QUSTom

D1.2 – Data Management Plan

Version 5.0

Document Information

Contract Number	101046475
Project Website	https://qustom-project.eu
Contractual Deadline	30/03/2024
Dissemination Level	P - Public
Nature	R - Report
Author	Eduardo Cabrera
Contributors	Gabriela Espinosa (BSC), Cristina Duran (FrontWave), Torsten Hopp (KIT), Josep de la Puente (BSC), [Almudena Maceda (VHIR), Gladys Lino (VHIR) and Nicole Ruiter (KIT) contributed in updates to original deliverable]
Reviewers	Nicole Ruiter (KIT), Susana Castel (FrontWave)

European Innovation Council

Funded by the European Union

UK Research and Innovation

The QUSTom project has received funding from the European Union's Horizon Europe research and innovation programme under the Grant Agreement Nº 101046475, and for the UK partner from the United Kingdom Research and Innovation ("UKRI") under the UK Government's Horizon Europe Guarantee with UKRI Reference No. 10038375.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Innovation Council. Neither the European Union nor the granting authority can be held responsible for them.

Change Log

Version	Description of Change
V1.0	Initial draft for internal review
V2.0	Comments about the first draft Nadia Tonello
V3.0	Data from KIT and other WP leaders
V3.5	Pre-final version
V4.0	Reviewers' comments
V5.0	Review Meeting suggestions implemented and additional formal review
V6.0	Added Annex II, regarding the execution of the Clinical Study. Added additional authors

Table of contents

Glossary of abbreviations and definitions

DMP: Data Management Plan

GDPR: General Data Protection Regulation

FAIR: Findable, Accessible, Interoperable and Reproducible

PACS: Picture Archiving and Communication System

VHIR: Vall d'Hebron Institut de Recerca

HUVH: Hospital Universitari de la Vall d'Hebron (affiliated with VHIR)

DICOM: Digital Imaging and Communication in Medicine, a standard format for radiological images.

B2Drop: A cloud-based storage environment to share data. In particular, we will refer to an instance hosted at BSC and available to partners (b2drop.bsc.es).

GitLab: A service to store git projects. In particular, we will refer to an instance hosted at BSC and available to partners (gitlab.bsc.es).

GPFS: Global Parallel Filesystem, in particular, we refer to the disk attached to BSC HPC services.

USCT: Ultrasound computer tomography, not to be confused with 3D USCT III, which is a particular USCT device design and background of the project.

CC: Creative Commons.

GA: Grant Agreement, specifically, we refer to QUSTom's grant agreement.

HPC: High-Performance Computing.

LSDF: Large Scale Data Facility of the Helmholtz Association.

BWDA: BW Data Archive.

1. Executive Summary

This document presents the data management plan (DMP) of the QUSTom project, which describes the data management life-cycle for all tasks and datasets to be collected, processed and/or generated along the project's lifetime. This document describes the contents and organisation of the DMP, considering that the actual data (and metadata) description will be furnished as a “live” Annex that will be periodically updated throughout the project. Concretely, this deliverable describes, among others:

- Which datasets will be used, generated, collected and processed for the development and execution of the QUSTom project research activities.
- Which methodology and standards will be applied to QUSTom datasets.
- How datasets will be stored and handled during the lifetime of the project and after the end of it.
- How the datasets will be made (openly) accessible.
- This deliverable also describes similar aspects (storage, accessibility, openness) of the source code and software used and developed in the project.

This deliverable is necessary to achieve all project objectives and milestones successfully.

2. Introduction

The QUSTom project aims to improve and maximise access and reuse of research data and to consider the need to balance openness and protection of scientific information, commercialisation and Intellectual Property Rights (IPR), privacy concerns, security, and related data management and preservation questions. This data management plan (DMP) describes how data will be managed along the project, but it is not a final closed document. The DMP described in this report (D1.2) will be updated during the project's development to account for the generation/acquisition of new datasets, implementation of consortium policies, and/or other external factors.

This DMP is aligned with Guidelines on FAIR Data Management in Horizon Europe, i.e., the data must be Findable, Accessible, Interoperable, and Reusable.

Throughout this DMP, the organisations participating in the QUSTom project will be referred to as “partners” and all partners collectively as the “Consortium”. Other subjects not considered partners are referred to as “third parties”.

3. Data Summary

QUSTom focuses on transforming ultrasound-based breast cancer diagnosis by means of quantitative, high-resolution images. The imaging will complement or even replace current techniques for breast cancer detection using X-rays, such as mammograms. To obtain the medical images, researchers will develop mathematical algorithms that can show not only the image of the patient's tissue but also its associated uncertainty, which shows, pixel by pixel, how reliable the information is.

QUSTom’s objective is to obtain full 3D images by analysing data acquired with 3D ultrasound devices, with simulation-based quantitative imaging algorithms run in HPC environments. The data management plan for the QUSTom project facilitates data flow and utilisation of the data between the parties, including third parties/public, where appropriate, and ensures proper data preservation for future use.

The project partners are aware of the special care needed when treating sensitive clinical data and related personal data. Due to data sensitivity, as well as exploitation and licensing needs, the project will keep certain data closed, according to the “as open as possible, as closed as necessary” principle.

The collected data will be primarily used for the purpose of further development of candidate medical devices, regulatory reporting and new submissions, IP protection, licencing and technology transfer, but also for results dissemination and communication to interested stakeholders, including the scientific community, patients, and the wide public.

A variety of data will be produced by the QUSTom project. Most of this data falls into one of the following three major groups of data: **administrative** data, **technical** data, and **image** data.

Type of Data	Kind of Information	Data Owners
Administrative Data	Personal and non-personal data, Project documents, Deliverables, Reports, Participants' data	All
Technical Data + Administrative Data	Findability of data (DOIs, Repositories)	All
Technical Data	Raw ultrasound data and models	All
	Interoperability	All
	Standardisation	All
	Health Record Security and Privacy	VHIR
	Software Code and Design	BSC, FrontWave, KIT, ARCTUR
Image data	Medical-grade images meant for diagnosis	VHIR, FrontWave, KIT
Data Protection & Ethical Aspects	Ethical statements Approvals by ethical committees	FrontWave, VHIR, KIT

Table 1. *Summary of the data created in the QUSTom project.*

3.1 Scope of the document

This DMP analyses the main elements of the QUSTom data management policy. It is intended to cover the complete life cycle of the research data gathered and obtained within QUSTom and will outline the following:

- the types of research data that will be generated or collected during the project.
- how the research data will be processed and preserved.
- which parts of the datasets will be shared for verification or reuse.
- the standards that will be used.
- the handling of research data after the end of the project.

The DMP, with its updates along the project execution, aims to monitor the generated data regarding their privacy and confidentiality and ensure that the legal and ethical standards for data generation, use, storage and sharing are applied throughout the project (see Section 5 for details). It will guarantee coherence with the overall management of the project, as foreseen in the grant agreement and consortium agreement, and that appropriate technical standards are applied for data representation.

Furthermore, this DMP will ensure that QUSTom activities are compliant with the Horizon Europe Open Access policy and the recommendations of the Open Research Data. The DMP will address measures the Project partners forming the Consortium will employ in order to cope with legal, ethical and privacy concerning personal data usage and to ensure the application of relevant national and EU regulations, primarily the GDPR.

4. FAIR Management of Research Data

The ISO defines the DMP as "recorded information" (ISO 22005:2007) and describes data management as the "process of keeping track of all data and/or information related to the creation, production, distribution, storage, […] use of e-media and associated processes" (ISO 20294).

The QUSTom Data Management follows the "Guidelines on FAIR Data Management in Horizon Europe"[1], released by the European Commission Directorate – General for Research & Innovation. According to these guidelines, the management and organisation of data should be based on four basic principles, which determine how research outputs should be processed to be more easily accessed, understood, exchanged and reused. This means that data must be findable, accessible, interoperable and re-useable by researchers interested in using the data in future medical research.

All QUSTom partners are committed to respecting the policies outlined in this DMP and ensuring that all data are created, managed and stored according to applicable legislation. The partners that generate or collect data oversee its *integrity, compatibility, backups, validation and registration* during the project's lifetime. Backing up data for sharing through open-access repositories is within the responsibility of the partner processing the data. All partners chairing a lead role for a specific project task outlined in the Grant Agreement must assume responsibility for the quality control of the data generated or processed during the work on that specific task.

4.1 Making data Findable

The Information for this deliverable was gathered by the QUSTom Consortium through a questionnaire (see ANNEX 1), which is based -as already stated- on the "Guidelines on FAIR Data Management" and corresponds to a template associated with them. Considering the different aspects of these

[1] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

recommendations, the project management and the partners agreed on procedures to map the specific details of the QUSTom project to the general structure provided by Horizon Europe.

Over the course of the data collection phase, regular data management activities will be performed to ensure that data quality standards are met and the database stores appropriately complete, accurate, and logically consistent data sufficient to achieve protocol objectives.

As stated above, a variety of data will be produced along the QUSTom project. The following tables provide additional details about the three content-related groups (administrative, technical, and image data). In accordance with the FAIR principles -already addressed- specific information on data created /collected and processed in the respective processes is provided in the following tables:

Type of Data	Data Production & Storage
Data Generated/Collected	Administrative Data: Reports, Mailing lists, Partner Contact Details, Meeting Minutes and Information Technical Data: Software Code, Software Design, Ultrasound data, Models Image Data: Interoperable images which can be of USCT origin or from other imaging modalities
Data Format	Administrative Data: Binary raw, PDF, ZIP, Docs, Sheets, Slides Technical Data: UBIware format, 3D USCT III format Image Data: DICOM
Reproducibility	Administrative Data: B2Drop Drive and local repositories Technical Data: GitLab and local repositories, restricted in some cases Image Data: No
Size of Data	Administrative Data currently is likely not larger than a few Gigabytes. Technical data: In the order of 20 TB estimated. Image data: Tens of GBs
Software tools for creating/processing /visualising data	Administrative Data: B2Drop Drive Technical Data: PSM software, UBIware software, 3D USCT III device, HUVH imaging equipment, MATLAB Image Data: HUVH PACS
Use of pre-existing data	Administrative Data: No Technical Data: Open-Source libraries, Gitlab of standardisation projects, pre-existing data from coding lists, nomenclatures, and software declared as background. Image Data: No

Table 2. *Production and storage of data.*

Type of Data	Organisation and metadata of data
Standards for Documentation and Metadata	Administrative Data: No specific standardisation protocols observed. Technical Data: For raw ultrasound data: enforcement of UBIware format, which includes its own metadata descriptors. The rest is not standardised but documented. Image Data: Enforcement of DICOM format, which includes its own metadata descriptors
Best Practices/Guidelines for Data Management	Administrative Data: The present document Technical Data: The present document Image Data: HUVH's best practices for PACS
Tools for Formatting Data	Administrative Data: No automatic tools currently used Technical Data: Under development Image Data: Python DICOM support libraries
Directory and File Naming Convention	Administrative Data: document tree established in B2Drop, free naming convention Technical Data: For ultrasound data, embedded in UBIware data format. Image Data: Following HUVH's best practices for PACS

Table 3. *Findable data.*

4.2 Making data openly Accessible

Data in QUSTom will be made openly accessible mainly through the following mechanisms: Deliverables and reports marked PU will be made available at the project website (www.qustom-project.eu). The open data website is under development and will include a set of repositories and registries to make the open software components, trained artificial intelligence models, and raw ultrasound data available for workflow developers and users.

All deliverables marked PU will be made available by means of the project's website. SEN deliverables will be restricted to access only internal to partners. Several developments (software, results) can have additional restricted access to few or no partners for the sake of their exploitation potential. This is particularly enforced for IP presented as background (e.g., 3D USCT III specifications, UBIware software, UQ-patent derived foreground).

A specific effort will be made to enable an open-access repository of project-related data. The repository will include a subset of raw ultrasound data and models together with the necessary metadata to develop medical technologies. We will use the platform to promote UBIware's format as a potential standard for ultrasound data interoperability, which is currently missing in the ultrasound community.

Alternatively, other means may be explored, such as Zenodo[2].

	Data Access
Risk	Administrative Data: unauthorised access Technical Data: Stealing of GitLab credentials and access to source code by an unauthorised person Image Data: unauthorised access
Procedures to Follow a Data Breach	Administrative Data: Will be defined at a later stage of the Project and documented in the next version of the DMP. Technical Data: Will be defined at a later stage of the Project and documented in the next version of the DMP. Image Data: Will be defined at a later stage of the Project and documented in the next version of the DMP

Table 4. *Data accessible.*

4.3 Making data Interoperable

The interoperability of the data will be guaranteed using different common formats to exchange the data. Samples of these formats are DICOM or UBIware format. On the one hand, DICOM is a well-established format for medical images, which is broadly used and allows information to be visualised and actuated upon with standardised tools. On the other hand, we will promote a novel standard for raw ultrasound data. In particular, FrontWave Imaging has developed the UBIware format based on HDF5, including all necessary information to operate with ultrasound data safely.

The actual use of one or another format may change according to the best practices in each pillar community.

We are also considering the Data Catalogue to include metadata such as access protocol, format and information about making data interoperable.

4.4 Increase data Reuse

The QUSTom project makes very little use of pre-existing data. Most relevant data will be generated explicitly during and for the project (see Fig. 1).

[2] https://zenodo.org/

The owners of the data used in the project are responsible for defining the license of the data. Therefore, the license may change from case to case, but open licenses offer most data. Another license considered reasonable for some partners is the CC BY-SA1.

Open licenses under Creative Commons will be promoted for the new data created in the project context. Open-source licenses such as Apache v2.0 or BSD will be promoted regarding the software code. It is worth mentioning that some software, PSM and UBIware, both have limited access because of their exploitation potential. However, a version of PSM2.0 has been opened to partners for its development, and the results obtained with UBIware will be made available for other project participants when necessary.

	Data Sharing & Reuse
Reuse of Data	Administrative Data: N/A Technical Data: No previous technical data will be used; the only potential exception of synthetic models, which would be accordingly cited if used. Image Data: No previous data will be used
Organisation/ Labelling of Data for Easy Identification	Administrative Data: N/A Technical Data: Not defined yet Image Data: N/A.
Data Sharing Requirements	Administrative Data: N/A Technical Data: Standardisation tools Image Data: N/A
Audience for Reuse	Administrative Data: N/A Technical Data: Anyone (Healthcare organisations, research organisations, public and private institutions) for publicly released data at an open website. Other data is restricted. Image data: Internal data, project only
Restrictions on Reuse of Data	Administrative Data: N/A Technical Data: Only data specifically added to the open website repository Image Data: Fully restricted
Publication	Administrative Data: Only PU deliverables on the website Technical Data: QUSTom Website. Academic publications will be open access, governed by Consortium Agreement rules. Image Data: Fully restricted, project only

Table 5. *Sharing and reuse of data.*

4.5 Resources for FAIR data preservation and archiving

Archiving will primarily be done at BSC servers, using B2Drop and GPFS systems. We remark that most data will be restricted and only used internally by the project. Public data on the open website will be advertised. We will enable a clear directory tree and instructions in order to help the findability of data. At a later stage, we will discuss using long-term open storage solutions such as Zenodo to ensure a long lifetime for open data.

	Data Preservation & Archiving
Archiving of Data for Preservation and Long-term Access	Administrative Data: B2Drop Drive, EC Website and local repositories Technical Data: GitLab, GPFS, LSDF / BW Data archive, and B2Drop, fully backed up Image Data: PACS of HUVH
Data Retention	Administrative Data: anticipated to be for at least five years Technical Data: at least five years Image Data: Will be stored in accordance with Article 17 GDPR[3]
File Formats	Administrative Data: .docx, .xls Technical Data: UBIware format primarily Image Data DICOM, preferably
Data Archives	Administrative Data: EC Portal Technical Data: Local repositories Image Data: According to national regulation
Long-term Maintenance of Data	Administrative Data: Not yet defined, will be presented in the next version of the DMP (potentially B2Drop) Technical Data: Not yet defined, will be presented in the next version of the DMP (potentially Zenodo) Image Data: Not yet defined, will be presented in the next version of the DMP (potentially PACS at HUVH)

Table 6. *Preservation and Archiving of Data.*

4.6 Responsibilities

Regarding the source code, each partner responsible for a given component is responsible for updating the source code in the corresponding git project repository.

[3] https://gdpr-info.eu/art-17-gdpr/

About data, each data owner is responsible for selecting the repository to store the generated data and its license. Each data owner is also responsible for the metadata and documentation of their datasets.

BSC, as coordinator of QUSTom, is responsible for maintaining the Data Management Plan.

5. Data Description

5.1 Administrative Data

The following table provides the characteristics and standards to be followed with respect to administrative data generated and processed during the Project. As mentioned before, the structure of the next tables follows the FAIR principle.

Type of Data	Data Production & Storage
Data Generated/Collected	Reports/Deliverables defined in the GA Templates Partner contact information Meeting/Web conference-related material (participants' list, agenda, presentations)
Data Format	pdf, doc, pptx, xlsx
Reproducibility	B2Drop maintains a history. All the deliverables will be uploaded to the Participant Portal website. The process is not replicable, but the partners produce several copies of the deliverables. Regarding contractual documents, copies are maintained on the Participant Project site.
Software tools for creating/processing /Visualising data	Deliverables, meetings/web conference-related material, effort data and contact details of partners can be visualised online through the Open office or offline through usual document readers.
Storage and Backup Strategy	The project management will keep the project data in B2Drop folder, Project website and EC Portal. Living versions of deliverables are stored in the collaborative workspace of the project repository on B2Drop, in the related WP/Task subfolder; the PC stores final submitted versions of deliverables in the "Submitted Deliverables" folder of the project repository.

Table 7. *Production and storage of administrative data.*

Type of Data	Organisation and metadata of data
Standards for Documentation and Metadata	pdf and doc format for reports
Best Practices/Guidelines for Data Management	The project repository provided through B2Drop is a collaborative space that is the main exchange means of documents among partners.
Tools for Formatting Data	No automatic tool. Validation by the PMO and the Project Coordinator.
Directory and File Naming Convention	Naming of the document in each deliverable: <ProjectName>_<Document_number>-<Document_Name>_<v.#> Naming emails: [QUSTom][issue]
Project Data Identifiers	Deliverables names/identifiers are agreed with the EC and stated in GA.
Automatic Creation of Metadata	No

Table 8. *Findable administrative data.*

	Data Access
Risk	Administrative Data: unauthorised access Technical Data: Stealing of GitLab credentials and access to source code by an unauthorised person. Image Data: unauthorised access
Procedures to Follow a Data Breach	Project Coordinator will be responsible for ensuring secure access protocols for administrative data.

Table 9. *Administrative data accessible.*

	Data Sharing & Reuse
Reuse of Data	Public deliverables are openly accessible on the project website. Other Project administrative data is not intended to be publicly shared or otherwise made available to third parties.
Audience for Reuse	Public deliverables are intended for anyone.
Restrictions on Reuse of Data	Consortium deliverables are published under a Creative Common License. QUSTom Deliverables are for public, restricted or confidential circulation, as stated in Part A of the GA.

Publication	The QUSTom project website makes available public information about the Project and the Consortium, disseminating objectives and outcome of the research to the general public and acting as a pathway for interested users to go deeper into details about project outcomes through the "contact us" section. The website is available at https://www.qustom-project.eu/. Public deliverables are published on the project website as soon as the EU Commission approves them. They may also be published as draft documents on the project website after being sent to the Commission. Released public software will be accompanied by reports following the same publishing process.

Table 10. *Sharing and reuse of administrative data.*

	Data Preservation & Archiving
Archiving of Data for Preservation and Long-term Access	Deliverables and effort/financial data are stored on the EC website. The Project Coordinator will archive all deliverables and project-related documentation on its internal cloud service for future audits until five years after the end of the Project.
Data Retention	At least five years
File Formats	The Project Coordinator archives copies of deliverables in the original format. All final documentation is in pdf format.
Data Archives	The archive location for project management data (deliverables and effort data reported in the Project Progress Reports) is the EC portal, which is considered an institutional archive that preserves the deliverables and other information submitted permanently.
Long-term Maintenance of Data	The Project Coordinator will archive the deliverables and project-related material (minutes of meetings, agendas, presentations) for five years after the end of the project.

Table 11. *Preservation and archiving of administrative data.*

5.2 Technical Data

This section addresses the technical data generated and used during the project. Figure 1 presents a general overview of the technical data used within the project and its interoperation. In Figure 1, data is classified in:

- Subjects: Use cases representing a single experiment, which can refer to a patient, or a particular synthetic phantom, either of physical nature or in-silico. This data contains mostly metainformation describing the experiment, sources of information and other characteristics relevant to the imaging process. For patient data, all information will be pseudo-anonymised (the data cannot be associated with an identified or identifiable person because the information that identifies that person has been replaced by a series of technical measures. This process is carried on by the staff of the women's imaging section of the Hospital Val d'Hebron.). We assume all use cases are based on USCT III devices. Hence geometry information about the transducers is known. Otherwise, the geometry will be fully specified or referenced here.

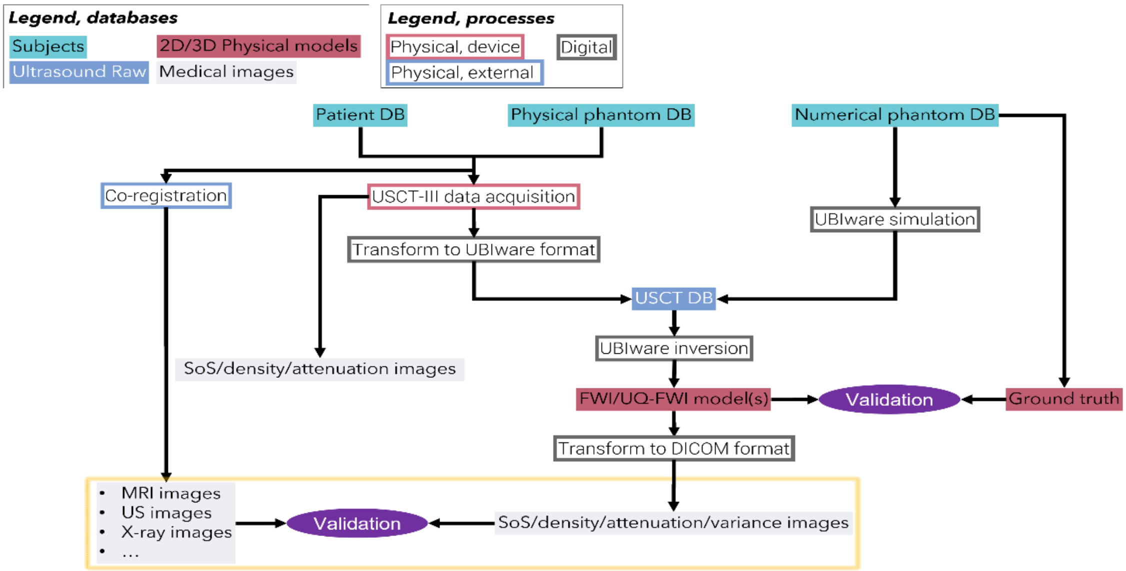


Figure 1. *Technical workflow, including the technical data (in coloured boxes) and processes (with coloured outlines) involved in QUSTom. In the graph, all data inherits vertical metadata from its predecessors for traceability.*

- Ultrasound raw: This data contains ultrasound recordings at a sufficient bit rate to capture the relevant information obtained at each source/receiver pair of the experiment. The data will always be stored in UBIware format for better interoperability.

- Physical models: These are gridded 2D or 3D datasets that contain for each pixel (voxel) the value of a certain physical property. Typically, physical models are the outcome of the inversion process. For in-silico cases, the ground truth models can be made available, which can be compared with the inversion result to check its validity. Models will be kept in UBIware format for better interoperability.

- Medical images: These images may have different origins but are all stored in DICOM format so that radiologists can compare them in a PACS. The inherited metadata from the Subject DB can be used to identify different image modalities for a single subject.

Processes that transform or augment data can happen at the hospital's existing facilities (i.e., co-registration of MRI images), at a USCT III device that may be installed at the hospital or digitally at the partners' servers. The validation process shown in Fig. 1 results in reports, documentation, and recommendations that can become administrative data. We remark that the validation marked in a yellow box happens within HUVH’s PACS and may require pseudo-anonymised data to validate results. In this case, only VHIR will be able to de-code patient data solely for the purposes of QUSTom’s studies.

5.2.1 Software Development

A special case of “technical data” is software. The principal software solutions used in the project are PSM and UBIware. Both have limited access because of their exploitation potential. PSM is the property of BSC, and its exploitation has been licensed to FrontWave Imaging. A “vanilla” version of PSM2.0 has been opened to partners participating in its development in Task 3.4 (ARCTUR, BSC, FrontWave). The package is stored at a Gitlab repository of BSC and containerised with Docker for interoperability at different systems. The other package, UBIware, is the property of FrontWave Imaging and will not be shared with other parties; results obtained with UBIware will be made available for other project participants when needed to proceed with the project’s activities.

5.2.2 Software Data

The following table shows the characteristics and standards to be followed with respect to software data generated and processed during the project. As stated above, the structure of the tables follows the FAIR principle.

Type of Software.	Data Production, Storage & Archiving
PSMv2.0 vanilla (open to project partners)	Python and C; Gitlab BSC;
UBIware (private)	Python and C; Internal Repositories FrontWave
[KIT software] (private)	MATLAB; KIT repositories (LSDF and BWDA)
Size of Data	GBs
Software tools for creating/processing /Visualising data	Any viewer capable of supporting DICOM

Table 12. *Production, storage and archiving of software data.*

5.3 Image Data

The following table shows the characteristics and standards to be followed with respect to image data generated and processed during the project. As stated above, the structure of the tables follows the FAIR principle.

Type of Data	Data Production & Storage
Data Generated/Collected	As per Fig. 1, either acquired from patients and physical phantoms at a 3D USCT III device or simulated from a numerical/in-silico phantom. Then inverted by means of UBIware or directly from 3D USCT III internal imaging packages.
Data Format	DICOM
Reproducibility	Fully reproducible, following software version and metadata of raw ultrasound
Size of Data	Tons of GBs
Software tools for creating/processing /Visualising data	Any viewer capable of supporting DICOM
Use of pre-existing Data	None

Table 13. *Production and storage of image data.*

Type of Data	Organisation and metadata of data
Standards for Documentation and Metadata	Metadata stored at DICOM following the standards of the format
Best Practices/Guidelines for Data Management	Those of HUVH PACS
Tools for Formatting Data	Any compatible with DICOM
Directory and File Naming Convention	As per HUVH PACS convention
Project Data Identifiers	Including metadata, the grant agreement nr (number), and the project name
Community Standard for Metadata/ Sharing/ Integration	As per DICOM standard
Automatic Creation of Metadata	As per DICOM standard

Table 14. *Findable image data.*

	Data Access
Risk	Inherited from HUVH PACS
Risk Management	Inherited from HUVH PACS
Correct execution of the Access process	Inherited from HUVH PACS
Procedures to Follow a Data Breach	Inherited from HUVH PACS

Table 15. *Image data accessible.*

	Data Sharing & Reuse
Reuse of Data	To be determined
Audience for Reuse	Potentially future clinical trials, internal to QUSTom partners and associates
Restrictions on Reuse of Data	Restricted to QUSTom partners and associates
Publication	Data may not be published. Academic results may be published as per the Consortium Agreement rules.

Table 16. *Sharing and reuse of image data.*

	Data Preservation & Archiving
Archiving of Data for Preservation and Long-term Access	Inherited from HUVH PACS; (5 years)
Data Retention	Inherited from HUVH PACS
File Formats	DICOM
Data Archives	DICOM
Long-term Maintenance of Data	Inherited from HUVH PACS

Table 17. *Preservation and archiving of image data.*

5.4 Clinical Study Data

Superseding the previous sections of this chapter, a complete analysis of the data involved in the clinical study is provided in Annex II. This encompasses all data transformation stages and all the technical and legal considerations required to fulfill the project’s goals

6. Data Protection & Ethical Aspects

Data protection is a central issue for research ethics. A fundamental human right enshrined in the EU Charter of Fundamental Rights provides all individuals with control over the way information about them is collected and used.[4] Article 8(1) of the Charter of Fundamental Rights of the European Union (the 'Charter') and Article 16(1) of the Treaty on the Functioning of the European Union (TFEU) grant everyone the right to the protection of personal data concerning him or her and the General Data Protection Regulation (GDPR) lays down rules relating to the protection of natural persons concerning the processing of personal data.

An identifiable natural person is one who can be identified, directly or indirectly, 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 (Art. 2(a) EU (GDPR)).

In the QUSTom project, all data processing will comply with EU law as well as national data laws and will follow the guidelines on "Ethics and Data Protection"[5]. It will be ensured that any partners, contractors or service providers that process research data at the QUSTom partners' request and on their behalf will comply with the GDPR and the H2020 ethics standards. Special attention will be given to a good balance between research objectives and the means used to achieve them.

As already discussed, the Project research will not include any specific clinical activity related to the QUSTom project. The partners of the QUSTom project will only collect data needed to meet the research objectives. Data sharing will be limited to a subset of health-related information strictly necessary for scenario testing.

This DMP is a living document, and further considerations will be made, especially with respect to imaging records.

7. Annex I

Data Management Plan Questionnaire

For each data category/data type (these two terms are used as synonymous) you plan to generate, collect and/or process, please provide a separate answer to the following questions. For example, if you are going to process medical diagnostic

[4] https://fra.europa.eu/en/eu-charter/article/8-protection-personal-data

[5] European Commission: Ethics and Data Protection, November 2018 https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-data-protection_en.pdf

data and generate software, you are managing two different data categories, and you need to answer the below questions for both data categories.

If you are collecting/processing or generating additional categories of data, we encourage you to add them and answer the questions below regarding those additional categories.
The questions concern data collected and processed during the lifetime of the Project only. This questionnaire does not address data management once the envisioned applications are on the market. However, please note that the questions address all data types or categories, not just personal data collected or generated during the Project.
If you are uncertain how to answer a question, please ask us: qustom_dmp@bsc.es

Your answers to this questionnaire will be seen in the Annex to the Data Management Plan deliverable.

1. **Data Types and Storage:** The following questions are intended to understand what data types will be generated and/or used in this project.

- What type of data will you produce or generate during the Project?
- What type of data will you collect during the Project?
- How will you collect the data? In what formats?
- How will you trace the collected data?
  - For example, how do you trace the provenance of the data collected or other metadata you maintain about the collected data?
- Will the process of data generation or production be reproducible?
- What would happen if collected data got lost or became unusable later?
- How much data will it be, and at what growth rate? How often will it change?
- Are tools or software needed to create/process/visualise the data?
- Will you use pre-existing data? From where?
- Storage and backup strategy?

2. **Data Organization, Documentation and Metadata**: The following questions are intended to understand the plan for organising, documenting, and using descriptive metadata to assure quality control and reproducibility of these data.

Answer the following questions only concerning the portion of data that you will publish (i.e., make it available to people external to the project).

- What standards will be used for documentation and metadata (e.g., Digital Object Identifiers)?
- No standards?
- Do you use any best practices/guidelines for managing the data to publish (i.e., make it available to third parties)?
- Do you use any tool for checking that the data are well formatted?
- What directory and file naming convention will be used?
- What project and data identifiers will be assigned?
- Is there a community standard for metadata sharing/integration?
- Can any metadata be created automatically?

3. **Data Access and Intellectual Property:** The following questions aim to identify any data access and ownership concerns.

- What are the major risks to data security?
- What steps will be taken to protect privacy, security, confidentiality, intellectual property or other rights?
- Have you prepared a formal risk assessment addressing each of the major risks to data security and potential solutions?
- Does your data have any access concerns? Describe the process someone would take to access your data.
- Who checks the correct execution of the access process (e.g., PI, student, lab, University, funder)?
- What procedures have you developed to safely transfer personal or sensitive data?
- Are there any special privacy or security requirements (e.g., personal or high-security data)?
- Any embargo periods to uphold?
- Have you implemented or outlined any procedures to follow in the case of a data breach?

4. **Data Sharing and Reuse:** The following questions are intended to clarify how the collected data will be released for sharing. Answer the following questions only with respect to the portion of data that you will publish (i.e., make available to people external to the project)

- If you allow others to reuse your data, how will the data be discovered and shared?
- List the categories of data that will be made reusable or openly accessible.

- Any sharing requirements? (e.g., funder data sharing policies often require that the digital data be released in machine-readable formats that supplement journal articles and presentations)
- Audience for reuse? Who will use it now? Who will use it later?
- Any restrictions on who can reuse the data and for what purpose?
- When will I publish it, and where?

5. **Data Preservation and Archiving:** The following questions are intended to clarify how the collected data will be preserved and archived.

- How will the data be archived for preservation and long-term access?
- How long should it be retained (e.g., 3-5 years, 10-20 years, permanently)?
- What file formats? Are they long-lived?
- Are there data archives appropriate for your data (subject-based? Or institutional)?
- Who will maintain the data for the long term?
- Who decides what data or what categories of data will be kept and for how long?
- The GDPR requires personal data not to be kept longer than necessary for the purpose for which it was stored. What protocol(s) will you implement to ensure you delete personal data that is no longer required to be stored?

6. **Ethical Aspects**

- What personal data do you intend to collect, generate or process?
- What types of sensitive data do you intend to collect, generate or process?
- Will any of the data subjects be children or vulnerable people?
- Will you be collecting personal or sensitive data from people who have not given their explicit consent to participate in the Project?
- If you collected personal data, as defined by the GDPR, which of the six Art.

6.1 Bases will you rely on for processing each category of personal data? http://www.privacy-regulation.eu/en/article-6-lawfulness-of-processing-GDPR.htm

- If you collect sensitive data, as defined by the GDPR, which of the ten Art. 9 bases will you rely on for the processing of each category of sensitive data? http://www.privacy-regulation.eu/en/article-9-processing-of-special-categories-ofpersonal-data-GDPR.htm

- Have you already gained consent for data preservation and sharing from any data subject(s)?

- How will you protect the identity of Project participants?
- Will you engage in large-scale or big-data processing?
- Will any entity (including any service provider) outside the E.U. have access to personal or sensitive data?
    - If yes, who?
    - For what purpose?
- Where is each of these entities located?

Participant no.	Participant organisation name	Part. short name	Country
1	Barcelona Supercomputing Center – Centro Nacional de Supercomputación	BSC	Spain
2	Karlsruhe Institute of Technology	KIT	Germany
3	FrontWave	FrontWave	Spain
4	Vall d'Hebron Institut de Recerca	VHIR	Spain
5	ARCTUR	ARCTUR	Slovenia
6*	Imperial College London	IMPERIAL	UK (associate)

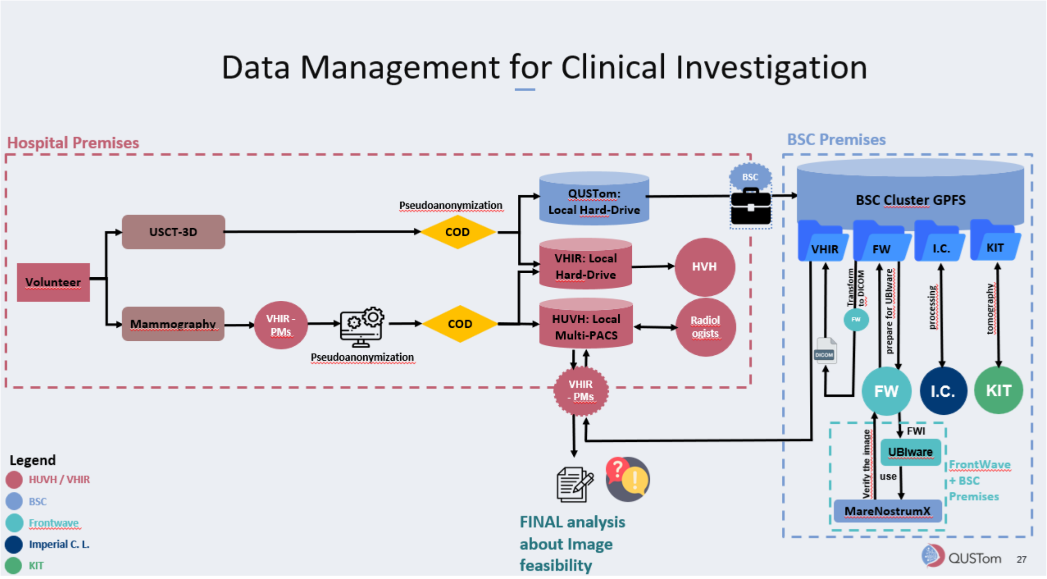
Table 18. *QUSTom's partners*

8. Annex II

Towards the activities of task 4.5 “Clinical feasibility as a medical imaging tool” we have prepared a complete description of the data handling involved in the process. The process must be robust and fully abide with the requirements of the clinical study. It is presented in the following.

Main data workflow

In the graphic below we present the foreseen data flow towards the clinical validation. In blue we depict data storage sites, in orange we depict devices and specific hardware and in grey different activities that need to be performed by specific project partners. The green box identifies the hospital's premises. The process starts with a volunteer being subject to data/image acquisition for different modalities. The mammograms will proceed to PACS as normal inside the hospital, on one hand. However, a copy will be anonymized and kept under custody. The data from the 3D USCT III device is extracted to a hard drive, anonymized and duplicated to a second drive. The first drive never leaves the hospital and is kept as a backup. The second drive is regularly transported physically to BSC premises where it will be uploaded to its GPFS for further use. BSC will provide with bsc21xxx accounts to all involved partners that will perform different tasks towards preparing the data towards the reconstruction. Once ready, FrontWave will run UBIware within BSC clusters and produce DICOM images. These will be uploaded to HUVH's PACS server or, if connectivity is an issue, physically sent via hard drive.

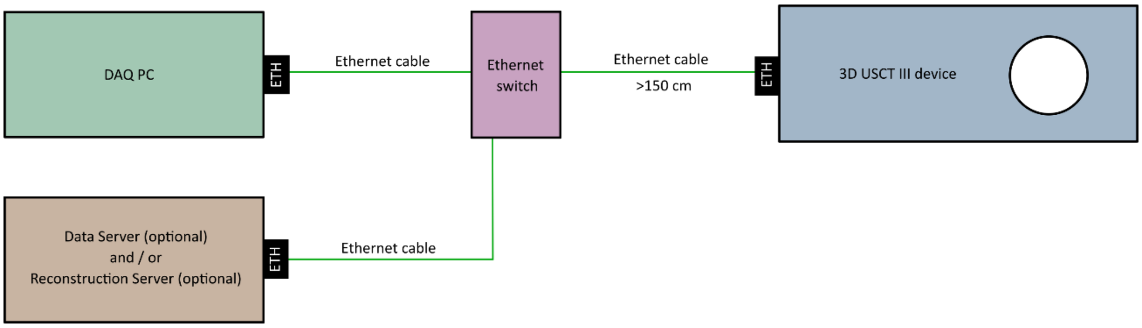


In particular, BSC will store the data at /gpfs/projects/bsc21/QUSTom_Clinical_Validation/ where only specific individuals with accounts at BSC and whitelisted can access (ACL). All data will be owned by the uploader/creator, giving only reading rights to the rest of the whitelisted QUSTom personnel. It is possible to FW, IMPERIAL or KIT to copy data outside from this system for the purposes of pursuing the project's clinical study.

Specific steps that are significant are further elaborated in the following.

Retrieval of USCT data and QC

The process of extracting patient ultrasound data needs a specific process involving access to USCT device via private network, adding data server optionally



The data server may be substituted with a large HDD-equipped DAQ PC. A NAS can play both functions: data server and switch.

Early estimates of data size are (FWI protocol: 4-rotation/translation-position case):

- **180 GB per breast**
- **45 GB per water shot**

This totals 405 GB per patient and **39.5 TB per 100** subjects (assuming that repetitions may be needed).

Estimates of device-to-DAQ/server transfer, using 1Gbps Ethernet, are approximately 54 minutes per patient. The device has a sufficient buffer to store a single volunteer's data (2 breasts and 1 water shot) but not two, hence the transfer process must be performed after each volunteer's data acquisition.

A basic check of the acquisition process will be performed directly after the data acquisition. This includes, e.g., the check for non-functional transducer arrays, incomplete measurement.

For further early data analysis in order to check if adaptions of the protocol during the study are necessary, KIT will use the data to reconstruct reflection and transmission images based on their established algorithms on their local infrastructure.

Pseudo-anonymization process and custody

The signal raw data will be acquired fully pseudo-anonymized at the hospital. For each patient only an identifier is stored in the meta data of the raw signal data set.

Pseudo-anonymization of existing images or data can be done with different programs and we still figuring out which one will be more suitable for us. The process will be performed through a configurable solution interface with multi-PACS.

We remark that the USCT data exits the hospital as fully anonymized data, and no one from BSC, or other participants with access to the BSC file systems, will have the possibility of identifying the data. HUVH/VHIR personnel may only have access to the DICOM final product, which will also be sent back to the hospital fully anonymized. The de-anonymization process will take place exclusively at the hospital, not at BSC servers, and only for the DICOM images. The mammograms or other data acquired by the hospital other than USCT data will never leave the hospital premises within the context of this study.

Data transfer in/out of HUVH

The transfer process to outside HUVH should be performed regularly, and only involving pseudo-anonymized data (see process above). A 5 TB disk seems reasonable to store data from 10 volunteers, which may be the throughput of a few days of clinical study.

A complete copy of the full data from 10 volunteers using USB 3.2 (approx. 300 Mbps) takes approximately 30 hours, for a process that is performed twice (from HUVH to external HDD and from external HDD to BSC) which may be the real bottleneck of the whole process. In a regular NAS, this process of copy should not interfere with other data transfer processes, such as extracting new USCT3D data.

Preparation for UBIware

The data from the USCT device needs to be reformatted and curated prior to being transformed into an FWI-ready UBIware format (FrontWave’s native format for US data). KIT provides a parser script and a data reader, both based on MATLAB, for this purpose. Reformatting will include selection of dead traces, chirp to pulse transformation, pre-filtering and trimming among others.

UBIware execution

The preprocessed and reformatted data will be sent to the “hot” disk of BSC to be ingested by UBIware. The reconstruction/inversion process will need to be queued and the result will take some time (TBD, still under evaluation). The resulting DICOM slices will then be ready for upload (or physical transfer) back to HUVH’s configurable solution interfacing with multi-PACS.

Radiologists’ comparative study

In order to save time to the radiologists, the responsibles from the trial in HUVH will access the PACS mammograms, pseudo-anonimyze the image and upload it to the configurable solution interface with multi-PACS, Senoiris, to be evaluated.

The FW-DICOMs from UBIware with the same codification of the patient will similarly be uploaded to the configurable solution interface with multi-PACS, Senoiris, to be evaluated. This program is the official one used by the radiologist, hence they won’t need to change their method of working and they all have access from their computers in the HUVH and, of course, it is configurable to create a project folder, define the checking structures to validate the images from each volunteer.

Persistent storage

The data will be kept at both HUVH and BSC for an additional amount of time after the project ends. The CEIm requires five years of storage of data after the study ends, keeping in mind that individual volunteers may force removal of their data, as their legitimate owners.

Recommended additional hardware

Besides the USCT device and controlling computer (or monitor and keyboard) the following items are *recommended* for the data management of the clinical study:

- 40 TB total storage at HUVH (in no less than 1TB individual units), with 1 Gbit ethernet and USB 3.2.
    - *e.g. QNAP TS-431X3-4G (700 EUR) + 4xWD Red Plus 3.5” 10 TB (4x260 EUR) for a total of 1800 EUR approx.*
- 1 x 5 TB HDD to transfer data BSC-HUVH, recommended USB 3.2.
    - *e.g. Seagate Game Drive FireCuda, (800 EUR).*
- 1x 1-Gigabit Ethernet (IEEE 802.3ae) switch with at least 4 connections
- 1 x (at least) 1.5 m Cat-5 ethernet cables (device to switch)
- 2 x short Cat-5 ethernet cables