MAESTRIA

Deliverable 5.1
Report on the feasibility of aggregating health data from different countries

Date: August 2021

HORIZON 2020 – RIA programme
Digital diagnostics – developing tools for supporting clinical decisions by integrating various diagnostic data

Grant Agreement number:	**965286**
Project acronym:	**MAESTRIA**
Contract start date:	**01/03/2021**
Project website address:	**www.maestria-h2020.com**

Due date of deliverable:	**31/08/2021 / month 06**
Dissemination level:	**Public**

Document properties

Partner responsible	IMT
Author(s)/editor(s)	Luis Pineda, Natalie Cernecka
Version	1

Abstract

MAESTRIA aims at developing digital diagnostic tools and personalized therapies for atrial cardiomyopathy. An essential phase to attain this objective is to aggregate the patients' cardiac data that will serve to train and validate the developed tools. This document presents a summary of legal and technical factors that may impact the feasibility to transfer and process data in MAESTRIA's central data repository, the Data Hub. After presenting the Data Hub and its hosting infrastructure, this report considers legal factors at the European level, ruled by the GDPR; at the French level (where the hub will physically reside), governed by the CNIL; and local regulations at each partner institution. From the technical perspective, the report focuses on three aspects: data diversity, data transfer, and data processing. A data access survey was carried out to understand the partners' concerns or requirements. The results of the survey, which came in support of our analysis, are also described in this report.

Table of Contents

1. Introduction

The MAESTRIA (Machine learning and artificial intelligence for early detection of stroke and atrial fibrillation) project is an 18-partner Research and Innovation action (RIA) with the objective of developing and validating the first integrative diagnostic digital platform for atrial cardiomyopathy diagnosis. This platform will be designed to provide support for improved diagnostic accuracy that increases effectiveness and efficiency of treatments, as well as prevention of the complications of atrial cardiomyopathy, such as atrial fibrillation and stroke.

The WP5 is led by Institut Mines-Télécom (IMT)[1], a French public higher education institution, federating a group of eight engineering and management schools. The objective of WP5 is to coordinate access and sharing of data between partners and to develop strategies of data integration. To this end, the involved partners will develop procedures, methods and computer models to integrate and validate new biological data and then to annotate and homogenise them. WP5 activities will run concurrently with the development of novel algorithms in WP1-3 and the data organised from WP4 as part of an iterative process.

The main outcome of this WP will be the implementation of the MAESTRIA demonstrator, a scalable and user-centric designed digital prototype platform, through the combination of a data hub, novel AI algorithms and an innovative digital atrial twin model to guide atrial cardiomyopathy diagnosis and therapeutic decisions.

1.1. Scope of the document

Most of MAESTRIA's medical data is considered as *personal data*. Moreover, medical data is generally catalogued as *sensitive*. This condition imposes a number of legal and technical constraints to transfer, store, aggregate, and analyse this data.

MAESTRIA's medical data will be consolidated in a central infrastructure, the Data Hub. This document presents a preliminary analysis of legal and technical aspects to take into consideration prior to transferring medical data into MAESTRIA's Data Hub, as well as during the processing of such data.

1.2. Structure of the document

The document is organized as follows: Section 2 describes MAESTRIA's Data Hub, the main data repository of the project. Section 3 summarizes the main regulations

[1] https://www.imt.fr/

applicable to data storage and processing at different scopes. Section 4 considers several technical aspects that apply to data transfer and storage in the context of MAESTRIA. Section 5 presents the results of an initial data access self-assessment submitted by the consortium members. Section 6 presents a roadmap for the following months in order to ensure legal and technical compliance for hosting and processing data in the Data Hub. Section 7 concludes the document.

2. The Data Hub

The Data Hub will be a common data repository for all the partners of MAESTRIA. It will provide a secure storage space, including regular data backups. In addition, the Data Hub will provide *workrooms*, isolated spaces to host partners' applications and algorithms, allowing them to analyse data directly in the Data Hub.

IMT will provide and administer MAESTRIA's Data Hub in its TeraLab[2] platform.

2.1. TeraLab

TeraLab is IMT's secure, sovereign platform for artificial intelligence and big data. It provides technological resources and a whole ecosystem of experts, which allow its users to accelerate experimentation and enable technology transfer.

For each of its hosted projects, TeraLab provisions an environment, called "workspace", which is a dedicated network of virtual servers. This network is customisable in terms of number and dimensioning of the virtual servers, interconnectivity between them, pre-installed software and tools, and security configuration.

TeraLab proposes three levels of autonomy for a workspace, depending on the skills of the user and the sensitivity of the data:

- **Autonomous workspace:** For projects where team members are skilled in system administration and/or the data has no confidentiality restrictions. The principle is to delegate a significant portion of the administration activity to these project members.

- **Administered workspace:** An intermediate offer targeted for projects with insufficient system administration skills. Operations and securing the workspace are fully performed by the TeraLab team. Users legally commit to

[2] https://www.teralab-datascience.fr/?lang=en

strictly follow defined procedures when extracting and exporting data from the workspace.

- **Highly secured workspace:** For projects requiring a high level of data security and control over data transfers. Beyond the legal commitment made by the users, TeraLab's experts implement means to technically prevent and trace the extractions of data.

Projects in the category of "Personal Health Data", such as MAESTRIA's Data Hub, are consistently allocated a Highly secured workspace.

TeraLab Data servers are physically located in metropolitan France; consequently, they are governed by both European and French regulations in terms of data privacy.

2.2. Data Hub Specification

The Data Hub will be hosted in a Highly secured workspace implemented by TeraLab. It will aggregate data from different MAESTRIA data providers, and enable controlled access for data users. Whenever possible, virtual servers, called *workrooms*, will be proposed to allow users to carry out their data processing directly on the Data Hub.

A generic schema of the Data Hub is provided in Figure 2.1. From a high-level perspective, two main profiles will interact with the Data Hub: data *providers* and data *users*. The first will have access to a secure channel to transfer their data into the Hub; for the latter, two scenarios are possible, as described below.

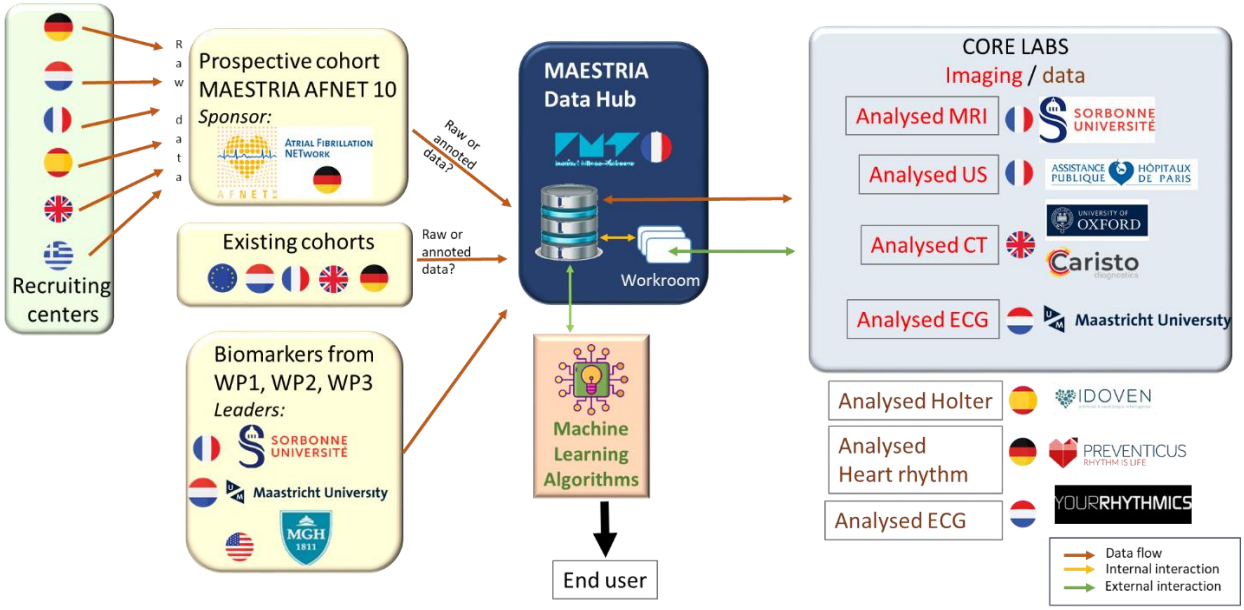


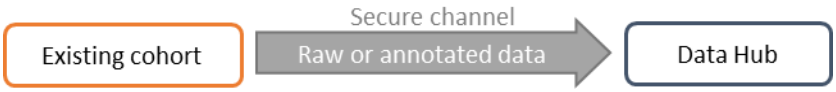
Figure 2.1. Schema of the Data Hub interactions

2.2.1. Data providers

The profile of Data Provider corresponds to any partner who will transfer data into the Data Hub, whether directly or indirectly. This data will be used for the development and validation of Machine Learning algorithms, as well as for imaging and data analysis carried out by the Core Labs. In some specific medical or legal scenarios, a direct transfer between a data provider and a Core Lab might happen, bypassing the Data Hub; however, this should remain exceptional.

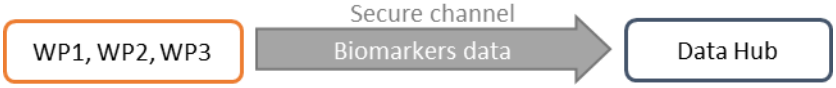
Data may come from different sources:

- **Data from existing cohorts:** This data will allow validating retrospectively specific parameters developed in the project. The raw or annotated data from the existing cohorts will be directly stored in the Data Hub. This means that the data owners will have the possibility to transfer their data themselves into the Data Hub through a secure channel.



- **Data from MAESTRIA's cohort:** A new cohort consisting of approximately 600 patients will be created. Clinical data entered directly into the MAESTRIA cohort by recruiting sites will be checked/cleaned. Once this process is completed, the clinical dataset will be locked and made available to the Data Hub. The recruiting centres will not transfer their data directly into the Data Hub; it will be the MAESTRIA cohort coordinator who will be in charge of transferring the cleaned data.

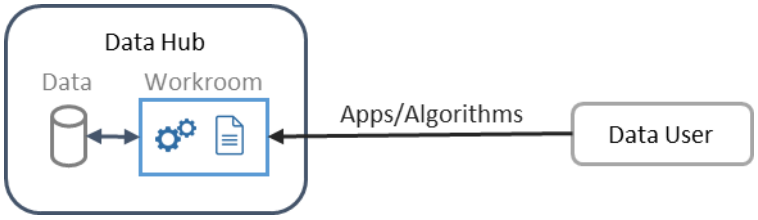
- **Biomarkers Data:** This data source corresponds to electrical, genetic, cardiac imaging, and other data produced in the context of WP1, WP2 and WP3 that can be used as biomarkers for atrial cardiomyopathy. The data owners will transfer this data directly into the Data Hub through a secure channel.



2.2.2. Data User

Any partner that will make use of data hosted in the Data Hub matches the profile of Data User. According to where the data processing takes place, two scenarios are identified for this profile:

- **Data processed directly in the Data Hub:** Processing and analysing data directly in the Data Hub will be the encouraged approach for MAESTRIA's Data Users. This ensures that the data is handled in a confined and controlled space. For this purpose, TeraLab will propose *workrooms*, secure, isolated working environments for the partners to execute their data processing. Workrooms are described in the section below.



- **Data processed at the partner's premises:** Only in the case that technical or legal constraints prevent a partner to bring their applications into the Data Hub, IMT will put in place a procedure to move data into the partner premises. This approach is discouraged by default, as it might compromise the data security and integrity. Every data transfer out of the Data Hub should be planned, authorised and logged. An authorisation procedure, whether manual or automated, should be defined for these transfers. A responsibility matrix is also recommended for each of these stages and for every data set. These elements are described in deliverable D7.3, the Data Management Plan.

2.2.3. Workrooms

Workrooms will be secure, independent working environments allowing data users to bring their algorithms and applications into the Data Hub. These Linux-based workrooms will bring several advantages to MAESTRIA data users; workrooms will:

- be customizable to the partner's technical needs in terms of storage capacity and processing power, they may consist of one or several virtual servers;
- allow the user to access data from the Data Hub in a secure and transparent manner;
- remove the overhead of data transfers whether out of or into the Data Hub; eliminating not only the actual transfer delays for big datasets, but also the waiting time for an eventual authorisation from data owners.

2.2.4. Technical Aspects

TeraLab provides Linux-based environments, with a customisable configuration in terms of CPU, RAM and storage, including backup. For the MAESTRIA Data Hub, this dimensioning will be determined by TeraLab's experts based on previous experiences with medical data use cases, and in accordance with the technical requirements informed by the partners through both a technical survey and individual meetings, if needed. Workrooms will be configured case by case. The TeraLab infrastructure consists of over 2000 vCPU, 16 TB RAM and storage capacity of nearly 1 Petabyte.

Data transferred into the Data Hub will not directly reach the shared area. Data will be transferred by means of an SFTP server (as explained in section 4.2.1) into a *staging* area. This staging space, which will be isolated, will serve as intermediate storage where data integrity can be verified before being transferred to the shared area. Analogously, a staging area will be used for outgoing data.

3. Legal Feasibility

Most of the data that will be handled by the MAESTRIA project correspond to patients' medical data. This type of personal data is considered as *sensitive* by the European Commission[3] and is subject to specific processing conditions. The MAESTRIA consortium should adhere to all applicable regulations that guarantee data protection and carefully apply them all along the duration of the project.

[3] https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_en

This section presents a summary of known data protection regulations in effect at the European and national scale, as well as locally at the member institutions.

3.1. General Regulations

Data privacy is a global concern; however, each country or territory establishes its own regulations with respect to data access and privacy. An increasing number of countries are requiring companies and organisations collecting personal data to inform users or members about it and to ask for their consent to do so. The notice can include, for example, which data will be collected, the purpose of such data, where and for how long will the data be stored, or even the fact that data will be shared with partners.

In the European Union, data protection is governed by the GDPR. In France, host country of MAESTRIA's Data Hub, additional regulations which apply to manipulating sensitive data are dictated by the CNIL.

3.1.1. GDPR

The General Data Protection Regulation (GDPR)[4] is a regulation on data protection and privacy in the European Union (EU) and the European Economic Area (EEA), addressing also the transfer of personal data outside of these areas. The GDPR enhances individuals' control and rights over their personal data. Enforceable as of May 2018, GDPR is a regulation, not a directive; it is directly binding and applicable but does provide a margin of manoeuvre for Member States to specify its rules, including for the processing of special categories of personal data ('sensitive data').

The MAESTRIA project is inherently linked to the GDPR, both from its European nature and because of the sensitivity of the data that it handles. Several concepts that are frequently used in the project and, more generally, in the medical domain are defined in the GDPR, including, but not limited to:

- **Personal data** means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.
- **Processing** means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or

[4] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN

otherwise making available, alignment or combination, restriction, erasure or destruction.

- **Pseudonymisation** means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data is not attributed to an identified or identifiable natural person.
- **Consent** of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her.
- **Genetic data** means personal data relating to the inherited or acquired genetic characteristics of a natural person which give unique information about the physiology or the health of that natural person and which result, in particular, from an analysis of a biological sample from the natural person in question.
- **Biometric data** means personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data.
- **Data concerning health** means personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status.

While the whole GDPR should be enforced for MAESTRIA's data, several principles are more tightly related to the project. A non-exhaustive list of these principles includes:

- **Processing of personal data (Article 5)**
  Expresses that personal data shall be: processed lawfully, fairly and in a transparent manner (lawfulness, fairness and transparency); collected for specific and legitimate purposes (purpose limitation); relevant and limited to what is necessary (data minimisation); accurate and up-to-date (accuracy); kept for no longer than is necessary (storage limitation); and processed in a manner that ensures security (integrity and confidentiality).

- **Conditions for consent (Article 7)**
  Explains that each partner providing data shall be able to demonstrate that the data subject (patient) has consented to processing of his or her personal data, and that the patient shall have the right to withdraw his or her consent at any time.

- **Processing of special categories of personal data (Article 9)**
  Regulates the processing of sensitive data. For the purpose of MAESTRIA and the medical data handled by the project, this principle establishes that sensitive data processing should be prohibited unless it becomes necessary for reasons of public

interest in the area of public health, such as *protecting against serious threats to health or ensuring high standards of quality and safety of health care*. It emphasizes the fact that the participating member states should provide for suitable and specific measures to safeguard the rights and freedoms of the patients.

- **Security of processing (Article 32)**
  Deals with technical security aspects concerning the processing of personal data. The concerned entity, namely IMT in the case of MAESTRIA, shall implement appropriate technical and organisational measures to ensure a level of security in accordance to the risk and the costs of implementation, including: pseudonymisation and encryption of the data; ensure the ongoing confidentiality, integrity, availability and resilience of processing systems; restore the availability and access to personal data in a timely manner; and regular testing and assessment of these measures.

Other sections of the GDPR are of equal importance but will not be summarized in this document, such as: obligations of the controller (legal entity responsible for the data processing), information to be provided to the subject (patient) for the sake of transparency, or the guidelines for preparing a code of conduct.

3.1.2. CNIL

The CNIL (Commission Nationale de l'Informatique et des Libertés) is an independent French administrative regulatory body whose mission is to ensure that the data privacy law is applied to the collection, storage, and use of personal data. Established in 1978, it is the national data protection authority for France.

The CNIL regulations have been adjusted to align with those of the GDPR since its publication. In what concerns MAESTRIA, the CNIL defines a procedure to follow when processing personal health data[5], which, generally speaking, places the responsibility of such formality to the actors, i.e. the MAESTRIA partners. In particular, the person responsible of the data processing should be able at all times to prove their compliance with GDPR (principle of *accountability*). Specific actions to realise in this framework are:

- Keep a record of all data processing
- Carry out impact analysis for the processing considered of “high risk” for the population
- Take care of informing the concerned people and make sure that their rights are respected (access, rectification, opposition, etc.)

[5] https://www.cnil.fr/fr/quelles-formalites-pour-les-traitements-de-donnees-de-sante-caractere-personnel (in French)

- Formalise the roles and responsibilities of the people in charge of the processing
- When applicable, appoint a Data Protection Officer
- List the actions performed to guarantee the security of the data

In light of the alignment of the CNIL with the GDPR and the consequent update of the French law of Computer technology and freedom ("loi Informatique et libertés"), the procedures before the CNIL have been simplified or suppressed under certain conditions, including:

- Processing for which the patient has given their express consent
- Processing needed for the safeguarding of human life
- Processing of personal data made public by the concerned person
- Processing needed for preventive medicine, medical diagnostic, etc. and carried out by a health professional or other person *with the imposition of professional secrecy*
- Processing with the exclusive use for internal research in a medical monitoring

This processing can be done without any formality before the CNIL. However, it is imperative to keep track of them in an activity record. Two types of processing remain subject to the CNIL's authorisation and require an impact analysis of their evaluation:

- Processing with a purpose of public interest
- Automated processing with the purpose of research studies in the health domain, as well as the evaluation or analysis of practices of care or prevention

The CNIL has two months to evaluate a request for authorisation after it is submitted. Past this period, if no response has been received the request is implicitly accepted. It is worth to mention that TeraLab has previous experience requesting a health data processing authorization from the CNIL, with a positive response.

3.2. Specific Regulations

On top of Europe-wide or countrywide regulations, other procedures and restrictions for data storage, transfer or processing may apply at each partner institution. IMT has been working with the WP5 partners, in particular with those who will serve as data providers, to determine what internal regulations apply to their data.

Overall, many MAESTRIA data providers agree that the first step is to obtain authorisation from their organisation's board/committee. Other important concerns include approval from an ethics committee or the signature of data protection agreements. On the other hand, some institutions have already obtained clearance from their local

authorities to transfer the data. Section 5.3.1 of this report presents the results of the data access survey and follow-up discussions with the data providers. It is important to point out that the provided figures remain preliminary at this stage of the project. Some authorisations or restrictions might not yet have been identified, defined or linked to the data at the time of submission of this document. Moreover, a large amount of data is yet to be produced and local regulations might apply in a case-by-case basis for this new data.

A consortium-wide Data Sharing Agreement is being drafted by Sorbonne University as lead institution. This document shall govern the transfer of datasets from the owner partners to the Data Hub; and from it to the user partners, if needed. The signature of this agreement is the starting point for further local data transfer authorisations.

In addition, if data is to be *processed* at a partner's premises, the institution must obtain clearance for sensitive data processing from their national authority responsible of enforcing GDPR.

3.3. Data Anonymisation

As defined by the GDPR, *pseudonymisation* means processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional data, a key. This key, containing the user's personal data, should be stored securely and separately from the pseudonymised data. Full *anonymisation* happens when the key is safely deleted, meaning that the data can no longer be traced back to an individual.

Anonymisation of data is a recurrent concern shared by all the project members who will act as data providers. From the earliest discussions, several partners have pointed out that it is critical for patients' data to be anonymized or pseudonymised when residing in the Data Hub. An argument to favour pseudonymisation over full anonymisation for some of MAESTRIA's datasets is to be able to trace back patients and ultimately allow them to benefit from the results of the project. Either way, neither raw personal data, nor de-anonymisation keys will be kept in the Data Hub.

Several actions are underway to guarantee data anonymisation. Some partners, like CARISTO, have already implemented pseudonymisation solutions. AP-HP Hôpital Saint-Antoine are working in defining a systematic procedure to produce anonymized images and files from commercial tools. Maastricht University has expressed the possibility of implementing an anonymisation solution for their provided data types, which could be shared with other partners.

To consolidate these initiatives and solutions, a working group on data anonymisation has been created within the project. It will involve all partners requiring anonymisation steps or anonymised data. The group will session over the next months with several goals:

- Define the anonymisation/pseudonymisation needs of each partner
- Identify the existing or implementable anonymisation solutions and classify them by data type
- Determine whether these solutions can be hosted at the Data Hub and/or shared with other partners (licences, portability, compatibility)
- Identify risks, workarounds and blocking points
- Propose a road map for anonymisation, including testing the solutions

4. Technical Feasibility

In parallel to the legal regulations, a number of technical aspects should be examined to determine the feasibility of data aggregation and processing at the Data Hub. This section presents three technical elements: medical data types, data transfer considerations, and data processing concerns.

4.1. Data Types

Data types can influence the way data is transferred, stored and processed in the Data Hub. Diverse medical data types are used in cardiology diagnostics, they differ in format and size, but also in the way they are collected, processed and visualized. Typically, these data types include:

- **Medical imaging:** Common imaging techniques in cardiology include computed tomography (CT) scans, magnetic resonance imaging (MRI), ultrasound, and positron emission tomography (PET). Data can be in the form of still images, 3D images, or video.

- **Electrocardiography (ECG):** Measurement and recording techniques that are not primarily designed to produce images, but rather to produce data susceptible to representation as a parameter graph over time. ECG signal recording come in specific binary formats (*e.g.* a .dat file) and require dedicated tools to be visualized, for instance Open ECG[6].

[6] https://uk.mathworks.com/matlabcentral/fileexchange/49822-open_ecg-ecg-dat-file-reader

- **Omics:** Omics technologies are defined as high-throughput biochemical assays that measure comprehensively and simultaneously molecules of the same type from a biological sample[7]. While omics data exchange formats follow several standards[8], they are orthogonal to the previous data types.

For medical images (X-ray, CT, MRI, ultrasound, etc.) and their related metadata, DICOM®[9], Digital Imaging and Communications in Medicine, has become the international standard. It defines the formats for medical images exchange by specifying the data and quality necessary for clinical use. DICOM is widely implemented in radiology, cardiology imaging, and radiotherapy devices. However, a disadvantage of the DICOM standard is probably the possibility for entering too many optional fields, which can result in data inconsistency due to fields left blank or filled with incorrect data. A second issue arising from DICOM's flexibility is that different manufacturers use different data segments to embed vendor-specific information; this might result in unreadable data between two manufacturers in the best case, or in data being overridden by another vendor, in the worst case. Finally, it has also been pointed out that the fact that DICOM allows to embed executable code and thus a malicious software could be inserted into the medical record[10].

The heterogeneity in cardiac data representation prevents to apply a single anonymisation solution to all data. Anonymisation for an image would require removing the visual embedded patient's data plus the associated metadata; whereas for a tabular or XML file it would mean the deletion of some columns or sections. Moreover, binary files like ECG might be unstructured and thus become more complex to anonymize.

A more detailed description of medical data types is provided by the MAESTRIA partners in the project's deliverable 7.3, the Data Management Plan.

4.2. Data Transfer

Two main issues linked to data transfer are discussed in this section: transfer securing and data volume.

4.2.1. Security

Any data transfer over the public internet is exposed to external attacks. Data can be stolen or corrupted in the way. Unsafe methods to transfer data include the use of unencrypted protocols, such as: HTTP, *e.g.* uploading documents through a web browser;

[7] https://www.nature.com/articles/s41597-019-0258-4
[8] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4152841/
[9] https://www.dicomstandard.org/about-home
[10] https://researchcylera.wpcomstaging.com/2019/04/16/pe-dicom-medical-malware/

FTP, unencrypted file transfer protocol; or simple email messages, which use the basic SMTP protocol without encryption or authentication.

A first strategy to secure data transfers is to use encryption, such as the one provided by the HTTPS protocol, which encrypts the communication using Transport Layer Security (TLS). However, this is not enough to prevent from data theft or corruption, particularly in a sensitive context like MAESTRIA's.

To transfer the partners' data into the Data Hub, a secure end-to-end channel is compulsory. The Secure Shell (SSH) Protocol is a protocol for secure remote login and other secure network services over an insecure network. On top of it, the SSH File Transfer Protocol (SFTP) enables secure file transfer capabilities between two hosts. SFTP additionally provides remote file-system management functionalities, such as resuming interrupted file transfers, listing the content of a directory, and deleting remote files. TeraLab will provide an SFTP server for data transfers in the Data Hub.

An advantage of SFTP is that MAESTRIA's partners would be able to use it regardless of their operating system, since command line and graphic interface clients exist for Windows, Linux and macOS. The ability for a user to copy data into the Data Hub or visualize content of a directory will be determined by the system administrators through access permissions.

4.2.2. Data Volume

Data transfers are also conditioned to the network bandwidth available between the hosts. Big volumes of data may require a channel to be open for long time. Connection might drop, causing an inconsistent state of the transferred data, and it might simply be unpractical and insecure to leave an ongoing transfer without supervision.

Some partners have indicated that their datasets reach the order of Terabytes. For these scenarios, it may be useful to create a *data transfer plan*. Depending on the volume of the data and the available transfer bandwidth, the data can be split and transferred in batches.

Once the data has reached the Data Hub, it is advised for data users to take advantage of the proposed workrooms to process the data *on site*. This will avoid additional time-consuming and potentially unsafe data transfers.

4.3. Data Processing

TeraLab is able to provide customisable workrooms (working environments) for partners to process data directly on the Data Hub. These workrooms will give the advantage of obviating data transfers and allowing data processing within the secure boundaries of the Data Hub. However, they are also prone to some constraints.

TeraLab services are granted only for Linux servers. The platform does not currently support Windows virtual machines and its security expert team is only specialized in Linux systems. Some utilities allow running a Windows application on a Linux system, but they come with no guarantees, as they are community versions. In the highly sensitive context of MAESTRIA, TeraLab is not able to host Windows-based applications in the Data Hub.

In addition, TeraLab does not provide tools that require a paid license. Any partner needing to host a paid application in the Data Hub should be in the capacity to transfer their license to the workroom or to purchase a new one.

In case that these technical constraints cannot be addressed, data processing should then happen on the premises of the data user. Any data extracted from the Data Hub will require an authorization from the data owner and the operation will be logged by TeraLab.

5. Data Access Form

This section presents a data access survey carried out in order to complement the legal and technical feasibility analysis with a real preliminary feedback from the project partners.

5.1. Summary

IMT has conducted a survey involving all MAESTRIA partners expected to interact with the Data Hub as data providers, data users, or both. The survey was sent as a Data Access Form, which is included as Annex to this document. The purpose was twofold: first, to enrich the current feasibility report with actual information coming from the partner institutions; and second, to understand and anticipate the actions to undertake in order to obtain the required authorisations to transfer partners' data into the Data Hub.

The survey was realised during the month of April 2021. A total of 12 answers were received. Table 5.1 summarizes the user profile of the partners.

Partner	Data Provider	Data User
AFNET	✓	✓
AP-HP INSERM		✓
AP-HP Hôpital St. Antoine	✓	
CNIC	✓	✓
Idoven		✓
Maastricht University	✓	✓
Owkin		✓
Preventicus		✓
Siemens Healthineers		✓
UD Essen	✓	
University of Birmingham	✓	✓
Yourrhythmics	✓	✓
TOTAL	**7**	**10**

Table 5.1. Partners' profile

5.2. Data Characterization

It is possible to break down the characterization of data into three categories:

- **Medical data types.**
  All data providers will transfer *medical imaging* data: MRI, ultrasound. One partner in particular specified that their medical images correspond to animal cardiac images. Interestingly, several partners will transfer *ECG* data, but not all of them in the same data format, which might pose a challenge for data anonymisation. One partner indicated to contribute with *omics* data from existing medical records. Another partner has also mentioned the possibility to include Implantable Loop Recorder (ILR) data, but this is yet to be confirmed.

- **Data formats.**
  - Unsurprisingly for *medical imaging*, DICOM format was mentioned by some partners, but also regular JPEG images.
  - For *tabular clinical data,* the partners use either XLS or CSV files.
  - For *ECG* one partner indicated that data is provided in CSV format, other in JPEG and for others the format was not specified.
  - For *omics* data, the format has not been specified.

- **Data size:**
  Data sizes vary considerably between partners and data types. One partner's full dataset is in the order of *tens of Megabytes*, while for another this is the size of *each* of their recordings. A partner explains that *one* DICOM exam *averages 2.5 GB* without specifying how many exams they will transfer.

Another partner estimates that their full dataset will be in the order of *hundreds of GB*, probably reaching 1 TB. On the large end, one partner reported to have data *exceeding 5 TB*.

Thanks to this information, IMT will be able to plan one-to-one technical meetings with the partners in order to understand the granularity of the data (size of a single record) and prepare a data transfer strategy, particularly for those partners providing large datasets.

5.3. Data Transfer Requirements

Different data transfer requirements were raised by the data provider partners. The restrictions can be classified according to the concerned authority.

5.3.1. From Data Provider

The most common answer from data providers in terms of internal authorisation involved their local authority. The data transfer approval shall come from a Board of Directors, Steering Committee or Managing Director of the organisation; or, more specifically, a Principal Investigator responsible for the datasets. A second authorisation emanating from the data provider is the approval of an Ethics committee, as indicated by two partners.

Some pre-requisites for these approvals to be granted include the signature of the consortium's Data Transfer Agreement, the guarantee that all data should be pseudonymised, or the patient's consent. Even if the institutions are the data owners, the patient should be consulted or informed before transferring their data elsewhere than the original repository for which they have given consent.ADD TODO

On the other hand, one partner has declared not needing additional authorisations on their medical data. Another partner will provide animal data, which is not considered as sensitive and is not concerned by the GDPR.

5.3.2. From MAESTRIA Consortium

A consensus from the partners is that, in order to start their internal procedures to obtain the approval for data transfer, the first step is to sign the projects' Data transfer agreement, which will govern the data sharing between members. In addition, a project-wide or local Data Protection Policy, including Technical organisational measures, is expected. These requirements have been discussed during the project's Data Management meetings and are work in progress.

5.3.3. From Hosting Institution

The last aspect of the Data access survey referred to requirements to be fulfilled by the Data Hub hosting institution, IMT. From a technical point of view, two partners requested a secure transfer service, or the use of secure protocols. Secure transfers will be granted via an SFTP server and user authentication. Legal requirements included a Data Protection Policy and the signature of a non-disclosure agreement. Some partners have also requested further information about the data hosting institution, which can be consulted in Section 2 of this document.

6. Roadmap

This section proposes a set of actions for the following months from the legal and technical points of view. These two lines of action can be performed in parallel.

6.1. Legal roadmap

As stated in Section 5.3, MAESTRIA partners agree that the first step towards obtaining local authorisations is the signature of the Data transfer agreement. Sorbonne University is currently working on the production of this document, which will be promptly shared with the consortium partners. Once the agreement is approved and signed by partners, which might require several iterations, then each partner will proceed to launch their internal authorisation procedure.

A consensus has been recently reached with respect to data anonymisation, which will facilitate the procedures at the Data Hub. All data transferred into the Data Hub will be *ready to be used* by the partners. This will mostly be achieved by pseudonymising the data before the transfer, but could also mean that explicit consent of the patient has been given for the use of their un-anonymised data. A next step in this direction is for institutions to define a procedure or implement a solution to pseudonymise their data. The consortium should also agree on a project-wide ID format for pseudonymisation of patients. The data anonymisation working group is overseeing these actions.

6.2. Technical roadmap

IMT has worked on a second survey to gather the technical requirements of the partners acting as data providers and data users. The form will be sent in September 2021 (M07) and, depending on the complexity of the partner's requirements, it might be followed by a technical discussion. A data transfer plan per data provider shall help organize the transfers better.

The technical survey will provide metrics to determine the dimensioning of the Data Hub. Once this dimensioning has been established, TeraLab will proceed to create the Data Hub and its various components, including setting up the SFTP server and the backup server.

A second stage will be the management of roles and users. TeraLab will aggregate the corresponding information from all partners, who will appoint members for each of the roles defined in the Data Management Plan. Afterwards, the credentials will be created and should be tested before any data is transferred to the Data Hub.

Finally, all transfer procedures must be tested with sample data, including: transfer of data into the Data Hub's staging area, approval of the integrity of the data and further transfer into the shared area, request and approval of data extraction. Different attempts of unauthorised operations must be tested as well, none of which should succeed.

7. Conclusion

Aggregating data from different medical sources into MAESTRIA's Data Hub is a complex task that should keep at its core the respect and guarantee of the patient's confidentiality. It requires aligning to local and international regulations and devising a secure technical environment.

From the legal point of view, the European regulation GDPR shall serve as MAESTRIA's standard to govern data privacy, with an emphasis on the *sensitive* character of medical data. At the local level with respect to the Data Hub, France's CNIL generally complies with GDPR but imposes additional procedures to follow in order to process medical data. Additionally, the crucial concern for data anonymisation is already being addressed by the consortium partners in a dedicated working group, where the consensus is that data should be pseudonymised before transferring it to the Data Hub. The pseudonymisation techniques and tools are still under discussion.

From the technical point of view, the standardization of medical data types, such as the use of DICOM for medical imaging, will favour data aggregation and use between partners. An important issue to address in this sense is the heterogeneity of DICOM data from different manufacturers. Data anonymisation will require also the technical involvement of expert partners to come up with solutions applicable to the different data types. Data transfer might be, technically speaking, a minor concern: a secure channel will be made available and a data transfer plan can be defined for very large datasets. On the other hand, on site data processing might be constrained if the required tools are not Linux-compatible or require licenses that cannot be transferred.

These legal and technical concerns were generally confirmed by our partners through their responses to a data access survey. The survey showed that their data is indeed very heterogeneous, but most importantly it highlighted the partners need for several authorisations before the data transfer can happen. Several actions will follow in the coming months to accelerate the obtaining of these authorisations. In particular, the next steps include the signature of a Data sharing agreement and technical discussions with the partners for the implementation of the Data Hub.

8. Annex

The Annex corresponds to the Data Access Form sent to WP5 partners in April 2021.

DATA ACCESS FORM

TeraLab (IMT)

TeraLab will host the Data Hub for MAESTRIA. The goal of this form is for TeraLab to initiate an inventory of regulations and restrictions applicable to data hosting. This will allow us to anticipate as much as possible any action to obtain clearance to transfer the data.

Partner information

- Organisation's name: _____

- Person responsible for filling this form:
  Name: _____
  Position: _____
  Email: _____

- What is your organisation's role with respect to MAESTRIA's Data Hub? *You may select several options.*

  ☐ Data Provider – we will feed the Data Hub with our data
  ☐ Data User – we will use the data stored in the Data Hub
  ☐ Other – Please specify _____

- Will you transfer data into MAESTRIA's Data Hub?
  ○ Yes ○ No

If you will NOT transfer data into the Data Hub, you do not need to go through the rest of the form.

Data characterisation

- What data will you transfer to TeraLab for the MAESTRIA Data Hub?
  *Please provide a summary of your data. You do **not** need to be precise at this time.*

  Example
  - Data type: Main data correspond to cardiac images, MRI scans, and related tabular records.
  - Data format: Data are in Excel and JPG formats.
  - Data size: The dataset size is between 500 GB and 1 TB.

Data transfer authorisations

- Is your dataset subject to any transfer/storage restriction or authorisation?
  ○ Yes ○ No

If there is NO restriction on your data (i.e. you own the data and you have permission to transfer it), you do not need to go through the rest of the form.

- Do you know which authorisations you need in order to transfer data into MAESTRIA's Data Hub?

  ○ Yes ○ No

- Please mention any ***internal authorisation from your organisation*** needed to transfer your data to TeraLab. ***List as many entries as needed.***
  - Authorisation:
  - Contact person details (name, position, email):
  - Pre-requisites to obtain the authorisation (if any):

  Example
  - *Authorisation:* ***Signature of the head of cardiology department***
  - *Contact person details:*
    - *Dr. John Doe*
    - *Head of cardiology of The European Research Centre*
    - *john.doe@researchcentre.eu*
  - *Pre-requisites to obtain the authorisation (if any):*
    - *the data have to be anonymised*
    - *IMT has to have signed an NDA*

- Please mention any ***regulatory*** authorisation needed to transfer your data to TeraLab. ***List as many entries as needed.***
  - Authorisation:
  - Contact person details (name, position, email):

- Pre-requisites to obtain the authorisation (if any):

Example

- *Authorisation:* ***GDPR compliance***
- *Contact person details:*
  - *Dr. John Doe*
  - *DPO at The European Research Centre*
  - *john.doe@researchcentre.eu*

- Please mention any ***additional*** authorisation/restriction/policy that TeraLab should comply with in order to transfer your data to the Data Hub. ***List as many entries as needed.***
  - Authorisation:
  - Contact person details (name, position, email):
  - Pre-requisites to obtain the authorisation (if any):

Example

- ***Restriction: The data should be transferred using secure protocols***
- *Contact person details:*
  - *M. John Doe*
  - *Quality Officer at The European Research Centre*
  - *john.doe@researchcentre.eu*

- Do you require anything from TeraLab (IMT) as hosting institution to get these authorisations? *Please specify.*