Ref. Ares(2023)8622203 - 15/12/2023

Project title: European Federation for Cancer Images

Project acronym: EUCAIM

Grant Agreement: 101100633

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

D*4.3*: First rules for participation report

Responsible partner(s): HULAFE

Authors: Irene Marín (HULAFE), Ignacio Gómez-Rico (HULAFE), Ana Miguel (HULAFE), Carina Soler (HULAFE), María Toboso (HULAFE), Pedro Mallol (HULAFE), David Vallmanya (HULAFE), Leonor Cerdá (HULAFE), Luis Martí Bonmatí (HULAFE), Ignacio Blanquer (UPV), Ricard Martínez (UV), Valia Kalokyri (FORTH), Tobias Kussel (DKFZ), Kurt Majcen (BBMRI-ERIC), Esther Bron (Health-RI), David Rodriguez Gonzalez (CSIC), Xavier Rafael-Palou (QUIBIM), Alejandro Tejada (QUIBIM), Alejandro Vergara (QUIBIM), Jose Munuera Mora (QUIBIM), Laure Saint-Aubert (MEDEXPRIM), Paris Laras (MAGGIOLI), Marcel Koek (Erasmus MC), Sara Zullino (EATRIS)

Contributors: Ana Blanco (QUIBIM), Maria Jose Alarte (QUIBIM), Ricardo Sánchez (ISCIII), Francisco Soriano (UV), Melanie Sambres (LIMICS)

Reviewers: Monika Hierath (EIBIR), Amelia Suárez (MAT)

Date of delivery: [30/09/2023]

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

DSA = Data Sharing Agreement

DTA = Data Transfer Agreement

DU-R = Data User-Researcher

EHR = Electronic Health Records

EHDS = European Health Data Space

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

OHDSI = Observational Health Data Sciences and Informatics

OMOP = Observational Medical Outcomes Partnership

RC = Research Community

RDA = Research Data Alliance

SQAaaS = Software Quality as a Service

SRAM = Surf Research Access Management

SQL = Structured Query Language

SNOMED CT = Systematized Nomenclature of Medicine-Clinical Terms

TP = Tool Providers

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 tools 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 and tool providers (TP). 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 give to such providers the 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 should 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 provider, 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 should 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).

1.2. EUCAIM: Providers and users

EUCAIM envisions to incorporate 3 different types of providers of data and solutions, as detailed in Deliverable 4.1 (Section 2 – User roles), namely:

1. **Data Holders (DH)** (also referred to as Data Providers or Data Controllers): 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 research purposes. Examples of DHs include data repositories, infrastructures, regional biobanks, clinical centres, cancer screening programs, public entities, pharmaceutical companies and data altruism initiatives among others. The current view is that DHs will contribute with data either by (a) becoming a federated node or (b) transfering de-identified data directly to the Central Repository.

2. **Tool Provider (TP):** entity (e.g. startups, enterprises, research institutions, government agencies, non-profit organisations) that would like to make their already developed data processing tools, 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. An example of a tool provider would be a start-up company that is willing to provide an Artificial Intelligence (AI) explainability platform that helps explain, analyse, and monitor the behaviour of AI models in real-time.

3. **Research Communities (RC):** groups or entities with a common research goal, typically formed through the course of already finalised, currently ongoing or newly emerging projects that would like to make use of EUCAIM’s research environment to continue the research their original project facilitated in the first place. With this, the community taking part in that project (e.g. consortium) will need to agree to transfer the data collected (together with the tools developed through the project lifespan where applicable) to EUCAIM’s Central Repository. In

this manner, work can continue to be done for this project, just in a different environment. The expectation is that the Research Community (RC) will remain connected via EUCAIM and will as a result be able to continue the work done in the scope of such a project via the EUCAIM platform. In return, EUCAIM will include the related datasets in its catalogue, providing the RCs with a secure and highly interoperable environment and enabling them to initiate new projects within the EUCAIM infrastructure, while establishing new collaborations with other partners connected to EUCAIM.

It is worth noting that depending on the type of provider identified, the on-boarding process to join EUCAIM may differ. While the steps for the on-boarding process of a Data Holder have been described in deliverable D2.1, any specifics on the on-boarding of each particular role where applicable have been added in the relevant section.

A EUCAIM end-user, herein referred to as "**Data User-Researcher (DU-R)",** is any person or entity that explores the public catalogue of available metadata and eventually requests access to data and processes them using either the tools available in the platform or their own AI tools. An example of a DU-R can be a Principal Investigator conducting a research project on prostate cancer, with one of its objectives being the prediction of the best treatment allocation based on the analysis of baseline Magnetic Resonance Images (MRI) at the time of diagnosis.

2. Rules for Participation for Data Holders and Research Communities

Facilitating the inclusion of new DHs within the EUCAIM network is one of the main project 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 and RCs, highlighting the specifics of the onboarding process for the latter.

2.1. Minimum requirements in terms of data

To allow the smoothest possible onboarding process for new DHs and RCs and to facilitate as much as possible their participation in EUCAIM, the proposed approach is based on the principle of minimising the requirements for the provision of their datasets, and based on the hypothesis that datasets will mainly come from research projects. DHs and RCs will be required to comply with all legal obligations defined in WP3, including signing of all necessary agreements with EUCAIM. Project metadata and related documentation will be necessary to facilitate the understanding of the origin and structure of the provided data. EUCAIM understands that this minimum set of documentation is necessary considering that these source repositories may not fully or even partially comply with the Data Federation Framework (DFF), whose compliance criteria have been described in D5.1. This DFF encompasses both the federated nodes and the Central Repository, which by itself, acts as another node of the European Federation for Cancer Images (also typically referred to as "The Federation").

EUCAIM understands this approach as a way to provide an environment that streamlines the integration of new DHs and RCs into the Federation, while maintaining the commitment to the highest standards of data quality and alignment with EUCAIM's objectives.

Considering this and taking into account that the adaptation of the data to a common Hyper-ontology may be challenging and a clear stopper, low compliance with the DFF defined by EUCAIM will not be a restriction or an obstacle for DHs and RCs to share their data with the Federation. Instead, EUCAIM will incentivize data curation and facilitate data transformation, if necessary, to align with the DFF, and will seek project funding to carry out these efforts for data contained in the Central Repository.

To avoid blocking the incorporation of new data sources to the platform even if their datasets do not fully comply with the EUCAIM DFF, three different technical tiers of data compliance with the EUCAIM DFF have been established. These tiers will be scalable for DHs and RCs since their datasets will be used for new research projects; and both DHs and RCs will be incentivized to upgrade from one tier to another. The following tier levels have been defined (Figure 1):

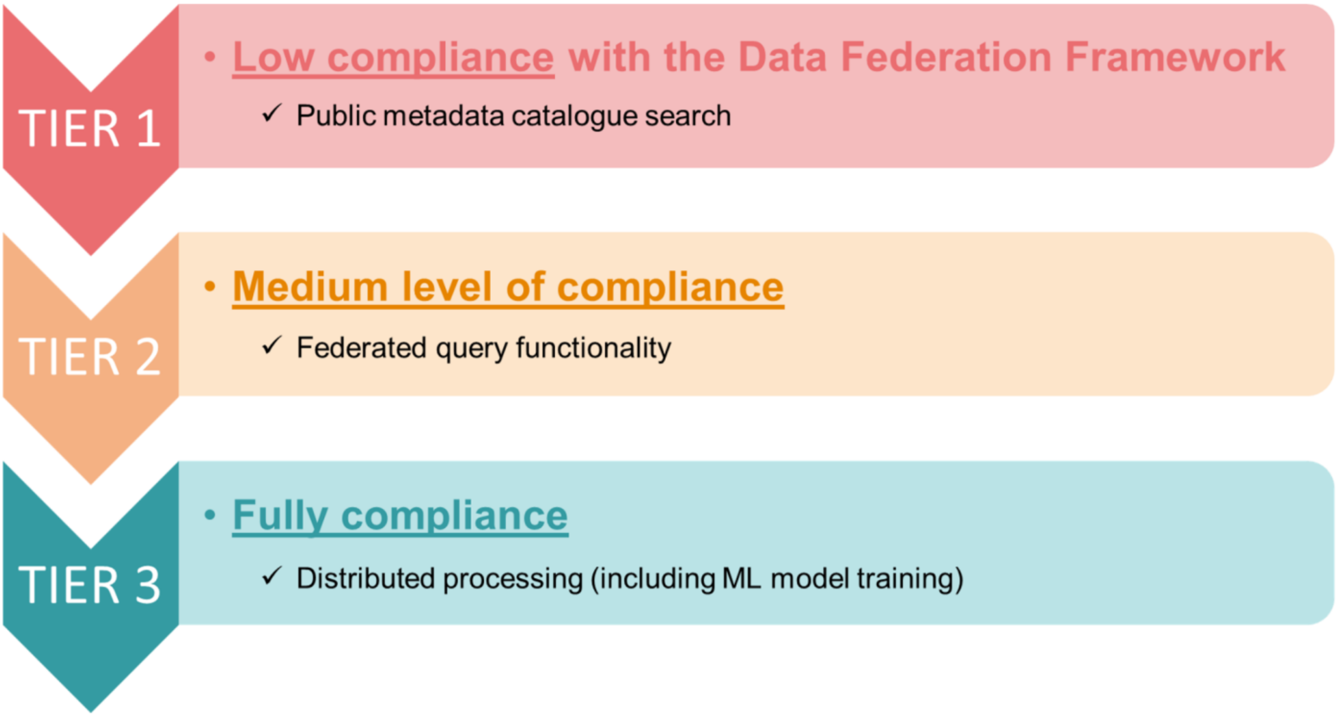
TIER 1
• Low compliance with the Data Federation Framework
✓ Public metadata catalogue search
TIER 2
• Medium level of compliance
✓ Federated query functionality
TIER 3
• Fully compliance
✓ Distributed processing (including ML model training)

Figure 1. EUCAIM Data Federation Framework compliance tiers and key functionality that can be offered at each level.

2.1.1. Tier 1: Low compliance with the Data Federation Framework

Within this first Tier, data will be accepted by the Federation with no additional technical requirements in terms of data compliance with respect to the source repository (mainly linked to an existing research project in the European framework) and, in case of a clinical environment, the data quality requirements established by the clinical centre of origin. For Tier 1 data, the functionalities offered by the EUCAIM platform will be limited: only the publication and visualisation of the dataset in the public metadata catalogue will be possible, allowing basic centralised filtering. For Tier 1 the data is not in compliance with the common data formats (EUCAIM's Hyper-ontology), hence no federated/distributed processing capabilities nor a homogeneous framework for research will be available. In other words, the datasets in the public catalogue will be listed and made accessible (under the defined data request process), and the DU-Rs- will be warned that the data use is under these conditions. Even with these limitations, 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.

There will be a mid-term commitment for adapting the data to EUCAIM's Hyper-ontology. To this end, for example, EUCAIM will encourage DHs to improve their level of compliance by flagging dataset compliance levels accordingly, which will incentivize both DHs and DU-Rs to achieve higher levels of compliance with the EUCAIM DFF, through the funding obtained from new research projects.

At present and in the case of the Central Repository, EUCAIM is responsible for adapting the Tier 1 data to the DFF using project funding. It is currently foreseen that if after a reasonable amount of time no interest is shown in a specific dataset, the long-term storage of such a dataset may be subjected to evaluation by the technical coordination team to ensure efficient use of resources. On the other hand, providers who have joined in a federated manner will be requested to propose an adaptation plan to the DFF, which will be subjected to revision and negotiation by both parties. EUCAIM will facilitate the execution of such adaptation as much as possible so as not to hinder their inclusion in the Federation. This mediation will be facilitated by the training, technical and FAIR implementation support teams established in WP2.

Partner scenarios will differ depending on the environment (research or clinical), as depicted in Figure 2. Based on the requirements for each specific scenario, different agreements will be established between EUCAIM and the DP/RC. The minimum agreements and documentation needs for each scenario are described below. To further illustrate these scenarios, examples of existing projects and their foreseen connection with EUCAIM have been used.

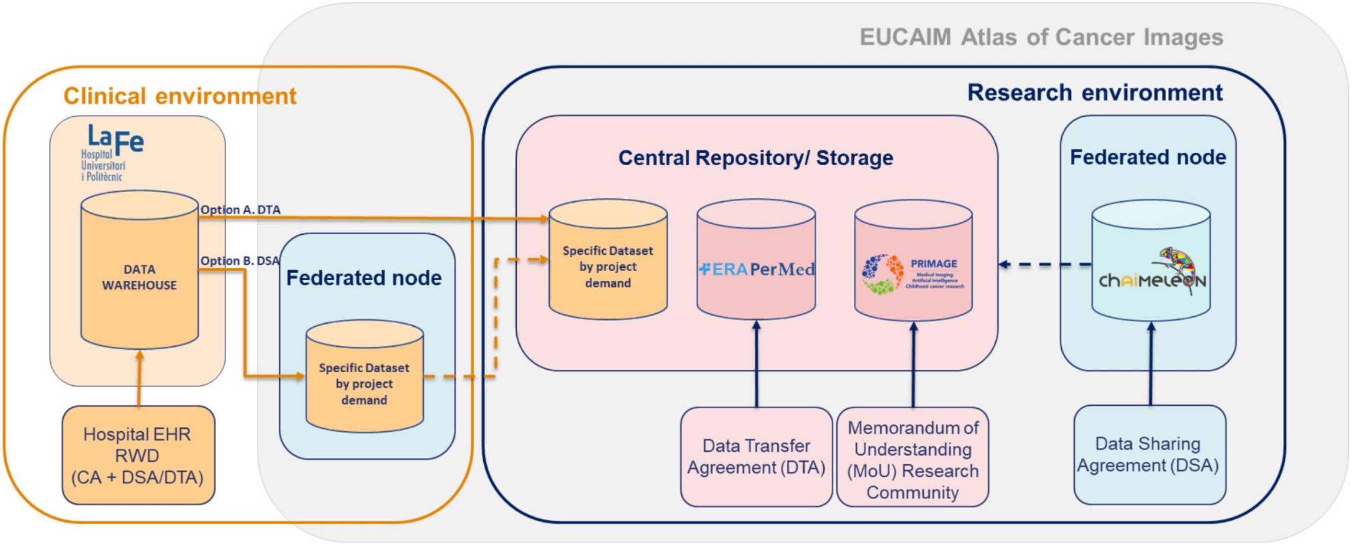
EUCAIM Atlas of Cancer Images
Clinical environment
La Fe Hospital Universitari i Politècnic
Option A. DTA
Option B. DSA
DATA WAREHOUSE
Federated node
Specific Dataset by project demand
Hospital EHR RWD (CA + DSA/DTA)
Research environment
Central Repository/ Storage
Specific Dataset by project demand
ERA PerMed
PRIMAGE
Federated node
chAImeleon
Data Transfer Agreement (DTA)
Memorandum of Understanding (MoU) Research Community
Data Sharing Agreement (DSA)

Figure 2. EUCAIM Data Provision Scenarios.

1. Research environment

a) <u>Central Repository providers:</u>

Finalised research projects without a data sustainability plan that would like to maintain their datasets openly available for research in the long term, but do not have the means to do it. Example: EraPerMed PerProGlio project[1], which was finalised in February 2022 and is currently seeking a way to maintain their data available for other researchers. In this case, the finalised project will directly transfer their data to the EUCAIM Central Repository upon project end. In this scenario, DHs will be asked for:

- The signature of a Data Transfer Agreement (DTA) between parties.
- Information about their research project, metadata catalogue and software.
- Data de-identification

b) <u>Federated providers:</u>

[1] Integrative Personal Omics Profiles in Glioblastoma Recurrence and Therapy Resistance — ERA-LEARN. (2019, March 1). ERA-LEARN. Retrieved September 14, 2023, from https://www.era-learn.eu/network-information/networks/era-permed/1st-joint-transnational-call-for-proposals-2018/integrative-personal-omics-profiles-in-glioblastoma-recurrence-and-therapy-resistance

Existing active repositories that would like to maintain their datasets in a federated node. Example: The Chaimeleon project[2], a 4-year funded project with expected end date in 2024. In this scenario, DHs will be asked for:

- The signature of a Data Sharing Agreement (DSA) between parties.
- Information about their research project, metadata catalogue and software.
- Information about local computational and storage capabilities, for the federated node.

Once the project concludes, it is envisioned that it will be moved to the Central Repository, under the conditions specified above. In the specific example of the Chaimeleon project, being an Artificial Intelligence for Health Imaging (AI4HI) project, it is expected that the project once it concludes becomes an RC.

c) Research Communities:

Finalised, currently ongoing or newly emerging projects that would like to maintain their datasets openly available for research purposes while maintaining alive the collaborative network of researchers that made them possible (i.e. project partners), to continue working with them. To this end, Research Communities will transfer their data to the Central Repository, and also keep the research collaboration (e.g. the partners) alive, with access to their generated results (data and fully operative tools), and even continue to apply for projects as a community with EUCAIM as a partner. An example is the Primage project[3],currently finalised and undergoing this transition. In this scenario, Research Communities will be asked for:

- The signature of a Memorandum of Understanding (MoU) between parties, which encompasses a DTA and a Collaboration Agreement (CA).
- Information about the research project, partners, metadata catalogue and related software.

More details on the MoU, within the specifics on the RCs onboarding process, are described in Section 2.5.

2. Clinical environment

Hospital providers:

This refers to the Real World Data scenario, where partner hospitals, such as the *Hospital Universitari i Politècnic La Fe*, have their own Data Warehouses populated via their Electronic Health Records (EHR). In this context, hospitals will prepare specific datasets for projects in which they have agreed to participate. To do so, they will prepare the specific dataset and decide whether to transfer it to the Central Repository or keep it within a federated node. Potentially, upon project completion, they may choose to transfer the dataset to the Central Repository. In both cases, the following agreements and related documentation will be requested:

- The signature of a DSA/DTA per project
- The signature of a ColA per project

[2] Bonmatí, L. M., Miguel, A., Suárez, A., Aznar, M., Beregi, J. P., Fournier, L., Neri, E., Laghi, A., França, M., Sardanelli, F., Penzkofer, T., Lambin, P., Blanquer, I., Menzel, M. I., Seymour, K., Figueiras, S., Krischak, K., Martínez, R., Mirsky, Y., … Alberich-Bayarri, Á. (2022). Chaimeleon Project: Creation of a Pan-European Repository of Health Imaging Data for the Development of AI-Powered Cancer Management Tools. Front. Oncol., 12, 742701. 10.3389/fonc.2022.742701

[3] Martí-Bonmatí, L., Alberich-Bayarri, Á., & Ladenstein, R. (2020). PRIMAGE project: predictive in silico multiscale analytics to support childhood cancer personalised evaluation empowered by imaging biomarkers. Eur Radiol Exp, 4, 22. https://doi.org/10.1186/s41747-020-00150-9

- Metadata catalogue
- Information about how the hospital data warehouse is structured: Common Data Model (CDM) (Observational Medical Outcomes Partnership (OMOP) /Fast Healthcare Interoperability Resources (FHIR)), High-Performance Computing (HPC) requirements, IT policies.

Below, a more detailed explanation of these minimum requirements is provided.

Data Sharing Agreement

A DSA is a legally binding contract that plays a pivotal role in formalising the sharing of data between entities (in this case between EUCAIM and the DP), defining the terms and conditions of data usage, and ensuring compliance with the General Data Protection Regulation (GDPR). Any DP that wishes to lawfully share specific datasets within the Atlas of Cancer Imaging as a federated node, will need to have this agreement in place.

These DSAs can take various forms depending on the context and the specific needs of the partners but should 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 DP. 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.

Data Transfer Agreement

A DTA is a legally binding contract used to justify the transfer of personal or sensitive data from one entity or jurisdiction to another. This agreement primarily focuses on the secure transfer of data, ensuring that data remains adequately protected throughout the process and that all relevant data privacy laws and regulations are adhered to, particularly in the context of international data transfers.

The GDPR further elaborates on the concept of "Data Transfer"[4], which involves the action of moving Personal Data from the Data Controller to a Contracted Processor. In the clinical environment, this would apply to hospital providers that want to transfer datasets from their Data Warehouse to the EUCAIM Central Repository, either because they do not have storage and/or processing resources to be 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 Central Repository 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.

[4] Data Processing Agreement. GDPR compliance. Retrieved September 14, 2023, from https://gdpr.eu/data-processing-agreement/

Collaboration Agreement

ColAs will be established between EUCAIM and both Research Communities and Clinical Providers, fostering symbiotic relationships built upon collaboration, co-governance, and long-term sustainability.

EUCAIM will facilitate the creation of dedicated RCs for EU-funded projects, such as the Primage RC. These communities will have their own dedicated spaces, complete with repositories, tools, and enhanced communication features, ensuring their sustainability and continuity over time.

These EU-funded projects share a common goal of developing cancer imaging and related data repositories, with a focus on sustainability and availability to the RC beyond project completion. Grant Agreements, like in the case of PRIMAGE, may specify that partners must take measures to ensure the utilisation of their project results and guarantee Open Access to the digital research data generated. To meet these objectives, datasets must be stored and shared in repositories designed for sustainability, extending their life cycle beyond the projects themselves.

Consequently, RCs will actively contribute to the Atlas of Cancer Images through the Central Hub. In return, EUCAIM will provide them with a secure and highly interoperable environment, enabling RCs to initiate new projects within the EUCAIM infrastructure. This strengthens the platform's role as an enabler for research and collaboration, facilitating the integration of diverse types of (meta)data and tools into the Central Hub.

Thus, EUCAIM will give the partners the opportunity to continue working with their data and tools, and they will be able to continue applying for projects with EUCAIM as a partner. In return, their data and tools will be usable by other researchers, upon request.

For this, a transition period will be established, in which the RCs will have a working space, but in exchange they will have to meet certain objectives to publish, apply for more projects, acknowledge EUCAIM in all their scientific publications, etc. proving to be an active community.

In the case of hospital providers, 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.

Project Documentation

Each data provision related to datasets elaborated in the context of a research project should include a complete and concise project description, encompassing its purpose, scope, and overall context, as well as key results and findings, including any insights or discoveries and information about any academic or non-academic publications resulting from the investigation.

If applicable, the documentation should include information about any standards used, as CDM adopted, ontology selected, or if it was necessary to adopt a custom implementation. Detailed documentation on data elements, including at least the variable names, data types and the description of each variable; and data quality, including validation, preprocessing, cleaning, and transformation processes applied to the data, is required. If available, an exploratory data analysis (EDA) of the data will be highly appreciated, with visualisations, descriptive statistics, and key findings.

Documentation demonstrating the appropriate consent to share the data, especially if it contains personal or sensitive information, must be provided. Additionally, it is essential to comply with European data protection regulations, such as the GDPR and clearly specify the licence under which the data is shared.

Finally, it is important to define how data access will be facilitated and establish access restrictions if necessary.

Metadata Catalogue

When exposing medical imaging and clinical information datasets, it is crucial to include a minimum set of common metadata elements to ensure proper findability and to enable understanding its potential for usage. This information will allow users to effectively evaluate whether a dataset aligns with their selection criteria. This importance is also highlighted in the EHDS proposal, which emphasises the need for a metadata catalogue to inform users about available datasets.

The selection of the elements to be included into the EUCAIM metadata catalogue have been the result of a systematic and collaborative process within the EUCAIM project. Initially, a bottom-up approach was adopted, collecting mandatory clinical and imaging information elements defined by the AI4HI projects. This step involved identifying common attributes across these projects. Additionally, insights were gathered from the European Network of Cancer Registries (ENCR) Recommendations document, which provided standardised dataset specifications on capturing essential clinical information. The collaborative effort also leveraged existing European initiatives (e.g. Biobanking and Biomolecular Resources Research Infrastructure - European Research Infrastructure Consortium (BBMRI-ERIC)) biobank metadata catalogue, EIBIR public catalogue metadata) and the IHE Radiology White Paper, incorporating data elements relevant to AI model development. As a result, the following metadata elements have been selected to ensure they meet the project’s objectives and align with best practices in cancer research and data sharing. Note, however, that this list is subject to change as the project progresses.

1. Dataset Creation and 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.
- Dataset Name: A clear and concise name 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:
    - 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
    - Annotated Dataset
    - Processed Dataset
- Dataset Access Rights: The accessibility level of the dataset, indicating how users can obtain and interact with the data. The following values clarify the access methods available:
    - Restricted Access: The dataset is accessible only to authorised individuals or organisations with specific permissions or roles. Users need to meet specific criteria or have approval to access the data.
- Data Access:
    - 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 Terms of Use: The terms and conditions that govern dataset usage. Possible values are based on the Data Use Ontology[5].
- Dataset Intended Purpose: The primary objective for which the dataset was created.
- Dataset Contact Point: Contact information using VCard format for dataset-related inquiries.
- Dataset Metadata Issued: The date when the dataset's metadata was generated.
- Dataset Last Modified: The most recent date when the dataset was updated.
- Dataset Version: The version number or identifier for the dataset.
- Dataset Provider: The entity responsible for providing the dataset.

2. Demographic and Clinical Information

The following information is mandatory for all datasets, except for “Image-Only” datasets. In that case, only the “Topography” and “Diagnosis” are 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: Sex distribution of subjects in the dataset.
- Topography: Anatomical sites specified using ICD-O3, SNOMED CT.
- Diagnosis: Diagnostic information using ICD-O3, SNOMED CT.

3. Image and Modality Information

The following information is deemed mandatory for all datasets:

- Image Modality: The imaging modality used (e.g., DICOM tag (0008,0060), Radlex).
- Image Body Part: Anatomical areas captured in the images using DICOM.
- Image Vendor: 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. This element is highly recommended rather than mandatory.

[5] Data Use Ontology. Retrieved September 14, 2023, from https://www.ebi.ac.uk/ols4/ontologies/duo?tab=classes

4. Dataset Statistics

- Number of Subjects: Total count of unique individuals in the dataset.
- Number of Studies: Total count of DICOM studies.
- Number of Series: Total count of DICOM series within the dataset.
- Image size (GB): The size of the dataset in gigabytes. This is not a mandatory element, but rather a recommended one.

Data de-identification

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 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.

The EUCAIM platform will securely store anonymized data within the Central Repository in accordance with standards defined by the EUCAIM Technical and Legal teams. Specifically, imaging data must undergo anonymization in line with a predefined de-identification profile before being part of the platform. If the data has already been anonymized, EUCAIM will conduct checks to verify its compliance with the established standards. Whenever the data does not meet the specified anonymization standards, the platform will offer a suite of tools designed to facilitate this process and ensure that all data stored is duly anonymized.

Given that the exposure of datasets to the Federation is subject to the generation of research 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.

It is important to emphasise that this possibility could only occur for the scenario in which the data is federated. As indicated above, within the EUCAIM Central Repository, only anonymized data will be available (where the validation process is implicit).

Data access negotiation

Depending on the data accessibility level indicated for the Tier 1 datasets and the DSA/DTA signed, the data access negotiation process, which includes several parties, such as the DU-Rs, the Access Committee, and may involve the DHs and RCs.

EUCAIM uses an access negotiator tool based on the BBMRI-ERIC Negotiator, available as version 2 in the EUCAIM demonstrator corresponding to Meta Milestone 2 (MM2). The Negotiator is undergoing intense modification which will result in version 3 during the year 2023 that allows more flexible use of the tool especially on the side of DHs and RCs. The Negotiator allows users to file requests (including project and request description with research question, inclusion criteria, positive ethics vote from the user's local ethical committee) for data to one or several providers as selected in a previous discovery step in the EUCAIM catalogue. The request can be made to several providers in parallel and the negotiation mechanism allows the Access Committee and providers, when applicable, (a) to obtain more information from the requestor to better understand the reason of the request and the requested data in this broadcast mode, (b) to enter negotiation with the requester, or (c) to step back from a request in case thinking of not being able to fulfil what was requested for whatever reason. Anyhow the requester and the asked DP stay in their setting with no other external users and providers stepping in and having access to this process and the information exchanged. At the point when the requester and a DP, with the Access Committee, have a common understanding that the request can/shall be fulfilled they can enter into bilateral negotiation on the details of transfer and necessary agreements. Of course the requester can step into several such negotiations as needed within the project for which data were requested.

To perform the above described process both the users and the providers need to have access to the Negotiator, which will be integrated in EUCAIM's Dashboard. The Negotiator requires login via Life Sciences Authentication and Authorization Infrastructure (LS AAI) for that purpose. Both the requesters and the members of DHs and RCs in charge of the negotiation process need to authenticate within LS AAI and then get access to the Negotiator tool through the Dashboard.

FAIR compliance

The FAIR principles[6] define a set of guiding rules and 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 strict adoption of the 41 indicators specified by the Research Data Alliance (RDA) Maturity Model Specification and guidelines by all the DHs and RCs given their diverse circumstances and backgrounds as explored above. On top of these "standard" indicators, EUCAIM is defining a set of cancer imaging specific ones in the form of the metadata catalogue.

In the spirit of the tiered classification of datasets, different targets will be set for data FAIRness for the different tiers. Given the sensitive nature of the data a full compliance with all RDA indicators is not to be asked even for Tier 3, so a tailored evaluation process is needed to take into account EUCAIM's requirements.

The FAIR EVA[7] tool (Evaluator, Validator & Advisor) will be used to evaluate the different indicators of their level of compliance. This automatic evaluation tool to monitor FAIR compliance, developed by the H2020 EOSC-Synergy project, not only evaluates the RDA indicators but provides a plugin mechanism that allows

[6] 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

[7] FAIR EVA. Retrieved September 14, 2023, from: https://github.com/EOSC-synergy/FAIR_eva

the project to extend its functionality in order to check the cancer imaging specific indicators defined by EUCAIM. Furthermore, its functionality can be extended with plugins, giving EUCAIM the possibility of adapting the evaluation to the tiers requirements.

FAIR compliance indicators are not all equal, 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 this lower level of FAIR compliance.

Findability is the most basic quality that can be asked from a dataset, and thus for this Tier 1 entry level it will be established requirements related only with this part of the FAIR principles:

- 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).

2.1.2. Tier 2: Compliance with the Federated Query

Compliance with EUCAIM’s Federated Query service is more involved from a DHs perspective but allows for improved visibility and usability of the data. The successful execution of federated queries requires providing (meta)data according to EUCAIM’s Hyper-ontology, and/or operating a local “mediator” service to execute the federated queries and report back the aggregated results.

Datasets at Tier 2 adhere to the standardised EUCAIM CDM, making it easier for researchers to use multiple datasets from Tier 2 in their projects compared to those from Tier 1. As previously discussed, EUCAIM will incentivize DHs and RCs to enhance their data quality and bring it to the next tier, potentially increasing their utilisation in more projects and thus gaining greater recognition within the community.

To facilitate this, by default, if a DU-R requests a dataset from the Central Repository that is at Tier 1, for its use in a research project, EUCAIM will be responsible for adapting the data model to meet Tier 2 standards. This implies that for the first project using that dataset, an adaptation process must occur, at least until Tier 2 compliance is achieved.

On the other hand, federated DP will be required to submit an adaptation plan when joining EUCAIM. Therefore, they will be responsible for adjusting their data model, funded through projects when necessary. In addition, EUCAIM will actively encourage federated DP to upgrade their data to higher compliance levels within their sustainability model. For example, EUCAIM may publish calls for projects specifically aimed at aligning with its sustainability model, offering financial support for the adaptation process.

In summary, in the case of the Central Repository, EUCAIM takes on the responsibility of improving the level of compliance. In the case of federated nodes, this responsibility is shared, and the mediator component is available as an option to reach Level 2 compliance.

The specific additional requirements for Tier 2 are detailed below.

CDM, standardised metadata compliance, data ontology

Compliance with a standardised framework is of paramount importance to ensure the interoperability, reusability, and consistency of the exposed datasets within the project. The adoption of a standardised framework facilitates that the datasets adhere to a common structure and semantics, promoting seamless integration and analysis across diverse data sources.

The project is committed to ensuring compliance with several prominent data standards, including the Observational Health Data Sciences and Informatics (OHDSI) OMOP-CDM and FHIR for exchange purposes, as well as aligning the dataset model with the World Wide Web Consortium (W3C) Data Catalogue Vocabulary (DCAT) model for enhanced discoverability.

Furthermore, the project recognizes the value of the Hyper-ontology as a means to query different local nodes participating in the project's data federation. The Hyper-ontology serves as a shared knowledge representation system that encompasses the domain's entities, attributes, and relationships. By adopting the Hyper-ontology, Data Holders and researchers gain a common language to query and explore the data distributed across various nodes.

1. OHDSI OMOP-CDM and FHIR

As part of the project's dedication to data standardisation, in Tier 2 it is required to transform the datasets to adhere to the OHDSI OMOP-CDM or FHIR standards. The OMOP-CDM standardises the structure and representation of healthcare data, enabling consistent analysis and cross-study comparisons. FHIR, on the other hand, provides a robust framework for sharing healthcare information across disparate systems. By complying with these standards, datasets become compatible with well-established data models used in the healthcare domain, promoting interoperability and facilitating the exchange of valuable data and insights.

2. Hyper-ontology

The Hyper-ontology's role goes beyond just defining data entities and relationships. It also acts as a valuable tool for querying data across the project's local nodes. Researchers and stakeholders can use the shared concepts and relationships defined in the Hyper-ontology to formulate queries that traverse and retrieve information from different nodes. This approach streamlines the process of aggregating information from disparate sources, enabling comprehensive analyses that would be otherwise challenging to perform.

3. W3C DCAT Model

To enhance the visibility and accessibility of the datasets, the project ensures that the dataset model aligns with the W3C DCAT model. DCAT provides a standardised way to describe and catalogue datasets, making it easier for stakeholders to discover and understand the available data resources. By conforming to DCAT, the project aims to promote effective data sharing and collaboration among participants, researchers, and stakeholders. However, since cancer related clinical and imaging information requires granularity in describing attributes such as tumour location, histological type and patient demographics, it has been adopted a more tailored approach to the DCAT model, as this was explained in the sub-section describing the metadata catalogue in Section 2.1.1.

More information on these topics is included in deliverable D5.1.

Mediator component

Executing federated queries requires the local operation of a lightweight “Mediator” component. This component is responsible for connecting to the central infrastructure, translating the search query to the site’s Structured Query Language (SQL) for sites providing OHDSI OMOP-CDM compliant data, Clinical Query Language (CQL) for sites providing FHIR compliant data), aggregating (and optionally obfuscating) the results, and finally returning the aggregated results to the central components. Operating this component requires server resources, allowed outbound internet connections to well-known URLs, IT personnel for site-specific configuration (although assistance will be provided), and IT personnel for ongoing maintenance. The Mediator component is designed to be easy to deploy and easy to maintain, Kubernetes and Docker-Compose based deployment packages will be provided.

DICOM metadata exploitation

The utilisation of DICOM metadata plays a key role in medical imaging and AI-related applications. By extracting and organising DICOM tags effectively, users can efficiently query and assess the suitability of available EUCAIM datasets for their specific needs. In this section, an overview of the key considerations related to DICOM metadata exploitation will be provided, with a particular emphasis on how these align with the objectives of the project.

1. DICOM Tag Extraction

DICOM is the standard for the exchange, storage, and management of medical images and associated data. DICOM files consist of a structured set of metadata attributes, referred to as DICOM tags, which contain essential information about the patient, imaging equipment, acquisition parameters, and image characteristics.

The first step in exploiting DICOM metadata is the extraction of these tags from DICOM files. This process involves parsing the DICOM header to identify and capture specific attributes of interest. Note that the set of DICOM tags that will be exploited depends on the EUCAIM anonymization profile to be employed. Currently, the DICOM tags that have been identified and kept for all AI4HI projects are described in the D5.1. Annex, with the most important DICOM tags for querying datasets to be: the modality, the body part imaged and the imaging equipment used, including the manufacturer.

2. Storage and Organization

Once DICOM tags are extracted, it is imperative to store and organise them effectively to support the objectives of the project. The OMOP radiology extension serves as a specialised framework for this purpose. It allows mapping DICOM tags to a structured, standardised data model, facilitating data integration and analysis. In addition, the FHIR resources for representing DICOM studies and series will also be employed.

Key features of the radiology model extension include:

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

FAIR compliance

Findability and accessibility to the metadata are considered essential characteristics to enable Tier 2 access, so more thorough checks should 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).
- Metadata identifier resolves to a metadata record (RDA-A1-03M).
- Metadata is accessed through standardised protocol (RDA-A1-04M).

On top of this, DHs should 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. 27 attributes are mandatory for the EUCAIM metadata catalogue. To check the presence of these attributes and validate their values an extension (plugin) for FAIR EVA will be developed by subtask 5.2.5. 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.

2.1.3. Tier 3: Fully compliance with the EUCAIM Data Federation Framework

Within the EUCAIM network, achieving compliance with data standards and the DFF across the three proposed technical tiers is an essential goal to ensure the seamless integration of diverse datasets. Tier 3 represents full data compliance, encompassing alignment with a wide set of requirements such as data harmonisation, annotation and quality assessment among others. The ultimate objective for DHs and RCs within EUCAIM is to attain Tier 3 compliance, thus maximising the usability and impact of their datasets within the Federation. The aim at this level is to not only meet the Federated Query capabilities of Tier 2 but also to enable federated processing, including Machine Learning (ML) and other advanced data processing techniques. Thus, this section covers the vision of making all EUCAIM functionalities available for Tier 3 datasets, ensuring that they become a valuable asset in driving innovation and insights within the EUCAIM network.

Once EUCAIM has officially become its own legal entity, it will actively participate as a partner in new research projects, facilitating tools to advance data compliance from one tier to another. The needed resources for this upgrade will come from funded research projects channelled through the EUCAIM ecosystem. Following this funding model, datasets stored in the Central Repository will be curated and transformed by EUCAIM. In the case of federated data, EUCAIM will actively encourage DHs to upgrade their data to higher compliance levels using the same sustainability model. The Technical and FAIR Implementation Support teams established in WP2, as well as the provision of pre-processing tools within the marketplace by EUCAIM, play a pivotal role in this regard, offering guidance, assistance and facilities to DHs and RCs throughout their compliance journey.

It is expected that EUCAIM will act as a key partner in numerous research projects, and this partnership will provide the mechanism to improve compliance of datasets with the DFF. The process will consist of several key steps, which will involve the following technical requirements for the data:

FAIR compliance

As mentioned for the Tier 1, 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:

- Metadata includes the identifier for the data (RDA-F3-01M).

- Data identifier resolves to a digital object (RDA-A1-03D).
- Data is accessible through standardised protocol (RDA-A1-04D).
- Metadata includes information about the licence under which the data can be reused (RDA-F1.1-01M).

Data harmonisation

Effective Data Harmonisation is absolutely crucial to ensure that a project's data fully complies with the EUCAIM DFF. For Tier 3 compliance, the medical imaging data needs to be susceptible of data harmonisation using the tools available in the EUCAIM platform, thus the following requirements are mandatory and must be rigorously followed:

- **Data Format**: Data provided for harmonisation must adhere strictly to the specified format requirements of each data harmonisation tool. For instance, image data must be in DICOM format, while numeric data will be structured as Comma-Separated Values (CSV) files. Compliance with these format standards facilitates seamless integration with data harmonisation tools and ensures data consistency.
- **Data Shape and Structure**: The data's shape must align with the correct number, type, and order of dimensions/fields as expected by the data harmonisation tool. Ensuring data is properly structured prevents compatibility issues and data misalignment, which can hinder the harmonisation process.
- **Modality and Target Alignment**: The modality of the input data (e.g. MRI T1, T2) must correspond accurately to the modality expected by the data harmonisation technique. Furthermore, the input data must resemble the type of organ or disease that the data harmonisation technique is designed for. This level of alignment is critical to optimise the harmonisation process for the precise medical context and organ system under consideration.
- **Sequence Consistency**: The number of sequences in the input data must precisely coincide with the expectations of the data harmonisation technique. This consistency ensures that the harmonisation process can be carried out efficiently and effectively.

For Tiers 1 and 2, while these requirements are recommended, they are not mandatory. However, DHs are strongly encouraged to follow these guidelines. By adhering to these data harmonisation requirements, the effectiveness of the data harmonisation tools can be enhanced and ensure that the data used in the project is well-prepared for harmonisation processes, ultimately contributing to more accurate and meaningful results.

Data annotation and labelling

Data annotation is a critical component of AI biomedical cancer imaging projects, with a primary focus on image segmentation tasks due to their time-intensive nature. An approach to data annotation is outlined below, with a specific focus on establishing a standard format for storing segmentation masks. For Tier 3 compliance, the following requirements are considered mandatory and must be strictly followed. For Tiers 1 and 2, while adherence to these requirements is highly recommended, they are not mandatory. However, DHs are strongly encouraged to ensure compliance with the DICOM SEG format as it significantly enhances the consistency and quality of data annotations within the EUCAIM framework.

1. Standardising with DICOM SEG Format

The chosen standard for segmentation in the EUCAIM platform is the DICOM SEG format. DICOM SEG offers a comprehensive and standardised approach to exchanging 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.
- **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.

2. Ensuring Compliance

It is imperative that the DICOM SEG format is adhered to in all annotation pathways within the EUCAIM framework, whether the annotation takes place in a local node or the Central Repository. The process varies slightly in each case:

- **Local Node Annotation**: the annotation is performed using in-house software and a rigorous evaluation is essential to verify compliance with the defined requirements. This quality check must be executed using the DIQCT tool[8] (from INCISIVE) before storing annotations in the local node or ingestion by the central repository.
- **Central Repository Annotation**: The annotation environment must strictly follow a "DICOM in - DICOM out" approach. This approach guarantees that the output adheres to the DICOM SEG format, obviating the need for an additional quality check.

3. Handling Non-Standard Datasets

For datasets that have been previously annotated outside the EUCAIM framework and are in a non-standard format, such as Neuroimaging Informatics Technology Initiative (NIfTI), a conversion to DICOM SEG will be performed if the original dataset is in DICOM format. Unfortunately, if the original dataset is not in DICOM format, conversion to DICOM SEG will not be feasible.

This meticulous adherence to the DICOM SEG format ensures consistency, interoperability, and the highest standards of quality in the data annotation efforts.

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

Data quality assessment

While using data from multiple sources is a strength as it better reflects the reality of clinical ground, it is often one of the greatest weaknesses when it comes to the quality of the datasets. In order for real-world data from various projects and sites to be exploitable by all end users (researchers, AI developers, etc.), it is crucial that data exposed are of good quality. This applies to both clinical and imaging data, and both at the single data as well as the dataset level.

The following is a general description of some of the aspects that need to be checked to ensure data quality. Although their assessment could be moved to lower levels of compliance, the curation and improvement of data quality corresponds to Tier 3. It should be noted that the rules for participation defined in this deliverable are preliminary and a more granular approach will be provided in future versions as the project definition progresses. Indeed, it would be ideal for WP5 to identify a list of data quality elements and discuss the definition of the tier of compliance at which each item should be checked, as well as how EUCAIM will ensure that each DP fulfils these requirements, but at the time of this deliverable, this has yet to be defined.

1. Data quality assessment

Quality assessment starts with making sure that datasets are complete. While complying to prominent data standards such as OMOP-CDM and FHIR will ensure some level of quality, data, when transformed to adhere to OMOP-CDM/FHIR standards, must remain as complete as possible. Imaging data should be in DICOM format and respect the DICOM standards.

At minimum, all imaging data should go through a visual check to ensure the image is not corrupted, contains the correct body part for the use case it addresses, and has a limited extent of artefacts and noise.

2. Data cleaning, enhancing and integrating

Following data assessment, corrections may be applied to the data to ensure their quality. For clinical variables, any missing information that can be retrieved in clinical or research files should be completed before made available to guarantee a complete dataset. The structuring of data is also crucial, so any unstructured dataset should go through a first structuring set. As for imaging data, sequences that are too noisy or artifacted may go through a cleaning process with any tool available at the site. The names of the tools used to the aim must be specified in the metadata.

3. Time coherence

It must be ensured that examination dates are always present at any time point, both in clinical data and in DICOM metadata. and coherent between each other (eg. a follow-up exam data cannot be anterior to the baseline date).

4. Clinical endpoints definition

It must be ensured that information on clinical endpoint (diagnosis, disease) is always present in the DICOM metadata and clinical data at any time point. The naming will follow the CDM standards.

By adhering to these data quality requirements, the effectiveness of the curation tools can be enhanced and make sure that, wherever the data come from, they reach a level of quality that can guarantee their future processing.

2.2. 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 technical tier of compliance of the datasets with the EUCAIM DFF discussed in the previous section:

a) Authorisation to download the datasets (available from Tier 1 of compliance)
b) Authorisation to access, view and process in-situ the datasets (available from Tier 1 of compliance), both for the federated and centralised repositories
c) Authorisation to remotely process the datasets from a federated node without the ability to access and visualise data, even remotely (available only in Tier 3 of compliance). Instead, DHs will need to provide synthetic or simulated data, ensuring data privacy and security while still allowing for meaningful analysis for DU-Rs.

2.3. Minimum requirements in terms of infrastructure

DHs need to have the appropriate software to upload data to the infrastructure and comply with the minimum requirements explained in this section.

Central Repository

EUCAIM has two different Central Repositories (described in the Technical Architecture Document and in D5.1):

- UPV-Universitat Politècnica de València’s Central Repository: based on the Chaimeleon Cloud Repository technology[9].
- The Medical Imaging Storage service of Euro-BioImaging ERIC: An XNAT (Extensible Neuroimaging Archive Toolkit) instance operated and supported by Health-RI and Erasmus MC, together with upload and ingestion support and a service desk[10].

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

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

- UPV’s Central Repository:
  - A valid EUCAIM user.
  - A standard DICOM client compatible with DICOM for Clinical and Hospital Environments (DCM4CHEE).
  - A REST-based client application for uploading the eForms with the clinical data.

- The 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) (users as described in chapter 4.1.).
  - DICOM receiver of XNAT is accessible via the Clinical Trial Processor[11] to ensure secure encrypted transport of DICOM data.
  - Application Programming Interface (API) is accessible for users to be able to download and upload data (xnatpy is a python library that could be used for this purpose[12].

[9] Chaimeleon D4.1 Initial Repository Deployment. Retrieved September 14, 2023, from: https://doi.org/10.5281/zenodo.7588625

[10] XNATExtensible Imaging Archive Toolkit. Retrieved September 14, 2023, from: https://www.health-ri.nl/services/xnat

[11] Wetting up XNAT. Retrieved September 14, 2023, from: https://www.health-ri.nl/setting-xnat

[12] XNAT Python Client. Retrieved September 14, 2023, from: https://xnat.readthedocs.io/en/latest/

The **non-functional requirements** are:

- Access to a dedicated machine for uploading and administering data in the central 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.

Federated Node

The functional requirements to be met by each DP federated node are as follows:

- Federated Node Infrastructure Procurement:
    - Each organisation must adhere to its local procurement procedures to timely obtain and set up the essential management and technical infrastructure, including Servers, Virtual Machines, and IaaS, necessary to host a federated node.
- Federated Node Processing Requirements:
    - To participate in EUCAIM's federated infrastructure, organisations must procure the infrastructure and hardware that meets the specific processing demands of the federated node. These requirements are categorised into various tiers of node participation, detailed in D5.1, spanning from strict data federation to GPU-aided edge computing.
- Federated Node Storage Requirements:
    - Organisations must procure and set up the storage infrastructure ensuring that every hosted federated node aligns with EUCAIM's data storage specifications. These needs, based on participation tiers, range from storing the organisation's contributed dataset to offering supplementary storage for localised data processing projects; more details can be found in D5.1.
- Federated Node Network Requirements:
    - Each federated node must be connected to the public internet via a wired connection. Relevant network infrastructural adjustments, such as firewall configurations should be made to enable specific network port inbound or outbound access to the public internet.
- Configuration of Federated Node Software:
    - Each organisation must install an operating system compatible with EUCAIm's software stack. It is recommended to install stable Linux distributions such as Ubuntu, CentOS, or Debian. Furthermore, organisations must install the services required to download, deploy and manage the EUCAIM software stack. Guidelines for installation of the required tools and the EUCAIM software stack are available in the online EUCAIM GitHub repository.
- Third parties temporal data transfer:
    - In the event that a DP wishes to federate its datasets (and therefore not move them to the Central Repository), but does not have the necessary computational resources, it may choose to transfer the data to a trusted third party (TTP) of its choice when some processing

needs to be performed on these datasets. The conditions of de-identification and temporality of the data will be agreed between the TTP and the DP.

The **non-functional Requirements** are:

- 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 team.
- Physical Infrastructure Management:
  - All procured physical infrastructure (processing, electrical, or network units), essential for the federated node, should be securely positioned, safeguarded from external hazards and detrimental environments. Adequate precautions must be taken against potential risks like food, liquids, or cleaning agents.
- Physical Infrastructure Access Management:
  - Physical infrastructure for federated node hosting should reside in a restricted access zone, allowing entry only to approved individuals. Any access events must be continually monitored and recorded.
- Federated Node Data Redundancy Configuration:
  - 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. Additionally, a comprehensive backup plan should be established to consistently safeguard the data's latest version.
- Management of Digital Access to Infrastructure:
  - For management of the infrastructure, digital access will have to be provided by the organisation to authorised persons. Authorised technical staff should be provided with user credentials and SSH keys, ensuring encrypted and secure access. Regular audits should be conducted to review access logs and ensure no unauthorised access attempts.
- Monitoring and Maintenance of Federated Node:
  - It is suggested that each organisation monitors the health and performance of their hosted federated node. Automation tools can be implemented to track metrics and set up alerts for anomalies in each metric.

2.4. Legal and Ethical Requirements

A first approach to legal and ethical requirements is presented based on the current regulation integrated by the GDPR, the Data Governance Act, biomedical ethical and legal practices and opinions, statements and or recommendations issued by the European Data Protection Board and/or national data protection authorities. Ethical requirements regarding AI are based on the Assessment List for Trustworthy Artificial Intelligence (ALTAI) developed by the High-Level Expert Group on AI[13] set up by the European

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

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, as specified in the Ethics Guidelines for Trustworthy AI[14,15].

The terms “Data Holder” and “Data Consumer” are being changed from Data Governance Act and future EHDS Regulation on to “Data Holder” and “Data User”.

2.4.1. Legal requirements

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

Documentary requirements

- 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 reports, independent audits, membership of codes of conduct or certification schemes or equivalent documentation.
  - A report issued by the Data Protection Officers regarding the adhesion to EUCAIM scheme.

Additional requirements

- In the case of providing access to pseudonymised data, it must be demonstrated whether exceptions to patient consent operate under national law. If not, it must be able to ensure that it is in a position to prove the existence of adequate consent for the purposes of EUCAIM 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.
  - Future requirements of data based on a permit issued by a Health Data Access Body, and/or ‘HealthData@EU infrastructure, under the EHDS should be considered.
- 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:
  - Ensure that there is a legal basis that allows the anonymisation and use of the data for EUCAIM purposes.
  - That the patient's transparency expectations have been met.

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

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

- That the risk analysis methodology derived from articles 24, 32, 35 and recital (26) of the GDPR has been applied.

Binding Agreements

During the process of accession or Data Sharing Agreement, 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. This may require the formalisation of one of the following documents:

- Joint Controllership agreement.
- Data Processor Agreement.
- Data Sharing or Data Transfer Agreement.
- Agreement or commitment to respect the security obligations in the use of EUCAIM's information systems.

Potential future requirements

EUCAIM might adopt the requirements envisaged in the EHDS Proposal for data holders in its own premises. It should mean that the Data Holders (Data Holders) providing data or access to data should collaborate in order to:

1. Cooperate in good faith with the health data access bodies, where relevant.
2. Communicate to the health data access body a general description of the dataset it holds in accordance with Article 55.
3. Inform about the existence of a data quality and utility label.
4. 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.
5. Shall make available a new dataset where a Data Holder 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.

2.4.2. Ethical requirements

This section provides the ethical requirements for Data Holders (Data Holders) 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.
- Ethically accept as valid the treatments authorised to a data access applicant by the Data Access Committee who will verify the provision of ethical guarantees.
- Inform about their availability to collaborate in ethical verification activities providing support to data access applicants if needed.

2.5. Specifics on the Research Communities onboarding process and organisational requirements

The on-boarding process for RCs is expected to be orchestrated through the creation of a MoU signed by the RC and EUCAIM. A template for such agreement is currently being defined between the leaders of WP8 and WP3 cooperatively together with the coordinators of the AI4HI projects. Even though it is anticipated that such template will be sufficient for the AI4HI projects in the first instance, it is expected that such agreement will need to be tailored for other RCs as they arise.

Data Provisioning Framework

- Data identification and ownership: Each RC provides its own data to create the corresponding project data repository in EUCAIM in the Central Repository, that will ensure that these data will be findable, accessible, sustainable, and reusable.
- Provisioning Model: Data Holders, as members of the RCs, establish the framework for DU-Rs to access their data repository stored on the EUCAIM platform once their project concludes. In any case, Data Holders can independently finalise the agreement with EUCAIM at any time. The MoU details that the project data repository and related tools can be accessed and used by Data Holders for a specified number of months after their project's conclusion, also granting EUCAIM permission to store, process, and use these data repositories for a defined period. Both grants will automatically renew unless one of the parties provides notification.

Access Policy Framework to the project data frame stored at EUCAIM

- Data Sharing model: EUCAIM will release data to DU-Rs for specific research purposes, ensuring full anonymization and compliance with Open Data rules or relevant national/EU legislation.
- EUCAIM's Repository Governance Model: Specifies data access, legitimate data use, and ethical compliance. The model, overseen by the Data Access Committee, notifies Data Holders on the requests and is shared with their Data Protection Officer upon request.

EUCAIM’s Research Community Framework

Data Holders will become EUCAIM RC members, which includes a dedicated space for their Project RC, available at no cost for Data Holders and for a specified duration after their project concludes.

Additionally, they will participate in scientific publications of the community and have access to the available tools associated with each project, if feasible.

Sharing principles and responsibilities

Data Holders grant EUCAIM a licence to use the provided data for research and development purposes while ensuring data protection compliance. EUCAIM will acknowledge contributions and handle data properly in compliance with applicable regulation. In addition, EUCAIM will not sublicense, sell or otherwise transfer the data to third parties without the Data Holders’ prior authorization. The agreement between both parties is in force indefinitely but can be terminated by mutual agreement.

3. Rules for Participation for Tool Providers

This section outlines the essential guidelines that entities aspiring to become EUCAIM Tool Providers should adhere to. Their objective is to make their tools, services, or applications accessible to users, enabling them to engage in federated processing or pre-processing of data sourced from the EUCAIM platform. Tool 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 Data Holders in the EUCAIM network.

In order to promote active engagement from Tool Providers within the EUCAIM ecosystem, smooth requirements have been defined.

3.1. Minimum Requirements in Terms of Tool Deployment

3.1.1. Technical Requirements and Guidelines

All the tools that are to be part of EUCAIM must be provided as containerized images similar to those offered by Docker. A container orchestrator such as Kubernetes is used to make accessible in the central node and any federated nodes. These requirements match with the ones defined in Chaimeleon H2020 Grant Agreement nº 952172[16].

- 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*.
- A dedicated volume will be made available at the Central Hub to store the results of the tool.

3.1.2. Minimum Requirements for Tool Inclusion

- Any tool 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 tool that is to be added to EUCAIM must provide in its documentation information regarding the possible use cases in which the tool 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 tool should not be used, must be clearly specified in this documentation.

3.1.3. User Support and Tool Maintenance

- All the tools that are to be part of EUCAIM must ensure a communication channel that allows users to contact the provider of the tool when any kind of incident or issue arises while using it. This channel shall rely on the EUCAIM Helpdesk.

[16] Instructions to develop applications in Chaimeleon. Retrieved September 14, 2023, from: https://github.com/Chaimeleon-eu/workstation-images#how-to-design-a-workstation-image-for-the-Chaimeleon-platform

- Each Tool Provider must offer long-term support for their tool 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 Tool Providers shall sign a Service Level Agreement ("SLA") in order to guarantee that the software of the tool is up to date and presents no known vulnerabilities.

3.1.4. Minimum Documentation Requirements and Benchmarking Information

To promote transparency and facilitate tool evaluation, TP are required to provide comprehensive documentation and benchmarking information:

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

- 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 tool handles data. It shall comply with EUCAIM's data privacy requirements.

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

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

- Compliance and Certification Documents:
  - Compliance certificates or documentation proving that the tool 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 tool.

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

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

Benchmarking Information

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

- Tool Description: A concise overview of what the tool does and its primary purpose.

- Training and Validation Dataset Description: Details about the dataset used for training the tool, including the number of cases, data distribution, image modalities (if applicable), and any relevant preprocessing steps.
- Tool Type: Specify the type of tool, such as preprocessing, AI model or analysis tool.
- Task: Describe the specific task(s) that the tool is designed to perform, such as segmentation, harmonisation, classification, etc.
- Performance Metrics: List the performance metrics used to evaluate the tool, which may vary based on the task and tool 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 tool.
- Licence Information: Clarify the licensing terms under which the tool is available for use, including any open-source licences or restrictions.
- Hardware Requirements: Detail the CPU and GPU requirements for running the tool effectively.
- RAM Requirements: Specify the amount of RAM (memory) required for optimal tool 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 tool is developed or can be extended.
- Relevant Keywords: Include keywords or tags that describe the tool's focus and functionality, making it easier for users to find relevant tools in search queries.
- Publications: If applicable, list any research papers, articles, or documentation related to the tool's development or validation.
- URL: Provide a link to the tool's official website, repository, or relevant page to access more information.

3.2. Traceability Mechanisms

- The tool shall register all the relevant actions that each of the users perform involving the tool itself.
- Each tool 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.

3.3. Monitoring Capabilities

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

3.4. Quality Control Measures

All the tools 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 tool.
- Functional validation of the tool 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.

3.5. Security and Privacy Compliance

- All the tools 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 tools 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 should 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.

3.6. Legal Compliance

All the tools that may want to join EUCAIM’s infrastructure shall 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, should be complied with - once in force - by TP.

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

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

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

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

- Compliance with GDPR may be evidenced by reports on data protection impact assessments, audits done by third parties or adherence to codes of conduct or certification schemes. The tool provider shall provide proof of such certification whenever it has implemented a certifiable safety standard.

- The location and storage of the data generated by the tool 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 tool in order to ensure data protection by design and by default.

Among the ethical requirements to be fulfilled by TP - whenever applicable - it shall be highlighted that TP will need to provide an AI risk analysis, preferably by using the ALTAI Tool for assessing AI-based technologies. AI-based tools 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 should 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).

3.7. Evaluation and Integration

EUCAIM will evaluate tools to determine their integration into the platform. This evaluation process will ensure that the tools meet the minimum requirements outlined in this section. Once integrated, tools will be assessed by Data Users Researchers 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.

4. Rules for Participation for Data Users-Researchers

Having discussed the rules of participation for all EUCAIM providers, both for data and tools, it only remains to define those of the Data Users-Researchers. These end-users of the platform should be able to explore the public catalogue of available metadata if they are not registered, or perform a federated query on Tier 2-3 datasets once is registered in the platform. If they wish, they can request access to and, if access granted, process the data using the tools available on the platform 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 DU-Rs to have an approved research project, which can range from final undergraduate projects to large funded grants.

4.1. User identity checking procedure

A DU-R needs to register on the platform via the Life Science AAI[17] (several options are available for this, like affiliation to organisations, using an ORCID, an LS Hostel account) and authenticate when using the tools and services of EUCAIM. 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 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:

- 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[18] and the General Terms of Use End Users of Health-RI[19], as well as other resources produced by well known and established distributed data research initiatives and biobanks. The access criteria and procedures from these

[17] Life Science Login. Retrieved September 14, 2023, from: https://lifescience-ri.eu/ls-login/

[18] 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

[19] General Terms of Use End Users. Retrieved September 14, 2023, from: https://www.health-ri.nl/terms-use-health-ri-services

resources cover topics such as eligibility, application process, review process, criteria for access, and terms of access.

4.2. Data access request process

A DU-R may submit a request for access to data from one or more datasets that they have selected using the User's view of the Catalogue, to be used only within a research project. To do so, they must provide the required documentation (e.g. the study protocol) and the explicit approval of an Ethics Committee in the country/countries where the research project is going to be conducted, when applicable. The request process will be orchestrated through the EUCAIM Negotiator tool which will allow the interaction of the Access Committee and the requester. The request will be evaluated by the Access Committee, and, depending on the agreements signed with the Provider that owns the data, it may also be evaluated by the provider's Access Committee (or an equivalent entity). The approval of the request will imply the granting of authorization to use the datasets.

4.3. Request form and specific requirements

- The EUCAIM Negotiator serves as a request form linking DU-Rs with the Access Committee and the DHs and RCs.

- The EUCAIM Negotiator service facilitates and allows administering the process for the data user to request selected dataset(s) from the Access Committee.

- The Access Committee verifies whether the desired analysis to be carried out meets the conditions of use of the requested dataset(s).

- The DU-R files the data request via the EUCAIM Negotiator including information like:
    - Data user identification and authentication
    - Mention of the desired datasets (inclusion criteria)
    - Research question to be answered, with applying the desired analysis
    - Analysis processing environment to be applied
    - Approval of the data user's local Ethics Committee
    - Optional: further specific information as required by the particularly involved Data Holder/s (may be added in a more flexible way with the future Negotiator version and adapted processes)

- Depending on the agreement, Ihe Data Holder(s) check(s), the request on the terms of use applicable to the requested data

- The EUCAIM Negotiator service links the result of the assessment back to the data user, after which necessary contracts can be arranged and/or the release of, or access to, the requested data.

4.4. Legal and Ethical Requirements

EUCAIM DU-Rs, 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.

4.4.1. Legal requirements for data users

EUCAIM might 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)[20] 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:

- 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 should provide:
    - Entity:
        - 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, Data Act or AI Act.
        - 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.
    - Users:
        - The security obligations should 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.

[20] European Commission. (2022). Regulation of the European parliament and of the council—On the European health data space.

4.4.2. Ethical requirements for data users

Data users should be required where needed to:

- Describe the ethical conditions for the data processing.
- Provide a certificate of ethics approval issued by an ethics committee.
- Provide an AI risk analysis preferably by using the ATAI Tool[21].

5. 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.

EUCAIM assure to adopt a compliance framework on GDPR based in this structure (Table 1):

Table 1. 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

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

	• 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, should be communicated to:

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

The ALTAI tool[22] 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[23]. 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 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.

[22] 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

[23] 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

- **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 that include 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 should be applied to the AI application development environment and to requests for access and use of information.

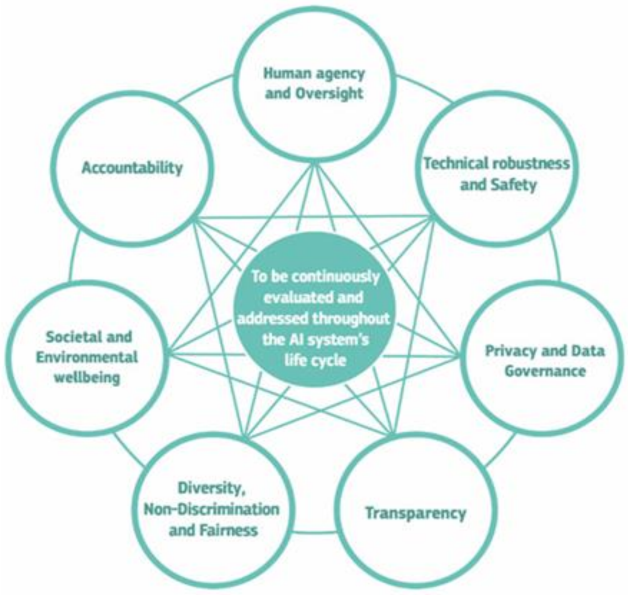
Human agency and Oversight
Technical robustness and Safety
Privacy and Data Governance
Transparency
Diversity, Non-Discrimination and Fairness
Societal and Environmental wellbeing
Accountability
To be continuously evaluated and addressed throughout the AI system's life cycle

Figure 3. Interrelationships between the seven ethical requirements[24].

6. Evaluation of applications and expected response times

The evaluation of applications and the expected response times (in terms of provision of data and tools, as well as data access requests) will undergo a structured process to ensure that participation to the EUCAIM infrastructure is granted based on rigorous scientific, technical, legal and ethical criteria. This process is an ongoing work by the time of this current deliverable and will be set up as per T7.1: Data incorporation use cases (Open Call). The EUCAIM Access Committee and the EUCAIM Management Board will oversee the entire process to maintain transparency and fairness. Comprehensive details of the evaluation procedure, including relevant timelines, will be made available through D7.1: Rules for the Evaluation and Prioritization of Use Case Applications from the Open Call (related to T7.1a), scheduled for release in December 2023 (M12).

7. Conclusions

This deliverable provides a first approach of the rules for participation for the main roles that will join the EUCAIM platform. This has led to the definition of minimum requirements for DHs and RCs 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 DFF defined in D5.1, as well as in terms of access to the data, the necessary infrastructure, and the ethical and legal requirements. A first approach of the minimum requirements to be met by TPs and DU-Rs has also been defined. This document also provides an initial overview of the compliance framework design.

The project consortium hopes this deliverable will set the first grounds on how to join the EUCAIM platform, both for providers and consumers. Looking ahead, it should be anticipated that this initial version of the rules for participation will undergo further updates by the end of project month 24, in the D4.4 Final rules for participation report.

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