EUCAIM
CANCER IMAGE EUROPE

Project title: European Federation for Cancer Images

Project acronym: EUCAIM

Grant Agreement: 101100633

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

D5.8 Set-up of Local Nodes for Data Federation

Author(s): Eirini Kaldeli (MAG), Gianna Tsakou (MAG), Valia Kalokyri (FORTH), Manolis Tsiknakis (FORTH), Mirna El Ghosh (LIMICS), Christel Daniel (LIMICS), Catherine Duclos (LIMICS), Laure Saint Aubert (MEDEX), Celia Martin Vicario (QUIBIM), Alejandro Vergara (QUIBIM), Jose Munuera (QUIBIM), Irene Marin (HULAFE), Olga Giraldo (DKFZ), Wahyu Wijaya Hadiwikarta (DKFZ), Enola Knezevic (DKFZ), Hanna Leisz (DKFZ), Clara Meinzer (DKFZ), Ignacio Blanquer (UPV), Esther Bron (Health-RI, Erasmus MC), Stefan Klein (Erasmus MC), Alexander Harms (Health-RI), Carles Hernandez-Ferrer (BSC), David Rodriguez Gonzalez (CSIC-IFCA), Victor Sóñora Pombo (BAHIA), Marco Aiello (SYNLAB), Alexandra Kosvyra (AUTH), Ioanna Chouvarda (AUTH), Dimitris Filos (AUTH), Dimitris Fotopoulos (AUTH), Ioannis Ladakis (AUTH) Alexandra Groth (Philips), Federica Cruciani (IFOM), Nuno Cruz, Santiago Frid (FCRB-HCB), Sebastiaan Huntjens, Katerina Nikiforaki (FORTH), Maciej Bobowicz (GUMed), Jose Alejandro Matute Flores, Dario Livio Longo, Heimo Müller (BBMRI-ERIC), Kurt Majcen (BBMRI-ERIC)

Reviewers Tobias Kussel (DKFZ), Michal Kosno (GUMed)

Date of delivery: 30.06.2025

Version: V0.1

Due date: Month 30

Type: Report

Dissemination level: Public

Table of contents

1. Introduction

1.1 Purpose and Scope

The current deliverable provides a description of the procedures to be followed and technical specifications to be met by data holders who wish to set up a local node compliant with the EUCAIM federation. A **Local Node (LN)** represents a technical infrastructure maintained by a Data Holder (DH), which stores and manages its data locally and is considered compliant with the EUCAIM ecosystem. A local node can support three different levels of compliance: three different levels of compliance: dataset cataloging (Tier 1), federated querying (Tier 2), and federated processing (Tier 3). A local node that supports federated queries and potentially also federated processing, i.e. a node that is Tier 2 or above, is called a **Federated Node**. It should be noted that the requirements and procedures to be followed by data holders who opt to transfer their datasets to one of the EUCAIM reference nodes are not covered by the current deliverable but are rather outlined in other documents, notably deliverable D5.6.

The objective of this document is to provide a concise but comprehensive overview of the various steps that local nodes should follow to meet the interoperability requirements following from the expected compliance level, while covering different cases regarding the status and format of the DH's primary data. To this end, it considers guidelines and methodologies produced as part of several WPs, including WP2 (about the helpdesk support and FAIR principles), WP3 , WP4 (about various aspects related to the rules of participation), WP5 (about the data interoperability requirements and supporting tools), WP6 (about the federated processing approach).

To avoid duplication, the current deliverable makes references to other documents, where relevant (e.g. with respect to the mandatory dataset metadata, the minimum clinical information etc). In this respect, it consolidates, updates and aims to present in a more concise way information included in previous deliverables, notably: **D5.6 - Minimum Data Federation and Interoperability Framework** and **D4.4 - Final rules of participation**. The information included in these deliverables is complemented with new insights and revisions that take into consideration the latest project developments, including new templates and documents produced in the period from February to June 2025. The most important developments concern:

- Updates with respect to the various data preparation tools and the search and federated processing components, accompanied with the publication of respective documentation in various formats. The latest versions of the tools are registered and documented on bio.tools and a Gitbook and the current deliverable refers to updated resources for more detailed information about each of the tools.
- The organisation of a (still ongoing) pilot study for the integration of a selection of 6 DHs of various Tiers. This process led to the update of important references, such as the Common Data Model, and the preparation of several types of material (mostly in the form of templates and forms), which can support the data holders especially at the initial data documentation and preparation steps of the setup procedure.
- The preparation of a Handbook, designed to guide DH through the onboarding process for sharing or transferring data to the EUCAIM infrastructure. Several sections of the current deliverables draw material from the handbook.
- An internal document outlining the technical requirements for data holders prepared under WP4 by the Technical Support team.

To outline the structure of this deliverable, Section 2 provides an overview of the local node architecture and some initial legal and organisational steps, which need to be followed by all DHs. The following core sections are structured along the EUCAIM levels of compliance. For each level, the following aspects are covered: the necessary steps for preparing the dataset metadata, clinical, and imaging data, so that DHs ensure alignment with the Tier-appropriate format and interoperability requirements (Section 3); the required hardware, network and software specifications (Section 4); and instructions for interconnecting with the EUCAIM federated search (Section 5) and processing components (Section 6). Finally, an overview of the pilot study is provided in Section 7.

2. Architectural overview and initial preparatory steps

2.1 Overview and basic concepts

The EUCAIM Data Federation is a decentralized infrastructure enabling the secure and privacy-preserving sharing and processing of cancer imaging and clinical datasets across Europe. As a decentralized system, it enables multiple institutions to collaborate while maintaining control over their own data. Unlike centralized data repositories, where all data is collected and stored in a single location, a federated network such as EUCAIM allows data to remain within the institutions that own it. Instead of transferring large amounts of sensitive information, federated networks provide structured and secure mechanisms for researchers and analysts to query, analyze, and process data remotely.

One of the primary functionalities supported by the EUCAIM federated network is federated querying, which allows external users to search datasets across multiple institutions without the need to have direct access to the actual data. Once the relevant datasets are identified, institutions may choose to transfer the requested data under strict governance policies. The EUCAIM federated network also enables federated data processing, where institutions collaborate on computational tasks without exposing or transferring their raw data. In this model, algorithms are sent to local datasets instead of centralizing the data itself. In this context, machine learning models can be trained locally on each institution's dataset and then aggregated to create a more robust global model, without any individual dataset ever leaving its original location. This method has proved to be a powerful approach for large-scale collaborative research, while ensuring privacy and security.

EUCAIM defines two ways of participation for data holders:

1. **Data holders transferring data to a Reference Repository**. This option is available for DH who wish to transfer their anonymised datasets to one of EUCAIM Reference Nodes. EUCAIM provides two reference nodes that can host data from data holders who do not wish to set up local services that support the search and processing of their data. Reference Nodes have secure processing environments where data can be processed safely.
2. **Data Holders who set up a Local Node**: This option is available for Data Holders who wish to maintain and manage their datasets on a local infrastructure deployed on their premises, which is compliant with the EUCAIM ecosystem. Various levels of compliance are supported, as explained below.

As already mentioned, the current document describes the setup procedure that data holders should follow to become a LN. Data holders who wish to participate in the EUCAIM federation by transferring their data to a Reference Node should follow a separate process, as detailed in Sections 4.3.3, 5.2.2 and 6.2.2 of D5.6 (covering the case of data holders who wish to become Tier 1, 2, and 3, respectively) and overviewed in Section 6 of the Handbook.

2.2.1 Compliance Tiers

Datasets are categorized into “Tiers”, according to their level of compliance with the EUCAIM Data Federation Framework. Different federated concepts apply to the different tiers, which implies different technical requirements. It is important to emphasize that in order to achieve a higher tier, you must meet the requirements of the previous tier, as each higher tier encompasses the requirements of the lower ones.

The capabilities supported by each compliance Tier are:

1) Tier 1: The datasets hosted by the local node are registered in the central catalogue. Users can explore the metadata of the datasets registered in EUCAIM’s platform.
2) Tier 2: The data of the federated node is searchable via the EUCAIM search system. The users can explore the actual number of subjects and studies fulfilling a number of search criteria defined by the user.
3) Tier 3: The federated node has a materialisation component that makes the data available to the federated processing, according to EUCAIM’s model. The user will be able to run processing actions on the actual data, if the access to them is granted.

2.1.3 Local Node architectural overview

Figure 1 provides a high-level overview of the Local Nodes’ architectural schema, showcasing its main local components and their interactions with the EUCAIM central services. Local data, including image data and, if relevant, clinical data should be processed following the workflows depicted in Figures 2, 3, and 4 in subsequent Sections of this document, depending on the Tier of the local node and whether it chooses to transfer its data to a Reference Node. After the data are processed so that they are in line with the EUCAIM standards, they have to be registered to the Public Catalogue (see Section 4.3). As already noted, the DH may choose to transfer their datasets to the EUCAIM Reference Node, in which case they do not have to set up a local node.

Tier 2 nodes need to connect with the Federated Search Beam broker to support federated queries. To this end, they need to deploy the Beam Proxy and Focus Docker components (see Section 5.2). Moreover, they potentially need to implement a Mediator component that maps the queries and the responses between the formats expected by the federated search and the ones produced by the local search implementation. In addition to the Tier 2 capabilities, Tier 3 nodes must deploy and configure a Federated Execution Manager (FEM) client, that connects to the FEM manager located at the EUCAIM’s central node, as well as the Data Materializer Tool (DMT), which allocates the right data through a Sandbox environment (see Section 6).

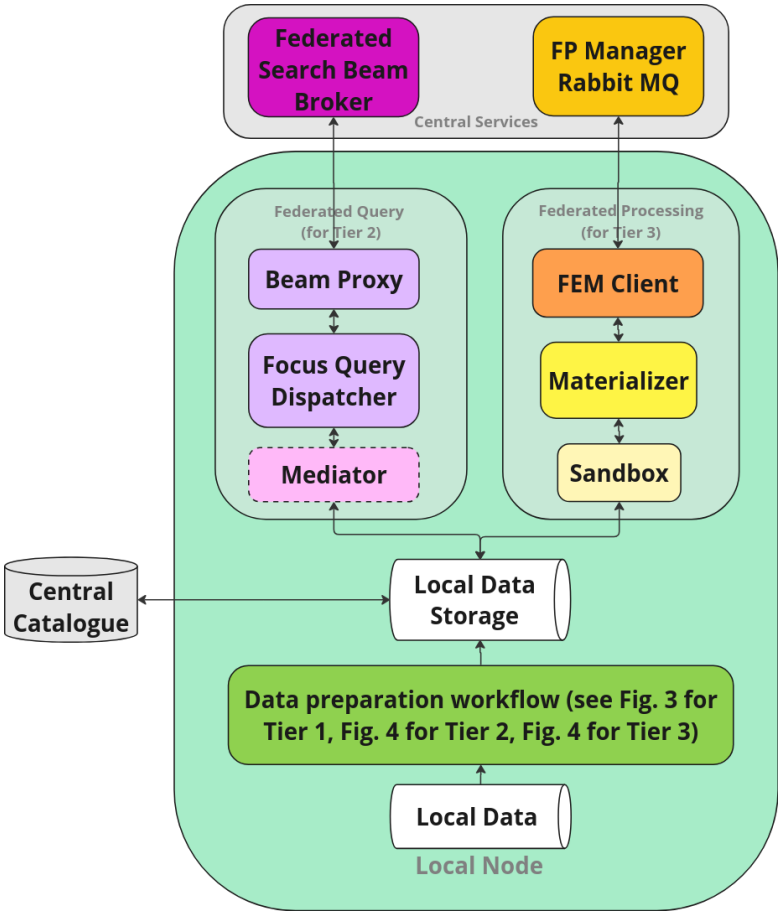

Figure 1: Architectural overview of a local node

2.2 Supporting resources and mechanisms

EUCAIM has produced various types of informative and capacity building resources that cover various aspects of the EUCAIM infrastructure and which can be useful for DH who wish to become a member of the federation. These resources come in various formats including dashboard profile pages; a set of GitBooks; documents; a Youtube channel; the EUCAIM training platform; and a HelpDesk. The following Table provides an overview of some of the resources that DH can consult to find out more details about various topics relevant to the onboarding process.

Support source	Purpose	Link (s)
EUCAIM Glossary	Glossary of the most commonly used terms in the field of Health Data Research	https://eucaim.gitbook.io/glossary/
Dashboard profile pages	Get general awareness of the platform functionality and purpose, and the user profiles.	https://dashboard.eucaim.cancerimage.eu

Support source	Purpose	Link (s)
Dashboard GitBook	Brief description of the components and functionality available from the Dashboard.	https://eucaim.gitbook.io/eucaim-dashboard
End User Guide GitBook	Information on how to use the platform being a Data User, Data holder or SW provider.	https://eucaim.gitbook.io/end-user-guide
Architecture GitBook	Details on the architecture, software dependencies, protocols and interactions of the components in EUCAIM.	https://eucaim.gitbook.io/architecture-of-eucaim
EUCAIM Common Data Model v3.0	Description of the EUCAIM Common Data Model (CDM) and Hyper-ontology	https://eucaim.gitbook.io/eucaim-common-data-model/ https://docs.google.com/spreadsheets/d/1ox9PdvfCDxpDmEnFzC1M6OFhUhXpjQzg/edit?gid=357336097#gid=357336097 https://www.google.com/url?q=https://zenodo.org/records/15558108&sa=D&source=docs&ust=1749645280452875&usg=AOvVaw0K-S-P8ZoZ3MTJ_D5jk2_3
YouTube Channel	Videos with interviews and demonstrations.	https://www.youtube.com/@EUCAIM
Training Platform	Provides access to comprehensive training materials that facilitate usage of the platform for data holders, data users and software providers.	https://training.eucaim.cancerimage.eu/
Helpdesk	When you face an issue you can submit a ticket to the Help Desk. Additionally, actions that require interaction with the user are encouraged to be done	https://help.cancerimage.eu

Support source	Purpose	Link (s)
	through the helpdesk.	

Table 1: Sources of documentation and support in EUCAIM

The Helpdesk is the main mechanism used to support DH in their onboarding process. DH who wish to participate in the EUCAIM federation are encouraged to register to the EUCAIM Helpdesk and use its ticketing system to submit their questions and issues. In particular, technical questions should be addressed to the Technical Support Team unit group of the Helpdesk. The Helpdesk can be contacted in two ways:

1. DH who already have a Life Science AAI account can access the Helpdesk user interface on https://help.cancerimage.eu/ by authenticating using their account. There, they can create a ticket describing the issue, assign it to the dedicated support team, and follow its status.
2. DH who do not have a Life Science AAI account should access the Helpdesk from the EUCAIM dashboard via a webform, under the "Helpdesk" menu. There, they will be requested to provide their contact information in the webform, in order to receive assistance. Once the form is submitted, it will automatically create a ticket in the Helpdesk instance, where the Technical support unit will be able to access the request and address it.

2.3 Initial preparation steps: legal, security, and organisational aspects

Before starting with the core technical workflow, DH should first make sure that they abide by legal, security and organisational requirements. DH who set up a local node have to demonstrate that the site implements good practices related to security and privacy preservation. Although they are not mandatory, a certification such as ISO/IEC 27001 and/or ISO/IEC 27701 would be appropriate to prove this capability. As a reference, DHs can consult the security measures followed by the UPV reference node and adapt the measures to their own organisational context. DHs should make sure that they assign and declare responsible persons and contacts in charge of the management of the local node, the monitoring, backuping and security incidents. The local node should implement periodic security audits. External security audits are also encouraged.

Local nodes have to guarantee that they commit enough resources to deal with the necessary level of service. This should be committed by signing a Service Level Agreement that declares the resources committed to the infrastructure, the Service support conditions, the committed Availability and Reliability and the contact points. As a basis, the UPV reference node has the Service Level Agreement publicly available here.

Before starting with the core technical workflow, DH should make sure that they complete the following preparatory steps that concern *organisational and legal compliance*:

1. Entities that are not EUCAIM partners and wish to join the EUCAIM data federation, should first complete the Expression of Interest. When a DH submits an application, the Access Committee initiates the review process. It coordinates with the Technical Board (TB) which evaluates whether the proposed infrastructure, anonymisation protocols, risk analysis and data quality controls are in line with EUCAIM's technical requirements. At the same time,

the Ethical and Legal Board assesses the legal documentation submitted as evidence for the technical aspects reviewed by the TB, verifying its compliance with data protection and ethical norms. Once all evaluations are completed, the Access Committee prepares a consolidated report which is sent to the Management Board and Steering Committee to make the final decision. Throughout the process, DH are expected to collaborate closely with the involved boards, provide documentation, and requests for clarification.

2. DH should register into the EUCAIM's Dashboard with a Life Science AAI (LS-AAI) account, following the guidelines available on the AAI in EUCAIM.
3. DH should complete the Tier Maturity Level Questionnaire, that allows the assessment of the readiness and compliance of datasets provided by data holders.
4. DH should ensure compliance with legal and ethical requirements. It is essential that DHs provide a contact person of its legal team to be in close communication with the legal team of EUCAIM. A set of legal agreements must be prepared and signed to clearly state the obligations and responsibilities of the parties involved, as overviewed in Table 2. More detailed information about the necessary legal documents that need to be prepared can be found in Section 3.2 of the Handbook.

Action	**Description**	**Documents**
Provide documentation	- Proof of legal representation and legal basis if necessary. - A report from the DPO confirming legal compliance. - Data Protection Impact Assessment (DPIA) - Documents demonstrating the security of the information system. - Any documents required under your national legislation.	D4.4 Final rules for participation report (see Sections 4.4.1 (Legal requirements) and 4.4.2 (Ethical requirements for Data Holders)
Data Sharing Agreement	Fill-in and sign the DSA	Draft DSA
Define special Access Conditions	A Document to be signed by the Data User that indicates the conditions under the Data User can access the data.	Draft Template
Contact point for the negotiation (Only in federated nodes)	The LS-AAI details of the data holder delegate who will interact with the Data User through the negotiator.	Registration of users in EUCAIM LS-AAI.

Table 2: Summary of legal compliance steps required by Local Nodes.

3. Data preparation and registration

3.1 Overview of data preparation process and tools

DH should take a number of actions to ensure that their data (including imaging and clinical data as well as dataset metadata) are compliant with the EUCAIM requirements depending on their Tier. The process starts with the annotation and de-identification of data, is followed by steps ensuring alignment with EUCAIM standards, complemented with re-identification risk assessments and quality checks, and concluded with registration to the EUCAIM catalogue. The steps vary depending on the Tier, as detailed in the following sections, but the overall outline covers the following steps:

- **Compliance with Tier 1** : Data has to be in the required format and de-identified. Quality and re-identification risks must be checked, and the dataset needs to be registered in the public catalogue. The process is detailed in Section 3.2.
- **Compliance with Tier 2**: Data required for the federated search must be de-identified and harmonized, so as to ensure semantic alignment with the EUCAIM common data model. The harmonization can take place either via a data transformation process or via the setup of a query mediator component. Quality checks and re-identification risk assessment must undergo a more comprehensive verification process. The process is detailed in Section 3.3.
- **Compliance with Tier 3**: The imaging data have to follow a certain structure and transformation to the EUCAIM common data model is required. Quality and de-identification risks must be verified at the highest level. The process is detailed in Section 3.4.

EUCAIM offers several tools that can support various aspects of the data preparation workflow. Table 3 provides an overview of the main tools selected for this phase. More details on the use of each of these tools are given in the following sections.

Tool	Data preparation step	Tiers	Documentation
MITK (Medical Imaging Interaction Toolkit)	Data annotation	All Tiers - recommended if relevant	https://bio.tools/mitk
NIfTI to DICOM-SEG converter	Imaging format alignment	All Tiers - recommended if relevant	https://hub.docker.com/r/mariov687/dicomseg
EUCAIM Anonymizer	Data anonymization	All Tiers - recommended if relevant	https://bio.tools/eucaim_dicom_anonymizer

Trace4Medical Image	Data anonymization and re-identification risk assessment	All Tiers - recommended if relevant (for 2D ultrasound and mammograms)	https://bio.tools/trace4medicalimagecleaning
DICOM File Integrity Checker	Data quality assessment (on imaging data)	All Tiers - recommended	https://bio.tools/dicom_file_integrity_checker_by_gibi230
Data Integration Quality Check Tool (DIQCT)	Data quality assessment (on both clinical and imaging data)	All Tiers - particularly recommended for Tier 2 and 3	https://bio.tools/data_integration_quality_check_tool_diqct
DICOM tags extractor	Data transformation to EUCAIM CDM; Preparation to feed the re-identification risk assessment	Recommended for Tier 2 and 3 to feed the ETL tool; Can be used by Tier 1 to feed the Wizard tool.	https://bio.tools/dicom_tags_extractor
Wizard tool	Re-identification risk assessment	Recommended for all Tiers: Tier 1 to assess the DICOM metadata; Tier 2 and 3 to assess both the clinical and imaging metadata in the EUCAIM CDM.	https://bio.tools/eucaim_wizard_tool
ETL tool	Data transformation to the EUCAIM CDM	Mandatory for Tier 3 nodes, if the data does not already follow the EUCAIM CDM; Recommended for Tier 2 nodes.	https://bio.tools/etl_toolset

Table 3: Overview of preparation tools.

3.2 Node setup for data preparation for Tier 1 nodes

Tier 1 Data Holders must ensure that their data satisfy the following minimum requirements:

- Imaging data should be in **DICOM format** and associated annotations and segmentations, when available, should be in **DICOM-SEG format**. Specific cases will be considered if a DH has diagnostic images in different data formats. In exceptional cases where NIfTI format images are provided, these will only be considered if the Data Holder (DH) is unable to retrieve the original DICOM files from which the NIfTI images were derived. In such cases, DHs are responsible for ensuring the minimum requirements for anonymization, risk analysis, and quality through their own tools and procedures.
- All images should include a **minimum set of imaging metadata** in accordance with the DICOM standard. The minimum imaging metadata are outlined in Annex 3 of D4.4, available in this document.

- Imaging data must be accompanied by a set of **minimum clinical metadata**, detailed in Annex 3 of D4.4 referenced above.
- Image and clinical data must be linked using the same patient identifier.
- No entity can exist more than once within a dataset.

The steps that Tier 1 nodes should follow to set up their local nodes in accordance with the EUCAIM requirements for the clinical and imaging data as well as the dataset metadata are overviewed in Figure 2 and explained in more detail in subsequent sections.

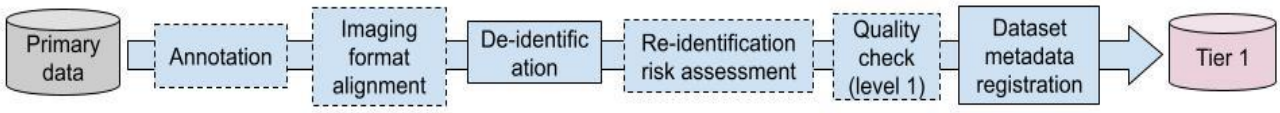

Figure 2. Data preparation workflow for Tier 1 nodes.

3.2.1 Setup for data annotation and imaging format alignment

Although it is not a requirement, datasets with associated annotations are highly desired and recommended into the EUCAIM federation as they enhance the datasets' value and usability. The availability of good-quality and unbiased annotations is important for training AI models, so that they deliver better outcomes, but also for reducing future manual annotation workload through pre-segmentations.

If annotations have not been previously generated, DHs can use their software of preference to support the annotation process or use the **MITK Workbench** provided by EUCAIM. The toolkit provides a user-friendly interface for viewing, processing, and segmenting medical images. It offers a comprehensive set of slice-based 2D and (semi-)automated 3D segmentation tools, that allows DH to enrich medical images with annotations. More information about the tool can be found in the tool's respective entry on bio.tools here.

To achieve standardized annotations, the development of annotation guidelines is underway, based on establishing a common process and criteria across different use cases, with a focus on standardization and homogenization. When ready, the annotation guidelines will become available as training materials in the EUCAIM Training Platform. Annotations should be associated with a set of mandatory annotation metadata as described in Annex 3 of D4.4, based on the DICOM SEG standard. Additionally, all mandatory DICOM tags as specified in Annex 3, must be present at all times in the DICOM SEG files.

DICOM SEG is preferred over NIfTI for medical image segmentation, due to its dedicated and standardized format. It is recommended that DHs convert NIfTI or other formats to a valid DICOM format, where possible. This will align the dataset with the rest of DICOM based EUCAIM datasets, facilitate the dataset preparation process and ease its future upgrade to higher Tiers. However, some datasets in NIfTI format might be accepted for Tier 1. Annotations are also preferred in DICOM SEG format, although this is not a requirement for Tier 1.

If annotations are in formats other than DICOM SEG (such as NIfTI or RTSTRUCT) and the respective original imaging data in DICOM format is available, the **EUCAIM Converter** can be used

to convert them to DICOM SEG. The tool handles the following conversion scenarios: One NIfTI to One SEG; Multiple NIfTIs to Multiple SEGs; Multiple NIfTIs to Multiple SEGs; Non-Overlapping SEG to NIfTI; Overlapping SEG to NIfTI. The tool is containerised and can be executed via Docker as explained here. An extension of the tool is currently under development to include support for the DICOM RT STRUCT format as a non-standard format that can be converted to DICOM SEG. DICOM RT STRUCT was chosen due to its widespread use in radiotherapy, where a significant number of annotations are expected to adhere to this standard.

3.2.2 Setup for data de-identification and re-identification risk assessment

To ensure the privacy and protection of patients in the data available in EUCAIM, it is imperative that the data undergo rigorous processing to eliminate any elements that could potentially lead to patient identification. Direct identifiers are elements of the original data that are explicitly unique to a patient, such as their name, social security number, or original patient ID. These elements should either be entirely removed or replaced by pseudonyms before the data is shared. Pseudonymisation is only allowed under certain conditions, e.g. in case of dedicated national legislation provisions, that need to be approved by the EUCAIM. Besides direct identifiers, indirect identifiers must also be addressed. Indirect identifiers refer to data points that, while not explicitly unique, can still be used in combination with other information to re-identify an individual. For example, attributes like a patient's date of birth or detailed clinical measurements could, when combined, uniquely identify a person.

It is also important to highlight that data privacy concerns extend beyond the individual patient level. Privacy risks can arise at a dataset level due to unique combinations of attributes that create outliers, which potentially expose individual identities when analyzed. For instance, a dataset might include a rare combination of medical conditions or treatment outcomes that uniquely identify a single patient. To mitigate this risk, datasets should be carefully reviewed for such combinations and appropriately processed by the DH.

Furthermore, medical imaging data also present unique privacy risks, which must be carefully managed to preserve patient confidentiality. Textual information embedded in images, such as ultrasound, may include sensitive patient details, such as names, dates of birth, or medical record numbers. Beyond textual information, medical images can pose additional risks of re-identification due to their inherent uniqueness. For example, certain anatomical features or abnormalities may be so distinct that they could inadvertently identify an individual. Moreover, medical images that include the head or facial structures also introduce privacy concerns. Such risks necessitate rigorous de-identification measures taken by the data holder.

In addition to imaging data, EUCAIM also includes clinical data from patients, making it essential to preserve the link between these datasets during the de-identification process. Specifically, if a patient ID in the DICOM metadata is replaced, the same new ID must be consistently applied to the corresponding clinical data to maintain the integrity and usability of the datasets. If any clinical data is also linked at a study level, the same procedure should be performed to guarantee that no relation is lost.

DH are responsible for ensuring that no identifiable information (direct or indirect) is present in the datasets they share. To this end, they are free to use any software of their choice or take advantage of a set of tools offered by EUCAIM:

1) **EUCAIM DICOM Anonymizer**: The Anonymizer can be used to de-identify the DICOM metadata. It supports both anonymization on a case-by-case scenario, meaning one DICOM Study at a time (single-mode) or on multiple cases concurrently (batch mode). Moreover, hashing is used to de-identify the identifiers of the clinical data, while preserving the linkage between image and clinical data. The tool can be applied if the data meets certain requirements as described in the tool's documentation, accessible here.

2) **Trace4MedicalImage**: The tool is recommended for datasets that include 2D ultrasound and mammograms. The tool analyses 2D ultrasound and mammography DICOM files to detect and remove encapsulated text, as it may contain potentially identifying information. It produces a list of processed files and corrected DICOM files.

The use of the **Wizard Tool** is recommended for Tier 1 nodes to assess the risk of re-identification of patients by analysing their **imaging metadata.** The necessary imaging metadata to feed the Wizard tool can be extracted by using the **DICOM tags extractor** tool. It should be mentioned that no automatic re-identification risk analysis on a combination of clinical and imaging metadata is possible for datasets that do not follow the EUCAIM Common Data Model. However, it remains a responsibility of the DH to ensure that no direct or indirect identifiers are present in their clinical data.

3.2.3 Setup for data quality assessment

All datasets added to the EUCAIM federation must comply with a data quality standard. This applies to all tiers, including Tier 1 data. The EUCAIM data quality framework addresses the following dimensions, taking into consideration the recommendations by the European Health Data Space (EHDS): Completeness, Uniqueness, Validity, Consistency, Accuracy, and Integrity. EUCAIM assesses these dimensions based on a set of quality metrics. More information about the EUCAIM quality dimensions and metrics can be found in Section 4.3.2 of D5.6.

The most important quality dimensions for Tier 1 nodes include validity, accuracy, and integrity. EUCAIM provides a set of optional tools that DH can make use of to assess their compliance with the EUCAIM data quality standards. The **Trace4MedicalImage** mentioned in Section 3.2.3 can be used for the purpose of conducting a validity check on 2D ultrasound and mammography DICOM files.

The **DICOM File Integrity checker** tool is recommended to perform a quality check in terms of accuracy and integrity. The tool performs various checks with respect to the correct number of files per series, the presence of corrupted files, the precise directory hierarchy, the identification of separated dynamic series for merging them, interest series filtering/selection by specific series description lists, and diffusion sequence identification by b-values. It generates a report containing information about corrupted files, missing files, merged series, the selected series and, optionally, copying the dataset while applying the desired changes, such as correcting folder names, fixing the

directory hierarchy, selecting specific series (without copying the non-selected ones), and applying the necessary DICOM tags changes to identify the previously unmerged series as joint ones. Note that corrupted files are also copied, and the user has to manually remove the full series (since it is useless in this state) or retrieve the original non-corrupted files and replace them if possible.

3.3 Node setup for data preparation for Tier 2 nodes

Tier 2 nodes should meet the requirements about the imaging and clinical data set out in Section 3.2 about the Tier 1 nodes, with the additional proviso that imaging data is in DICOM format and annotations, if applicable, in DICOM SEG. Thus, the step for ensuring alignment with the required imaging format is mandatory (see Figure 3).

Moreover, Tier 2 nodes should ensure that at least the minimum imaging and clinical attributes are structured in a way that supports querying. The mandatory and "mandatory if available" query criteria used by the EUCAIM basic federated queries are defined in Tables 14 and 15 of D5.6. Additional desired query criteria are presented in Table 16 of D5.6.

Tier 2 nodes can support compliance with the aforementioned search criteria via two ways:

a) **Transformation to the EUCAIM Common Data Model**: For nodes lacking a pre-existing database, EUCAIM encourages that datasets are directly transformed to meet the EUCAIM Common Data Model (CDM). In this case, the DH does need to implement a custom mapping component, since the datasets adhere to the CDM structure, the federated query can be directly executed through **Focus** without additional mapping. The transformation can be made through an ETL (Extract, Transform, Load) process supported by EUCAIM's tools (see below). This transformation also ensures automatic compliance with Tier 3 data model requirements. The ETL process is described in Section 3.3.2 below.
b) **Implementation of a Mapping Component:** A mapping component must be implemented, to translate the requested minimum set of clinical and imaging attributes from the DH's local structure to the EUCAIM concepts. This mapping ensures that queries expressed using the EUCAIM concepts (e.g. "imaging modality", "age at diagnosis") can be executed seamlessly across all participating nodes. In addition, this mapping component must be responsible for translating the search query (AST syntax), into the local node's query language. More information about how to implement a custom mapping component is provided in Section 5.

Tier 2 nodes should follow the steps for data annotation, alignment of imaging data, and de-identification as described for Tier 1 nodes in Sections 3.2.1 and 3.2.2. However, for Tier 2 nodes that opt to transform their data to the CDM, the use of the **Wizard Tool** for the assessment of the re-identification risk should take place after the transformation to the CDM, since several checks by the Wizard tool assume that the data follow this model. The steps that Tier 2 nodes should follow to set up their local nodes in accordance with the EUCAIM requirements are overviewed in Figure 3.

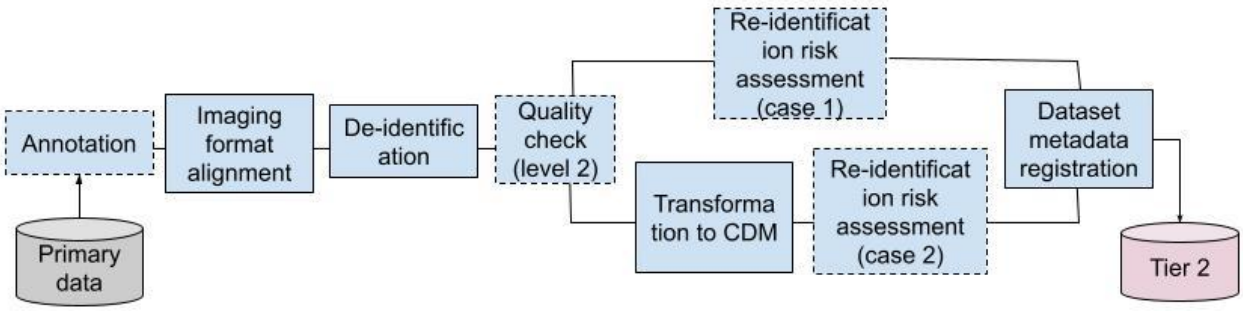

Figure 3. Data preparation workflow for Tier 2 nodes.

3.3.1 Setup for data quality assessment

The EUCAIM data quality framework applies to Tier 2 datasets, with Tier 2 nodes being required to address all quality dimensions of the framework, i.e. completeness, uniqueness, validity, consistency. The **Trace4MedicalImage** and the **DICOM File Integrity Checker** mentioned in Section 3.2.3 are recommended for use by Tier 2 nodes as well. In addition to these, the use of the **Data Integration Quality Check Tool (DIQCT)** tool is highly recommended. The tool applies both to clinical and imaging data. It checks the clinical metadata quality (validity, completeness), the integrity between images and clinical metadata provided, the de-identification protocol applied, the imaging analysis requirements, and informs the user on corrective actions prior to data upload. The DIQCT tool also includes a module dedicated to annotations, which enables the validation of DICOM SEG files against predefined requirements specified in an Excel file. A detailed report is generated, highlighting issues such as missing, invalid or conditionally required attributes, including file paths and affected DICOM tags.

3.3.2 Transformation to the EUCAIM Common Data Model and de-identification risk assessment

Transformation to the ECAIM CDM is mandatory for Tier 3 nodes and recommended for Tier 2 nodes, as explained above. The EUCAIM CDM defines a standardized structure for representing clinical and imaging metadata across the EUCAIM platform, ensuring that data contributed by different partners can be understood and used in a consistent way. It supports multimodal data (i.e. imaging and clinical) and facilitates efficient querying, tool compatibility, and federated analysis and learning. The CDM is based on the conceptual model of mCode specification and its most recent version can be found here.

A terminology-binding process is followed to ensure that the data elements represented in the EUCAIM CDM are semantically aligned with standardized biomedical concepts and data properties as described in the EUCAIM hyperontology. The hyper-ontology ensures semantic interoperability and consistent harmonisation of heterogeneous information, including different types of data (clinical, biological, and imaging), into a common semantic meta-model. It provides rich context, making it easier for users and tools to interpret, search, and reason over the data. This ensures a coherent interpretation and understanding of data between the hyper-ontology and CDM.

As a DH, understanding the CDM and hyperontology is essential for:

- **Mapping your data correctly**: Ensuring your local dataset aligns with EUCAIM standards.
- **Using tools effectively**: Tools in the EUCAIM ecosystem rely on the CDM to operate correctly.
- **Supporting reproducibility and scalability**: Harmonized data makes it easier to run federated analysis and integrate new tools.

In order to transform their data in accordance with the EUCAIM CDM, DH should follow an Extract-Transform-Load (ETL) process with the support of appropriate tools. The ETL process consists of the following steps:

1) *Preparation for the imaging dataset*: The DICOM metadata can be extracted with the support of the **DICOM tags extractor**. The tool scans the imaging DICOM files of a defined directory at the series level, and produces as output a single JSON file containing all the DICOM tags for each detected series. If a list of selected DICOM tags is provided, the tool can also produce a single CSV file containing the listed DICOM tags. This CSV file with the DICOM tags including the header can then be passed to the ETL tool (see step 3).
2) *Preparation for the clinical dataset*: DHs can make use of the **ETL tool** to transform their data in line with the EUCAIM CDM. In this case, they should follow a preparation process and provide a tabular file that describes how each clinical variable in their local dataset should be mapped to the semantically equivalent data element of the EUCAIM CDM. Moreover, the tabular mapping file should provide appropriate information about the values of the local variables and the use of standards or custom value sets. A **template** with instructions and examples has been prepared to guide DHs through this mapping process. The clinical data should be provided as CSV, JSON or XLSX files, with files corresponding to different datasets stored under separate dedicated folders. The combination of folder name and file name will be used by the ETL to identify the dataset. One dataset may comprise a single file or several complementary files, as long as all files belonging to the same dataset keep the same schema (if JSON) or header (CSV, XLS).
3) *Actual transformation*: Taking as input the domain-specific mapping rules following from the previous step, the **ETL tool** can be applied on the clinical data and imaging tags and transform them to a structure compatible with the EUCAIM CDM. The ETL is offered in the form of a Docker container, providing a modular, extensible and customizable pipeline that allows a step-by-step approach. The converts the entry files and runs quality checks to ensure proper transformation, invoking another tool (also dockerized) the Tabular Data Curator as part of these quality checks. It allows data holder users to fully accept the automatic quality checks, review them, or correct them manually. The result of the transformation process is stored in a Postgres database (local to the node, also dockerized) that can be queried by the EUCAIM federated search and processing components. Additionally an output as CSV, JSON or XLS format but compatible with the EUCAIM CDM can be generated, for easier usage of other EUCAIM tools, such as the federated processing component. DHs are free to follow their own custom process and make use of tools of their choice to ensure that their data is provided in a format aligned with the EUCAIM CDM. Dedicated transformation pipelines to support DHs who maintain their clinical data in the FHIR and OMOP healthcare data standards may be considered.

The **Wizard tool**, already mentioned above, can be used to assess the re-identification risk. For Tier 2 nodes, the Wizard tool considers both the available imaging and clinical information, provided in the EUCAIM CDM format.

3.4 Node setup for data preparation for Tier 3 nodes

Figure 4 depicts the data preparation steps that DH should follow to become Tier 3 nodes. The process is the same as for Tier 2 nodes, with two additional requirements: (a) that imaging data is provided following a specific structure; and (b) the transformation to the EUCAIM CDM is mandatory.

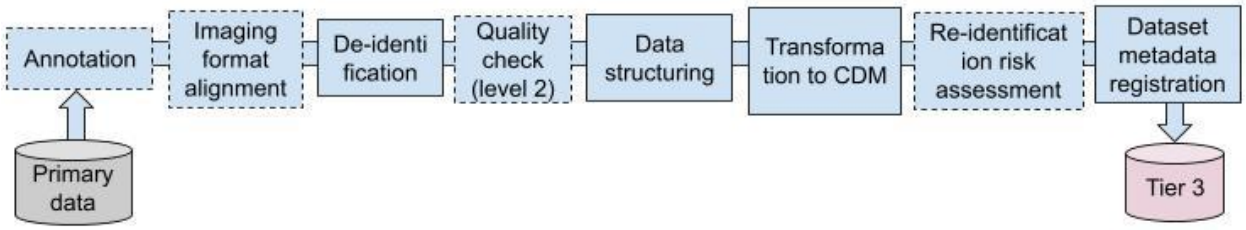

Figure 4: Data preparation workflow for Tier 3 nodes.

3.4.1 Setup for image data structuring

Tier 3 datasets should allow the federated processing EUCAIM software to operate autonomously across all nodes. Therefore, it is crucial to provide imaging data in a well-structured manner with precise mapping of each dataset, patient, study, and series. Hence, DHs should consider the following aspects when compiling their Tier 3 cohorts, additional to the data requirements for Tier 2 nodes:

- **Imaging Data Shape and Structure**: The imaging data should be organized according to the designated hierarchical folder structure described in Figure 20 of D5.6. This can be achieved through a custom procedure by the DH or by using the **DICOM File Integrity Checker**, which includes a functionality to process the input imaging dataset and save it in the described structure.
- **Annotations Managing**: When annotations are included on a Tier 3 dataset, i.e. a Tier 3 A+ dataset, they must be in DICOM format and must be added as an additional series in the same series hierarchy level, as shown on Figure 3 of D4.6.
- **Series Identification and Tagging**: The identification of relevant series is crucial for both efficient visualization and automated processing. Medical imaging studies often include a mix of series, many of which are irrelevant for secondary use or difficult to identify due to non-standardized naming conventions (e.g., the SeriesDescription DICOM tag). In Tier 3 datasets, normalizing these names is essential to enable tools to autonomously identify and process the required series within a federated environment. DHs should therefore consult EUCAIM's standardized naming conventions and either update the SeriesDescription tags in their DICOM files or apply normalized names to the directory paths where Tier 3 files are stored for processing. The **DICOM File Integrity Checker** tool can support this process by identifying and tagging relevant series based on a predefined list of SeriesDescription tags provided by the DH.

Data holders may choose to host a local imaging data repository to store the DICOM imaging data. This is not mandatory, but could facilitate local data management of imaging data for multiple

projects. The setup of the local imaging data repository could be based on XNAT[1], similar to that used in the Health-RI Euro-BioImaging reference node. The DatMat tool[2] can be used to facilitate the mapping of the data to the Imaging Data Shape and Structure described above while it provides an out-of-the-box plugin to work with XNAT. In case a data holder uses a local imaging data repository other than XNAT, a custom DatMat plugin can be written in Python.

The remaining steps for the data preparation setup of a Tier 3 node are the same that apply for Tier 2 nodes, with the additional requirement that the mapping to the EUCAIM CDM is mandatory for Tier 3 nodes.

3.5 Data FAIRification and registration to the EUCAIM catalogue

To achieve interoperability at a dataset cataloguing level (Tier 1), EUCAIM adopts established standards for the definition, documentation and exchange of aggregated dataset metadata across the federation. The EUCAIM DCAT Application Profile, defined in deliverable D5.2, is based on the widely used DCAT-AP and the recently published HealthDCAT-AP standards for describing health-related datasets that also comply with the European Health Data Space Regulation. The EUCAIM DCAT-AP, as an application profile, re-uses terms from one or more base standards, adds more specificity by identifying mandatory, recommended and optional elements, and requires the use of specific controlled vocabularies to guarantee interoperability.

EUCAIM supports two ways for DH to register a dataset in the central catalogue:

1. Follow a manual process supported via the Helpdesk, which requires the completion of the necessary dataset information in a template sheet, which is compliant with the EUCAIM DCAT Application Profile. The steps that DHs have to follow in this case are described in Table 4.
2. Setup a FAIR Data Point ((FDP), from which the metadata is then harvested and added to the catalogue. The harvester is still under development, with its latest status described in D4.6 - Final Federated Core services.

Action	Description	Support
Complete the dataset's metadata in the spreadsheet template (Data Holder Template sheet)	The dataset schema can be downloaded from this link. In case of doubts with the terminology, use textual descriptions.	Open a helpdesk ticket under the category "catalogue".
Make a request of registry upload	Create a helpdesk ticket on the category catalogue, providing the spreadsheet file with the metadata	Same procedure.

[1] https://wiki.xnat.org/documentation/xnat-installation-guide
[2] https://pypi.org/project/datmat/, https://gitlab.com/radiology/infrastructure/data-materialisation/.

	information. The helpdesk team will contact you back, informing whether the dataset has been properly registered or requesting more information.	
Verify the entries in the catalogue	Access the registry in the catalogue at the URL: https://catalogue.eucaim.cancerimage.eu/#/collection/<<identifier>>	Same procedure.

Table 4: Steps for the manual registration of a dataset to the EUCAIM catalogue.

Experts in the technical committee of EUCAIM will check the metadata and manually register them in the public catalogue. This superuser has the permissions to add entries to the tables for datasets, dataset series and persons, and can use a graphical user interface to add catalogue entries manually. During the registration process the property values are validated against the configured schema, matching the EUCAIM metadata model. It is recommended to register the dataset in the catalogue once all the steps of the data preprocessing are completed.

DH should ensure that their datasets are aligned with a set of FAIR principles. The FAIR principles establish a framework of guidelines and best practices designed to facilitate the Findability, Accessibility, Interoperability, and Reusability of data and metadata by both machines and humans. Before publishing a dataset to the EUCAIM catalogue, DH should ensure that their dataset complies with the EUCAIM FAIR compliance level corresponding to the Tier as defined in Deliverable D4.4 Annex 6. These definitions outline the mandatory indicators - a subset of the full list of indicators specified by the Research Data Alliance - that are required by a dataset at a given Tier and are enforced by the EUCAIM DCAT-AP standard and the cataloguing process.

Among the Tier 1 mandatory indicators, RDA-F1-01M and RDA-F1-01D (indicating the Dataset Identifier) are linked to the assignment of persistent identifiers for metadata and data respectively. EUCAIM can't provide the data identifiers and, thus the DHs must ensure their proper assignment before submitting the metadata to the EUCAIM catalogue. On the other hand, metadata identifiers can be assigned by EUCAIM when adding the dataset record to the EUCAIM catalogue. Assignment of the metadata identifiers at this point also enforces the adoption of a coherent PID policy that ensures compliance with the PID requirements set for Tier 2 and 3.

In order to check a dataset's compliance to a certain level, the FAIR EVA tool will be used. This automatic evaluation tool helps to monitor FAIR compliance following the RDA indicators. A EUCAIM plugin for FAIR EVA is under development to check the presence of EUCAIM-defined mandatory metadata attributes, incorporating the Tier-level compliance checks. This plugin is meant to interact with FDPs, including the central EUCAIM Catalogue. The latter will be the main point of FAIR compliance checking, as it will maintain the metadata for all datasets in EUCAIM. However, for DH that set up a FDP for facilitating the registration of datasets to the EUCAIM catalogue, FAIR EVA (with the EUCAIM plugin) can be used to test the datasets FAIR compliance in their FDP before importing them into the EUCAIM Catalogue.

3.5.1 Data registration via a FAIR Data Point

The Public Catalogue has been designed and developed to harvest information from other catalogues, and to allow its metadata to be harvested by other catalogues as well. This implementation is being designed around a FAIR data point , which will act as an intermediary to manage and share metadata between catalogues. The full functionality of the EUCAIM FDP is expected to be realized and presented in deliverable D4.11 - Final version of the Central Core Infrastructure due in M36. The current status is presented in D4.6 - Final Federated Core services.

After Data FAIRification, DH can choose one of the following options to register their datasets via a Fair Data Point:

1. *Implementing a metadata expose pipeline*

Metadata can be exported automatically from the local system via an exposed pipeline. If the feature to directly expose metadata is not already supported by your local system, automated exports by the DH's local system should take place through a script or workflow. In this case, the data to be exposed is stored within a specific application (e.g., CHAIMELEON, XNAT, Grand-Challenge, Castor) used by the DH. To automate the export process, the raw data that lies beneath this application needs to be accessed. This can typically be done either by accessing the data through an API provided by the application, or by directly connecting to the database where the data is stored (with read access). A script should be written that queries the local system, aggregates the data if necessary, and generates EUCAIM FDP compliant metadata in RDF format. To ensure up-to-date representation of the metadata in the catalogue, this script to expose the local metadata should either run when the data in the local system is updated, or on a regular basis, e.g., weekly or daily, depending on the update frequency of the data. An example application that allows the automatic export of metadata from the DH's local system to a FAIR Data Point is img2catalog, developed by Health-RI. This tool queries an XNAT instance and generates DCAT-AP 3.0 metadata. It can register image collections directly at a FAIR Data Point which can then be linked to the EUCAIM catalogue. More details about this process can be found on https://github.com/Health-RI/img2catalog.

2. *Implementation of a FAIR Data Point*:

There are several approaches to implementing a FAIR Data Point:

   a. Exposing the local system

Establishing a direct pathway that enables the DH's local system to directly expose the dataset metadata as an FDP creates the strongest connection between the original data and publicly available information. This setup means that responsibility for maintaining this metadata is kept at the source. Achieving this requires either a system already supporting FDPs or deep knowledge of the source system and the availability of software engineering capacity to extend the functionality of the system. The systems at the DH's site must support functionality for displaying metadata from the existing data sources. We recommend using Molgenis and the UI of the catalogue as used in EUCAIM. Deployment can be done through a Docker container or a Kubernetes manifest. The UI of the catalogue as used in EUCAIM is available in GitLab, including the Dockerfile of the catalogue container and the Docker Compose File. The Molgenis catalogue software stores its data in a PostgreSQL database, which is included in the Docker Compose deployment.

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

Running a standalone FDP to expose their metadata is another option for DHs, if the necessary resources or capacity and knowledge are present in the institute. The FDP reference implementation is available as a Docker Compose distribution. Detailed instructions on how to use the reference implementation can be found in the official documentation at https://fairdatapoint.readthedocs.io.

c. Publishing dataset metadata via a Central FDP

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

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

To harvest the exposed metadata, the DH should contact the EUCAIM helpdesk with an onboarding request and include the details of the FDP to harvest. EUCAIM then performs the harvesting. The metadata is harvested into a testing environment where a check of the data is performed by EUCAIM and the DH, before reaching the central catalogue. The harvester for the EUCAIM catalogue is currently under development.

3.5.2 Setup of a local catalogue

Besides?? registering to the EUCAIM public catalogue, Data holders can decide to host a local instance of the catalogue to store the description of the datasets following the required metadata schema. This is not mandatory, but could facilitate further update of datasets when the automatic harvesting is implemented. The setup of the local catalogue could be based on Molgenis and the Catalogue application developed by ErasmusMC[3]. Deployment can be done through a Docker-compose or a Kubernetes manifest available in the Github repository of EUCAIM[4]. A database project should be created in Molgenis, using the Excel template available in the repository. The information about the dataset's metadata is defined in section 3.5, but DHs should also define:

- The network, as the framework in which the data has been collected (e.g. a project, regional initiative, network of centre, unique centre, etc.).
- The biobank, as the umbrella that gathers several datasets (e.g. different versions or splits).

The Excel spreadsheet contains sample information for all the necessary fields.

4. Hardware, network, and software requirements

Given that there are no mandatory requirements for Tier 1 local nodes with respect to their integration with certain EUCAIM components and in view that requests for data access in this case are served in-situ, via the local node's own service, there are no particular hardware, network or software requirements that Tier 1 nodes need to comply with, besides the ones following from the

[3] https://gitlab.com/radiology/infrastructure/studies/eucaim/molgenis-emx2-eucaim
[4] https://github.com/EUCAIM/k8s-deployments/tree/main/emx2

needs of their local services. In case Tier 1 nodes choose to make use of one of the recommended EUCAIM tools presented in Section 3.2, they should check for specific hardware and software requirements outlined in the respective tool's documentation (see Table 3 in Section 3.1).

4.2 Requirements for Tier 2 nodes

The hardware requirements for Tier 2 local nodes are described in Table 5. These requirements were selected according to the expected workload demands by the federated data search. A RAID 1 Configuration that mirrors the data on a second disk is highly recommended for improved data integrity and fault tolerance.

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

Table 5: Hardware requirements for Tier 2 local nodes.

Network requirements

Each Tier 2 local node must be connected to the public internet via a stable (outgoing) connection, with a minimum symmetrical bandwidth of 200 Mbps to avoid performance bottlenecks. Firewall and network policy exceptions must be made to allow the local node access to the following online resources (required for Tier 2 nodes and above):

- https://github.com (for access to code projects)
- The EUCAIM official repository of software artifacts (library project) at https://harbor.eucaim.cancerimage.eu/harbor/projects/1/repositories, the DKFZ repository https://docker.verbis.dkfz.de or the official docker hub (for access to pre-built docker images). If you choose the latter, you should make sure that the appropriate URLs are added to your firewall allowlist to ensure Docker images can be pulled properly from Docker Hub within your organization (see https://docs.docker.com/desktop/setup/allow-list/).
- https://broker.eucaim.cancerimage.eu (for access to the EUCAIM federation Beam broker)

Relevant network infrastructural adjustments, such as firewall configurations, must be made to allow outbound access on port 443 (HTTPS) to allow for the interactions required by the federated search component (see Section 5.1). Organisations which use added security protocols (e.g.,VPN, packet monitoring), must notify and collaborate with EUCAIM's Technical Support Team via the Helpdesk so as to address potential issues.

Software requirements

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

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

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

After ensuring the installation of the aforementioned dependencies, Tier 2 nodes can proceed with the installation of the necessary modules to enable interaction with the EUCAIM federated search, as described in Section 5.

4.2 Requirements for Tier 3 nodes

Hardware requirements

The hardware requirements for Tier 3 local nodes are described in Table 6. These requirements were selected according to the expected increased workload demands of data access for federated processing. The real-world processing workload may vary depending on the set of processing tools that the local node will run locally and additional overheads from any non-EUCAIM-specific services running in parallel on the local node.

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

Hardware	Requirement	Notes
CPU	Minimum: ● Option 1: 16 Cores >=1.8GHZ ● Option 2: 12 Cores >=3.0Ghz Recommended: 32 Cores /64 Threads 3.0Ghz	● If a GPU is not present, a server-grade, high core-count CPU is necessary for the Second Prototype. ● If not comparable by cores, the ideal thread count is 24+.
RAM	Minimum: 64GB Recommended: 128 GB ECC	● DDR5 is ideal. ● ECC memory is highly recommended for stability.
Motherboard	4+ RAM Slot	● Make sure to double check the compatibility of selected CPUs with the Chipset of the motherboard. ● In the case of DDR5, double check motherboard compatibility with DDR5.

Storage	A 1x(Dataset size) is the minimum requirement, 2x(Dataset size) is recommended. Examples include: ● 521 GB SSD Drive for Operating System (Either NVMe M.2 PCI Gen4 or SATA III) ● 1TB++ SATA III Drive (SSD or HDD) for local storage of medical data	● M.2, NVMe, Gen4 Drives are suggested for the OS ● For data storage size, Data Holders (DH) are expected to plan their purchase depending on the size of the Data they will provide. 1TB is a minimum, with some DHs already planning for 2 TB + datasets. ● For data storage, SSD are preferred for speed but are not mandatory.
GPU	>150 Tensor Cores 16GB VRAM	● Maximizing the amount of Tensor Cores is a priority, most recent GPUs will generally have higher Tensor Core counts. ● Ampere and Volta architectures are preferred.
Power Supply	-	● Each DH must make calculations depending on the hardware setup that will be selected to make sure that needed wattage is covered and ideally exceeded to prepare for any future upgrades to the machine.

Table 6: Hardware requirements for Tier 3.

Network requirements

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

Software requirements

Tier 3 nodes must deploy and configure the Federated Execution Manager (FEM) component in order to allow federated processing of data and images. The FEM daemon supports deployment within a wide range of infrastructure types, including a highly complex Kubernetes cluster, a SLURM Workload Manager, and containers from both Singularity and Docker. It also can deal with a custom API managing an intricate job queue system like UPV's Jobman. The setup is entirely up to the local node as long as some basic formatting and return codes (e.g., standard HTTP codes and JSON responses for execution IDs and similar metadata) are respected. More detailed instructions about how to deploy and interact with the FEM component are provided. More information about the setup of the FEM component is provided in Section 5.2

5. Setup for federated search (Tier 2)

Integration with the EUCAIM central search component is the last essential step that Tier 2 nodes should take care of and assumes that all the data preparation steps, as described in Section 3.3,

have been completed. Through the federated search functionality, EUCAIM users can retrieve the number of subjects and studies provided by each DH that fulfil certain search criteria (see also Section 3.3). To ensure secure and efficient communication between the central components and the local node, the **Samply.Beam** network communication middleware is employed. After completing all the necessary data preparation steps detailed in Section 3.3, Tier 2 local nodes should take care of the deployment of the following components, as depicted in Figure 1 of Section 2:

- **Beam Proxy:** The component is responsible for handling the communication with the central task broker of the EUCAIM federated search component, taking care of authentication, encryption, and signatures. It is made available as a Docker container (https://hub.docker.com/r/samply/beam-proxy).

- **Focus query dispatcher**: The component retrieves tasks via the Beam Proxy and transforms the data structure containing the query, and sends the query to a defined endpoint (the Mediator or directly to the data store). It is also responsible for returning the results to the central component via the Beam Proxy, executing them, and returning the results. It is offered as a Docker container (https://hub.docker.com/r/samply/focus/). The Beam Docker image and the Focus service can be consolidated within a single docker-compose.yml file. Specific instructions about how to configure and deploy both services are provided in Section 6.3.2.1 of the Gitbook User Guide for DHs.

- **Mediator (mapping component):** It is a custom component that needs to be implemented by DH whose local data structure does not comply with the EUCAIM CDM. In such cases, a mapping component is responsible for translating the abstract syntax tree containing the query into the query language of the local data store and for mapping the requested minimum set of clinical and imaging attributes (described in Section 4.2) into the respective local concepts. For DHs whose datasets adhere to the EUCAIM CDM, the federated query can be directly executed through the Focus query dispatcher, without the need for a mediator. The deployment of a mediator component can be done as a Docker container. Section 5.2.1 Dataset in a Federated Node of D5.6 provides an example of a query in the abstract syntax tree format and how it can be translated into an SQL query on an OMOP-CDM PostgreSQL database schema (with a radiology extension).

Once the aforementioned components are set up, a ticket in the Helpdesk, under the category “federated search” should be created with the request to register a new federated search provider. A Certificate Signing Request (CSR) needs to be created and signed by the central node manager as explained in Section 6.3.2.1 of the Gitbook User Guide for DHs. After the CSR is signed, the DH will be able to connect those components to the central services.

Once the aforementioned steps have taken place DHs should verify that their node has been correctly added to the Explorer and can respond meaningfully to queries. All communications are performed using encrypted protocols (TLS 1.3).

6. Setup for federated processing (Tier 3)

Once the data preparation process, ensuring that their data is compliant with the EUCAIM CDM format and required structure (see Section 3.4), and integration with the search component have

been completed, Tier 3 nodes should take the extra step of integration with the EUCAIM federating processing component. The Federated Processing architecture uses a pull model to fit restricted environments in which services are not exposed to the outside world (incoming connections), but can connect to external services (outgoing connections). The central services for Federated Processing manages an execution queue, one per node, that is populated and managed by the Federated Execution Manager (FEM orchestrator) backend and consumed by the local nodes (using FEM client). The processing job description includes details regarding the reference to the datasets, and the execution parameters. Jobs are pulled by the FEM-client, deployed at each node, and run locally. The federated processing architecture on the local node's side is depicted in Figure 1. It consists of following main components:

a) **FEM (Federation Execution Manager) client**: The component is running as a daemon that periodically queries the RabbitMQ of the FEM orchestrator for pending jobs involving the local node. The processing job description includes details, such as the reference to the datasets and the execution parameters. Jobs are pulled by the clients at each local node and run locally. In order to deploy the FEM client, data holders must fill in a configuration file which documents specific details related to the connection of the FEM client with the local infrastructure. The FEM client supports deployment within a wide range of infrastructure types, including a highly complex Kubernetes cluster, a SLURM Workload Manager, and containers from both Singularity and Docker. It also can deal with a custom API managing an intricate job queue system like UPV's Jobman. The setup is entirely up to the local node as long as some basic formatting and return codes (e.g., standard HTTP codes and JSON responses for execution IDs and similar metadata) are respected. More detailed technical instructions about how to deploy and interact with the FEM component can be found here.
b) **Data Materializer Tool (DMT)**: After a new request is issued, the FEM client triggers the local execution of the Data Materializer Tool, which allocates the requested data locally. The DMT also performs a validation of the dataset ids that were requested, filtering out the ones that are not present in the current local node. It is important to state that the execution environment does not allow downloading data and uses a secure proxy to access the Virtual Environment that permits the access to the data. Therefore, the required data is made available to the job in a **Sandbox** environment. Clinical data is provided as a JSON file extracted from the database. If the clinical data is not fully compliant to the CDM of EUCAIM, then it may not be properly consumed by the processing services. Blind execution through a federated model will restrict visualization of the data. The Materializer is offered in a containerized form. In order to make this process work, each federated node must provide a local configuration file for the DMT, as described in this document.

DHs should perform the following preparation steps before setting up the federated processing components:

1. Instructions assume that the software will be installed on a single host (or Virtual machine) that is isolated from the internal network at the site and able to run Docker containers (and pull Docker images from UPV's Harbor or provide an image repository in the FEM client configuration file). Other setups will require adaptations.
2. The FEM client requires only outbound connections to the RabbitMQ message broker and to the FEM orchestrator. Connections are encrypted using node-specific credentials.

3. No inbound connections or connection to other nodes are required (unless required by a specific tool).
4. Raw data never leaves the host machine. Only results (e.g., model weights) are shared.
5. During installation, you'll be required to define a read-only $DATA_PATH that will hold to your local datasets (formatted according to EUCAIM requirements), and a writable $SANDBOX_PATH that tools will use for temporary and final outputs.
6. Tools will be executed as Docker containers. Docker images will be available from EUCAIM central registry (UPV's Harbor), and will follow EUCAIM's agreed security requirements.

DHs should follow this step-by-step procedure to deploy the FEM client (see the basic guide here), the component responsible for connecting a node to the EUCAIM's federated network:

1. **Express Your Interest**
    - Start by submitting a ticket to the Helpdesk, addressed to the Technical Support Team, expressing your interest in joining the federated system.
2. **Initial Guidance**
    - A member of the UB/BSC team will respond with a link to the FEM client repository: https://gitlab.bsc.es/fl/fem-client
    - The README includes key information, especially in the "Prerequisites" and "Getting Started" sections.
3. **Credentials Delivery**
    - Once you're ready to deploy, confirm with the team.
    - The technical team will then send you a separate email containing your FEM-client credentials.
4. **Final Setup & Testing**
    - After setup, we'll run some tests to verify: 1) Network connectivity; 2) FEM-client's ability to access local infrastructure and trigger container executions; and 3) materialization of data for EUCAIM (link).

7. Pilot for integrating a selection of DHs into the EUCAIM federation

In order to test, operationalise, and improve the processes and associated tools involved in the overall workflow for the setup of local nodes, we organised a pilot study focused on integrating a selection of data holders across various Tiers. Through the pilot study, which is currently ongoing, we expect to identify gaps, challenges, refine the process and gather useful feedback about what needs to be improved in order to scale up with the integration of data from all EUCAIM DH and use cases. It should be noted that the pilot study is running while parts of the technical infrastructure are still being developed in parallel with the data onboarding. This enables us to incorporate elements resulting from the pilot testing and collected feedback into the overall preparation and technical setup process.

The selection of the pilot participants has been decided based on the information provided during the internal call application, taking into consideration the relevance and maturity level of the DH's data as well as the desired end Tier. The list of the pilot participants is provided in Table 7.

Organisation	Tier	Datasets
SCIENSANO	Tier 1	Histopathological images (Genomic Data; Molecular profiling data; digitized HE-slides; tissue biopsy; imaging data)
KAROLINSKA INSTITUTET (KI)	Tier 2	Breast (Mammo)
UNIVERSITY OF ATHENS (UoA)	Tier 3	Prostate; Brain; Sarcoma (MRI)
SERVICIO ANDALUZ DE SALUD (SAS)	Tier 3	Prostate (MRI); Breast (PET & CT); Lung (CT)
ASSISTANCE PUBLIQUE HOPITAUX DE PARIS (APHP)	Tier 3	Prostate (MRI)
ARISTOTELIO PANEPISTIMIO THESSALONIKIS (AUTH)	Tier 3	Breast (Mammo), Thyroid (Scintigraphy)

Table 7: Overview of DHs participating in the pilot study.

The implementation of the pilot use cases demonstrated the complexity arising from the diversity of the datasets maintained by the various DHs. The information provided by the pilot DHs via the **TIER Maturity Level Questionnaires** - covering aspects related to the current status of DHs' data, the technologies that have already been employed etc. - constituted the starting point in the process of defining the data requirements and tools that are relevant in each use case. It became clear that different variations of the general onboarding workflow needed to be adopted, depending on the specific characteristics of each DH. For instance, in cases where clinical data preparation was still underway, such as the case of AUTH, DHs were advised to begin representing their data in the format required by the EUCAIM Common Data Model (CDM) from the outset. KI on the other hand, a Tier 2 node with clinical data already stored in a database were advised to implement a custom mediator to integrate directly with the search component, rather than converting their data to the EUCAIM CDM. The steps followed by each pilot participant and the status of implementation until June 30 is shown in Figure 5.

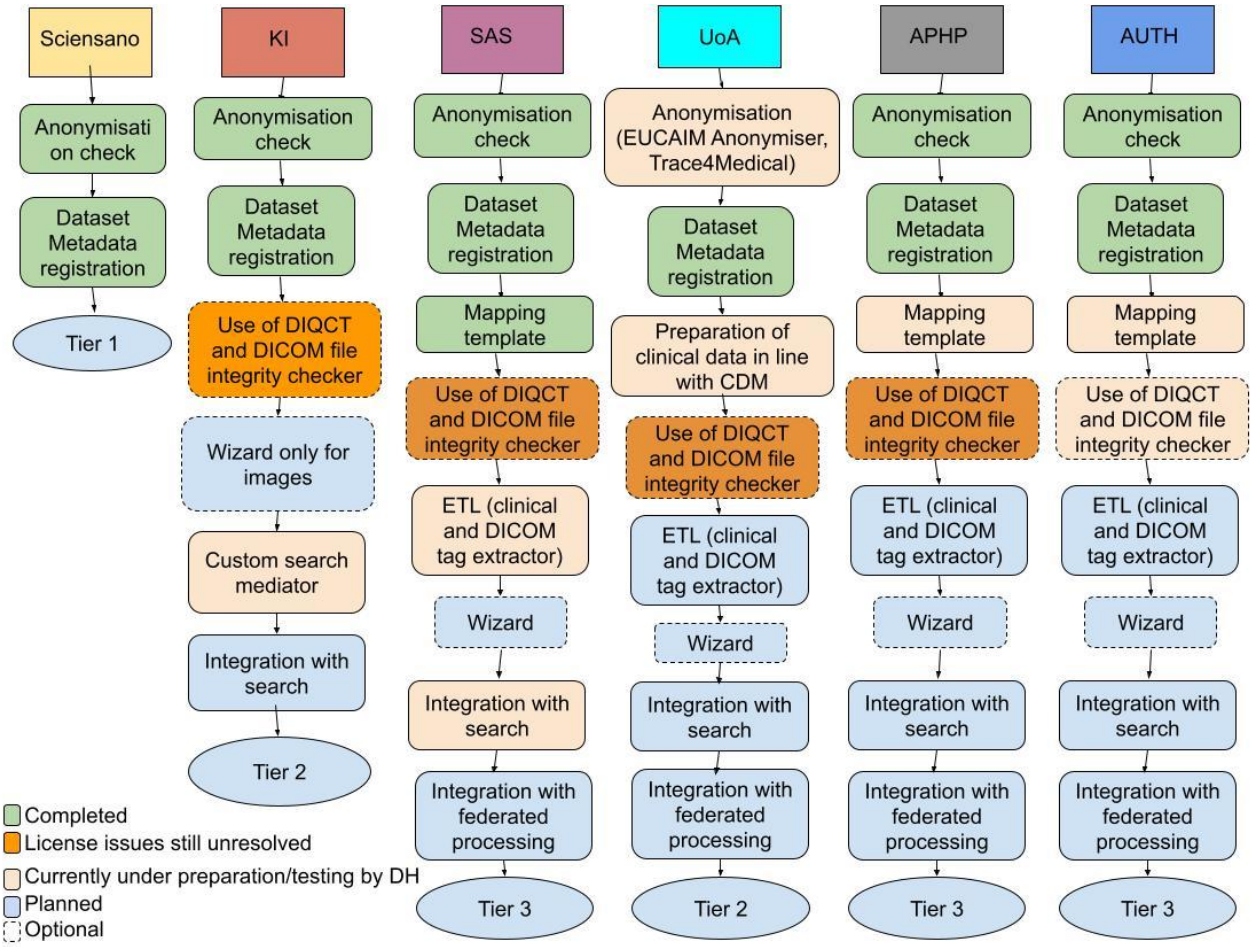

Figure 6. Overview of the steps followed by each pilot DH and progress until June 30 2025.

The piloting process also highlighted several limitations, gaps, and special cases or needs, leading to appropriate revisions and mitigation actions. For example, the review of the diverse local data models and formats used by DHs, along with the new cancer types introducing new types of information for inclusion in the hyper-ontology, led to several updates and extensions of the Common Data Model to support additional encoding formats and types of information.

Overall, the pilots demonstrated that data transformation to the CDM is a time-consuming process that requires thorough inspection of local data structures, a deep understanding of the underlying semantics, significant manual effort, and the involvement of various domain experts, including semantic, clinical, and technical experts. A key document developed and refined during the piloting phase is a template that guides DHs in mapping their local data structure to the CDM, thereby helping to systematise the ETL process.

Additionally, the piloting phase revealed licensing issues with certain tools. These concerns were escalated to both the tool providers and the legal team to ensure the tools could be used without barriers by the DHs. Testing also revealed some minor limitations in certain tools, which were communicated to the providers, prompting improvements in tool deployment and documentation.

8. Conclusions and Future Work

This document provides a concise yet comprehensive overview of the necessary steps for data holders to establish a local node integrated within the EUCAIM federation. By defining clear onboarding procedures and requirements aligned with established standards, EUCAIM promotes consistent access, semantic alignment, and technical interoperability across diverse and distributed data sources.

The onboarding framework is supported by a suite of resources and mechanisms designed to assist DHs throughout the process. Importantly, the framework allows DHs to select the level of interoperability that best aligns with their institutional objectives and available technical or organisational resources. The design of the process supports a gradual path toward more advanced levels of interoperability, fostering both flexibility and long-term integration.

Based on insights gained from testing and pilot activities, we anticipate adjustments to the procedures and technical specifications required for setting up local nodes at various tiers. As more DHs implement the setup workflow in practice, their feedback and experience will further refine and streamline the onboarding process. This will enhance operational efficiency, reduce time-to-integration, and support the scalability of the federation.

All updates and enhancements to the DH integration process will be continuously reflected in the EUCAIM Data Holders Federation Handbook, which is maintained as a living document. These updates will also be captured in Deliverable D5.7 – Definition of the Maximum (Version B) Data Federation and Interoperability Framework, ensuring that the latest procedures and standards are consistently documented and shared.