EUCAIM
CANCER IMAGE EUROPE

Project title: European Federation for Cancer Images

Project acronym: EUCAIM

Grant Agreement: 101100633

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

D4.4: Final rules for participation report

Responsible partner(s): HULAFE

Authors: Irene Marín (HULAFE), Carina Soler (HULAFE), Pedro Miguel Martínez-Gironés (HULAFE), Patricia Serrano (HULAFE), Luis Martí Bonmatí (HULAFE), Ignacio Blanquer (UPV), Valia Kalokyri (FORTH), Celia Martín (QUIBIM), Laure Saint-Aubert (MEDEX), Carles Hernández (BSC), Esther Bron (Health-RI), Ana Miguel (MAT), Kurt Majcen (BBMRI), Konrad Lang (BBMRI), Alejandro Vergara (QUIBIM), Xavier Rafael (QUIBIM), Jose Munuera (QUIBIM), Eirini Kaldeli (MAG), Mirna El Ghosh (LIMICS), Gianna Taskou (MAG), David Rodriguez (CSIC-IFCA)

Reviewers: Hanna Leisz (DKFZ) and Linda Chaabane (EUBI & CNR)

Date of delivery: 30/01/2025

Version: 1

Table of contents

List of abbreviations

AI = Artificial Intelligence

AI4HI = Artificial Intelligence for Health Imaging

ALTAI = Assessment List for Trustworthy Artificial Intelligence

API = Application Programming Interface

BBMRI = Biobanking and Biomolecular Resources Research Infrastructure

ColA = Collaboration Agreement

CDM = Common Data Model

CQL = Clinical Query Language

CSV = Comma-Separated Values

DCAT = Data Catalogue Vocabulary

DCM4CHEE = DICOM for Clinical and Hospital Environments

DFF = Data Federation Framework

DH = Data Holder

DICOM = Digital Imaging and Communications in Medicine

DPIA = Data Protection Impact Assessment

DPO = Data Protection Officer

DSA = Data Sharing Agreement

DTA = Data Transfer Agreement

DU = Data User

EHR = Electronic Health Records

EHDS = European Health Data Space

EHDSR = European Health Data Space Regulation

EOSC = European Open Science Cloud

ENCR = European Network of Cancer Registries

ERIC = European Research Infrastructure Consortium

EU = European Union

EVA = European Variant Archive

FHIR = Fast Healthcare Interoperability Resources

GDPR = General Data Protection Regulation

HLEG = High Level Expert Group

HPC = High-Performance Computing

ICDO-3 = International Classification of Diseases for Oncology, 3rd Edition

LS AAI = Life Sciences Authentication and Authorization Infrastructure

ML = Machine Learning

MM2 = Meta Milestone 2

MoU = Memorandum of Understanding

MRI = Magnetic Resonance Images

NIfTI = Neuroimaging Informatics Technology Initiative, an image file format

OHDSI = Observational Health Data Sciences and Informatics

OMOP = Observational Medical Outcomes Partnership

RC = Research Community

RDA = Research Data Alliance

RWD = Real World Data

RWDH = Real World Data Holder

SP = Software Provider

SQAaaS = Software Quality as a Service

SRAM = Surf Research Access Management

SQL = Structured Query Language

SNOMED CT = Systematized Nomenclature of Medicine-Clinical Terms

UPV = Universitat Politècnica de València

W3C = World Wide Web Consortium

WP = Work Package

XNAT = Extensible Neuroimaging Archive Toolkit

1. Introduction

1.1. Aim and scope of the deliverable

EUCAIM's ambition is to incorporate the largest possible array of data and software for them to be made available to users in its infrastructure. To allow this, EUCAIM plans to guide and facilitate as much as possible all aspects pertaining to the on-boarding of data holders and software providers. This is planned to be articulated via the support teams established in the context of Work Package 2 (WP2), namely evangelization, training, technical support and FAIR implementation support teams respectively. Throughout all the phases of the on-boarding process, these teams will ensure that such providers receive any necessary assistance required for their data and solutions to join the infrastructure.

This deliverable aims to describe the rules that the different players identified in this interaction must adhere to throughout the on-boarding process, with special focus on the requirements at the functional, technical and legal level on both ends. These requirements are expected to be dependent on the specifics of each role, which are described below.

The present deliverable also describes the technical and organisational constraints imposed on data users once the data is available for use on the platform.

It should be noted that the terminology used herein for the different roles, is descriptive of their nature. In the future, this terminology will refer to concepts legally defined by the Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space (EHDS).

It is worth noting that this deliverable outlines the rules of participation for each user role in high-level, abstract language understandable by each of them. Additionally, this document is accompanied by:

- The deliverable *D5.6 Minimum Data Federation and Interoperability Framework*, which serves as a supporting technical document that addresses each of the rules defined here (when applicable) for the Data Holders, detailing the steps to be followed and the tools to be used or installed from a technical perspective. Therefore, while D4.4 focuses on the overall purpose and general guidelines, D5.6 provides the technical details necessary for their implementation.
- The deliverable *D5.2. The EUCAIM CDM and Hyper-Ontology for Data Interoperability* which offers a detailed overview of the key features and contributions of the initial version of the EUCAIM Common Data Model (CDM) and Hyper-ontology.
- The WP6 internal document *EUCAIM software on-boarding guideline* which consolidates the work done by Work Packages 4, 5, and 6 of EUCAIM toward unifying a single strategy, enabling EUCAIM's partners to onboard their software into the project seamlessly and efficiently.
- The deliverable *D2.3: Requirement analysis of Real World Data Holders* which specifically focuses on the analysis of the RWD partners currently involved in the EUCAIM project to understand their present status and to assess their readiness to generate datasets dynamically in alignment with the Federation's requirements.

1.2. EUCAIM: Providers and users

As detailed in deliverable *D4.2 Final EUCAIM Operational Platform*, EUCAIM envisions to incorporate 3 different types of User Roles namely:

1. **Data Holder (DH)**: any natural or legal person (including entities, agencies and research organisations in the healthcare sector) who has the right, obligation, or capability to make certain data available for secondary uses under European Health Data Space Regulation (EHDSR), including registering, providing, restricting access, or exchanging the data. In the EUCAIM Framework, DHs can come from both the Research and Innovation Environment (existing research repositories from European projects, initiatives and infrastructures, as well as clinical trials or other secondary-use data repositories) and the Real World Data (RWD) Environment (such as hospitals and research institutes with access to primary health data, cancer screening programs and data altruism initiatives). A specific data provision model applies based on the environment: the data push model for the Research and Innovation Environment and the data harvest model for the RWD Environment (for further details, refer to D4.2). In both cases, they will be able to contribute with data either by (a) becoming a federated node or (b) transferring anonymized data directly to the Reference Nodes.

2. **Software Provider (SP):** any entity (e.g. startups, enterprises, research institutions, government agencies, non-profit organisations) that would like to make their already developed data processing software (SW), services, or applications available in EUCAIM's marketplace for EUCAIM users to utilise them for federated processing or data pre-processing purposes of the platform.

3. **Data User (DU):** any person or entity that explores the public catalogue of available dataset-level metadata and eventually requests access to data and processes them using the SWs available in the platform and/or their own AI tools.

Additionally, it is important to highlight that any of these roles will have the possibility to join or create a **Research Community (RC)**, a virtual organisation in EUCAIM within the platform. Research Communities are formal multidisciplinary teams composed of researchers, innovators, clinicians, data scientists, engineers, and AI specialists, dedicated to exploring, innovating and advancing a specific field or topic, to improve the role of imaging in healthcare, fostering collaborations and improving outcomes. This virtual organisation encompasses the previously described roles and will apply their rules for participation accordingly, based on the functions the team members wish to perform. Additionally, some specifics on the RCs onboarding are detailed in Section 7.

Finally, it is worth noting that depending on the type of role identified, the on-boarding process to join EUCAIM may differ. While the steps for the on-boarding process of a Data Holder was initially described in deliverable *D2.1 Onboarding invitation package*, and a recent study of the status of some of them has been carried out in *D2.3 Requirement analysis of Real World Data Holders*. Based on this analysis, the on-boarding of each particular role, has been specified in the relevant section where applicable.

2. Tiers of Compliance with the EUCAIM Data Federation Framework

Facilitating the population of the Atlas of Cancer Images with new data is one of the main objectives of the project. To achieve this, WP5 has developed the EUCAIM Data Federation Framework, as outlined in *D5.1 Early Release of the Data Federation Framework*, to ensure the interoperability and integration of cancer imaging and clinical data across Europe. This framework establishes the foundational elements needed for sharing, accessing, and analysing federated datasets while ensuring compliance with EU standards, enabling DUs to effectively conduct their research and innovation projects with high-quality, interoperable data.

Therefore, the DFF, released in *D5.1*, aims to facilitate integration of distributed data sources without physically centralising the data, maintaining autonomy of the individual databases. The DFF encompasses both the Local & Federated Nodes and the Reference Nodes, which act as federated nodes of the European Federation for Cancer Images (also typically referred to as "The Federation").

However, we recognize that having existing datasets already aligned with the DFF, especially aligning clinical data with the common hyper-ontology and the EUCAIM Common Data Model (CDM), may be challenging and a clear obstacle in the incorporation of new data to the EUCAIM Atlas of Cancer Images. To mitigate this issue, and to prevent blocking the integration of new data sources, even if the datasets initially do not fully comply with the EUCAIM DFF, three different technical Tiers of data compliance and available functionalities have been established. Thus, EUCAIM facilitates enhancing the quality and security of datasets and ensures processing guarantees, adding value as an effective instrument to address this challenge.

It is important to note that while imaging datasets with minimal clinical information are accepted, enriched datasets that include clinical data are more valuable for populating the EUCAIM Atlas. Consequently, these Tiers focus on both imaging and clinical data, while considering that enriched clinical data are not mandatory.

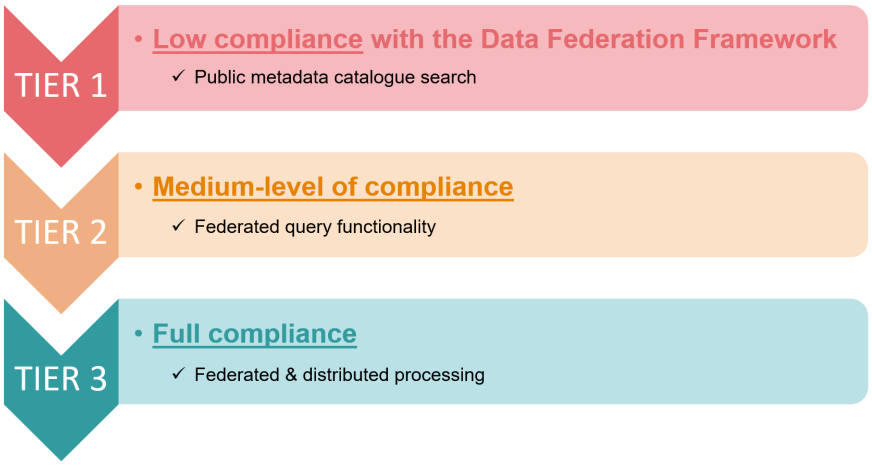

Figure 1. EUCAIM Data Federation Framework compliance Tiers and key functionality offered at each level.

Consequently, the minimum requirements requested have been adapted according to the datasets' level of compliance with the EUCAIM DFF. These Tiers will be scalable, as the datasets will be used for new research projects, and both providers and users will be incentivized and supported to upgrade the datasets from one Tier to another. It is important to emphasise that, in order to achieve a higher Tier, datasets must meet the requirements of the previous Tier, as each higher Tier encompasses the requirements of the ones below. That is, to comply with Tier 2, they must meet the requirements of Tier 1, and to comply with Tier 3, they must meet the requirements of both Tiers 2 and 1. Nonetheless, access to datasets across all tiers will be granted to Data Users upon request, subject to the specific limitations defined for each tier.

This tier based approach has a direct impact on almost all the functionalities that will be available at the Central Services level for both Data Holders and Users, as explained in the next sections.

2.1. Functionalities at each Tier

Tier 1. Low level of compliance with the DFF: Public Catalogue

Data will be accepted by the Federation with minimum requirements concerning the source repository (mainly linked to existing research projects in the European framework) and the data quality specifications established by the clinical centre of origin (in case of a RWD environment). The functionalities offered by the EUCAIM platform will be limited, but even so, incorporation of datasets as Tier 1 data is highly relevant as an entry point to ensure the participation of partners with valuable data but low resources.

TIER 1: Functionalities

- Only publication and visualisation of the dataset in the public metadata catalogue will be possible, allowing basic centralised filtering.
- Limited EUCAIM SW can be executed in the datasets and only for some pre-processing purposes. For example, at Tier 1, imaging data can be visualized and annotated using the EUCAIM DICOM viewers (local and reference nodes versions). However, other tools, such as the EUCAIM Anonymizer, require higher-tier compliance due to the need for standard data formats and predefined input/output structures, which are better achieved in Tiers 2 and 3.
- As data does not comply with the common data formats (EUCAIM's Hyper-ontology, CDM and folder structure hierarchy), neither federated/distributed processing capabilities nor a homogeneous framework will be available.
- The datasets in the public catalogue will be listed and made accessible (under the defined data request process), DUs will be informed that the data use is limited and EUCAIM will act as a contact point between the DUs and the DHs.
- The EUCAIM Access Committee manages contact points for the negotiation of the access requests, in cooperation with the DHs when needed.

Tier 2. Medium level of Compliance with the DFF: Federated Query

The second level of compliance with EUCAIM's DFF allows improved visibility and usability of the data, and requires a stronger involvement of the DHs. This is an improved version of Tier 1 datasets.

TIER 2: Functionalities

- The Federated Query service is available, allowing DUs to retrieve the number of cases fulfilling a specific criteria.

Tier 3. Full Compliance with the DFF: Federated Processing

Full data compliance entails alignment with a wide set of requirements. Tier 3 compliance is the ultimate goal for holders and users in order to achieve the best possible usability and impact of the datasets within the Federation. It goes beyond the Federated Query capabilities of Tier 2 and enables all EUCAIM functionalities, including federated processing.

TIER 3: Functionalities

- Federated and distributed processing, including Machine Learning (ML) and other advanced data processing techniques, are available.
- All EUCAIM SW can be executed in the datasets, for both pre-processing and AI analysis techniques

2.2. Tier upgrades in the context of the EUCAIM EDIC

The main dimension of value of the EUCAIM infrastructure is providing access to large, homogenised, standardised, and de-identified cancer imaging datasets together with curated, highly structured clinical data for secondary use. For this reason, while it is understood and accepted that the data made accessible through EUCAIM will in many instances initially only be Tier 1-compliant, the achievement of Tier 3 will be the ultimate end goal for all the data collections contained in the infrastructure.

The upgrading of the tier level also offers the advantage of achieving a higher level of fairness. This serves as a potential incentive for DHs involved in European or publicly funded projects, where alignment with FAIR principles is often a mandatory requirement. Figure 2 shows how compliance with a higher tier inherently increases adherence to the FAIR data principles.

In the context of the future EUCAIM EDIC, this will be facilitated through active seeking of new funding and collaboration opportunities that ultimately turn into new projects, whose scope involves, among others, the quality enhancement of the EUCAIM collections. To ensure such opportunities are pursued, the Central Hub core services will include an "external relations and project management unit", whose goal will be to seek regular R&D funding to sustain and further improve the infrastructure as a whole, including the datasets made accessible through it. More information on this unit and its operation can be found in *D8.3*.

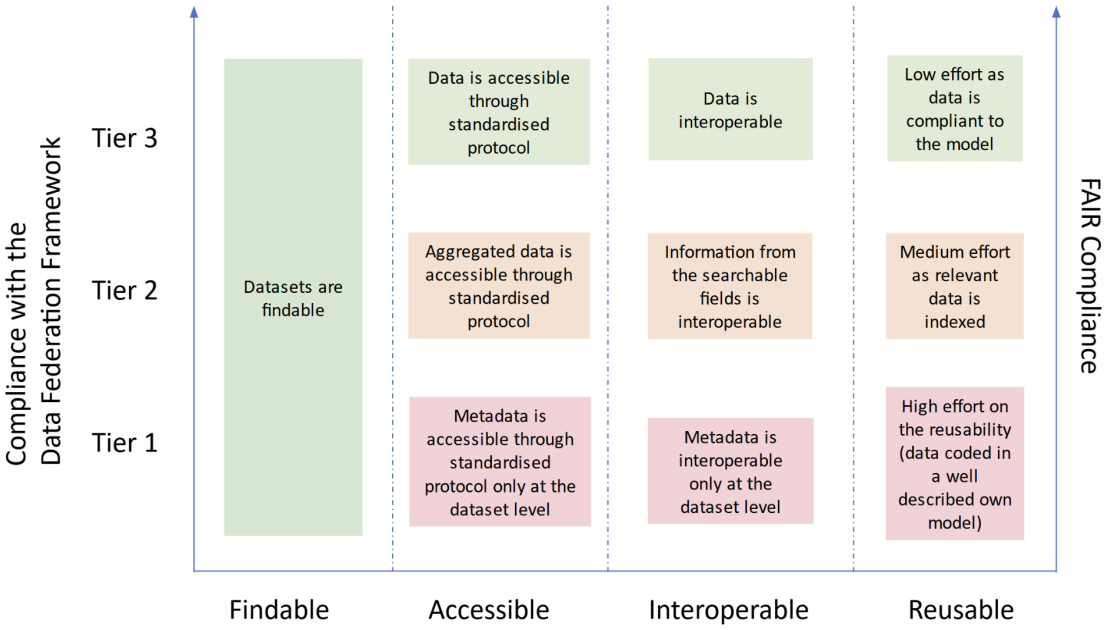

Figure 2. Tiers and FAIR compliance.

3. Common Rules for Participation for all User Roles

Regardless of the User Role, two requirements apply to all users of the platform (DHs, SPs, DUs and RCs):

- All users must have a valid EUCAIM user account.[1]
- All users must successfully complete the EUCAIM mandatory legal and ethical training via the Moodle platform (*refer to D2.4 Training Evaluation: Guidelines, Best Practices, Lessons Learned*).

4. Rules for Participation for Data Holders

To allow the smoothest possible onboarding process for new DHs and to facilitate as much as possible their participation and interaction with the other user roles in EUCAIM, the proposed Rules for Participation detailed below are based on the principle of minimising the requirements for the provision of their datasets. The Consortium understands this approach as a way to provide an environment that streamlines the integration of new DHs into the Federation, while maintaining the commitment to the highest standards of data quality and alignment with EUCAIM's objectives.

This section presents both the minimum technical requirements, in terms of data, data access and infrastructure, as well as the legal and ethical requirements for DHs.

[1] https://u.i3m.upv.es/m44jv

It is important to note that the following sections outline the rules of participation for each Tier in high-level, abstract language. Additionally, this document is accompanied by *D5.6 Minimum Data Federation and Interoperability Framework*, which serves as a supporting technical document that addresses each of the rules defined here (when applicable), detailing the steps to be followed and the tools to be used or installed from a technical perspective. Therefore, while D4.4 focuses on the overall purpose and general guidelines, D5.6 provides the technical details necessary for their implementation.

As mentioned before, it is important to note that in order to achieve a higher Tier, datasets must meet the requirements of the previous Tier, as each higher Tier encompasses the requirements of the ones below.

4.1. Minimum requirements for Tier compliance

To comply with the EUCAIM tier model, Data Holders (DHs) must meet mandatory requirements specific to each tier. Additionally, regardless of the tier, it is highly recommended, though not mandatory, that datasets follow certain best practices to enhance their overall quality and utility:

- **Annotations associated with the datasets (annotations-enriched (A+))**: Including annotations, whether manually created or AI-generated (with post-validation by expert clinicians), adds critical value to the dataset. Fully automated annotations will not be stored as they can be generated on demand *(refer to ANNEX 9)*.

- **Extended clinical data associated with the imaging datasets (clinical data-enriched (C+))**: Beyond the mandatory minimum clinical data, providing extended clinical information significantly enhances the value and usability of imaging datasets, allowing them to support various purposes and use cases. For detailed information about the clinical data currently supported by the EUCAIM CDM, please refer to *D5.2: The EUCAIM CDM and Hyper-Ontology for Data Interoperability: Initial Version.*

When these recommended practices are followed, certain additional requirements within the tier model become **mandatory if available**. In other words, these criteria are only mandatory if the dataset includes extended clinical data and/or annotations, and in this case the Tier is enriched as Tier **N**A+ and/or Tier **N**C+ (where **N** can be 1, 2, or 3).

Table 1. Summary of the Tier model nomenclature.

Minimum requirements	Tier 1	Tier 2	Tier 3
Annotations-enriched	Tier 1A+	Tier 2A+	Tier 3A+
Clinical data-enriched	Tier 1C+	Tier 2C+	Tier 3C+
Both-enriched	Tier 1A+C+	Tier 2A+C+	Tier 3A+C+

4.1.1. Tier 1 (Dataset Cataloging)

TIER 1: Mandatory requirements

- Imaging data must be provided in DICOM format[2], as it is the interoperable standard within the EUCAIM infrastructure, although some exceptions can be considered[3]. The tools provided by EUCAIM to ensure minimum requirements in terms of anonymization and data quality are based on DICOM format.
- In the reference nodes, clinical data is accepted in JSON and CSV formats, although it is preferable in JSON format. If the data is shared from the DH's node, there is greater flexibility regarding the clinical data format (e.g., JSON, CSV, XLS, Avro, Parquet)
- Datasets must follow an anonymization process and have a risk analysis performed *(refer to ANNEX 1)*
- The Patient ID must follow the EUCAIM specifications. Regardless of the format and structure of the data, the Patient ID for clinical and imaging data, corresponding to the DICOM tag (0010,0020), must be identical and must be listed as the first variable in the clinical dataset (*refer to ANNEX 1*)
- Dataset metadata must be registered in the Public Catalogue according to the metadata specification for the datasets, in alignment with the European Health Data Space Regulation, when applicable for the context of EUCAIM *(refer to ANNEX 2)*
- The minimum set of clinical and imaging attributes at record level/patient level must be provided[4] *(refer to ANNEX 3)*
- Detailed documentation on data elements must be provided to facilitate the understanding of the structure of the provided data (*refer to ANNEX 4)*
- The minimum requirements in terms of Data Quality must be ensured *(refer to ANNEX 5)*
- The minimum requirements in terms of FAIR must be ensured *(refer to ANNEX 6)*
- The requirements in terms of infrastructure must be met:
    - For DHs transferring their data to the reference nodes, the requirements described in *Section 4.3.1* must be in place.

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

[3] Specific cases involving NIFTI format will be considered if the DH actually cannot retrieve the original DICOM files from which the NIFTIs were derived. These NIFTI-based datasets will be deprioritized during ingestion due to their lack of interoperability. DH must assure the minimum requirements in terms of anonymization, risks analysis and quality by own tools/procedures. Nonetheless, if the DH can convert these images to DICOM and supplement them with all minimum and relevant clinical information that could have been extracted from the original DICOM (EUCAIM encourages that), the dataset will be able to upgrade to higher tiers. This will enhance its interoperability, bringing it in line with other datasets. If there is no other option then uploading non-dicom formatted data, the required metadata needs to be supplied in DICOM JSON format so the findability and interoperability through metadata use can still be met.

[4] Specific cases of only-imaging datasets, with only imaging attributes, will be case-by-case considered and accepted in the platform, specially if they attend to specific use cases where associated clinical data is not relevant for the intended purpose.

- For DHs sharing their data, the local node infrastructure is required and must be in place *(refer to Section 4.3.2)*

- The legal and ethical requirements must be followed *(refer to Section 4.4)*

TIER 1A+: Highly recommended if available

- When the annotations provided contain segmentations, they are preferred in DICOM SEG format. If they are provided in other formats (such as NIfTI or RTSTRUCT) and the related original imaging data in DICOM format are available, it is highly recommended to convert them to DICOM SEG via the EUCAIM Converter tool, especially when they are transferred to the Reference Nodes *(refer to ANNEX 9)*.

TIER 1A+: Mandatory

- Annotations must be accompanied by a set of minimum annotation metadata *(refer to ANNEX 3)*

4.1.2. Tier 2 (Federated Query)

TIER 2: Mandatory requirements

- Fulfil requirements of tier 1.
- Integration into the federated query by installing the Mediator service that is accessible from the central services of EUCAIM.
- To extract and retrieve the aggregated data requested for the Federated Query from the minimum set of clinical and imaging attributes at record level/patient level *(refer to ANNEX 7)*:
    - Minimum metadata must be structured in a way that supports querying (e.g. JSON in a MongoDB/PostgreSQL database) for the Federated Query.
    - The mapping component (within the Mediator service) to the minimum hyperontology concepts must be in place or a directly transformation of the requested minimum set of clinical and imaging attributes at record level/patient level to the EUCAIM CDM structure must be perform, so the minimum metadata can be queryable.
- Second level of FAIR compliance must be ensured *(refer to ANNEX 6)*
- The requirements in terms of infrastructure must be met:
    - For DHs transferring their data to the reference nodes, the requirements described in *Section 4.3.1* must be in place (same as in Tier 1).

- For DHs sharing their data, the Tier 2 federated node infrastructure is required and must be in place *(refer to Section 4.3.2)*

TIER 2A+ Mandatory if available

- When the annotations provided contain segmentations, they must be provided in DICOM SEG format. If they are provided in other formats (such as NIfTI or RTSTRUCT) and the related original imaging data in DICOM format is available, they will be converted to DICOM SEG via the EUCAIM Converter tool *(refer to ANNEX 9)*

4.1.3. Tier 3 (Federated Processing)

TIER 3: Mandatory requirements

- Fulfil requirements of tier 2.
- Configuration of the Data Materializer Tool, which copies and prepares on the fly the local data, with enough space to make the data accessible *(refer to ANNEX 10)*. The materialised data is expected to be in the EUCAIM CDM.
- The minimum set of clinical and imaging attributes at record level/patient level must follow the hyperontology concepts and be structured according to the CDM specification *(refer to D5.2 for extended documentation of the first version of EUCAIM hyperontology and CDM and to D5.6 for the minimum interoperability requirements)*
- The imaging dataset structure/hierarchy according 'Patient/Study/Series/Images' and the series identification/tagging must be followed to ensure the correct automated federated processing of EUCAIM software (refer to ANNEX 8)
- Third level of FAIR compliance must be ensured *(refer to ANNEX 6)*
- The requirements in terms of infrastructure must be met:
    - For DHs transferring their data to the reference nodes, the requirements described in *Section 4.3.1* must be in place (same as in Tier 1).
    - For DHs sharing their data, the Tier 3 Federated node infrastructure is required and must be in place *(refer to Section 4.3.2)*

TIER 3A+ Mandatory if available

- For imaging data, if associated annotations are provided they must follow the imaging Structure hierarchy according '*Patient/Study/Series & Annotation Series/Images'*

TIER 3C+ Mandatory if available

- If extended associated clinical data is provided, it must follow the hyper-ontology concepts and be structured according to the CDM specification *(refer to D5.2 for extended documentation of the first version of EUCAIM hyper-ontology and CDM and D5.6 for the minimum interoperability requirements)*

SUMMARY TIERS MINIMUM REQUIREMENTS

Table 2. Summary of the Minimum requirements in each Tier.

Min. Requirements	TIER 1	TIER 2	TIER 3
Metadata Catalogue registration	Yes	Yes	Yes
Federated Query (FQ) integration	No	Yes	Yes
Federated Processing integration	No	No	Yes
A+ (annotations-enriched)	Highly recommended	Highly recommended	Highly recommended
C+ (clinical data-enriched)	Highly recommended	Highly recommended	Highly recommended
Dataset metadata	Yes	Yes	Yes
Minimum imaging metadata attributes	Yes	Yes + available in Federated Query	Yes + following CDM/Hyper-ontology
Minimum clinical attributes	Yes**	Yes** + available in Federated Query	Yes** + following CDM/Hyper-ontology
Minimum annotation metadata attributes in A+	Yes	Yes + available in Federated Query	Yes + following CDM/Hyper-ontology
Dataset metadata standard	EUCAIM DCAT-AP	EUCAIM DCAT-AP	EUCAIM DCAT-AP

Min. Requirements	TIER 1	TIER 2	TIER 3
Imaging Data Format	DICOM, NIFTI*	DICOM	DICOM
Minimum imaging metadata format	DICOM-Tags, DICOM-JSON, JSON	DICOM-Tags	DICOM-Tags
Minimum clinical attributes format	Any	Partially in CDM/Hyperontology or at least mapped in the FQ	CDM/Hyperontology
Segmentation Format in A+	DICOM SEG, NIFTI, RTSTRUCT, others	DICOM SEG	DICOM SEG
Minimum Annotation Metadata Format in A+	DICOM-Tags, JSON	DICOM-Tags	DICOM-Tags
Mediator	No	Yes	Yes
CDM/Hyper-ontology compliance	Partially (only dataset metadata level)	Partially (minimum required attribute) or at least mapped in the FQ	Partially (minimum required attributes)
CDM/Hyper-ontology compliance in C+	Partially (only dataset metadata level)	Partially (minimum required attributes) or at least mapped in the FQ	Complete
Mapping component within Mediator	N/A	Only if no CDM/Hyper-ontology compliance	Not needed with CDM/Hyper-ontology compliance
Metadata available in the federated query	N/A	Minimum required attributes	Minimum required attributes
Metadata available in the federated query in C+	N/A	Minimum required attributes	All attributes provided in the dataset
Data Materializer	No	No	Yes

Min. Requirements	TIER 1	TIER 2	TIER 3
Tool			
Imaging structure hierarchy	No	No	Yes
Anonymization + Risk analysis	Yes (In the Reference Nodes data is kept in quarantine until it is checked)	Yes (In the Reference Nodes data is kept in quarantine until it is checked)	Yes (In the Reference Nodes data is kept in quarantine until it is checked)
Legal & ethical requirements	Yes	Yes	Yes
Legal and Ethical training	Yes	Yes	Yes
Data Elements Documentation	Yes	Only if no CDM compliance	Not needed
Data Quality	Yes	Yes	Yes
FAIR	Datasets are findable and metadata is accessible	Datasets are searchable and partially interoperable	Improved interoperability and reusability
Data sharing (Node infrastructure)	Local node	Federated node	Federated node
Data transfer	Local environment + Reference Node with quarantine	Local environment + Reference Node with quarantine	Local environment + Reference Node with quarantine (faster to release)

* Specific cases will be accepted in NIFTI
** Specific cases will be accepted without the minimum clinical attributes

4.2. Minimum requirements in terms of data access

DHs may offer various access conditions, once the access request has been granted, which will depend on the usage conditions of the dataset and the technical Tier of compliance of the datasets and services with the EUCAIM DFF discussed in the previous section:

a) Already public and irreversibly anonymised datasets registered in EUCAIM catalogue, potentially mirrored in the reference node which can be downloaded to the users' premises (available for all tiers).
b) Datasets that cannot be downloaded but can be accessed, viewed and processed *in-situ* in a secure environment (available for all tiers), both for the federated and centralised repositories.
c) Datasets that can be processed remotely on a federated node without the ability to access and visualise data, even remotely (available only in Tier 3 of compliance). In this case, data must be very well described and standardised. In any case, it will be advisable to have access to a small representative sample or even synthetic or simulated data if possible, to tune up the analytic applications.

4.3. Minimum requirements in terms of infrastructure

DHs need to have the appropriate software and resources to share or transfer data and comply with the minimum requirements explained below. It should be noted that the first section, corresponding to Reference Nodes, applies to all DHs physically transferring their datasets regardless of the tier. The second section, corresponding to the Setup of Local Nodes, applies to all DHs sharing their datasets, with specific requirements depending on the tier.

4.3.1 Data Transfer to the Reference Nodes (any Tier)

EUCAIM has two different Reference Nodes (described in the *Technical Architecture Document* and *in D5.6*):

- UPV-Universitat Politècnica de València's Reference Nodes: based on the Chaimeleon Cloud Repository technology[5] and using the QP-Insights application (both web-based and API-based), offered by Quibim, which provides the interface for the ingestion of DICOM objects and clinical data through an electronic Case Report form (eCRF).
- The Medical Imaging Storage service of Euro-BioImaging Node: An XNAT (Extensible Neuroimaging Archive Toolkit) instance operated and supported by Health-RI and Erasmus MC, together with upload and ingestion support, including advice and tooling, and a service desk[6].

They are being set up to cover a wider range of DH´s functional requirements, geographical location reference and service offer.

The **functional requirements** for DHs to use each one, are:

- UPV's Reference Node:
    - A valid EUCAIM user account[7]

[5] Chaimeleon D4.1 Initial Repository Deployment. Retrieved September 14, 2023, from: https://doi.org/10.5281/zenodo.7588625
[6] XNATExtensible Imaging Archive Toolkit. Retrieved September 14, 2023, from: https://www.health-ri.nl/services/xnat
[7] https://drive.google.com/file/d/1EsFYxbzqpyYKggyeKrKKw3FkVecDby8P/view

    - A standard DICOM client compatible with DICOM for Clinical and Hospital Environments (e.g. DCM4CHEE).
    - A REST-based client application for uploading the eForms with the clinical data. A sample client will be provided by UPV.

- The EuroBioImaging Medical Imaging Storage:
    - Linking to a federated Identity Provider or broker that authenticates users based on their institutional user accounts (like LS AAI or Surf Research Access Management (SRAM).
    - Local IT support to upload DICOM data based on provided software
    - Possibility to run python scripts to upload and download data (e.g. in DICOM or NIfTI format (including DICOM JSON files containing DICOM header information), and also eCRFs).

The **non-functional requirements** are:

- Access to a dedicated machine for uploading and administering data in the Reference Node.
- Allow outgoing network connection over HTTPS in order to connect to the central services of the Federated Learning platform.
- A technical contact point in the staff for technical support in case of technical issues.

Datasets ingested in the reference nodes will enter in a revision status until they are registered in the EUCAIM Public Catalogue. This period of "**quarantine**" will facilitate the developers and the technical team of EUCAIM to revise the data introduced and validate the Tier level, before being published.

When extended associated clinical data adheres to Tier 2, ensuring the required attributes are available in the federated query, but does not meet Tier 3 requirements, it is possible for the imaging dataset to be released while the clinical data remains under quarantine. This is because validating clinical data that does not comply with the CDM demands additional effort, while for images in the DICOM standard format is more straightforward. In such cases, the imaging data will be published in the Public Catalogue with a notification stating that extended associated clinical data may be available upon review.

Please, for more detailed information about the reference nodes, refer to D5.6 Section 4.3.3 Dataset to be transferred to the EUCAIM Reference Node.

4.3.2 Setup of Local Nodes for data sharing (Tiers 1, 2, 3)

Below, the technical and operational requirements that need to be met by each local DH node, depending on the tier, are outlined. In this case, data will not be physically transferred to the reference nodes, but will be shared from the DH's instances. The requirements are based on a more summarised and updated version of the requirements set out in deliverable *D5.11 Interim set-up of local nodes for data federation*.

To distinguish between the terms 'local node' and 'federated node', it is important to note that a local node refers to the infrastructure established by a Data Holder across all three tiers.

Local nodes are considered federated nodes when they meet the requirements of Tier 2 or Tier 3, enabling federated query capabilities in Tier 2 and adding federated processing capabilities in Tier 3.

Basic common requirements for local nodes for all Tiers

A) Infrastructure Procurement:

Organisations must procure and set up the storage infrastructure ensuring that every local node aligns with EUCAIM's data storage specifications. These needs range from storing the organisation's contributed dataset to offering supplementary storage for localised data processing projects, to promote redundancy measures (such as Redundant Array of Independent Disks, RAID[8]) and temporary storage for federated processing if needed.

Throughout this process, each organisation must adhere to its local policies and procurement procedures to obtain and set up the essential management and technical infrastructure, including Servers, Virtual Machines, and IaaS, necessary to host a federated node. To participate in EUCAIM's federated infrastructure, organisations are required to procure the specified infrastructure and hardware that meets the exact processing demands of the system, which are differentiated based on the local node's Tier of participation, as described in the requirements outlined below.

B) Network requirements

- Each local node must be connected to the public internet via a stable connection, with a minimum symmetrical bandwidth of 200 Mbps to avoid performance bottlenecks. This requirement is necessary for Tier 2 and 3 nodes, while less demanding bandwidth requirements are required in case of Tier 1 nodes that maintain their data locally.
- Relevant network infrastructural adjustments, such as firewall configurations must be made to enable specific network port inbound or outbound access to the public internet, specifically outbound on port 443 (HTTPS). This requirement is necessary for Tier 2 and 3 nodes.
- Firewall and network policy exceptions must be made to allow the local node access to the following online resources (required for Tier 2 nodes and above):
    - https://github.com (for access to code projects)
    - https://docker.verbis.dkfz.de (for access to pre-built docker images)
    - https://broker.eucaim.cancerimage.eu (for access to the EUCAIM federation Beam broker).

[8] Red Hat Documentation. Retrieved January 28, 2025, from: https://docs.redhat.com/en/documentation/red_hat_enterprise_linux/7/html/storage_administration_guide/s1-raid-levels

- Network Infrastructure Management: Organisations which use added security protocols (e.g.,VPN, virtual, reverse proxy networks, packet monitoring), must notify and collaborate with EUCAIM's Technical Support Team via the Helpdesk.

C) Infrastructure requirements

- All procured physical infrastructure (processing, electrical, or network units), essential for the local node, must be securely installed and positioned, safeguarded from external hazards and detrimental environments.
- The physical infrastructure for the local node hosting must reside in a restricted access zone, allowing entry only to approved individuals.
- Any events of physical access to the physical space where the local nodes are installed must be continually monitored and recorded with records being kept in accordance with organisational regulations.
- Multi-layered redundancy of data is recommended at both the infrastructure and organisational levels. An initial recommendation includes the deployment of a RAID disk configuration.
- A comprehensive backup plan should be established to consistently safeguard the data's latest versions.
- For the management of the infrastructure, digital access will have to be provided by the organisation to authorised persons. Authorised technical staff must be provided with user credentials and SSH keys, ensuring encrypted and end-to-end secure access. Two-factor authentication measures must be applied wherever possible. Regular audits must be conducted to review access logs and ensure no unauthorised access attempts.
- Each organisation should monitor the status and performance of their hosted local node. Automation tools can be implemented to track metrics and set up alerts for anomalies in each metric.
- DHs are recommended to pursue the highest degree of infrastructure elasticity whenever possible when procuring and setting up this infrastructure. This recommendation for elasticity is future-facing, as DHs might be required to scale their infrastructure either horizontally (by adding more local nodes) or vertically (by increasing the storage or processing capabilities of specific local nodes).

D) Operating system requirements

- Each local node above Tier 1 must use an operating system that is compatible with EUCAIM's software stack, including stable Linux distributions such as Ubuntu, CentOS, or Debian. For Tier 1 nodes the use of a Linux operating system is recommended to enable the setup of EUCAIM's local node services. Alternatively, Windows or Mac operating systems can be used, as long as a Docker environment or, other, similar virtualization engines, is installed, in order to deploy the EUCAIM services as containers.

TIER 1 Local Node requirements:

A) Hardware requirements

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

B) Software requirements

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

It is expected that Tier 1 datasets may not already comply with the respective minimum requirements concerning data formats and fairness. Therefore, it should be noted that DHs may wish to deploy software for Data Preparation provided by EUCAIM within their Local node. In that case, they will need to set up and deploy locally the EUCAIM software components that can be used to facilitate alignment with the data preparation requirements.

TIER 2 Federated Node requirements

Tier 2 federated nodes need to comply with all requirements for the Tier 1 nodes as well as with the additional ones outlined below.

A) Hardware requirements

Tier 2 nodes need to fulfil minimum hardware requirements, concerning CPU, RAM, and storage, that allow them to deal with the expected workload for data searching. More details are provided in *D5.6 Section 5.2 Guidelines for Federated Query support.*

B) Software requirements

To install, interact and configure the EUCAIM software components required to provide the federated query through Tier 2 federated nodes, DHs are required to configure the operating system of their local node to locally operate the Mediator component, and to further install basic Linux packages (e.g. wget, Docker etc) which are required for the functioning of the EUCAIM federated search components.

For Tier 2 participation, several software dependencies must be installed on the local node, which is planned to locally deploy the necessary EUCAIM software that allows them to interact with the Explorer component. The software requirements are listed in *D5.6*.

It must be noted that depending on the storage systems and structure of their data, DHs may be required to implement a mapping component within the local Mediator service, to perform the necessary mappings to the minimum hyper-ontology concepts for the federated query.

TIER 3 Federated Node requirements:

A) Hardware requirements

Tier 3 nodes need to fulfill more demanding hardware requirements following from the expected increased workload demands for federated processing. The actual processing workload entailed for Tier 3 nodes may vary depending on the set of processing software that the federated node will run locally and additional overheads from any non-EUCAIM-specific services running in parallel on the local node.

B) Software requirements

To provide data access through Tier 3 participation, software dependencies must be installed on the federated node which is planned to locally deploy the necessary EUCAIM software that allows them to interact with the EUCAIM federation and enable data access for data processing and storage of data processing projects.

The software requirements for Tier 3 participation are partially covered through the software requirements for Tier 2 Participation. Extending those requirements for the integration of federation processing capabilities for local nodes participating at a Tier 3 requires the local deployment of the dedicated EUCAIM Federated Execution Manager (FEM) component.

If a DH wishes to federate its datasets but lacks the necessary computational resources, they can choose to transfer the data to a trusted third party (TTP) of their choice when some processing needs to be performed on these datasets. The conditions for data de-identification and temporality of the data will be agreed between the TTP and the DH. Therefore, the project is currently looking into the development of services for temporarily transferring local node data to the reference nodes or secure processing infrastructures for the execution of federated processing when local node resources (storage, processing power) do not cover the needs of federated processing execution. Specific hardware and software requirements related to this central infrastructure service are yet to be defined by WP6 and WP4. Once these are available, the respective Tier 2 and Tier 3 hardware and software requirements will be updated accordingly.

4.4. Legal and Ethical Requirements

A first approach to legal and ethical requirements is presented based on the current regulation.

- General Data Protection Regulation (GDPR)[9]
- Data Governance Act[10]
- Data Act[11]
- Cybersecurity (NIS 2)[12]

Moreover, EUCAIM will operate as a key platform in ensuring the regulatory compliance framework related to the development of medical devices, particularly those based on AI or integrating embedded AI resources:

- Artificial Intelligence Act[13]
- Medical Diagnostic Device Regulation[14]
- In Vitro Medical Device Regulation[15]
- Health Technology Assessment Directive [16]

Finally the European Health Data Space (EHDS) will be a foundational element of the European Health Union[17] and represents the EU's first specialized data space as part of its European strategy for data[18]. In the spring of 2024, the European Parliament and the Council reached a political consensus on the European Health Data Space Regulation (EHDS-R)

[9] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).
[10] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act).
[11] Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act) .
[12] Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive)
[13] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).
[14] Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC
[15] Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU
[16] Regulation (EU) 2021/2282 of the European Parliament and of the Council of 15 December 2021 on health technology assessment and amending Directive 2011/24/EU (Text with EEA relevance)
[17] The European Commission is building a strong European Health Union to enhance collaboration among EU countries in preparing for and responding to health crises. This initiative ensures that medical supplies are accessible, affordable, and innovative while promoting joint efforts to improve disease prevention, treatment, and aftercare, particularly for conditions like cancer. The European Health Union aims to protect citizens' health, equip the EU and its Member States to effectively prevent and manage future pandemics, and strengthen the resilience of Europe's healthcare systems.
[18] The European data strategy aims to establish a single data market to strengthen Europe's global competitiveness and data sovereignty. This strategy includes creating Common European Data Spaces, which will make more data accessible for economic and social use while ensuring that companies and individuals who generate the data retain control over it.

proposal from the European Commission[19] and it has recently been published [20]. The European Health Data Space (EHDS) regulation aims to improve individuals' access to and control over their personal electronic health data, while also enabling certain data to be reused for research and innovation purposes for the benefit of European patients. It provides for a health-specific data environment that will ensure cross-border access to digital health services and products within the EU.

Ethical requirements regarding AI are based on the AI Act, and the Assessment List for Trustworthy Artificial Intelligence (ALTAI) developed by the High-Level Expert Group on AI[21] set up by the European Commission to help assess whether the AI system that is being developed, deployed, procured or used in order to support the compliance the seven requirements of Trustworthy AI[22], as specified in the Ethics Guidelines for Trustworthy AI[23]. These requirements also have to fulfill the relevant upcoming laws such as the European Health Data Space Proposal, as it will be approved in the next future.

These requirements must be followed by Data Holders **regardless of the Tier of compliance with the EUCAIM DFF**. However, different agreements will be established between EUCAIM and the DH based on the requirements of each specific data provision scenario as detailed below.

We can distinguish three types of participants in EUCAIM according to their respective roles:

A) Participants willing to share datasets that are made available to third parties.

B) Participants requesting access to datasets for legitimate purposes such as health research or innovation.

C) Participants who may share software and in particular artificial intelligence tools or systems.

In all three cases, legal requirements apply:

a) General Data Protection Regulation and/or national laws.

b) National laws on research and/or research ethics.

c) National laws applicable to the health/healthcare sector.

d) Intellectual property laws.

e) National laws or standards relating to the security of processing environments.

---

[19] All references to EHDS-R are based in the document: "Proposal for a Regulation on the European Health Data Space
- Analysis of the final compromise text with a view to agreement" (No. Cion doc.: 8571/22 ADD1-8, Brussels, 18 March 2024). Available at https://www.consilium.europa.eu/media/70909/st07553-en24.pdf

[20] REGULATION (EU) 2024 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847 - https://data.consilium.europa.eu/doc/document/PE-76-2024-INIT/en/pdf

[21] High-level expert group on artificial intelligence. Retrieved September 14, 2023, from: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

[22] Ethics Guidelines for trustworthy AI. Retrieved September 14, 2023, from: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[23] Assessment List for Trustworthy AI (ALTAI). Retrieved September 14, 2023, from: https://altai.insight-centre.org/

f) Specific EU legislation (Artificial Intelligence Act, Data Act (data from wellness applications with patient consent), Data Governance Act (Data Altruism) and the future European Health Data Space Regulation (admissible secondary uses, obligations of data holders, obligations of data users, etc.)

4.4.1. Legal requirements

4.4.1.1 Legal requirements for data holders

This section presents the legal requirements for Data Holders in order to provide data or to provide access for federated data processing activities and adhere to the EUCAIM Project.

The expected end result in this area will be the sharing of an electronic health dataset with EUCAIM. Notwithstanding the fact that the legal requirements and evidence provided may be the same or very similar, it is necessary to distinguish possible scenarios. Initially two types of datasets should be differentiated.

1. Datasets which originated from a research project or activity prior to EUCAIM:
    a. Research consortia are instrumental to the research objectives.
    b. Other research projects of any kind.
    c. Datasets generated in screening programmes.
2. Datasets specifically developed for the purpose of integration into EUCAIM.

Documentary requirements

Documentation demonstrating a legal basis for sharing the data, must be provided. The legitimation must be based in conditions stated by the article 6 of GDPR related with the cases where article 9.2 legitimates data processing based on European or national laws. Finally, it is important to define how data access will be facilitated and establish access restrictions if necessary.

- Legal representation (certification and identification of the persons who can legally bind the entity such as Power of administration). Participation shall be requested by the person who can prove that they have adequate legal representativeness that enables him/her to express the will to join.
- Evidence must be provided on:
    - The lawful origin of the data and the existence of a lawful basis for the processing in accordance with the provisions of Articles 6 and 9.2 of the GDPR, and national law.
    - The existence of a legally valid controller decision allowing it to integrate into the EUCAIM scheme by providing data.
    - The due diligence in complying with the GDPR. This may be evidenced by:

        - Data Protection Impact Assessment (DPIA) reports, of the federated infrastructure (summary or report signed by the Data Protection Officer (DPO)) (only in case of Tier 3 compliance). In all other cases, it is appropriate to share this information if the DPIA was carried out in accordance with the requirements of GDPR or the positive list of cases of the national data protection authority. If the DPIA was not legally required and was not performed, it should be expressly stated in the DPO's report.
        - Duly risk assessments, independent audits, membership of codes of conduct or certification schemes or equivalent documentation.
    - A report issued by the DPO regarding the awareness and the lawfulness of the adhesion to EUCAIM scheme. This report shall justify if necessary the existence of limitations on the use of the dataset based on national legislation. Restrictions regarding the data use must be expressly included:
        - Prohibitions on commercial use of the data.
        - Restrictions due to intellectual property rights.
        - Restrictions on use depending on the conditions of the expression of consent by the data subject.
    - Report of the chief security officer and/or ISO 27001 security certification (only if the applicant indicated Tier 3 compliance). The IT team will need a complete description in terms of security, interoperability and cataloguing.

Additional requirements

A. Cases where national law lays down conditions or limitations for the processing of data.

In the transitional period until full applicability of the EHDS-R, any existing rules, obligations or limitations in the applicable national law should be indicated.

In particular, the cases in which the processing is carried out must be documented and justified. These include when:

- requires the patient's consent,
- are subject to prior communication to or prior authorisation by a public authority,
- is limited to specific purposes or subjects such as commercial uses

B. Provide evidence of the conditions for anonymisation of the data.

In the case of processing of anonymised data, evidence must be provided that adequate anonymisation is ensured. In addition to meeting the technical requirements, the following legal requirements must be met:

- There is a legal basis that allows the anonymisation and use of the data for EUCAIM purposes.
- The patient's transparency expectations have been met.
- The risk analysis methodology derived from articles 24, 32, 35 and recital (26) of the GDPR has been applied.

C. Provide evidence of the conditions related to the use of pseudonymised data.

In the case of providing access to pseudonymised data:

- It must be demonstrated whether exceptions to patient consent operate under national law.
- If consent was needed, it must be able to ensure evidence on the existence of an adequate consent issued for the purposes of EUCAIM. Depending on National law provisions the data holder should give evidence that
    - there is an exemption for consent in this case;
    - the law allows for the provision of broad consents enabling the re-use by the users of the EUCAIM infrastructure;
    - and/or that it has a procedure in place to ensure that the required consent is obtained for each new processing activity required form an EUCAIM Data User and approved by its Data Access Committee.

D. Future requirements after full application of the EHDS-R.

Once the final version of the EHDS-R is published, EUCAIM may extend the requirements to facilitate the integration of datasets provided by data holders in areas such as:

- Exclusions or prohibitions of use.
- Additional categories of available data.
- Strengthened safeguards on the use of specific data or data sources
    - human genetic, epigenomic and genomic data;
    - other human molecular data, such as transcriptomics, proteomics, metabolomics, lipidomics and other omics data;
    - data from welfare applications;
    - health data from biobanks and associated databases;
- Obligation to keep the datasets affected by the exercise of the patient's right to opt out up to date.

- Uses specifically defined by national law
- Intellectual property and rules concerning the exploitation of data that have been enriched during use.
- Transfers of data to third countries.
- Payment of fees.
- Future requirements of data based on a permit issued by a Health Data Access Body, and/or 'HealthData@EU infrastructure, under the EHDS will be considered.
- Any other obligation resulting from authorisations granted to states or defined by implementing acts of the European Commission.

Potential future requirements on EHDS-R for Data Holders

EUCAIM might adopt the requirements envisaged in the EHDS Proposal for DHs in its own premises. It will mean that the DHs providing data or access to data will collaborate in order to:

- Cooperate in good faith with the health data access bodies, where relevant.
- Communicate to the health data access body a general description of the dataset it holds in accordance with Article 55.
- Inform about the existence of a data quality and utility label.
- Put the electronic health data at the disposal of the health data access body within 2 months from receiving the request from the health data access body. In exceptional cases, that period may be extended by the health data access body for an additional period of 2 months.
- Shall make available a new dataset where a DH has received enriched datasets following a processing based on a data permit, unless it considers it unsuitable and notifies the health data access body in this respect.

Binding Agreements

During the process of accession, the envisaged processing activities will be analysed and, in particular, the anonymised or personal nature of the data and the information systems that will support the processing, whether they are the applicant's own or those belonging to EUCAIM.

The relationship between the Data Holder and EUCAIM will be established by the following types of agreement relating to properly anonymised dataset:

A. **Data Sharing Agreement (DSA):**

This document regulates data access scenarios using federated processing methodologies. In this case the datasets remain in the DH's information systems.

These DSAs can take various forms depending on the context and the specific needs of the partners, but will at least take into account the types of data to be shared, the relationship between the parties and each party's insight into the other party's activities, as well as whether the exchange could include sharing with parties in a third country.

The creation of a DSA template will be undertaken by WP3, as part of their responsibilities in developing the legal operating framework for the EUCAIM platform, and its generation and completion will depend on the conditions and terms set by each DH. The DSAs already defined in active research projects, such as Chaimeleon, can be used as a guide, which include clauses addressing essential aspects such as purpose, data provision, rules for access and use of anonymised data, data subject commitments, licence terms, confidentiality, responsibilities, applicable subsidiary criteria, integration with other platforms, validity, possible modifications, low compliance procedures and jurisdiction.

The creation of the DSA undertaken by WP3 is essential to formalise the sharing of data between entities - EUCAIM and DHs, while ensuring compliance with legal and ethical standards such as GDPR. It includes clauses that define the responsibilities of DHs and DUs, as well as technical requirements and security measures.

- **Purpose:** it regulates the rights and obligations regarding the provision of access to datasets via the federated EUCAIM infrastructure
- **Data Holder Commitment:** they are required to maintain the technical standards to the EUCAIM platform and ensure proper anonymization or pseudonymization.
- **Security and interoperability:** DHs must meet interoperability and security standards outlined in the DSA.
- **Licence Terms:** Data use must respect intellectual property rights of both the data and any results derived from it.
- **Confidentiality and Liability:** DHs must ensure confidentiality, and penalties are in place for any breaches.

B. Data Transfer Agreement (DTA):

This document regulates the transfer of data to EUCAIM information systems.

In the RWD environment, this would apply to DHs that want to transfer datasets from their premises (Data Warehouse) to the EUCAIM Reference Nodes, either because they do not have storage and/or processing resources to become a federated node or because the project has been completed. In the research environment, the DTA shall be required for projects wishing to transfer their data to the Reference Nodes so that they can be processed and used by EUCAIM even if they have ended, or for researchers who want to keep their communities alive after the project lifetime.

The DTA is crucial when data needs to be transferred between entities or jurisdictions. It governs the secure transfer of data.

- **Purpose:** To regulate the transfer of data from a DH to EUCAIM's Reference Nodes when federated processing is not feasible.

- **Anonymization and Pseudonymization:** Data must be fully anonymized before transfer, ensuring compliance with GDPR. EUCAIM will verify this before releasing the data for use.
- **Technical Requirements:** These include ensuring that the data format complies with the technical requirements of EUCAIM.
- **Legal Commitment:** The DH must ensure all legal obligations are met before transferring the data.

Some extra agreements are needed related to the processing of personal data or pseudonymised data:

- **Data Processor Agreement**. It shall apply in cases where the DH entrusts EUCAIM with tasks such as data anonymization or technical support to support its activity as DH or trusted DH in different areas such as hosting of data, implementing quality of data or security of the information systems.

- **Joint Controllership agreement**. This agreement shall only apply to data for which conditions of responsibility and liability are established in accordance with Article 26 of the GDPR.

C. Collaboration Agreement (ColA)

ColAs will be established between EUCAIM DHs, fostering symbiotic relationships built upon collaboration, co-governance, and long-term sustainability.

In the case of RWDHs, ColAs will be tailored and specified for each individual centre, taking into account their own requirements, resources, and objectives. This approach ensures that each centre's ColA aligns precisely with their specific circumstances and EUCAIM's goals.

The ColA establishes a formal partnership between research communities, hospitals, and EUCAIM.

- **Purpose:** To formalise collaboration between EUCAIM hospitals, and other RWDHs. It ensures long-term engagement with EUCAIM.
- **Sustainability:** The RWDH will contribute actively to research, participating in new observational studies within EUCAIM.
- **Specific obligations:** RWDHs must design their agreements based on individual resources and goals, ensuring their alignment with EUCAIM's objectives. ColAs are instrumental documents whose final objective is the formalisation of a DTA or a DSA. They can only legitimise the use of the respective DH's data in EUCAIM under the following conditions:
    - When the person or entity acting on behalf of the research consortium, cross-border data registry, research infrastructure or any other entity proves to have the necessary legal representation to formalise the corresponding DTA or DSA and to legally bind the entities concerned.
    - When it includes the corresponding DTA or DSA as an annex.

Otherwise, the legal relationship must be established individually with each DH.

4.4.2 Ethical requirements for Data Holders

This section provides the ethical requirements for DHs in order to share data and adhere to the EUCAIM Project.

- When required under national legislation, the entity must provide a certificate of approval validly issued by an ethics committee.
- In countries where ethical approval is not required, document the existence of an ethical risk analysis and provide a corresponding document if necessary.
- The DH shall transfer to the EUCAIM Data Access Committee the ability to verify that data access requests meet ethical requirements and accept as valid the treatments authorised to a data access applicant by the Data Access Committee who will verify the provision of ethical guarantees. In other cases, it must inform about their availability to collaborate in ethical verification activities providing support to data access applicants if needed and compromise its participation on implementing this task.

5. Rules for Participation for Software Providers

This section outlines the essential guidelines that entities aspiring to become EUCAIM Software Providers must adhere to. Their objective is to make their software, services, or applications accessible to users, enabling them to engage in federated processing or pre-processing of data sourced from the EUCAIM platform. Software providers participating in EUCAIM will benefit from increased visibility within the scientific community and the opportunity to improve or refine their tools through their use in research projects and valuable feedback from researchers and DHs in the EUCAIM network.

The cross-work package discussion on software on-boarding and packaging defined an internal document[24] depicting the main three stages for it: 1) validation of the software including ethical and technical description; 2) risk assessment of the software using an internal risk grid from EUCAIM and ALTAI; 3) packaging and deployment, where the software needs to be packaged, integrated in EUCAIM infrastructure and tested by the federated nodes.

Moreover there are a series of elements not included in the document that are mentioned in 5.1 as follows.

5.1. Minimum Requirements in Terms of Software Deployment

4.1.1. Technical Requirements and Guidelines

All the software that is to be part of EUCAIM must be provided as containerized images similar to those offered by Docker. A container orchestrator such as Docker image repository is used to make the software accessible in the reference nodes and to any federated nodes. Although, each federated node may deploy their own Docker

[24] EUCAIM software on-boarding guideline: Software on-boarding

image repository locally. These requirements match with the ones defined in Chaimeleon H2020 Grant Agreement nº 952172.

- All of them shall comply with the specifications for inputs and outputs described in the EUCAIM technical documentation, defined in the *D5.1 Early release of the Data Federation Framework*.

5.1.2. Minimum Requirements for Software Inclusion

- Any software that is to be added to the project must comply with the technical guidelines and terms of usage provided by EUCAIM, defined in the *D5.1 Early release of the Data Federation Framework*.
- Any software that is to be added to EUCAIM must provide in its documentation information regarding the possible use cases in which the software can be used and the expected output in terms of performance from each of them. Any possible contraindication, meaning any specific case in which the software should not be used, must be clearly specified in this documentation.

5.1.3. User Support and Software Maintenance

- All the software that is to be part of EUCAIM must ensure a communication channel that allows users to contact the provider of the software when any kind of incident or issue arises while using it. This channel shall rely on the EUCAIM Helpdesk.
- Each Software Provider must offer long-term support for their software to ensure a secure and stable behaviour through the whole lifespan of EUCAIM. Unless duly justified, this support will last, at least, until the end of the EUCAIM piloting stage, planned to last up until December 31, 2026.
- The Software Providers shall sign a Service Level Agreement (“SLA”) in order to guarantee that the software is up to date and presents no known vulnerabilities.

5.1.4. Minimum Documentation Requirements and Benchmarking Information

To promote transparency and facilitate software evaluation, SP are required to provide comprehensive documentation and benchmarking information (as described in the internal document):

- Product Documentation:
  - User manuals, installation guides, and configuration instructions for the software.

    - Library dependencies (internal ones and 3rd party ones), including documentation when required.

- Licence Agreement:
    - A copy of the software licence agreement that outlines the terms of use.

- Data Usage and Privacy Policy:
    - Documentation that explains how the software handles data. It shall comply with EUCAIM's data privacy requirements.

- Security Documentation:
    - Information on the security measures implemented in the software, such as encryption protocols, access controls, and vulnerability management.

- Software version control:
    - All the software must implement a clear and concise version control mechanism that outlines relevant changes and additions to each version of the software.

- Compliance and Certification Documents:
    - Compliance certificates or documentation proving that the software adheres to industry standards, legal requirements, and best practices (e.g., GDPR compliance).

- API Documentation:
    - API documentation that outlines how to interact programmatically with the software.

- Technical Support and Maintenance Agreement:
    - Details about the entity's technical support availability, response times, and the terms of maintenance and updates for the software (SLAs).

- Instructions for use:
    - Access to training materials or resources that can help to understand the software and how to use it.

Benchmarking Information

Software providers must communicate the following information about their software. It is strongly recommended to provide a set of tests to verify the software benchmarks:

- Software Description: A concise overview of what the software does and its primary purpose.

- Training and Validation Dataset Description: Details about the dataset used for training the software, including the number of cases, data distribution, image modalities (if applicable), and any relevant preprocessing steps.
- Software Type: Specify the type of software, such as preprocessing, AI model or analysis software.
- Task: Describe the specific task(s) that the software is designed to perform, such as segmentation, harmonisation, classification, etc.
- Performance Metrics: List the performance metrics used to evaluate the software, which may vary based on the task and software type. Examples include DICE score, AUC ROC, precision, recall, F1-score, etc.
- Input Requirements: Specify the input data requirements, including image modality, format, and any important consideration.
- Output Description: Explain the type and format of the output generated by the software.
- Licence Information: Clarify the licensing terms under which the software is available for use, including any open-source licences or restrictions.
- Hardware Requirements: Detail the CPU and GPU requirements for running the software effectively.
- RAM Requirements: Specify the amount of RAM (memory) required for optimal software performance.
- Processing Time: Provide an estimate of the time required to process a typical input, which can help users plan their workflow.
- Programming Language: Indicate the programming language(s) in which the software is developed or can be extended.
- Relevant Keywords: Include keywords or tags that describe the software's focus and functionality, making it easier for users to find relevant software in search queries.
- Publications: If applicable, list any research papers, articles, or documentation related to the software's development or validation.
- URL: Provide a link to the software's official website, repository, or relevant page to access more information.

5.2. Traceability Mechanisms

- The software shall register all the relevant actions that each of the users perform involving the software itself. This register must be accessible from the outside of the software for auditing purposes.

- Each software that is part of EUCAIM has to be able to provide relevant logs that allow it to monitor their usage. Additionally, these logs must allow to identify unequivocally any incidence that may arise through proper error codes.

5.3. Monitoring Capabilities

- The software shall be able to provide information that allows EUCAIM to properly monitor its status.

5.4. Quality Control Measures

All the software must present quality control measures in the following regards:

- Code-related quality controls in the form of unit tests in the codebase that conforms to the software.
- Functional validation of the software by a designated person.
- Appropriate registries showcasing that the two previous points have taken place and their outcomes.
- In the case of SW under Open Source Licences, an external assessment of the SW Quality would be recommended. For example, achieving at least a Bronze Badge (Silver Badge preferably) in the EOSC SQAaaS (Software Quality as a Service https://sqaaas.eosc-synergy.eu). This tool evaluates SW code according to a baseline of metrics related to scientific SW. SQAaaS provides this evaluation “as a service” and automatically, providing a detailed report on the strengths and weaknesses of the SW.

5.5. Security and Privacy Compliance

- All the software that handle sensitive data must comply with all the specifications stated in the GDPR guidelines along with the ones described in the Legal and Ethical Requirements section of this document.
- The software shall pass a vulnerability analysis through software such as SonarQube.
- The containerized images shall pass an analysis related to the data privacy assessment.
- Breach procedure or contact point for information, patches or updates must be established. All AI act software tools must be updated and maintained. Any found possible vulnerability must be immediately communicated and action must be taken, which could include deactivation of the compromised tool. If software/libraries from third parties are used (like plugins) it will be necessary to maintain a list of these libraries.

5.6. Legal and ethical requirements for software providers and/or developers

5.6.1 Legal requirements

All the software that may want to join EUCAIM's infrastructure need to comply with current applicable European and national legislation. In addition, upcoming regulations that will - presumably - enter into force during the development of the EUCAIM Project, such as the EHDS or the AI Act Proposals Regulations, must be complied with - once in force - by SPs.

All software that handle personal data within EUCAIM's infrastructure must comply with the provisions established under Regulation (EU) 2016/679 GDPR, national privacy legislations and various legal and ethical recommendations, guidelines or opinions issued by national data protection authorities and European Union (EU) bodies (i.e., the European Data Protection Board).

Whenever a software is to be onboarded to EUCAIM's infrastructure and involves processing personal data, the SP must ensure beforehand (and be able to provide proof) that it has obtained all necessary approvals for the specific processing activity performed by the software. Additionally, the SP must inform users of the software's suitability for working with anonymized, pseudonymized, or both types of data.

Regarding compliance with provisions or limits set by the Access Committee, while software themselves don't have a direct link to data access, it's important to ensure that the software used within EUCAIM align with the access policies and guidelines set by the Access Committee. Therefore, SP will cooperate with EUCAIM's DHs, DUs, and national health data access bodies as needed to facilitate the responsible and ethical use of their software within the EUCAIM environment.

The processing of data by the software to be on-boarded into EUCAIM's infrastructure, whether anonymized or pseudonymized, may require executing different legal agreements, such as a data processing agreement or any other commitments to be undertaken.

From a privacy perspective, the SP shall be able to provide information on various aspects regarding the processing of data done by the software. For instance -and among others - the following aspects:

- Compliance with GDPR must be evidenced by reports on data protection impact assessments, audits done by third parties or adherence to codes of conduct or certification schemes. The SP will need to provide proof of such certification whenever it has implemented a certifiable safety standard.
- The location and storage of the data generated by the software and information on the safeguards in place in case data processing entails international data transfers.
- Accrediting the training of its staff in personal data protection, confidentiality and security systems.

- Information on the safeguards in place in case data processing entails international data transfers.
- Information on the methodology implemented throughout the development of the software in order to ensure data protection by design and by default.

In the case of sharing software resources, information systems, artificial intelligence systems, **legal risks and liabilities may arise** for the provider, for EUCAIM or for any user.

- It should be stated whether the development is proprietary, open source, or subject to reuse conditions such as a Creative Commons License or equivalent.
- In the case of experimental software, potential risks arising from its use should be disclosed.
- Transparency about the code should be provided, or when this is not possible, the legal basis or reason for maintaining business secrecy about the code should be stated.
- Sufficient information (AI Literacy) should be provided to ensure adequate knowledge about the nature, conditions, mode of use and risks in the use of an artificial intelligence system.
- If the software is subject to prior authorisation requirements (Medical Device Regulation), supporting documentation should be provided

In the case of providing software to EUCAIM or proposing the development of tests on any type of software, including the verification of algorithms evidence must be provided on:

- Ownership (proprietary, licensed) & IP rights declaration including any restriction for authorising access/or use to/of the product by DUs with an EUCAIM permission.
- In case you request to develop a software or any related technology or electronic product: Status of the software (marketed, research, prototype pending approval (medical device).
- In case the purpose of the software is the processing of personal data:
    - Report/self-declaration from the DPO certifying that the organisation is GDPR-compliant and the future EUCAIM's users will have been duly trained in this topic.
    - Impact assessment carried out if required
        - AI impact assessment (ALTAI) and AI Fundamental Rights impact assessment (FRIA).
        - Data Protection Impact Assessment (DPIA)
- Any supporting documentation of the product development conditions:
    - Security risk analysis including risks related to IA.

- Security measures applied.
- Documentation related to the software (requirements, code, etc.).
- Technical documentation on development conditions required by law (AI ACT technical annexes, Data protection By Design and by Default in GDPR).
- Legal representation (certification and identification of the persons who can legally bind the entity such as Power of Attorney).
- Signature of Terms & Conditions by the legal representative in case of software test or developing on EUCAIM infrastructure.
- Signature of security obligations and non-re-identification commitments for each user.

5.6.2 Ethical requirements

It should be noted that in the case of software development as a research or innovation activity, these must be provided:

- An ethical approval.
- In the event that the national legislation imposes some kind of ethical self-assessment process or ethical self-reporting process, a copy of these should be included. It could be stated in the DPO's report.

Among the ethical requirements to be fulfilled by SP- whenever applicable - it shall be highlighted that SP will need to provide an AI risk analysis, preferably by using the ALTAI Tool for assessing AI-based technologies. AI-based software shall comply with the ethical requirements established within the ALTAI Tool developed by a group of experts appointed by the European Commission to provide advice on its AI strategy.

The ALTAI Tool shall be used to assess whether AI-based technologies are developed, deployed, procured or used to comply with the seven requirements of Trustworthy AI established under the Ethics Guidelines for Trustworthy AI. Said guidelines establish that Trustworthy AI must be lawful (respecting all applicable laws and regulations), ethical (respecting ethical principles and values), and robust (both from a technical perspective while taking into account its social environment).

5.7. Evaluation and Integration

EUCAIM will evaluate software to determine their integration into the platform. This evaluation process will ensure that the software meets the minimum requirements outlined in this section. Once integrated, software will be assessed by Data Users and Data Holders (in the case of preprocessing tools), offering visibility and use among projects. Ratings and user feedback will serve as internal benchmarking for EUCAIM, further enhancing the quality and utility of the platform.

6. Rules for Participation for Data Users

After establishing the rules of participation for all EUCAIM providers, including those supplying data and software, the next step is to define the guidelines for Data Users. This group encompasses both researchers and innovators, as specified in the user roles outlined in deliverable *D4.2*. These end-users of the platform will be able to explore the public catalogue of available metadata without being authenticated. Additionally, once authenticated, they can perform federated queries on Tier 2-3 datasets. If they wish, they can request access to datasets through the Negotiator component and, if access is granted, they will be able to process the data using the software available on the platform and/or their own AI tools. To do so, they have to follow the procedures detailed in this section and comply with the corresponding requirements. It should be noted that all of them are based on the main prerequisite for DUs to have sufficiently documented and previously approved projects in the overall context of Research and Innovation projects, or alternatively, any necessary commercial agreements in place.

6.1. User identity checking procedure

A DU needs to register on the platform via the Life Science AAI[25] (several options are available for this, like affiliation to organisations, using an ORCID, an LS Hostel account). If the users' institutional IdP is supported (e.g. academic and research institutions affiliated to EduGAIN) this should be the preferred choice. As part of this authentication process, the user has to accept several usage conditions as defined in the Acceptable Use Policy[26] of EUCAIM services:

Common Conditions of Use:

The specific Conditions of Use will be included in the documents defined by WP3. Some of the Common Conditions of Use will be:

- The User agrees to be a bona fide researcher with (1) an intention to generate new knowledge and understanding using rigorous scientific methods, (2) an intention to publish the research findings and share the derived data in the scientific community, ideally without restrictions and with minimal delay, for wider scientific and eventual public benefit, and where (3) the intended activities are not inconsistent with legal and ethical requirements or widely recognised good research practice.
- The User will avoid any attempts to reverse privacy enhancing technologies (i.e., pseudonymization, anonymization) applied to the data.
- If possible, any incidental findings will be reported back to the corresponding body within the EUCAIM consortium.

Service-Specific Conditions of Use:

[25] Life Science Login. Retrieved September 14, 2023, from: https://lifescience-ri.eu/ls-login/
[26] EUCAIM Platform Acceptable Use Policy and Conditions of Use Retrieved December the 11, 2024, https://dashboard.eucaim.cancerimage.eu/eucaim_usage_policy.pdf

The Service-Specific Conditions of Use of EUCAIM for accessing data will be based in resources such as the Harmonised Access Procedure to Samples and Data[27] and the General Terms of Use End Users of Health-RI[28], as well as other resources produced by well known and established distributed data research initiatives and biobanks. The access criteria and procedures from these resources cover topics such as eligibility, application process, review process, criteria for access, and terms of access.

The registration process is described in the Registration of Users document[29]. It involves two main steps:

- The creation of a LS-AAI account linked to a personal account of the user.
- The request for membership to the EUCAIM group in LS-AAI.

The creation of a LS-AAI account can be performed through the “My Profile” area of the Dashboard (https://dashboard.eucaim.cancerimage.eu). The details of the account can be modified in https://profile.aai.lifescience-ri.eu/profile later on.

The creation of a LS-AAI account is a prerequisite to access EUCAIM restricted resources, but it does not grant access to them. In order to access the EUCAIM restricted resources, the user has to request membership to the EUCAIM Group. This process is manually verified by the security team of the platform, checking that the account requests are not fake and requesting any additional information if needed. It is important to outline that users in the EUCAIM Group will only have access to the aggregated data but they will not be able to access actual data. Data access is granted through the negotiator service. The request for membership to the EUCAIM Group can be performed through the enrollment form[30] or directly through the “My Profile” area in the dashboard, at the first access to the private area.

6.2. Data access request process

A DU may submit a request for access to data, whether it is already available from existing research repositories exposed in the Public Catalogue or they want to generate on-demand datasets within the RWD environment under a proposal for a new observational study. This request process is orchestrated through the EUCAIM Negotiator service which allows the interaction of the Access Committee and the requesters. The request is evaluated by the Access Committee, and, depending on the licences signed with the Data Holder that owns the data, it may also be evaluated by the DH’s Access Committee (or an equivalent entity). The approval of the request implies the granting the usage of the datasets.

[27] Georges Dagher, Petr Holub, Michael Hummel, Anu Jalanko, Outi Törnwall, Kaisa Silander, Marialuisa Lavitrano, Kurt Zatloukal, Mats Hansson, Michaela Th. Mayrhofer, Maimuna Mendy, Philip Quinlan, & Irene Schlünder. (2016). Harmonised Access Procedure to Samples and Data. Zenodo. https://doi.org/10.5281/zenodo.823013
[28] General Terms of Use End Users. Retrieved September 14, 2023, from: https://www.health-ri.nl/terms-use-health-ri-services
[29] Registration of Users in EUCAIM https://u.i3m.upv.es/m44jv
[30] EUCAIM Group enrollment form: https://signup.aai.lifescience-ri.eu/fed/registrar/?vo=lifescience&group=communities_and_projects:EUCAIM

Data access negotiation (BBMRI)

The Negotiator allows users to request access to data from one or several DHs as selected in a previous discovery step in the EUCAIM catalogue. The request can be made to several DHs in parallel and the negotiation mechanism allows the Access Committee

A. to obtain more information from the requestor to better understand the reason and other details of the request and the requested data in a broadcast mode;

B. to enter into bilateral negotiation with the requester;

C. or to step back from a request in case the requester is not eventually capable of fulfilling what was requested for any reason.

The information exchanged in this process as well as the related interactions is restricted to the Access Committee and requesters within the negotiator interface. The requester may step into several such negotiations as needed within the context of the project for which the data were requested.

To perform the above described process the DUs need to have access to the Negotiator, which is integrated into the EUCAIM Dashboard. The Negotiator requires login via Life Sciences Authentication and Authorization Infrastructure (LS AAI) for that purpose. Furthermore, it is important that all roles involved in the Negotiation are represented and can react timely to the request negotiation procedures, especially local DH and their Access Committee if being defined as part of the process.

6.3. Request form and specific requirements

The following are the information required from Data Users in the Negotiator component and the mandatory documents they must provide in order to submit a data access request[31].

6.3.1 Application documents for DUs requesting access for already available datasets

- Title
- PI details: name, surname and orcid or equivalent id
- Institutional letter of approval: a letter from a responsible person of the organisation of the PI declaring the acceptance to conduct the project
- Proponents: centres, researcher team and short CV with previous experience (no more than 50 words per CV)

[31] Although these are the documents currently requested for DU, it should be noted that this information was found to be too focused for research and not so much for innovation. For this reason, in deliverable D2.3 Requirement analysis of Real World Data Holders, where the requirements of end-users have also been analysed, a new proposal for application documents has been made, adapted to the profile of innovators within the role of DU. Both contents will be taken into account and the corresponding integration will be made in a new version of the Negotiator after its revision and updated in the Dashboard.

- Cover letter: provide a cover letter (main motivational objective, previous experience, key, contributions excepted) for your application (no more than 500 words)
- Hypothesis to be developed with its clinical impact (no more than 300 words)
- Objectives of the application, intended usage, study design and work plan (no more than 500 words)
- Expected results and applicability (no more than 100 words)
- Materials and methods: necessary material and methods, such as target population, type of data (image modalities, case report forms), datasets (filtering criteria, recruitment period), number of cases, annotations, tools, computational resources and temporary storage (no more than 600 words)
- Project timeline: duration and activities
- Sources of funding (no more than 200 words)
- Supporting documentation:
    - Favourable Ethics Committee report (pdf)
    - Approved project (pdf)

6.3.2 Application documents for DU willing to build observational studies with RWD

- Title
- PI details: name, surname and orcid or equivalent id
- Institutional letter of approval: a letter from a responsible person of the organisation of the PI declaring the acceptance to conduct the project
- Cover letter provide a cover letter (main motivational objective, previous experience, key, contributions excepted) for your application (no more than 500 words)
- Proponents: centres, researcher team and short CV with previous experience (no more than 50 words per CV)
- Consortium: check for EUCAIM partners experience based on your specific needs (such as hospital data holders, research centres, partners with local computational resources, partners with expertise in data preparation, data curation, AI and software development, legal and ethical aspects, project management, among others)
- Hypothesis to be developed with its clinical impact (no more than 300 words)
- Objectives of the application, study design and work plan (no more than 500 words)
- Materials and methods: necessary material and methods, such as target population, type of data (image modalities, case report forms), datasets (filtering criteria, recruitment period), number of cases, annotations, tools, computational resources and temporary storage (no more than 600 words)

- Expected results: expected results and applicability (no more than 100 words)
- Supporting documentation: Favourable Ethics Committee report (pdf)
- Research call: research Call being applied to

6.4. Legal and Ethical Requirements

EUCAIM DUs, as well as the other roles explained in this deliverable, must comply with a number of legal prerequisites and ethical obligations, to ensure responsible and secure processing of sensitive health data. For this role, the requirements to be fulfilled are detailed in the following subsections.

A data access application involves participants requesting access to datasets for legitimate purposes such as health research or any other purposes stated by future EHDS-R:

a) public interest in the area of public and occupational health, such as activities for protection against serious cross-border threats to health and public health surveillance or activities ensuring high levels of quality and safety of healthcare, including patient safety, and of medicinal products or medical devices;

b) policy making and regulatory activities to support public sector bodies or Union institutions, agencies and bodies, including regulatory authorities, in the health or care sector to carry out their tasks defined in their mandates;

c) statistics, such as national, multi-national and Union level official statistics defined in Regulation (EU) No 223/2009 related to health or care sectors;

d) education or teaching activities in health or care sectors at the level of vocational or higher education;

e) scientific research related to health or care sectors, contributing to public health or health technology assessment, or ensuring high levels of quality and safety of health care, of medicinal products or of medical devices, with the aim of benefitting the end-users, such as patients, health professionals and health administrators, including:

   i) development and innovation activities for products or services;

   ii) training, testing and evaluating of algorithms, including in medical devices, in-vitro diagnostic medical devices, AI systems and digital health applications;

f) improving delivery of care, treatment optimization and providing healthcare, based on the electronic health data of other natural persons.

In all cases, legal requirements apply:

a) General Data Protection Regulation and/or national laws.

b) National laws on research and/or research ethics.

c) National laws applicable to the health/healthcare sector.

d) Intellectual property laws.

e) National laws or standards relating to the security of processing environments.

f) Specific EU legislation (Artificial Intelligence Act, Data Act (data from wellness applications with patient consent), Data Governance Act (Data Altruism) and the future European Health Data Space Regulation (admissible secondary uses, obligations of data holders, obligations of data users, etc.)

The accountability principle **does not only apply in the area of GDPR**. **All EUCAIM operations governed by law and for which liability may arise must be supported by documentary evidence**.

6.4.1. Legal requirements for data users

EUCAIM will adopt the requirements envisaged in the EHDS Proposal, regarding the data access application requirements:

- a detailed explanation of the intended use of the electronic health data, including for which of the purposes referred to in Article 34(1) access is sought;
- a description of the requested electronic health data, their format and data sources, where possible, including geographical coverage where data is requested from several Member States;
- an indication whether electronic health data should be made available in an anonymised format;
- where applicable, an explanation of the reasons for seeking access to electronic health data in a pseudonymised format;
- a description of the safeguards planned to prevent any other use of the electronic health data;
- a description of the safeguards planned to protect the rights and interests of the data holder and of the natural persons concerned;
- an estimation of the period during which the electronic health data is needed for processing;
- a description of the tools and computing resources needed for using EUCAIM secure environment.

Procedural activities and documentary evidences:

- In the event that the proposal for participation may involve the contribution of the data access requester's own datasets, they must provide the documentary evidence required in the previous section for data holders.
- If the data access requester's ask for the integration of software, they must comply with the requirements of the software providers.

- Persons or entities requiring access to data shall register on the platform. Registration shall be of two types:
    - Individual: can be done by any interested person and only offers access to EUCAIM, available datasets, processing tools, collaboration forums, newsletters and other social tools. Joining EUCAIM implies acceptance of the individual registration terms.
    - Corporate. The request for access to data will only be valid when submitted or ratified by a person with sufficient legal power to contract.
- In case of a data permit the authorised entity and user/s must provide:
    - Entity:
        - Legal representation (certification and identification of the persons who can legally bind the entity such as Power of Attorney)
        - Must accept the terms and conditions for the use of data. These general conditions may be supplemented by additional obligations or safeguards, taking into account specific legislation in force such as the Data Governance Act and future EHDS[32], Data Act or AI Act.

[32] Article 41a
Duties of health data users
1. Health data users may access and process the electronic health data for secondary use referred to in Article 33 only in accordance with a data permit pursuant to Article 46, a data request pursuant to Article 47 or, in situations referred to in Article 45(3), a data access approval of the relevant authorised participant.
2. When processing electronic health data within the secure processing environments referred to in Article 50, the health data users are prohibited to provide access to or otherwise make the electronic health data available to third parties not mentioned in the data permit.
2a. Health data users shall not re-identify or seek to re-identify the natural persons to which the electronic health data which they obtained based on the data permit, data request or access approval decision by an authorized participant in Health Data EU relate.
3. Health data users shall make public the results or output of the secondary use of electronic health data, including information relevant for the provision of healthcare, within 18 months after the completion of the electronic health data processing in the secure environment or after having received the answer to the data request referred to in Article 47.
This period may in justified cases related to the permitted purposes of the processing of electronic health data be extended by the health data access body, in particular in cases where the result is published in a scientific journal or other scientific publication.
Those results or output shall only contain anonymous data.
The health data users shall inform the health data access bodies from which a data permit was obtained and support them to also make the information related to the results or output provided by the health data users public on health data access bodies' websites. Such publication on the health data access bodies website shall be without prejudice to publication rights in a scientific journal or other scientific publication.
Whenever the health data users have used electronic health data in accordance with this Chapter, they shall acknowledge the electronic health data sources and the fact that electronic health data has been obtained in the context of the EHDS.
4. Without prejudice to paragraph 2, health data users shall inform the health data access body of any significant findings related to the health of the natural person whose data are included in the dataset.
5. The health data users shall cooperate with the health data access body when the health data access body is carrying out its tasks.

- Provide a data protection Impact Assessment if needed. When necessary according to the conditions of the processing, additional documents such as a Data Protection Impact Assessment may be required EUCAIM may provide documentation related to its processing environment in order to facilitate the performance of these DPIAs.
- Identify the authorised users.
- National Law specific requirements such as a prior authorisation issued by a public body.
- In case of processing personal data including pseudonymised data:
    - Duly justification of the need.
    - Report/self-declaration from the Data Protection Officer (DPO) certifying that the organisation is GDPR-compliant and the future EUCAIM's users have been duly trained in this topic.
    - Impact assessment carried out if required
        - AI impact assessment (ALTAI) and AI Fundamental Rights impact assessment (FRIA).
        - Data Protection Impact Assessment (DPIA)

- Users:
    - The security obligations must be accepted by any individual user related with the entity at his/hers first connection to the platform.
    - The non-re-identification binding commitment if required.

6.4.2. Ethical requirements for data users

Data users will be required where needed to:

- Describe the ethical conditions for the data processing.
- Provide a certificate of ethics approval issued by an ethics committee.
- In the event that the national legislation imposes some kind of ethical self-assessment process or ethical self-reporting process, a copy of these should be included. This could be stated in a DPO's report.
- Provide an AI risk analysis preferably by using the ATAI Tool[33].

[33]ATAI Tool. Retrieved September 14, 2023, from: https://altai.insight-centre.org/Identity/Account/Login

7. Rules for Participation for Research Communities

The Research Communities (RCs) within EUCAIM function as virtual organizations, bringing together groups of users with common permissions, goals, and access to relevant datasets and software. These communities operate as overarching collaborative projects without predefined time constraints, providing a more flexible structure tailored to specific objectives and research areas.

Research Communities are established to enable focused collaboration and data sharing. They are formed through a formal creation request submitted via the Dashboard. This request must involve a consortium, providing detailed information about each member and designating a responsible person to lead the community.

Once approved, a dedicated space is allocated on the reference node, where members can securely share and access datasets and software relevant to their area of interest. RCs will also have the capability to initiate new projects within their space.

Membership in an RC is restricted to individuals who have completed the access request process. After its creation, other EUCAIM users can request to join the RC, subject to approval.

Periodic evaluations of the community's impact will be conducted to track progress.

Communication will take place through a dedicated channel within the Helpdesk or another designated forum to ensure smooth collaboration and support among members.

Legal requirements

The integration of datasets into EUCAIM shall meet the same safeguards as those set out in this document for DHs. The Memorandum of Understanding (MoU) is a statement expressing this commitment to integration.

From a legal point of view, the sharing of datasets with and through EUCAIM can take place in two ways:

a) Research consortia and any other form of aggregation of DHs (e.g. cross-border data registries, research infrastructures, etc.) may only agree directly to share data with EUCAIM when they can demonstrate ownership of the datasets or have been granted sufficient legal representation to enter into an agreement with EUCAIM.
b) If the above circumstances do not apply, data sharing shall be agreed individually with each single DH.

The sharing must necessarily be based on a DSA (federated data processing) or on a DTA in case the data are physically transferred to the EUCAIM premises.

8. Compliance framework design

Although the design of the compliance framework is not a rule of participation per se, this section addresses the common legal, ethical and data protection framework in which the rules of participation must be established. This section describes the compliance methodologies that apply to the design and development of the EUCAIM platform or secure

processing environment and any other data processing. It is a methodology based on a risk-based approach, focused on ensuring the rights of individuals, and governed by the principle of data protection by design and by default.

When it comes to security measures, the compliance framework ensures that security protocols are in place at both the platform and dataset levels.

EUCAIM assures to adopt a compliance framework based on GDPR in this structure (Table 3):

Table 3. Workflow for GDPR compliance.

Prerequisites	
Relations between project partners	• Definition of a framework for trust relationships/coordination of data protection officers • Definition of GDPR relationships (controller-processor-processors-responsible parties) and relationships with the use of de-identified data (data sharing and transferring agreements)
Description and contextualisation of treatment	
Notification process	• Description of treatment • Purpose • Data, persons concerned, uses...
Risk analysis	
Risk-based approach	• Data protection impact assessment • Risk analysis • Risk management (measures, residual risk...)
Data protection by design and by default	
Technical design	• Data minimisation. • Adequate implementation of applications and procedures in accordance with the provisions of the GDPR
Actions to adapt the legal design of the processing to the requirements of the GDPR	
Legal design	• Relations with third parties • Competent data protection authority • International transfers • Transparency • Register of processing activities

Specific requirements	• Cookies • Social networking • De-identification

Proactive accountability	
Evidence	• Risk analysis reports • Data protection impact assessment reports • Technical documentation in application development (data protection from design, functionality, security) • Legal documents • Audit reports

Additionally, EUCAIM will define the procedures for assessing the ethical impact of data analytics or the use of AI in its environment. This methodology, and the guarantees and obligations it entails, will be communicated to:

- Technology providers targeting platform design, data analytics software and/or APIs
- Applicants (users)

The ALTAI tool[34] could be used to carry out the assessment. This tool translates the principles defined by the European Commission's High Level Expert Group (HLEG) in its Ethical Guidelines for Trustworthy AI[35]. The reliability of AI relies on three components that must be satisfied throughout the life cycle of the system:

1. the AI must be lawful, so as to ensure that all applicable laws and regulations are respected

2. it must be ethical, i.e. ensure compliance with ethical principles and values

3. it must be robust, both technically and socially, as AI systems, even if well-intentioned, can cause accidental harm

To this end, AI systems must be human-centred, underpinned by a commitment to their use in the service of humanity and the common good, with the aim of enhancing human well-being and freedom.

The HLEG conceives of four main ethical principles in terms of "ethical imperatives". These are:

- **Respect for human agency.** The distribution of functions between humans and AI systems should follow human-centred design principles, and leave ample opportunities for human choice. This means ensuring human oversight and control

[34] Ala-Pietilä, P., Bonnet, Y., Bergmann, U., Bielikova, M., Bonefeld-Dahl, C., Bauer, W., & Van Wynsberghe, A. (17 jul 2020). The assessment list for trustworthy artificial intelligence (ALTAI). European Commission. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

[35] High-level expert group on AI. Ethics guidelines for trustworthy AI (08 April 2019. Retrieved September 14, 2023, from): https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

over the work processes of AI systems. AI systems can also fundamentally transform the world of work. They should help people in the work environment and aim to create useful jobs.

- **Prevention of harm.** AI systems should not cause harm (or aggravate existing harm) or otherwise harm human beings. This entails the protection of human dignity, as well as physical and mental integrity.
- **Equity (fairness).** Fairness implies that AI practitioners should respect the principle of proportionality between means and ends, and carefully consider how to strike a balance between different interests and competing objectives.
- **Explainability.** Explainability is crucial for gaining and maintaining user trust in AI systems. This means that processes need to be transparent, that the capabilities and purpose of AI systems need to be openly communicated, and that decisions need to be explained - as far as possible - to parties who are directly or indirectly affected by them. It is not always possible to explain why a model has generated a particular outcome or decision (or what combination of factors contributed to it). Such cases, which are referred to as "black box" algorithms, require special attention.

The ALTAI methodology integrates seven ethical requirements that are in practices dominions including a control checklist:

1. **Human action and monitoring.** Including fundamental rights, human action and human oversight.
2. **Technical soundness and safety.** Including resilience to attacks and security, a fall-back plan and security, accuracy, precision, reliability and reproducibility.
3. **Privacy and data management.** Including respect for privacy, data quality, data integrity and access to data.
4. **Transparency.** Including traceability, explainability and communication.
5. **Diversity, non-discrimination and equity.** Including freedom from unfair bias, accessibility and universal design, as well as the involvement of stakeholders.
6. **Social and environmental well-being.** Including sustainability and respect for the environment, social impact, society and democracy.
7. **Accountability.** Including auditability, minimisation and reporting of negative effects and trade-offs.

The following figure (*Figure 3*) depicts the interrelationship between these seven ethical requirements, which are all of equal importance and mutually supportive.

The HLEG establishes a non-exhaustive list of AI trustworthiness assessment (pilot version) to implement trustworthy AI. Depending on the nature of the project, this ethical impact assessment must be applied to the AI application development environment and to requests for access and use of information.

Figure 3. Interrelationships between the seven ethical requirements[36].

9. Evaluation of applications and expected response times

The evaluation of applications from potential new partners received through the open call for use cases (both for data provision and data access requests) followed a structured process to ensure that participation in the EUCAIM infrastructure is granted based on rigorous scientific, technical, legal, and ethical criteria. Applications were assessed by the EUCAIM Access Committee using the established criteria outlined in Deliverable D7.1.

Each application was independently reviewed by three Access Committee members. The Committee then convened twice to discuss all applications collectively, ensuring consistent evaluation metrics across all members and addressing any specific issues encountered during the review process. In addition to technical, scientific, legal, and ethical aspects, other factors were considered when shortlisting the applications, such as geographical coverage, the availability of the requested data on the platform, and the level of readiness of the legal documentation.

The applications shortlisted through this evaluation process were submitted to the European Commission for final approval. Successful applicants will become new beneficiaries of the EUCAIM project through an amendment to the Grant Agreement (GA) by the end of the year. The implementation of the use cases is expected to begin in early 2025, immediately after all required legal documents and agreements are finalized.

[36] HLEG (2019). Ethics Guidelines for Trustworthy AI. European Commission. p. 15. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

10. Conclusions

This deliverable provides the rules for participation for the main roles that will join the EUCAIM platform. This has led to the definition of minimum requirements for Data Holders to join the Federation and expose their data. These requirements have been established in terms of the data itself, with three technical levels in accordance with the Data Federation Framework defined in the deliverable D5.1, as well as in terms of access to the data, the necessary infrastructure, and the ethical and legal requirements. The minimum requirements to be met by Software Providers, Data Users and Research Communities have also been defined. This document also provides an overview of the compliance framework design.

The project consortium hopes this deliverable will set the grounds on how to join the EUCAIM platform, both for providers and data users. This deliverable will be further validated and developed through the implementation of use cases provided by EUCAIM partners.

ANNEX 1. Anonymisation (All Tiers)

Ensuring GDPR compliance, whenever applicable, is a fundamental element of the EUCAIM platform. Nevertheless, it shall be highlighted that anonymized data is not considered personal data, as defined under Article 4 of the GDPR, which states that personal data "*means any information relating to an identified or identifiable natural person*". An identifiable natural person (or data subject) is one who can be identified, directly or indirectly, by way of the name, an identification number, location data, or any other factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that data subject.

Therefore, given that the principles of data protection established under GDPR only apply to data concerning an identified or identifiable natural person (as stated in Recital 26 GDPR), GDPR shall not apply with regard to anonymized data. In EUCAIM, the data will follow a risk-optimization based anonymization approach. This not only refers to the proper data anonymization but it will also require that:

1. The data holder ensures that the data processing is legitimate and has been authorised by the data controller and that it will be shared under a transfer agreement with EUCAIM.
2. A legal-technical governance framework for data access must be established by the EUCAIM platform, with measures such as registration process by the users before accessing the data, ensuring they know the rules of the platform, the terms and conditions and their obligations which includes the non-re-identification commitment.
3. A secure processing environment preventing improper downloading or manipulation of data by the data users, excluding the risk of patient re-identification by third parties and the combination of data with other sources of information.

The EUCAIM platform will securely store clinical and imaging data within the Reference and Local Nodes in accordance with standards defined by the EUCAIM Technical and Legal teams. Regarding imaging data, they must undergo a process of de-identification, ensuring that any personal data is removed or changed to an impersonal item. The data holder may de-identify it using its own tools and services before uploading the data. It is important to highlight that in these cases, the Patient ID provided in the data should not match any records such as health cards or medical history records, it must be assigned during the de-identification process and it is preferred to be a hash (output of a hashing function) in the cases of pseudonymous data or a random id for anonymous data. However, EUCAIM will provide a set of tools (i.e. DICOM Anonymizer and Wizard Tool) for properly anonymizing data (i.e. implying anonymization rules and risk estimation / minimization from DICOM anonymizer and Wizard tool respectively) as long as the images are in a valid DICOM format and the clinical data follows the EUCAIM Common Data Model. Given that the exposure of datasets to the Federation is subject to the generation of research and innovation projects, full compliance with the GDPR is assumed on their part, regardless of the type of data (pseudonymized or anonymized) that has been processed in the project.

Therefore, there is the possibility that the data stored in the federated nodes is in a pseudonymized regime, which will depend on the conditions in which the project (and therefore the dataset/s) has been constructed, in addition to the agreements signed both at the intra-project/consortium level, as well as from the project with EUCAIM. EUCAIM will

conduct checks to verify its compliance with the established standards regarding pseudonymization.

The final design of EUCAIM will take into account the obligations deriving from the future European Health Data Space Regulation and in particular those relating to:

- Design of a secure processing environment

- Obligations of data users

- Prohibition of re-identification of data

- Traceability, obligations, legal and technical safeguards and binding commitments for data users.

ANNEX 2. Public Metadata Catalogue (All Tiers)

Dataset General Information

The following information is deemed mandatory for all types of datasets to be included in the EUCAIM metadata catalogue:

- **Dataset Identifier:** A unique identifier for the dataset. e.g. the URI or other unique identifier in the context of the Catalogue.
- **Dataset Name/Title:** A clear and concise name/title for the dataset.
- **Dataset Description:** A detailed description of the dataset's content, purpose, and scope.
- **Dataset Collection Method:** This attribute defines the scope of data aggregation within the dataset. It specifies how data records are organised based on different criteria, allowing users to understand the context in which the data was collected. Possible values (Each dataset can have more than one value):
    - Patient-based: Data records are organised individually based on patients. Each data entry corresponds to a single patient's information, not necessarily specific to a clinical use case.
    - Cohort: Data records are grouped according to specific medical studies or research projects. This grouping includes all relevant data elements such as imaging scans, clinical assessments, lab results, etc., related to a particular study or clinical use case.
    - Only-Image: Data elements in this category exclusively consist of imaging data and associated metadata. Clinical information is not included; only metadata present in the Digital Imaging and Communications in Medicine (DICOM) headers is provided.
    - Longitudinal: Data elements are structured to cover multiple time points for either a particular patient or study. This structure enables the analysis of changes over time, making it suitable for longitudinal studies.
    - Case-control: Data records are divided into two distinct groups: cases and controls. Cases encompass subjects with the disease or condition under study, while controls include subjects who do not have the disease or condition.
    - Disease-specific: Data records are gathered from subjects who have already developed a particular disease. This category is particularly focused on subjects with the specified condition.
- **Dataset Type:** The categorization of the dataset. Possible values include:
    - Original Dataset: The unmodified, raw data without any additional processing or labeling.
    - Annotated Dataset: A dataset where specific regions, features, or characteristics (e.g., tumors, lesions) in the imaging data are labeled or marked, often by experts like radiologists.
    - Processed Dataset: The dataset derived from the original data after applying transformations such as normalization, noise reduction, resizing, or feature extraction to make it suitable for analysis or model training.
- **Dataset Access Rights:** Information that indicates whether the Dataset is publicly accessible, has access restrictions or is not public. Possible values are:

- Public[37]: Publicly accessible by everyone. Usage note: Permissible obstacles include registration and request for API keys, as long as anyone can request such registration and/or API keys.
- Restricted[38]: Only available under certain conditions. Usage note: This category may include resources that require payment, resources shared under non-disclosure agreements, resources for which the publisher or owner has not yet decided if they can be publicly released.
- Non-public[39]: Not publicly accessible for privacy, security or other reasons. Usage note: This category may include resources that contain sensitive or personal information.

Please note:

1. According to Article 41 of the EHDS Regulation proposal, "*data holders of **non-personal electronic health data**[40] shall ensure access to data through trusted open databases to ensure unrestricted access for all users and data storage and preservation*". For non-personal electronic health data, it is mandatory that the property “Access Rights” takes the value Public.
2. In instances where the dataset is categorised as **personal electronic health data**[41], it is mandatory that the property Access Rights takes the value “Non-public” of the Access Rights Named Authority List. Data holders are further obliged to declare that the dataset contains personal data. Data holders are also further encouraged to detail the sensitive nature of the dataset. Data holders are also further encouraged to detail the sensitive nature of the dataset by providing a list of key elements that represent an individual in the dataset (e.g. Age, Birth Date, Drug Test Result, Ethnicity, Family Health History etc. For a list of possible values please refer to Data Privacy Vocabulary (DPV) Specification "Extended Personal Data categories for DPV" (DPV-PD)[42]).

- **Dataset Access Conditions:** A statement that concerns all rights regarding the use of the dataset.
  - Authorisation to download the datasets
  - Authorisation to access, view and process in-situ the datasets
  - Authorisation to remotely process the datasets without the ability to access and visualise data, even remotely.
- **Dataset Publisher:** An entity (organisation) responsible for making the Dataset available. Note: In case the dataset is transferred to one of EUCAIM reference nodes EUCAIM will be the publisher. If the dataset remains at local premises, the entity that retains responsibility for maintaining or owning the dataset should be considered the publisher of the dataset.

---

[37] http://publications.europa.eu/resource/authority/access-right/PUBLIC
[38] http://publications.europa.eu/resource/authority/access-right/RESTRICTED
[39] http://publications.europa.eu/resource/authority/access-right/NON_PUBLIC
[40] **Non-personal electronic health data** means electronic health data other than personal electronic health data, encompassing both data that has been anonymised so that it no longer relates to an identified or identifiable natural person and data that has never related to a data subject.(https://healthdcat-ap.github.io/#nonpersonalelectronichealthdata)
[41] **Personal electronic health data** means data concerning health and genetic data as defined in Article 4, points (13) and (15), of Regulation (EU) 2016/679, processed in an electronic form. [EUR-Lex - 32016R0679 (Art.4(13)(15))]
[42] https://w3c.github.io/dpv/2.0/pd/#dpv-classes

    - Dataset publisher type: A type of organisation that makes the Dataset available. The possible values are: Research Institute, Hospital of Healthcare System, Repository, European project, Cancer screening program, Patient association, Data altruism organization, ERIC and EDIC.
- **Dataset Contact Point:** Contact information that can be used for sending comments about the Dataset. Contact information is limited to the contact email and/or the contact page. At least one of the two MUST be provided. In the case of the datasets transferred to the reference nodes, a person from the Data Access Committee could be designated as the contact point.
- **Dataset Keywords:** A keyword(s) or tag(s) describing the Dataset. This attribute is mandatory in case the dataset contains sensitive health data (optional for protected or open health data).
- **Dataset Geographical Coverage:** A geographic region(s) that is covered by the Dataset.(e.g. Country name)
- **Dataset Applicable Legislation:** The legislation that mandates the creation or management of the Dataset. The value must include the ELI (European Legislation Identifier) of the EHDS Regulation. As multiple legislations may apply to the resource the maximum cardinality is not limited.
- **Provenance:** A statement about the lineage of a Dataset, including information about how the data was collected, methodologies, tools, and protocols used.
- **Sample:** At least one sample Distribution of the dataset should be available. This rule ensures meaningful use and interpretation of non-public datasets as described in the HealthDCAT-AP specification. These samples could be a synthetic subset or representative examples of the dataset to facilitate evaluation and understanding or even solely exhibit the dataset's structure, i.e. human-readable structural metadata providing the properties or columns of the dataset schema. A sample distribution of the dataset is mandatory in case the dataset's access rights is "Authorization to remotely process the datasets without the ability to access and visualise data, even remotely." Providing such a sample as a downloadable file can offer insights into the data's format, structure and set of values, aiding in understanding and utilisation while ensuring privacy and security.
- **Dataset Intended Purpose:** A free text statement of the purpose of the processing of data or personal data.
- **Dataset Legal Basis:** The legal basis used to justify processing of personal data or use of technology in accordance with a law.
- **Dataset Version:** The version for the dataset either using version numbers (e.g. v1.0,) or using calendar versioning - the date of dataset release (e.g. 2023-11-07) .
- **Number of Subjects:** Total count of unique individuals in the dataset.
- **Number of DICOM Studies:** Total count of DICOM studies.
- **Number of DICOM Series*:** Total count of DICOM series within the dataset. This is not a mandatory element, but rather a recommended one.
- **Dataset Distribution:** An available Distribution for the Dataset, which is a physical embodiment of the Dataset in a particular format. When a health Dataset is categorised as personal electronic health data, in accordance with the Data Governance Act - National Single Information Points [NSIP] requirements[43], datasets

[43] https://data.europa.eu/sites/default/files/course/v1.2_ESAP_Technical%20recommendations%20for%20member%20states_Harvesting%20guidelines.pdf

MUST include at least one distribution with the following essential properties:

- Access URL: A URL that gives access to a Distribution of the Dataset. The access URL may contain information about how to get the Dataset.
- Applicable Legislation: The legislation that mandates the creation or management of the Distribution. This MUST be the ELI URI of the EHDS Regulation proposal.
- Byte Size: The size of the distribution (imaging data) in gigabytes. The size can be approximated.
- Format:
  - Imaging data: the imaging format of the images in your dataset (e.g. DICOM, Nifti),
  - Annotation data: the format of the annotations (e.g. DICOM SEG, Nifti), if available.
  - Clinical data: the format of the available clinical data (e.g. csv, xls, json, parquet).
- Compression format (if applicable): The format of the file in which the data is contained in a compressed form. (e.g. .zip file containing the images and clinical data)
- Packaging format (if applicable): The format of the file in which the data files are grouped together, e.g. to enable a set of related files to be downloaded together.
- Rights: A statement that specifies rights associated with the Distribution.

Note: When a health Dataset is categorised as non-personal electronic health data, implementers MUST provide descriptions for, at least, one distribution of the dataset according to Article 41 of the EHDS Regulation.

Clinical and Imaging Information

The following information is mandatory for all datasets. In case of an “Image-Only” dataset, “Topography” and “Diagnosis” are not mandatory:

- **Age Low:** The minimum age of subjects in the dataset.
- **Age High:** The maximum age of subjects in the dataset.
- **Age Median:** The median age of subjects in the dataset (if available).
- **Sex:** The set of distinct sex values of subjects (i.e. sex assigned at birth) in the dataset (e.g. male, female).
- **Topography:** Anatomical sites specified using ICD-O3, SNOMED CT (e.g. breast).
- **Condition:** Diagnostic information of the subjects in the dataset using ICD-10, SNOMED CT (e.g. Malignant neoplasm of breast).
- **Image Modality:** The imaging modality used (e.g., DICOM tag (0008,0060)).
- **Image Body Part:** Anatomical areas captured in the images using DICOM.
- **Image Equipment manufacturer:** Manufacturer of the imaging device (DICOM tag (0008,0070)).
- **Image Creation Year(s):** A year range that the actual (DICOM) images were created/acquired (if this has not been changed in the anonymization process). If this is not available, an estimation should be added if possible.

* Recommended but not mandatory.

ANNEX 3. Minimum set of clinical and imaging attributes

Minimum imaging attributes from DICOM metadata*:

Table 4. Minimum imaging attributes.

VARIABLE	EXPLANATION	CLASSIFICATION	EXAMPLES
Patient ID	DICOM tag : 0010,0020	Mandatory	X123456
Image modality	DICOM tag : 0008,0060	Mandatory	CT
Image body part	DICOM tag : 0018,0015	Mandatory	Chest
Image manufacturer	DICOM tag : 0008,0070	Mandatory	Siemens
Date of image acquisition (YYYYMMDD)	DICOM tag : 0008,0022	Mandatory	20240101

* If images are in NIFTI Format, these metadata must be supplied in DICOM JSON format

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

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

Minimum clinical attributes:

- **Positive or diagnostic cases**

Table 5. Minimum clinical attributes for positive or diagnostic cases.

VARIABLE	EXPLANATION	CLASSIFICATION	EXAMPLES
Patient ID	A unique identifier for the patient. This should match the patient ID DICOM tag (0010,0020) and the anonymization processes	Mandatory	X123456
Population	Categorization of the subjects in the dataset based on their status.	Mandatory	Patient with Cancer; Patient with lesion not being a malignant tumor.
Sex	Biological sex at Birth	Mandatory	Female, Male, Unspecified
Date of radiology	Date when the tumor or	Mandatory if	January 1, 2024

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

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

the dataset purposes.

- **Negative screening and control groups**

Table 6. Minimum clinical attributes for negative screening or control groups.

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

Table 7. Minimum annotation metadata.

Name	Description	Level	DICOM tag	Type	Example
Segment number	Unique identification number of the segment	Imaging	Segment number (0062, 0004)	Mandatory	1,2,...
Segment label	User-defined label identifying the	Imaging/ Dataset*	Segment Label (0062, 0005)	Mandatory	“PZ (peripheral zone of prostate)”, or “CZ (central zone of

	segment.				prostate)”
Segment description	User-defined or ontology-defined description for the segment.	Imaging/ Dataset*	Segment Description (0062, 0006)	Mandatory. In the case the segmentations are made in the context of EUCAIM, the Segment Description should have specific terms from the ontology.	“Prostate Central Zone”, or “Prostate Peripheral Zone”
Segmentation method	Type of algorithm used to generate the segment	Imaging/ Dataset*	Segment Algorithm Type (0062, 0008)	Mandatory	Manual, semiautomatic, automatic
Algorithm name	The name(s) and version of the algorithm(s) used to generate the segment.	Imaging/ Dataset*	Segment Algorithm Name (0062, 0009)	Mandatory if Segment algorithm type (0062, 0008) is semiautomatic.	“Prostate segmentation Tool v1.0.0”
Number of annotators	Number of annotators involved in the annotation process	Dataset		Mandatory	1,2,...
Annotator type	List with the specific role(s) of the expert annotator(s).	Dataset		Mandatory	Radiologist, imaging technician, etc.
Experience	List with the years of experience of the annotator(s)	Dataset		Mandatory	1,2,3,5,10…
Sequence(s) used for segmentation	Modality(s)/ Submodality(s) used to perform the segmentation	Dataset		Mandatory	T2w, ADC, CT, CT+PET

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

ANNEX 4. Data elements documentation (Tier 1 & 2)

Detailed documentation on data elements, including at least the structure, format and the description of each variable, must be provided to facilitate the understanding of the origin and structure of the provided data. This documentation will include information about:

- Any standards used, such as standard CDM adopted, ontologies/terminologies selected, or if it was necessary to adopt a custom implementation, the description of the data model should be provided.
- Details on data elements/variables, including at least the variable names, data types and the description of each variable (if a standard CDM is not adopted).

If available, an exploratory data analysis (EDA) of the data will be highly appreciated, with visualisations, descriptive statistics, and key findings.

Additionally, it is highly recommended for each data provision related to datasets elaborated in the context of a research project to include a complete and concise project description, encompassing key results and findings, including any insights or discoveries and information about any academic or non-academic publications resulting from the investigation.

ANNEX 5. Data Quality (All tiers)

Data quality is defined as the measurement of how well a dataset serves its intended purpose. However, providing a universal definition for data quality is challenging since it is inherently tied to the specific characteristics of the dataset being evaluated. In general, data quality is measured by comparing the current state of the dataset to a desired or ideal state. Although there is no universally accepted framework for defining data quality dimensions, the literature offers extensive insights into commonly used metrics and their interpretations. The selection of dimensions for evaluation depends on the nature of the data and the specific problem being addressed.

For this specific health data repository, with the purpose to serve in the development or validation of various AI models focusing on cancer imaging (but not built specifically for the generation of a unique AI model with very narrow requirements), the following dimensions were identified as the most relevant after analyzing the data's characteristics: **Completeness, Uniqueness, Validity, Consistency, Accuracy, Integrity**.

As a result, it was established that the data provided by DHs should follow these rules:

1. **Accuracy:** Data must be following the announced structure, and must contain the same number of files as announced in the catalogue.
2. **Completeness**: All mandatory data and metadata must be present in the dataset.
3. **Uniqueness**: Data must be unique within the dataset and across all datasets from the DH.
4. **Validity**: Imaging and clinical data must follow the minimum requirements for the Tier they belong to. For all tiers, imaging and clinical data must be properly de-identified. Imaging data must be provided in DICOM format, and their metadata following the DICOM standard. Only specific cases of imaging data only available in NIFTi format will be accepted, and the associated metadata following DICOM standard must be provided as well.
5. **Integrity**: Imaging data files can't be corrupted
6. **Consistency**: for all datasets, information between imaging and clinical datasets must be consistent; also, clinical data can't contain inconsistencies (e.g outlier values, unknown data type, etc.)

Specific (non mandatory) tools will be made available for DHs to assess the quality of their data, and curate their data.

Data quality metrics may be defined as part of the EUCAIM data quality utility label to further help classify the data and datasets according to their level of quality (low quality, medium quality, high quality). Those levels shall be defined as part of EUCAIM alignment with the QUANTUM initiative. QUANTUM consortium partners are currently working on **the specification of the datasets' quality and utility label,** which should soon be made available as part of a deliverable of the project.

ANNEX 6. FAIR compliance (Depending on the tier)

The FAIR principles[44] define a set of guiding rules and best practices that enable both machines and humans to find, access, interoperate and re-use data and metadata. However, in the context of medical imaging research in general, and of EUCAIM in particular, it would not be realistic to expect a strict adoption of the 41 indicators specified by the Research Data Alliance (RDA) Maturity Model Specification and guidelines[45] by all the DHs and RCs given their diverse circumstances and backgrounds as explored above, and the sensitive nature of the data.

In the spirit of the Tiered classification of datasets, different targets will be set for data FAIRness for the different Tiers, so a tailored evaluation process is needed to take into account EUCAIM's requirements. Given the sensitive nature of the data a full compliance with all RDA indicators is not to be asked even for Tier 3. FAIR compliance indicators are not all equally important, and in the aforementioned Maturity Model a classification of their indicators into: essential, important and useful can be found. This will be taken into account when defining the FAIR compliance requirements for each Tier, which are cumulative, i.e. the higher Tiers must also comply with the requirements of the lower ones. We identify the requirements using the RDA indicator's codes.

Tier 1

Findability is the most basic quality that can be expected from a dataset; thus for the entry level Tier 1 the presence of persistent identifiers, that is a must for such findability, constitute the first two requirements, while we also set requirements to the metadata availability in order to guarantee a minimal functionality:

- Metadata is identified by a persistent identifier (RDA-F1-01M).
- Data is identified by a persistent identifier (RDA-F1-01D), that is already expected for the entry to be added to the Metadata Catalogue (Dataset Identifier).
- Metadata is offered in such a way that it can be harvested and indexed (RDA-F4-01M).
- Metadata includes information about the licence under which the data can be reused (RDA-F1.1-01M).

Tier 2

Findability and accessibility to the metadata are considered essential characteristics to enable Tier 2 access, so more thorough checks must be applied for those characteristics at this level. In particular, on top of the Findability attributes asked for Tier 1, DHs would need to comply with the following:

- Metadata is identified by a globally unique persistent identifier (RDA-F1-02M).
- Data is identified by a globally unique persistent identifier (RDA-F1-02D).
- Rich metadata is provided to allow discovery (RDA-F2-01M).
- Metadata includes the identifier for the data (RDA-F3-01M).

[44] Wilkinson, M., DumonTier, M., Aalbersberg, I., & et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

[45] FAIR Data Maturity Model Working Group. (2020). FAIR Data Maturity Model. Specification and Guidelines (1.0). Zenodo. https://doi.org/10.15497/rda00050

- Metadata identifier resolves to a metadata record (RDA-A1-03M).
- Metadata is accessed through standardised protocol (RDA-A1-04M).

On top of this, DHs will provide, through the EUCAIM hyper-ontology/metadata catalogue, information that will allow users to localise datasets with data that would be relevant to their research questions using the 27 mandatory attributes defined for the EUCAIM metadata catalogue. Given that providing metadata according to EUCAIM's hyper-ontology is a must for supporting the federated query functionality, the presence and validity of those attributes is necessary for a positive evaluation.

Tier 3

As aforementioned, FAIR indicators can be classified by different priorities (essential, important and useful). So while for Tier 3 it can not be expected full compliance of all indicators, higher expectations will be placed for the indicators that are considered essential by the RDA. In this level of compliance indicators that evaluate FAIR for data and not just metadata will be taken into account. On top of the ones listed for Tier 1 and 2, the following are considered essential:

- Data identifier resolves to a digital object (RDA-A1-03D).
- Data is accessible through standardised protocol (RDA-A1-04D).

ANNEX 7. Federated Query (Tier 2 & 3)

Executing federated queries requires the local operation of a lightweight "Mediator" component, which performs the following tasks:

- Connects to the central infrastructure.
- Translates the search query:
    - To the site's Structured Query Language (SQL) for sites providing CDM-compliant data.
    - To Clinical Query Language (CQL) for sites providing FHIR-compliant data.
- Aggregates the results and optionally obfuscates them.
- Returns the aggregated results to the central components.

In Tier 2, **the mandatory attributes outlined in Annex 3 must be accessible for the federated query.** There are two options for ensuring the minimum attributes can be queryable:

1. **Use a Mapping Component**:
    - If the dataset does not adhere to the standardized EUCAIM CDM and does not follow the hyper-ontology, a mapping component must be implemented, at least for the requested minimum set of clinical and imaging attributes.
    - This component performs the necessary mappings to the minimum hyper-ontology concepts for the federated query.
2. **Direct Transformation:**
    - If the dataset undergoes a direct transformation to meet the EUCAIM CDM structure and hyper-ontology concepts, or at least for the requested minimum set of clinical and imaging attributes, the mapping component is not needed, as the federated query can be directly executed through the mediator.

Further details on the mandatory query criteria are provided in D5.6.

ANNEX 8. Imaging Dataset Structure/Hierarchy and Series Identification/Tagging (Tier 3)

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

- **Data Shape and Structure**: The imaging data should be organized according to the designated hierarchical folder structure for tier-3 compliance (show *Figure 4*). Ensuring data is properly structured prevents compatibility issues ensuring that each software/tool can work with any EUCAIM dataset as input.
- **Annotations Managing**: When annotations are included on a Tier 3 dataset, i.e. a Tier 3 A+ dataset, they must be in DICOM format as EUCAIM software is prepared to use DICOM images as input[46] and must be added as an additional serie in the same series hierarchy level (also shown on *Figure 4*).

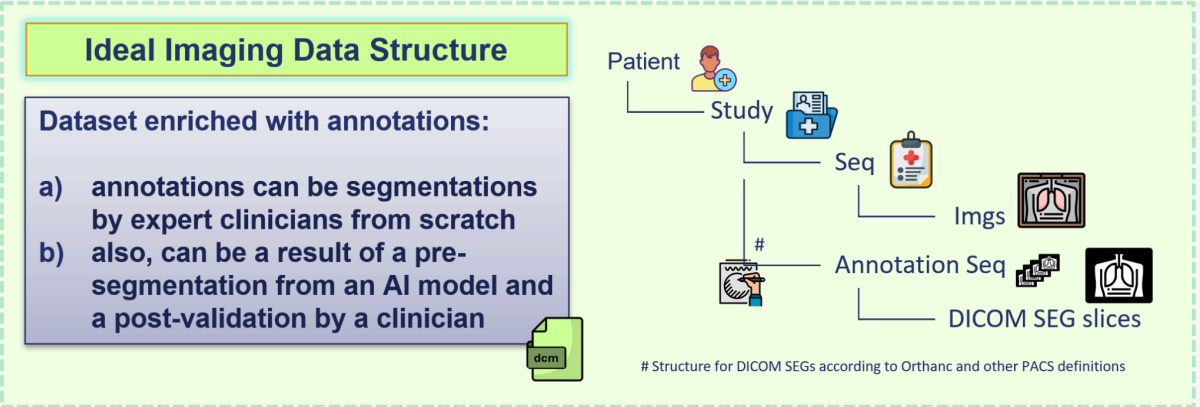

Figure 4. Ideal Imaging Data Structure in Tier 3 allowing federated processing capabilities enriched with annotations.

- **Series Identification/Tagging**: The identification of relevant series is of vital importance not only for their quick and efficient visualization but also for their automatic selection for processing. Medical imaging studies often include a mix of series, some of which are not relevant for secondary use, while others—though important—are not easily identifiable due to non-standardized naming conventions. These names are often written in the local language of the image's origin and may vary across protocols (e.g., the SeriesDescription DICOM tag).

  In Tier 3 datasets, it is particularly important to ensure that these names are normalized, enabling tools/softwares to autonomously identify and process the required series within the federated processing environment. To achieve this, DHs must consult EUCAIM's standardized naming conventions and either replace the SeriesDescription tag in their DICOM files or apply the normalized names to the specific directory paths where these tier 3 files are stored and exposed to federated

[46] EUCAIM software are mainly designed for DICOM format images and the ones initially designed for NIFTI format have been modified including a "DICOM to NIFTI converter" module.

processing services. This ensures that DUs can accurately target the relevant series for their analysis.

For example, if a tool needs to process all T2-weighted (T2W) MRI sequences in a dataset, it should be able to navigate through each patient and study, loading only the relevant T2W series, avoiding duplicates or other non-T2W series. Proper normalization and tagging at the dataset preparation stage are crucial to facilitating this seamless, automated processing capability.

ANNEX 9. Data annotation and labelling (All Tiers)

Data annotation is a critical component of AI biomedical cancer imaging projects. An approach to data annotation is outlined below, with a specific focus on establishing a standard format for storing segmentation masks. Each dataset could be on a specific Tier version from 1 to 3 and, on each Tier, associated manual/semi-automatic annotations give more value to that imaging dataset. Therefore, when annotations are included within the imaging dataset, a label/stamp will accompany it, indicating: **"This dataset is enriched with annotations (A+)✅"**, regardless of the tier (i.e. Tier 1 A+, Tier 2 A+, or Tier 3 A+). This designation enhances the dataset's value and visibility across catalogues.

The segmentation task has been prioritised because it is the most time-consuming for clinical experts, offering significant potential for improvement through the integration of semi-automatic and fully automatic tools into the workflow. Additionally, the standardisation effort required for this task is considerably higher in comparison with others such as detection.

Standardising segmentation annotations with DICOM SEG Format

When available annotations are segmentations, they must be converted to the standard format. The overall standard image format for EUCAIM is DICOM, specifically DICOM SEG for image segmentations, which aligns with the standard practice of hospitals and health centres. Another accepted annotation format is Neuroimaging Informatics Technology Initiative (NIfTI), as long as it contains the original non annotated images in DICOM format, making the conversion from NIfTI to DICOM SEG possible. Additionally, images in DICOM RTSTRUCT format will also be accepted, as this format contains the metadata required to facilitate conversion to the standard format. EUCAIM provides a converter from NIFTI and RTSTRUCT formats to DICOM SEG format which is available for the DHs[47]. EUCAIM encourages DHs to apply this converter or others before transferring or federating the annotation data.

DICOM SEG offers comprehensive and rich header data as well as an standardised approach to exchange information about image segmentations, representing and communicating spatial coordinates, and labelling segmented regions. Some key attributes of the format are:

- **Structured Reporting**: DICOM SEG enables the storage of segmentations alongside essential metadata, ensuring a complete record of annotations.
- **2D and 3D Compatibility**: It supports both 2D and 3D data, accommodating various medical imaging scenarios.
- **Segmented Structure Information**: DICOM SEG includes details about segmented structures, such as labels, colours, and descriptions, providing valuable context.
- **Spatial Mapping**: This format precisely maps segmented regions to the coordinates of the source image, preserving spatial accuracy.
- **Original DICOM Image Reference**: It references the original DICOM image series, ensuring traceability and consistency.

[47] This DICOM SEG converter requires the associated imaging data in DICOM format to extract the necessary metadata.

- **Imaging Modality Agnostic**: DICOM SEG accepts various imaging modalities, fostering interoperability between different medical imaging systems.
- **Additional Information**: It has the capacity to store supplementary data, such as measurements or qualitative assessments related to segmented regions.

Ensuring Compliance

In EUCAIM, two general annotation pathways are considered, depending on where the annotation occurs at a local node or within the Reference Nodes. The process varies slightly in each case:

- **Local Node Annotation**:
  The annotation is performed using in-house software. Besides, MITK Workbench tool[48] will be provided locally to DH as an alternative annotation environment. This viewer includes manual annotation tools and integrates several state-of-the-art automatic AI models (e.g. TotalSegmenator[49], nnUNet[50]) to streamline the annotation procedure. The resulting local segmentations must undergo a quality check to ensure compliance with the segmentation standard format. Additionally, a minimum metadata set will be required to better document the annotation process. This will include radiological details, such as the number of lesions and their locations, as well as annotation-specific DICOM attributes like "Segment Algorithm Type" and "Segment Algorithm Name." The extended list of metadata required can be found in *D5.6. - Minimum Data Federation and Interoperability Framework.*

- **Reference Nodes Annotation**:
  - Quibim DICOM Web Viewer is integrated into the EUCAIM platform as the annotation environment, maintaining a DICOM-in - DICOM-out approach. This viewer comprises a user interface with tools for image manipulation and manual annotation, as well as a backend that provides access to images and metadata, and handles security issues. Additionally, the annotation tools provided by EUCAIM partners will be integrated into the viewer, allowing for automatic annotation of the images. The segmentation outcomes are stored in DICOM SEG format.
  - The Euro-BioImaging Medical Imaging Storage Service provides the XNAT-OHIF viewer for creating annotations on the stored Medical Imaging data. The data needs to be stored as DICOM for this viewer. This viewer offers a range of tools to create Regions-of-Interest (ROI's), both contour and mask based. It can visualize the data using overlays, fractional segmentation mappings and surface meshes. Furthermore it is able to present the clinical data stored in eCRF's in side panel.

[48] https://github.com/MITK/MITK
[49] https://github.com/wasserth/TotalSegmentator
[50] https://github.com/MIC-DKFZ/nnUNet

Annotation storage

Annotations can be generated in manual, semi-automatic or fully automatic processes. In manual and semi-automatic approaches clinical experts are involved in the process to annotate from scratch or refine the AI-based generated annotation. Since in fully automatic scenarios clinical experts are not involved in the annotation procedure, the results require subsequent validation to make sure the outcomes are satisfactory. Therefore, it is worth highlighting that only annotations validated by experts will be stored persistently, since fully automatic ones can be generated on demand. This way a balance among storage to save the annotations and computational resources to generate them is kept. Also, it acknowledges that these automatic AI models are continuously evolving, making it impractical to save outputs that are subject to ongoing improvement, as these outputs could quickly become outdated or obsolete compared to the improved results generated by newer versions of the models.

ANNEX 10. Federated Processing (Tier 3)

Federated Processing in Tier 3 nodes of the EUCAIM infrastructure ensures that distributed data analysis can be performed while respecting data sovereignty and interoperability. This relies on the **Data Materializer Tool (DMT)** as the key component for preparing and managing datasets locally in a standardized format suitable for federated experiments. It ensures that data is accessible yet protected, enabling effective and secure use and analysis by EUCAIM users. The process operates uniformly across different federated nodes while respecting individual infrastructure choices. Key functions include:

- Validating dataset requests and filtering for the local node.
- Interfacing with local Federated Data Node (FDN) services to retrieve and materialize datasets.
- Preparing datasets already according to the EUCAIM CDM requirements, using the configuration file provided by data holders to ensure the correct allocation of datasets to each experiment and their accessibility in the local execution environment (refer to EUCAIM Data Materialization definition v2).

Additionally, the **Federated Execution Manager (FEM)** provides the necessary job orchestration within the broader EUCAIM architecture. Key functions include:

- Retrieving jobs from a central execution queue.
- Running experiments locally in a secure, sandboxed environment.
- Reporting job statuses to EUCAIM's central services and ensuring a seamless integration of federated processing workflows.

Thus, when a federated experiment is initiated, the DMT validates and materializes the requested datasets locally, filtering those relevant to the node and ensuring their compliance with EUCAIM's CDM. Once materialized, the FEM orchestrates the execution of jobs in a secure environment, leveraging local resources to process the data. Results from these local experiments are aggregated and returned to EUCAIM's central infrastructure for further analysis, enabling efficient and secure distributed data analysis.

In conclusion, to enable federated processing, Tier 3 nodes must deploy and configure the DMT to materialize datasets in compliance with the EUCAIM CDM and ensure their accessibility for local execution. Furthermore, the FEM component must be integrated to manage job orchestration and maintain interoperability within the broader architecture. Additionally, the needed infrastructure requirements for Tier 3 federated nodes are described in Section 4.3.2 of this deliverable, and more technical information is detailed in D5.6.