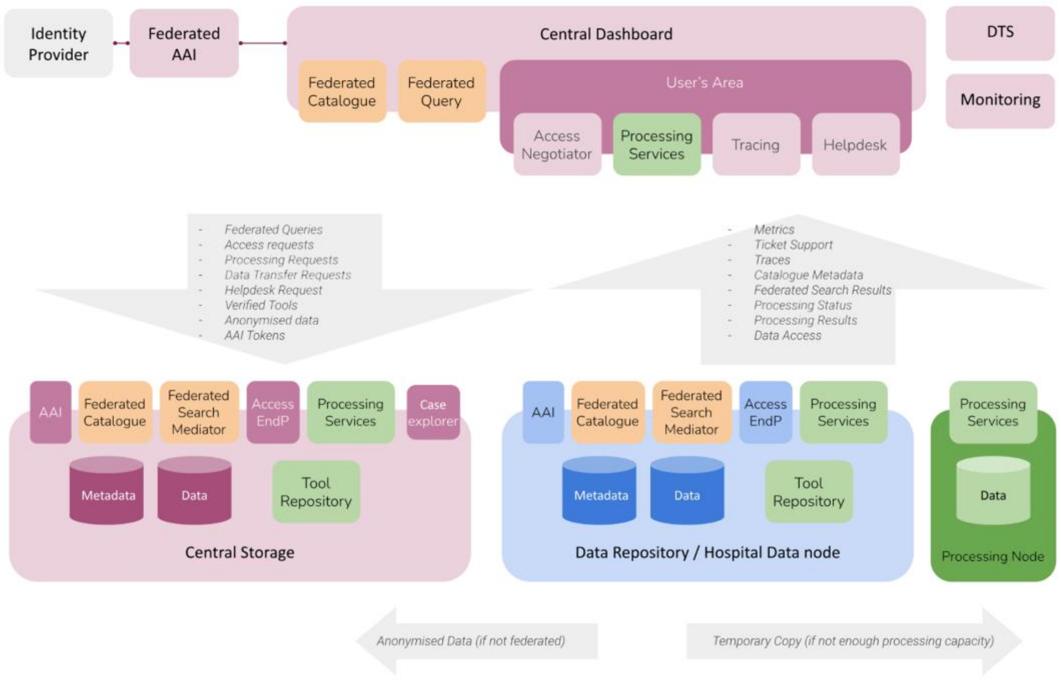


*Figure 1. High-level architectural view of EUCAIM.*

# Why becoming a data provider for EUCAIM

## Types of data providers invited to participate

EUCAIM welcomes different types of data providers, including:

1. Repositories with data storage and processing capabilities. These repositories will be federated, and data processing will be performed in a distributed way, respecting the access conditions of each repository and also supporting federated learning.
2. Repositories with data storage but no processing capabilities. Data storage will be federated and intensive data processing on anonymised data will be performed in the Central Hub or external computational infrastructures or through third party agreements and temporary data transfer.
3. Repositories without data storage nor processing capabilities. This is the case of imaging repositories created in the framework of a research project that is terminated or near termination. In order to make their data collections sustainable, they will upload anonymised data in the Central Storage.
4. Clinical data providers that are willing to set up a federated node that will connect to the federation. In case that they can offer data processing capabilities, data processing may either be performed in a distributed way, supporting federated learning, or on anonymised data in the Central Hub or external computational infrastructures or through third party agreements and temporary data transfer.
5. Clinical data providers that are not willing or do not have resources to set up a federated node and will upload anonymised data in the Central Storage.

6. Research infrastructures (RIs) providing distributed federated services that link data providers (holders-controllers) with existing anonymised and minimally annotated research data, which could be linked to the EUCAIM infrastructure in a federated mode. RIs should expose or register their catalogues and interact with EUCAIM regarding data access requests. Ideally, RIs should also support data discovery through federated query and distributed processing, if possible.

7. Data altruism organisations and patient associations that will either connect to the federation or upload anonymised data in the Central Storage.

Data providers shall ensure appropriate conditions of ethical and regulatory compliance required by the European Union law, and by national laws where applicable (see subsection Data holder requirements).

Finally, computing processing infrastructures will also be part of the EUCAIM research infrastructure. As they are not data providers, they are not identified in the previous list. In any case, computing infrastructures should provide means to temporarily store and process data from the EUCAIM federation.

For more information on the options for the relationship with EUCAIM, please refer to Annex Data Provision Agreement (options for the relationship with EUCAIM)”.

## Benefits for participation

As a data provider you will benefit from participating in EUCAIM in different ways:

- **Facilitated compliance with EHDS regulation:** According to the current regulatory proposal, data holders would be required to make health data available to third parties, for secondary use for the permitted purposes. EUCAIM provides you with the functional, operational and legal framework to facilitate compliance with this new regulation.

- **Engagement in large-scale, multicentric cancer-fighting studies**: EUCAIM, in coordination with Member States (Ministries of Health) and other stakeholders, will organise extensive studies to address new challenges in cancer research and innovation.

- **Enhanced visibility and prestige**: As an active participant in EUCAIM, your institution becomes an integral part of Europe's Beating Cancer Plan. This affiliation increases your institution's recognition, attracting additional funding and fostering partnership opportunities.

- **Access to a vast, diverse dataset**: By contributing your data to EUCAIM, you gain access to an extensive and varied database of cancer images. This resource significantly enhances the quality of research, AI tool development, clinical prediction models, and pharmaceutical product innovation.

On the one hand, you are promoting your data collections and collaborating with European research networks, thus obtaining benefits from the research outcomes (especially in the cases of data enrichment, reuse of data and data for basic science that are not related with commercial purposes or data exclusivity). On the other hand, EUCAIM could facilitate research promotion within your institution (e.g. oriented to physicians, researchers, etc.) in terms of networking and data access with the goal of developing improvements and/or capacity building.

- **Opportunity for revenue generation**: Data holders can benefit from sharing their data through partnerships with data users and stakeholders within the European Data Governance Act (DGA) and EHDS framework.

- **Facilitated cross-border collaboration and knowledge exchange**: EUCAIM connects data providers throughout the EU, promoting collaboration and knowledge sharing among institutions. The platform integrates a social network component, enabling direct communication and interaction between researchers, industry experts, and policymakers within the initiative.

- **Utilisation of EUCAIM's data tools**: Take advantage of the data management tools seamlessly integrated into the EUCAIM platform.

- **Advanced AI and imaging training**: EUCAIM supports scientists and physicians in their daily practice by providing cutting-edge training on AI and imaging tools.

- **Streamlined data readiness:** EUCAIM offers comprehensive technical support and guidance to assist you in implementing efficient data management processes. This ensures the standardisation, accessibility, and interoperability of your data for research and analysis purposes.

- **Comprehensive support for data managers**: EUCAIM assists data managers in collecting, consolidating, and presenting their datasets, ensuring that human, technological, and financial investments yield the best possible outcomes.

EUCAIM brings also relevant long-term socioeconomic impacts:

- **Prioritizing citizens' benefits**: EUCAIM amplifies the visibility and value of your data, enabling citizens to derive greater advantages from health research.

- **Enhanced healthcare and public health services**: EUCAIM's contributions to research and development facilitate more precise diagnoses, superior treatments, and effective preventive measures against various cancer types.

- **Informed resource allocation**: High-quality data and AI tools empower public authorities to make well-informed decisions regarding healthcare resource distribution and policy-making.

- **Strengthened regulatory oversight**: EUCAIM's infrastructure aids in standardising data practices and supports regulatory agencies in supervising data usage in research and clinical environments.

- **Fostering transparency and trust:** EUCAIM encourages responsible data use and collaboration, building trust among stakeholders and the public. It ensures easy, secure, and transparent access to health data, enhancing the quality of care and patient support.

- **Minimised data acquisition costs**: Unifying data in the Atlas of Cancer Images reduces expenses related to obtaining and managing extensive datasets.

- **Accelerating research and development:** Access to standardised, annotated, and readily available data accelerates the research process, promoting swifter discoveries and innovations.

- **Expanded personalised medicine opportunities:** The comprehensive dataset and AI algorithms support the development of personalised medicine, customizing treatments to cater to individual patients' needs.

- **Heightened awareness and education:** By establishing a unified cancer image resource, EUCAIM raises awareness and educates the public about various cancer types.
- **Ethical data utilisation:** EUCAIM's emphasis on anonymised and standardised data safeguards patients' privacy while enabling their data to contribute to significant research and innovation.

## Open calls for data providers

The EUCAIM consortium already includes several pre-existing data providers that will incorporate their data into the infrastructure. The internal call (for consortium partners only) is expected to be launched in Autumn 2023.

By the beginning of the second project period (expected by 2024–2025) and based on the feedback of the procedure for the internal call, the EUCAIM project will also launch an external open call for new beneficiaries to join the consortium, which will serve as a basis for including new cancer image databases in the federation. The new beneficiaries will receive funding under the same co-funding conditions as consortium partners (i.e. 50% of the budget). This open call will follow the guidelines stated in the call with respect to publication and openness and will pursue: i) the onboarding of new data providers, increasing the geographic dimensions, data modalities or cancer targets, and ii) the uptake of new trustworthy AI algorithms trained on the data of the repository. Supported by the internal governance bodies on ethics and legal compliance, the Access Committee (established in June 2023) will set up the rules for application and evaluation in alignment with the European Commission, receive and evaluate the proposals and finally prioritise them. The final acceptance will be taken by the Management Board taking into consideration the indications from the Access Committee and Steering Committee. The main objectives for this open call include:

- Establishing long-term collaboration with cancer image data providers, including addressing the legal issues, elaboration of data processing agreements, joint controllers' agreements, etc.
- Support for the upgrading of the necessary technical infrastructure to connect cancer image data sources at the national level to the federated European cancer image data infrastructure.
- Applying common protocols for cancer image annotation and curation in line with the infrastructure requirements.
- Training of trustworthy AI algorithms and prediction models of outcomes using the cancer imaging data available in the infrastructure.
- Validation of trustworthy AI algorithms and prediction models of outcomes using the cancer imaging data available in the infrastructure.
- Relevant awareness raising, including targeted up-skilling activities, for example necessary to join the network or for healthcare professionals to maximise the uptake of data, tools and services in clinical settings.

## Types of data and metadata to be provided

The types of data that the EUCAIM infrastructure is interested in collecting are:

**Imaging data:**

- Radiological and nuclear medicine cancer images (any modality): Digital Imaging and Communication In Medicine (DICOM) format.

- Segmentation masks with the annotations made (e.g. DICOM Segmentation object; DICOM-SEG).
- Histopathological images (whole-slide imaging; WSI) and any metadata collected from the images (optional, only under agreement).

**Other clinical data:**

- Clinical information accompanying the images, including demographics, any relevant history of the patient, diagnosis made, treatments (e.g. surgeries performed, medications administered, treatment response, follow-ups, metastatic episodes, Tumour, Node, Metastasis (TNM) stage group, cancer stage group, tumour markers, histopathology report, radiology report).
- Mutations status.
    - Multi-omics database information (e.g. genomics, transcriptomics, proteomics, metabolomics, radiomics).
- Biological sample results (e.g. tumour tissues, lab shipment, Next-Generation Sequencing (NGS) processing).
- Quality of Life (QoL), quality of care (PREMs & PROMs) and Health costs.

## Functional and technical requirements for data providers

In order to become a data provider, the following functional and technical requirements should be met.

**Technical requirements:**

1. **Data Provider Node Case**

In case the data provider agrees to set up their own data node, the following technical requirements should be met:

- Procurement or ownership of the needed infrastructure.
    - Each organisation is expected to obtain the technical and management infrastructure needed to host a federated node.
- Acquire and install dedicated hardware.
    - If the local nodes will support federated processing, at least a large server equipped with graphics processing units (GPUs) allowing for intensive learning processes[1].
    - Data redundancy measures (e.g. redundant array of independent disks; RAID) will have to be implemented to mitigate the risks associated with data loss, deletion or corruption, both during and after the project's lifespan. This includes the establishment of regular data backups, replication processes, and robust failover systems to guarantee the continuity and reliability of data storage and retrieval.

[1] Alternatively, data sites may set their own agreements with third-party trusted institutions (e.g., regional or national computational facilities, commercial providers) to provide the indicated computational infrastructure on their behalf. In those cases, data providers will be responsible to set the necessary legal and technical agreements with such institutions.

- Install operating system and deploy image preprocessing and EUCAIM repository connectivity software provided in the form of software containers.
- Local technical support.
  - Staff availability for technical support is required for installation of software and hardware.
  - Compliance with the technical guidelines through the deployment and execution of various EUCAIM tools (e.g. tools for data anonymisation, image annotation tools, data quality/cleaning tools, clients for federated analysis) and transformation of local data to the common data model of the EUCAIM infrastructure (e.g. Extract Transform Load (ETL) tools).
  - Application programming interfaces (APIs) for data sharing should be available according to EUCAIM specifications.
- Network configuration.
  - Site should allow outgoing connections to selected ports. The possibility of opening inbound ports is desirable but optional.

**Non-Functional Requirements:**

- Security measures should be enforced so that access is regulated.
  - All communications among services will be performed using Hypertext Transfer Protocol Secure (HTTPS) and Representational State Transfer (REST) APIs.
  - Authentication and Authorisation will use OpenID and Security Assertion Markup Language (SAML).
- Integration with traceability and auditing mechanisms provided by EUCAIM.
- Minimum uptime should be ensured for data providers.
- Provide the Service Level Agreement (SLA) agreed in terms of availability, scalability, and maximum number of requests per second for the federated data search and processing.

2. **Central Storage Case**

Organisations that are unable to acquire or accommodate a Node in their infrastructure can utilise the Central Storage of EUCAIM for data sharing (expected to start by December 2023). The Central Storage will be established and managed by the EUCAIM project, and will serve as a GDPR-compliant [1] centralised data storage solution that will enable the platform and cloud services to access the data within, in order to share them with potential data reusers. In this case, the following technical requirements should be met:

- Access to a dedicated machine provided by EUCAIM for uploading and administering data in the Central Storage.
- Network configuration: allow outgoing network connection over HTTPS in order to connect to the central services of the Federated Learning platform.
- Technical support: Staff availability for technical support is required for installation of software and hardware.

- Compliance with the technical guidelines through the deployment and execution of various EUCAIM tools (e.g. tools for data anonymisation, image annotation tools, data quality/cleaning tools) and transformation of local data to the common data model of the EUCAIM infrastructure (e.g. ETL tools).

**Functional Requirements:**

- Minimally-annotated and anonymised data availability.
- Compliance with EUCAIM's data sharing mechanism.
- Data mapped and transformed to the common data model.

Once these requirements have been met and the technical procedures to link the data provider to the EUCAIM infrastructure have been completed, the entity is ready to start with the data provision.  In the following paragraphs, we explain this workflow by the hospitals engaged with the EUCAIM project, by the Artificial Intelligence for Health Imaging network (AI4HI) projects, and by other research infrastructures.

**Scenario 1: Hospitals engaged with EUCAIM**

The following workflow will be applied:

- Data collections linked to EUCAIM will only be those previously accepted by the hospital and by the EUCAIM Management Board for use on approved research projects. Data collections will be visible in the EUCAIM Dashboard during the lifetime of the research project. As the research project is terminated, data collections may or may not be incorporated to the Central Storage, under agreement.
- Data will be de-identified using the EUCAIM de-identification tools or other de-identification tools available to the hospital, according to specific requirements provided by EUCAIM regarding de-identification procedures.
- Data will be annotated using the EUCAIM annotation tools or other annotation tools available to the hospital (optional).
- Data will be transformed through ETL to conform to the common data model agreed by the EUCAIM infrastructure.
- Data quality and cleaning tools will be applied to the data for addressing quality issues such as correcting data that is incomplete, incorrect, inaccurate or biased, etc.
- A local hospital node will be set up afterwards either at the hospital (if the hospital has available resources) or at the EUCAIM infrastructure (if the hospital has not enough resources for this). In both cases, data will be fully anonymised.

**Scenario 2: existing data repositories (e.g. AI4HI projects with EUCAIM within their sustainability plan)**

For the AI4HI projects, a slight variation of the aforementioned workflow will be applied.

- Data already available by the data repository will be mapped to the common data model agreed by the EUCAIM infrastructure.
- Appropriate APIs will be set up to ensure interoperability with the EUCAIM Dashboard, exposing the EUCAIM's query functionality and enabling data discovery by interested data reusers.

**Scenario 3: research infrastructures (e.g. European Research Infrastructure Consortium; ERIC)**

- Appropriate APIs will be set up to ensure interoperability with the EUCAIM Dashboard, exposing the EUCAIM's query functionality and enabling data discovery by interested data reusers.

## Ethical and legal aspects

EUCAIM will operate in full compliance with GDPR [1] and national laws. This means that personal data may only be used under strict conditions like irreversible anonymity or legal grounds for pseudonymised use. To ensure full compliance with data privacy principles, EUCAIM will be built under the privacy by design and by default approach. There are currently two scenarios for the implementation of the infrastructure:

1. Data in the local research repositories are already anonymised. Privacy is preserved.

2. Data in the local research repositories are pseudonymised (e.g. Primary Use Clinical Repository and Hospital, Data Warehouse architecture not exposed to the EUCAIM federated processing services). Since EUCAIM is pursuing a federated learning approach, data remain stored at local sites but are distributed with all the protective privacy measures in place.

## EUCAIM: a trustworthy environment

- High-level standard on GDPR compliance [1] and AI ethics risk-based approach.
- Alignment with future regulations such as EHDS Proposal and Artificial Intelligence Act [2].
- A model of governance in technical and legal aspects.
- Clear information about the dataset content and authorised uses.
- Data sharing/transfer agreements.
- Legal terms and conditions for data users.
- Governance procedures for the concession of data access permits.
- Secure environment that provides traceability on the activities of processing of the datasets.
- Transparency and public engagement on research activity placed at EUCAIM.
- Strong collaboration with data holders, Data Protection Officers (DPOs) and Chief Security Officers (CSOs).
- Compliance model that assures fulfilment at both national and EU level and promotes transborder activities of research.
- Technical environment specially addressed to provide legal and technical security, in the scenario of pseudonymisation.

## Data holder requirements

- Ensuring the legitimate origin of data.
- Providing evidence of the fulfilment of both EU and national laws (e.g. DPO statement).

- In case of federation with the EUCAIM data space, ensure interoperability, availability and security.
- In case of storage at the Central Hub, providing duly/properly anonymised datasets.

## Rules for participation: rights and obligations

The rights and obligations of the data providers for EUCAIM will be defined in the Data Provision Agreement. Different agreement models (or a common model with different annexes) may be needed to adapt to the specific conditions of different types of data provider nodes (e.g. hospitals, existing repositories, research infrastructure, or others).

By signing this agreement, the data provider commits to follow EUCAIM guidelines for good practices, embrace and endorse the vision of the EUCAIM consortium, fully subscribe to its statutes and respect the decisions taken by EUCAIM governing bodies.

The incorporation of a new data provider is subject of evaluation by the Access Committee according to objective criteria, corresponding to the assessment of their submitted Expression of Interest and Interoperability Checklist (see Annex Expression of Interest letter and Annex Questionnaire for Data Warehouse architectures (interoperability checklist)).

Data access requests for both the Central Storage and the Federated Nodes are also evaluated by the Access Committee, supported by the EUCAIM governance bodies on ethics and legal compliance. The Access Committee will evaluate the applications and provide recommendations to the Management Board for final decision of acceptance or rejection within 60 days, with additional 30 days if needed.

Researchers with an associated research project will have to submit a data access request by filling out a dedicated form available on EUCAIM webpage. In this case, the federated nodes (note that hospitals will always be asked for this decision) will have a period of 14 days to decide whether they accept or not the use of their data regarding a specific data access request, depending on their agreement with EUCAIM.  In case that the data access request is based on the laws of the European Union concerning the re-use of health data, the refusal of the hospital may only be based on the absence of adequate requirements or safeguards in the request or on the existence of explicit restrictions on the use of the data laid down in its national law. In cases of legal obligation of data sharing under European Union or national Law, the federated nodes will be notified.

## The onboarding process

### Steps

The onboarding process for data providers is the following (see Figure 2):

1. Initial contact (e.g. you will contact us directly at data@cancerimage.eu, or you will be contacted by our Evangelisation Team).
2. You will receive the onboarding invitation package, the **form for the Expression of Interest** and the **Interoperability Checklist** (described below, see Annex Expression of Interest letter and Annex Questionnaire for Data Warehouse architectures (interoperability checklist)). The Engagement Team will follow your status and activities.

   The form for the Expression of Interest asks for your contact information and to describe in detail your organisation, experience and capabilities. You will fill out the form and send it back to us.

The functional and technical requirements of EUCAIM's infrastructure will be compared to yours through an Interoperability Checklist that you will also fill out and send back to us.

3. Once your application is accepted, you will receive and sign a **Data Provision Agreement** (see Annex Data Provision Agreement (options for the relationship with EUCAIM)), and the Training Team will provide you with a training module on legal and ethical aspects for data provision (see Annex List of Training Modules).

4. The Training Team will provide you with a training module for technical provisions. Your infrastructure and EUCAIM's will start their connection. The Technical Support Team and the ELSI (Ethical, Legal and Social Issues) Team will audit your onboarding process, helping you with any issues that may arise via a Helpdesk (e.g. adaptation or development of the necessary infrastructure and software, clinical Common Data Model (CDM) and data protection guidelines; see Annex Helpdesk overview).

5. Once your infrastructure is connected, the Training Team will provide you with a set of additional modules on tools for data preparation activities (e.g. tools for data de-identification, data segmentation/annotation, data quality checking, etc.), data provision (including FAIR data) and general platform use (see Annex List of training modules). Additionally, the FAIR implementation Support Team will assist you in the adoption of FAIR data principles (see Annex Guidelines on data quality and FAIR principles).

6. Once the training and data preparation (i.e. preprocessing, cleaning, harmonisation, de-identification, annotation, etc.) are finalised, you will be able to provide EUCAIM with your first dataset. You will be in iterative contact with the Engagement Team that will supervise your activities, including the data compliance with FAIR principles.

7. The Engagement Team will periodically review the actions taken by your institution. The EUCAIM consortium (including all new data providers, among others) will publish papers promoting EUCAIM in the Q1 relevant literature (e.g. European Radiology, Insights into Imaging, etc.). The Evangelisation Team may also ask you to participate in online interactive workshops to explain the impact that your data has had on the development of new radiomics and AI tools in cancer research.

8. Periodically, the Data Monitoring Team will update your data provision statistics and will show them on the EUCAIM webpage, with your permission, as part of the dissemination and communication strategy.

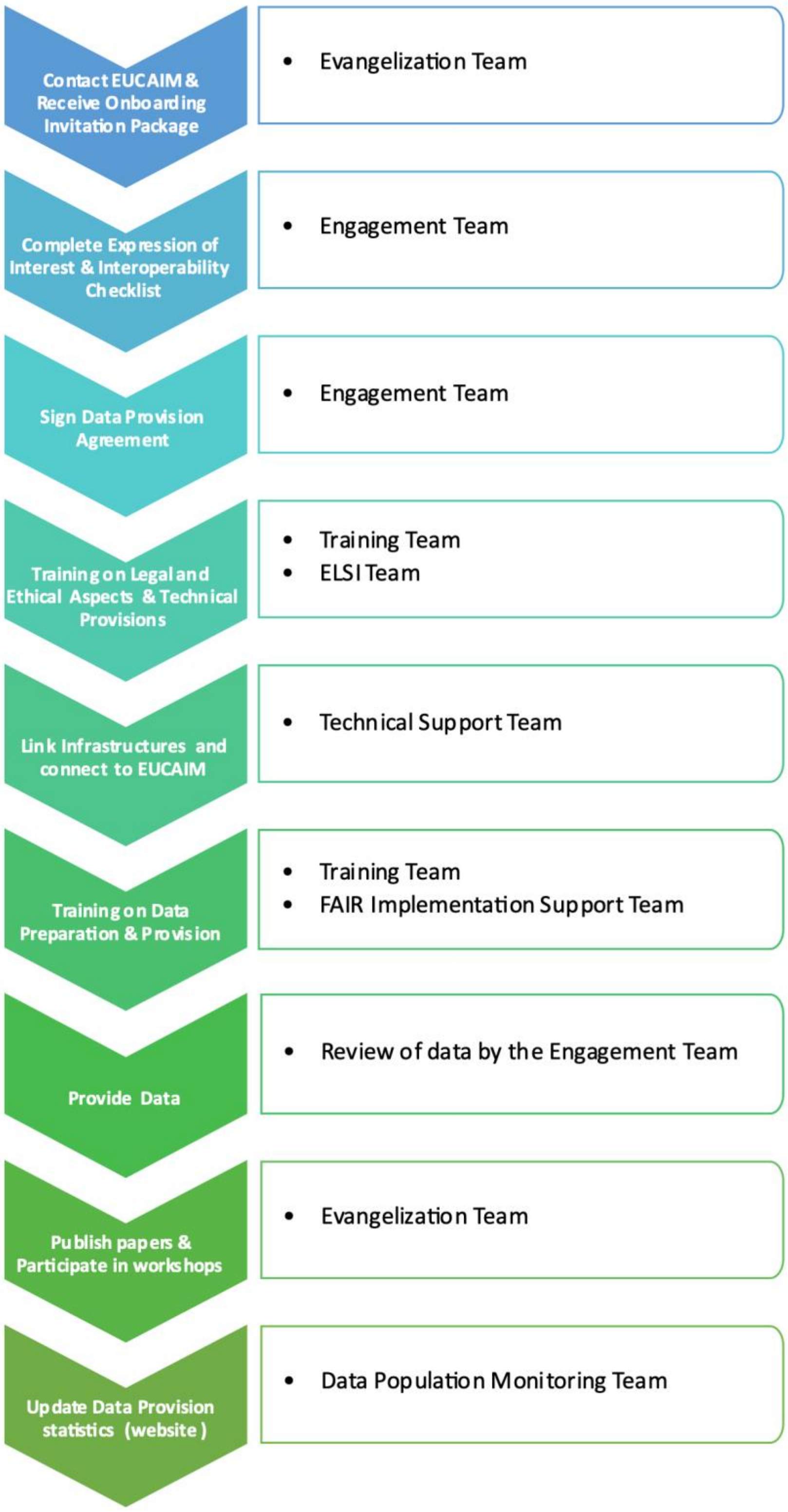

Contact EUCAIM & Receive Onboarding Invitation Package
Evangelization Team
Complete Expression of Interest & Interoperability Checklist
Engagement Team
Sign Data Provision Agreement
Engagement Team
Training on Legal and Ethical Aspects & Technical Provisions
Training Team
ELSI Team
Link Infrastructures and connect to EUCAIM
Technical Support Team
Training on Data Preparation & Provision
Training Team
FAIR Implementation Support Team
Provide Data
Review of data by the Engagement Team
Publish papers & Participate in workshops
Evangelization Team
Update Data Provision statistics (website)
Data Population Monitoring Team

*Figure 2. Onboarding process for data providers.*

## Main contacts

- General information for data providers: data@cancerimage.eu

For troubleshooting at different levels, please send an e-mail to the corresponding team:

- Technical issues: support@cancerimage.eu
- Training issues: training@cancerimage.eu
- Legal, ethical and operational issues: legal@cancerimage.eu

# References

1. General Data Protection Regulation (GDPR), Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016. https://eur-lex.europa.eu/legal- content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=es
2. Proposal for a Regulation of the European Parliament and of the Council Laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. COM/2021/206 final of 21 April 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206
3. Implementing FAIR Data Principles: The Role of Libraries. Association of European Research Libraries (LIBER). https://libereurope.eu/wp-content/uploads/2020/09/LIBER-FAIR-Data.pdf

# Annexes

## Expression of Interest letter

Please, fill out this form and send it back to Mr. Peter Gordebeke and Mr. Herrero Ramiro. It asks for your contact details, and for more specific information on your capabilities, experience and how your organisation could join the EUCAIM project.

**Contact details**

| | |
|---|---|
| Date: | YYYY/MM/DD |
| **Organisation, Full Legal Name:** | |

| | |
|---|---|
| Honorific: | Mr./Ms./Dr./Prof. |
| **Contact Person for the Work/Project, Full Name:** | |
| Job Title: | |
| ORCID/Other IDs | |
| Gender: | |
| Telephone Number: | |
| E-mail Address: | |
| Department name: | |
| Street: | |
| Town: | |
| Postcode: | |
| Country: | |

**Main role** (select one or more)

1. Data Provider (Primary Use Clinical Repository and Hospital, Data Warehouse architecture)
2. Repository (Secondary Use Research Repository)
3. Technical Resources and Expertise (Storage / Computation / Services / Interoperability)
4. AI Tools and Solutions
5. ELSI - EHDS
6. Dissemination and Communication

**Capabilities and functionalities** (overall description) (if your main role is Data Provider or Repository, add details on your infrastructure, common data model, number of studies, etc.)

**Experience** (main activities, ongoing and completed projects)

**Network of Collaboration** (main partners and Institutions you work with)

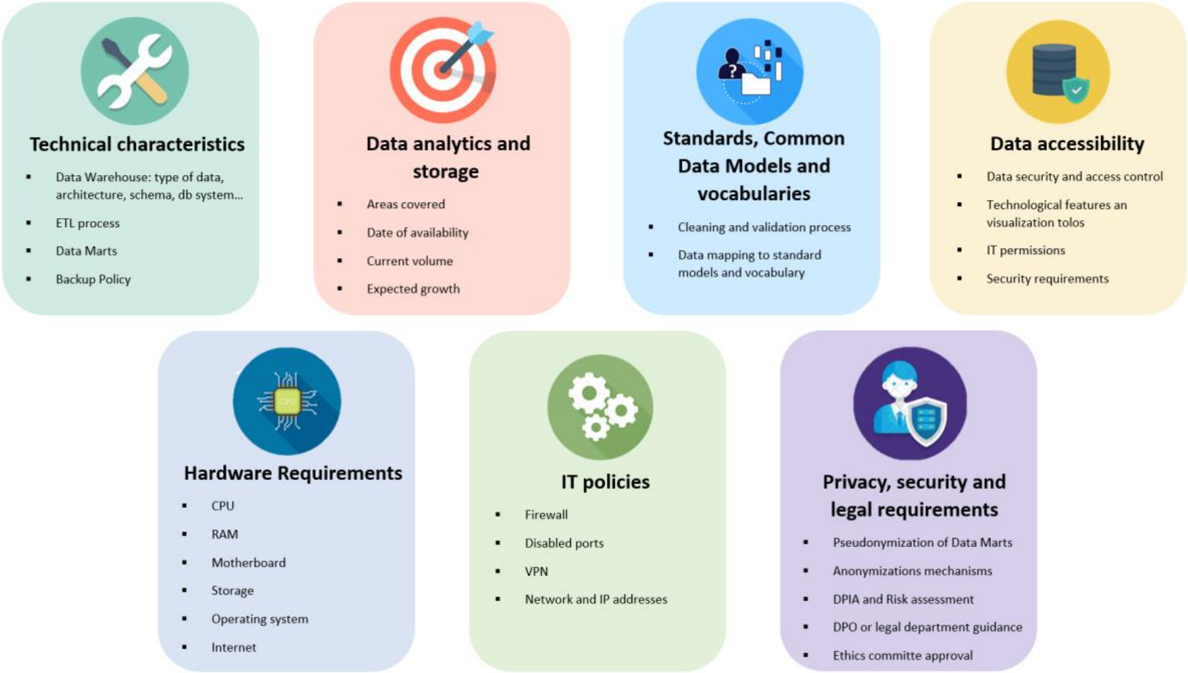
**Contributions to EUCAIM** (data, tools, technical resources, human resources)

## Questionnaire for Data Warehouse architectures (interoperability checklist)

The objective of this document is to conduct a self-assessment of the current state of the hospital's data warehouse, to determine its preparedness and maturity to be part of a federated European data infrastructure for research (EUCAIM project).

A hospital data warehouse contains clinical data of patients over time, from various operational systems, and stored in an integrated way for analysis and generation of information relevant for decision making. A well-designed data warehouse provides reliable, consistent, accurate and timely data, allowing researchers and clinicians to gain valuable insights to improve health outcomes. For a data warehouse to be used at the federal level, it is essential that it complies with interoperability standards, has high levels of data cleaning and quality, and is structured and documented in a way that is understandable and accessible. This will allow various actors to access and use the information for their research, while maintaining the privacy and security of patient data. The benefit of having a mature data warehouse lies not only in the possibility of sharing information through the EUCAIM federated infrastructure, but in the ability of any researcher attached to the hospital, with proper permissions, to exploit the data to generate knowledge applicable to improving healthcare.

This questionnaire seeks to evaluate aspects related to data modelling, governance, processing, interoperability, and accessibility of the data warehouse, to determine strengths and opportunities for improvement that allow the hospital to move towards a mature and reliable data infrastructure, enabling participation in the EUCAIM federation. It focuses on several key areas: technical characteristics, data analytics and storage, standards, common data models and vocabularies, data accessibility, hardware requirements, IT policies, privacy, security, and legal requirements.



**Technical characteristics**

- Data Warehouse: type of data, architecture, schema, db system...
- ETL process
- Data Marts
- Backup Policy

**Data analytics and storage**

- Areas covered
- Date of availability
- Current volume
- Expected growth

**Standards, Common Data Models and vocabularies**

- Cleaning and validation process
- Data mapping to standard models and vocabulary

**Data accessibility**

- Data security and access control
- Technological features an visualization tolos
- IT permissions
- Security requirements

**Hardware Requirements**

- CPU
- RAM
- Motherboard
- Storage
- Operating system
- Internet

**IT policies**

- Firewall
- Disabled ports
- VPN
- Network and IP addresses

**Privacy, security and legal requirements**

- Pseudonymization of Data Marts
- Anonymizations mechanisms
- DPIA and Risk assessment
- DPO or legal department guidance
- Ethics committe approval

## ***Questionnaire***

### Technical Characteristics

- Does your healthcare organisation utilise a data warehouse to store and analyse the data?
- Which of the following data types are incorporated into the data warehouse?
  a. Structured data (e.g. laboratory results)
  b. Semi-structured data (e.g. clinical-pathological reports, treatment plans)
  c. Unstructured data (e.g. clinical notes)
  d. Imaging data and associated reports
- Please describe the process by which data is extracted from source systems, transformed into the appropriate format, and loaded into the data warehouse (commonly referred to as ETL).
- Which data modeling techniques are employed in the design of your data warehouse (e.g. star schema, snowflake schema)?
- Which database management system is utilised to support the data warehouse?
- Is the data warehouse architecture designed to incorporate data marts for individual departments, or is all departmental data integrated into a single comprehensive data warehouse schema?

- Are the data marts deployed on-premises or in the cloud?
- How frequently is new data incorporated into the data warehouse?
- Please describe the process for maintaining and updating the data warehouse over time as source systems evolve. How are changes to source systems identified and tracked?
- Please describe the data backup policies and procedures for your organisation's data warehouse.

### Data Storage and Analytics

- What data domains are currently covered in your organisation's data warehouse?
    - Genetics
    - Pathology
    - Laboratory
    - Imaging
    - Treatment
    - Others (please specify)
- For each data mart available, please provide:
    - The date range of available data
    - The current data volume
    - The expected data volume growth over the next 12 months

### Standards, Common Data Models, and vocabularies

- Please describe the data cleaning and validation processes performed prior to loading data into the data warehouse.
- Please describe common use cases supported by the data warehouse and the types of data collected to support those use cases.
- Common Data Model and Terminologies: Are any of the data marts mapped to the Observational Medical Outcomes Partnership (OMOP) common data model or are you using FHIR (Fast Health Interoperability Resources)? If so, please indicate which data marts.
- Are any data marts mapped to other standard data models? If so, please indicate which data marts and which data models.
- Which standard vocabularies and terminologies are employed in your data models (e.g. SNOMED-CT)?
- Does your organisation store any type of data annotations (e.g. image segmentations)? If so, please indicate the format (e.g NIFTI, DICOM-SEG…)

### Data Accessibility

- Please describe how data security and access control are managed within the data warehouse environment.

- What types of reporting and data visualisation tools are utilised to access and analyse data within the data warehouse?
- Data Availability through APIs and Metadata Catalog: Are services available to provide access to data within the data warehouse?
- What technological capabilities and security requirements are necessary to utilise these data access services?

### Federated Node hardware

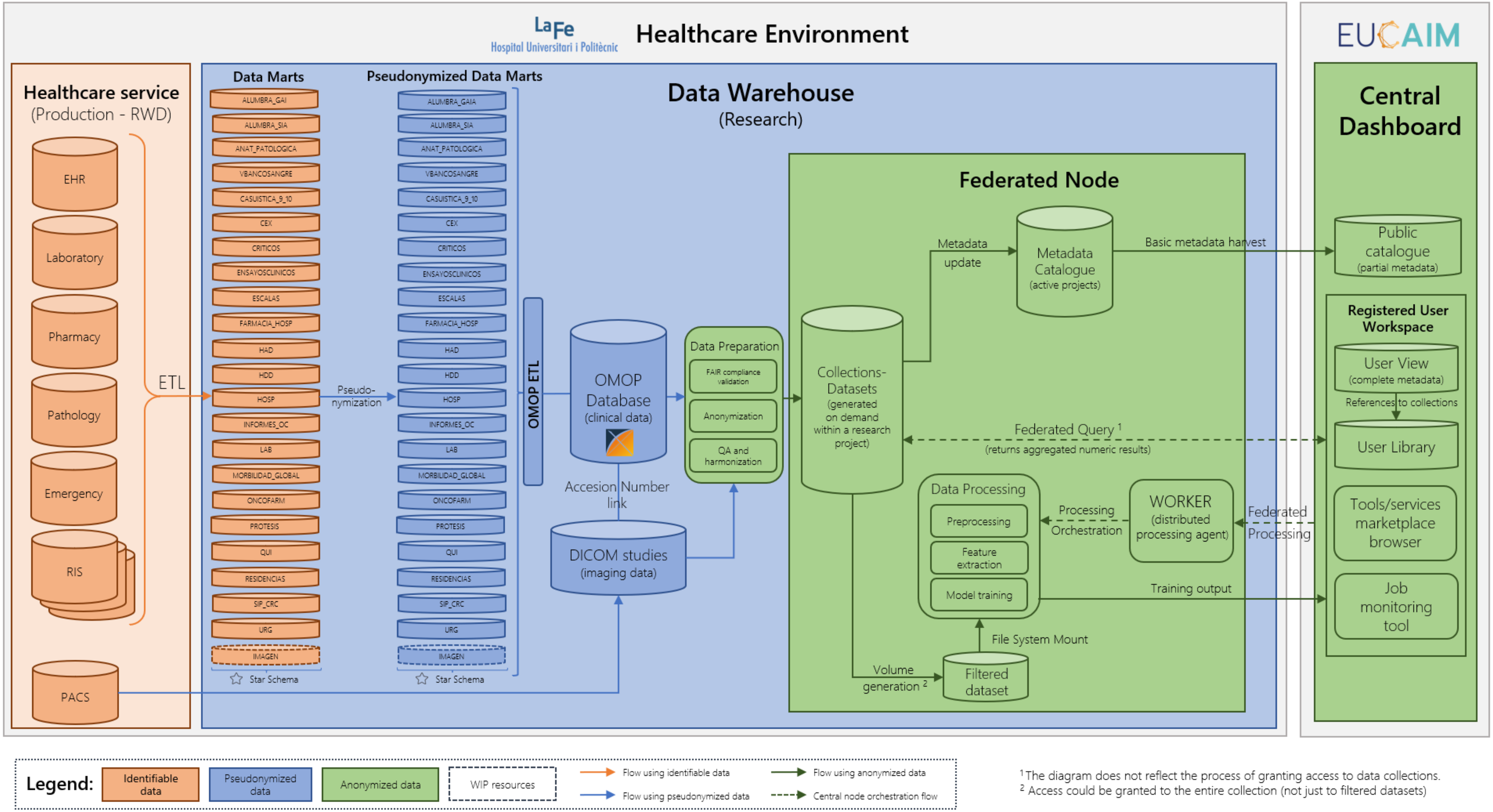
- Please provide details regarding the make, model, and specifications of the servers that will host your federated data node, including: CPU, RAM, GPU, storage, motherboard, server provider and model.

### IT policies

- Who within your organisation is responsible for governing access permissions and enabling secure remote connections for the purposes of configuring the federated data node?
- Does your organisation have formal policies regarding the use of virtual private networks (VPNs) and personal devices to establish remote access to the organisational network and federated data node?
- Has your organisation implemented specialised firewall policies to safeguard the network and federated data node?
- Are there any network ports that are restricted from external access?

### Privacy, Security and Legal requirements

- Is it possible to pseudonymise the data within your data marts? Will it be possible to re-identify pseudonymised data if required?
- How are data provenance and traceability maintained to enable re-identification of pseudonymised data when needed? Who governs access to re-identification keys and for what use cases (e.g. prospective studies, data correction/enhancement)?
- Has a risk assessment been performed for your data warehouse and computing infrastructure (e.g. Data Protection Impact Assessment)?
- Please provide contact information for your organisation's Data Protection Officer (DPO).
- Do you have ethics committee approval for your data warehouse and related uses of data?



[1] The diagram does not reflect the process of granting access to data collections.
[2] Access could be granted to the entire collection (not just to filtered datasets)

*Figure 3. La Fe Hospital (HULAFE; Spain, Valencia) Data Warehouse (EUCAIM Use Case).*

### *Annex I: Federated Node Hardware requirements for Federated Learning processing*

To effectively run federated learning algorithms over a decentralised network of hospitals, substantial computing power and specialised hardware are required. Federated learning involves repeatedly training machine learning models on sensitive data distributed across locations, while keeping the data in place. This requires significant processing capacity to iterate over training epochs in a reasonable time frame. The use of graphics processing units (GPUs) in federated learning systems is particularly impactful, as they can accelerate training time by handling compute-intensive matrix multiplications.

For a hospital federation, high-powered servers with GPUs are recommended to train models on site without interrupting normal computing operations. Models are trained iteratively on batches of local data, and the parameters are aggregated at a central server to update the global model. If underpowered, this process can slow hospital IT systems. Robust, high-memory servers ensure that federated learning can operate smoothly without hampering productivity or service.

Since the federated learning process cannot afford downtime, infrastructure must provide redundancy and ensure high availability. If servers go offline during model training, it can disrupt the synchronised learning across sites. Hospitals would need back-up servers, storage, network connections and power supplies to avoid single points of failure. With rigorous security measures and reliable IT systems in place, federated learning can be deployed in hospitals to build predictive models and gain insights, while keeping sensitive data safe within institutional firewalls.

**Federated Node Hardware Requirements**

Below you will find a table outlining the indicative Hardware requirements for a single Node as they have been defined so far in the context of the project. The requirements below are suggested to guarantee that the majority of EUCAIM use cases will be supported in terms of Node performance, and that no critical performance bottlenecks will occur during platform operations. These requirements are subject to update during the EUCAIM lifetime.

| **Hardware** | **Option 1** | **Option 2** | **Notes** |
|---|---|---|---|
| CPU | Minimum Cores: 16 >=1.8GHZ | Minimum Cores: 12 >=3.0Ghz | • If a GPU is not present, a server-grade, high core-count CPU is necessary for the Second Prototype<br>• If not comparable by cores, ideal thread count is 24+ |
| RAM | 64GB | 64GB | • DDR5 is ideal<br>• ECC memory is highly recommended for stability |
| Motherboard | 4+ Ram Slots | 4+ Ram slots | • Make sure to double check compatibility of selected CPUs with the Chipset of the motherboard<br>• In the case of DDR5, double check motherboard compatibility with DDR5 |

| | | | |
|---|---|---|---|
| Storage | 521 GB SSD Drive for Operating System (Either NVMe M.2 PCI Gen4 or SATA III)<br>1TB++ SATA III Drive (SSD or HDD) for local storage of medical data | | • M.2 NVMe Gen4 Drives are suggested for the OS<br>• For data storage size, DPs are expected to plan their purchase depending on the size of the Data they will provide. 1TB is a minimum, with some DPs already planning for 2 TB + datasets<br>• For data storage, SSD are preferred for speed but are not mandatory |
| Graphics card | NVIDIA Quadro | NVIDIA RTX 3XXX | • 12GB RAM+ is preferred<br>• Maximizing the amount of Tensor Cores is a priority, most recent GPUs will generally have higher Tensor Core counts<br>• Ampere and Volta architectures are preferred |
| Operating System | Linux | | • Latest version of any mainstream Linux distribution is acceptable Ubuntu, Alpine or other<br>• Windows is NOT acceptable, unless absolutely impossible for a DP to setup a Linux environment |
| Power Supply | - | | Each DP must make calculations depending on the hardware setup that will be selected to make sure that needed Wattage is covered and ideally exceeded to prepare of any future upgrades to the machine |
| Internet | 100mbps (baseline) | | Each DP must make best efforts to provide the best possible connection to their Node. Network performance will directly affect node stability and can invalidate AI training or prevent successful demonstrations of the platform |

## Data Provision Agreement (options for the relationship with EUCAIM)

There will be three main agreement options between data providers and the EUCAIM infrastructure.

1. To maintain the data within the original repository, the EUCAIM infrastructure will have access to the research repository metadata catalogue. The Data Provider shall provide the Data Receiver with this information (e.g. this solution envisions the lifetime of research projects).

2. A further step is to provide EUCAIM with federated access to the research repository dataset for Federated AI training after access is granted.

3. If agreed by both Parties (or at the termination of a research project, for example), anonymised data could be transferred from the Data Provider repository to the EUCAIM project cloud-based Atlas of Cancer Images. There might be an embargo period. Data to be transferred can be either the whole dataset or partial data from those partners who agree. Data completeness (images, and

related clinical information such as type of tumour, molecular profile, treatment or follow-up) will be defined upon agreement.

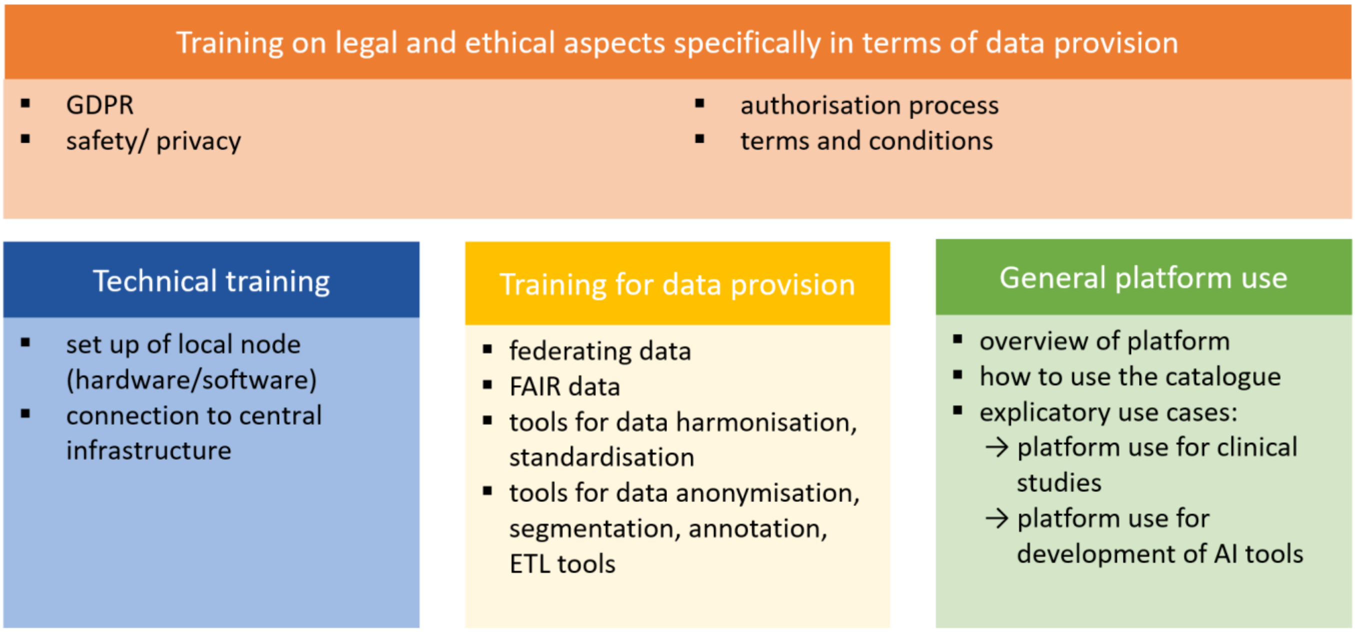
## Helpdesk overview

To assist the main stakeholders of the EUCAIM platform and receive the proper level of support, the project relies on EGI.eu to provide IT support to the users of the distributed and federated infrastructure.

The EUCAIM Helpdesk will act as a single point of contact for all users of the EUCAIM platform for requesting help, support and other requests. It provides ticket management and allows to track the inquiries related to EUCAIM services, resources, projects and general questions.

## List of Training Modules[2]

The training for data and tool providers will encompass one training module on legal and ethical aspects and three training modules, each consisting of a collection of training materials focusing on (1) technical provisions, (2) data provision and (3) general platform use.



Training on legal and ethical aspects specifically in terms of data provision

- GDPR
- safety/ privacy
- authorisation process
- terms and conditions

Technical training

- set up of local node (hardware/software)
- connection to central infrastructure

Training for data provision

- federating data
- FAIR data
- tools for data harmonisation, standardisation
- tools for data anonymisation, segmentation, annotation, ETL tools

General platform use

- overview of platform
- how to use the catalogue
- explicatory use cases:
  → platform use for clinical studies
  → platform use for development of AI tools

*Figure 4. Overview of the Training Modules.*

## Guidelines on data quality and FAIR principles

FAIR principles (Findable, Accessible, Interoperable and Reusable) are well described and summarised in document [3] as follows:

“FINDABLE: Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

ACCESSIBLE: Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.

INTEROPERABLE: Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

REUSABLE: Data and collections have a clear usage licenses and provide accurate information on provenance.”

[2] This overview of training modules contains the current status in the development of the training plan as of June 2023. The Training Plan (Deliverable 2.2) is due in December 2023 and may be subject to further revision.