SOLUS | SMART OPTICAL AND ULTRASOUND DIAGNOSTICS OF BREAST CANCER

Project title: Smart Optical and Ultrasound Diagnostics of Breast Cancer

Grant Agreement: 731877

Call identifier: H2020-ICT-2016-1

Topic: ICT-29-2016 Photonics KET 2016

Deliverable 8.3: Updated Data Management Plan

Leader partner:	Beneficiary 1, POLIMI
Author(s):	Andrea Farina
Work Package:	8
Estimated delivery:	M24
Actual delivery:	31 October 2018
Type:	Report
Dissemination level:	Public

Table of Contents

Abbreviations

DMP: Data Management Plan

US: Ultrasound

SW: Shear Wave

TRL: Technology Readiness Level

1. PREAMBLE

This document is the updated version of the Data Management Plan (DMP) submitted at the six-month (M6) as Deliverable 8.2. Main updates are related to:

- Dataset #2 Clinical Data, saved analysis.
- Detailed description of the database.
- POLIMI warehouse.

2. DATA SUMMARY

Provide a summary of the data addressing the following issues:

- **State the purpose of the data collection/generation**
- **Explain the relation to the objectives of the project**
- **Specify the types and formats of data generated/collected**
- **Specify if existing data is being re-used (if any)**
- **Specify the origin of the data**
- **State the expected size of the data (if known)**
- **Outline the data utility: to whom will it be useful**

This Data Management Plan (DMP) rules data management within the H2020 EU funded project "Smart optical and ultrasound diagnostics of breast cancer" (SOLUS – n.731877).

The project foresees the development of a novel combined optical / ultrasound / shear-wave instrument for discrimination of suspect breast lesions. Collected data will be exploited to understand the capability of the SOLUS prototype to discriminate malignant from benign breast lesions, using the optical data (absorption and reduced scattering coefficient of the suspect lesion) in combination with the ultrasound (US) information on lesion location and morphology and the elastography properties provided by shear waves (SW). More specifically, the collected data will be used to address two detailed objectives of the SOLUS project, that are:

1. Assess the multi-modal system performance in terms of sensitivity, spatial resolution and quantitation in laboratory settings (TRL4).
2. Validate the multi-modal system, demonstrating the feasibility and the advantages of the proposed multi-modal solution in real clinical settings (TRL5).

These two objectives will lead to two independent datasets that are:

- Phantom study (leading to deliverable D4.7 – Performance assessment of DOT with US priors – M36)
- Clinical study (leading to D2.5 – Database of lesion properties and composition – M48)

Data will be mainly originated by the SOLUS prototype, in terms of optical, US and SW measurements. Clinical information will be collected by the clinicians in charge of the study. Further, some data will be input by the operator related to the phantom properties or the measurement conditions.

In general, the data (in total or in part) could be re-used by partners of the SOLUS project during or after the project or external researchers for the following aims:

- validate novel analysis tools on phantoms and/or in vivo data
- exploit database of breast tissue optical properties
- exploit database of breast lesion optical properties.

Although the two datasets will pertain to different experiments, and could be reused in different context, yet we will establish a unified framework for the collection of data and their classification under a unique database. The key origin of data will be the SOLUS instrument, although addition of laboratory derived data is feasible. Details on the database structure are provided under section 3.1. In the following, we report the two tables describing the two datasets:

2.1. Dataset #1 – Phantom Data

SUBSET	ITEMS	TYPE	FORMAT	TOT SIZE
optical data				**2GB**
	raw data	histograms	XML+Base91	
US data				**1GB**
	US images	image	DICOM	
phantom data				**-**
	geometry	parameters	XML	
	nominal optical properties	parameters	XML	
instrument data				**-**
	Instrument Response Function	histogram	XML+Base91	
	responsivity	number	XML	
	power	number	XML	
	geometry	coordinates	XML	
measurement settings				**-**
	protocol settings (e.g. total acquisition, routing time)	parameters	XML	
analysis				**1.2GB**
	DOT tomographies	3D maps of μ_a, μ_s'	XML+Base91	
	US segmentation	region of interest	XML	
	recovered lesion properties	μ_a, μ_s' values	XML	

2.2. Dataset #2 – Clinical Data

SUBSET	ITEMS	TYPE	FORMAT	TOT SIZE
optical data				**4GB**
	raw data	histograms	XML+Base91	
US data				**2GB**
	US images	image	DICOM	
	lesion properties	descriptors	XLM	
SW data				**2GB**
	SW images	image	DICOM	
	lesion properties	descriptors	XLM	
lesion data				-
	clinical data (e.g. morphology, histology)	parameters / text	XML	
	classification	classifier	XML	
patient data				**-**
	demographic and personal data	parameters / text	XML	
instrument data				**-**
	Instrument Response Function	histogram	XML+Base91	
	responsivity	number	XML	
	power	number	XML	

	geometry	coordinates	XML	
measurement settings				**-**
	protocol settings (e.g. total acquisition, routing time)	parameters	XML	
analysis				**2.4G**
	DOT tomographies	3D maps of μ_a, μ_s’	XML+Base91	
		3D maps of chromophore concentrations and scattering parameters	XML+Base91	
	US segmentation	region of interest	XML	
	recovered lesion properties	μ_a, μ_s’ values	XML	
		chromophore concentrations and scattering parameters	XML	

3. FAIR DATA

3.1. Making data findable, including provisions for metadata:

- **Outline the discoverability of data (metadata provision)**
- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?**
- **Outline naming conventions used**
- **Outline the approach towards search keyword**
- **Outline the approach for clear versioning**
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how**

The two datasets will be identified using two unique identifiers (DOI) by uploading them onto a public repository. Data discoverability will be facilitated by adding a data description with keywords related to potential users (e.g. developers of new analysis tools), as described above.

For the Phantom dataset, different updated measurement sessions are possible depending on updated versions of the prototype. Conversely, for the Clinical dataset a single measurement session is foreseen, since there is no provision to recall back the same patient. Different versions of analysis are possible, depending on the update of the analysis tools. Therefore, the versioning will foresee a first number for the raw data acquisition (only for phantoms) and a second number for the analysis.

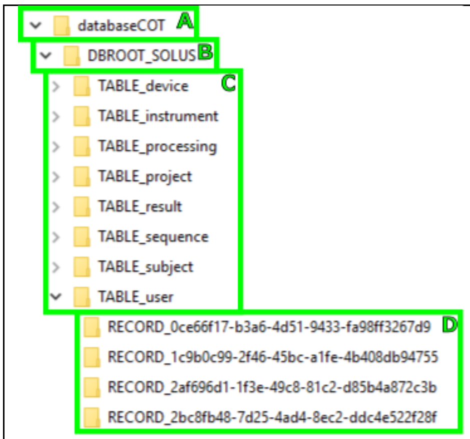
Apart from clinical images (e.g. US images) for which the DICOM standard is usually adopted, there are no specific standards for optical data. In general, we will create metadata files in XML, embedding large binary data in XML with Base91 encoding.

The software developed for the SOLUS project generates data from experiments. These data are the results of a sequence that acquires data on an instrument and then performs signal processing on them. An experiment takes place in a context that must be attached to the result to allow for a proper data management. All these data are categorized into eight classes:

Class	Description	Example
Project	Name of the project or the sub-project for which the experiment is done	SOLUS, SOLUS_characterization, SOLUS_clinical
User	Name of the operator during the experiment	Researcher during the characterization, radiologist during the clinical tests
Subject	The acquisitions are performed on the subject	Multimodal phantoms during the characterization, patient during the clinical tests
Instrument	Equipment used to acquire the data	Prototype #1

Device	Subsystems of an instrument	Bimodal probe #1, bimodal probe #2
Sequence	Set of instructions mixing acquisition and data processing	Sequence for phantom characterization, sequence for clinical tests
Processing	Algorithm for signal analysis developed during the project and available in the sequence	Ultrasound image segmentation, optical parameter estimation
Results of experiments	Set of acquisitions and processed signals	Results for characterization of phantom #3, results for patient ID31

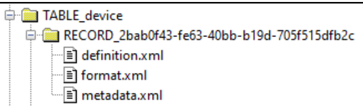
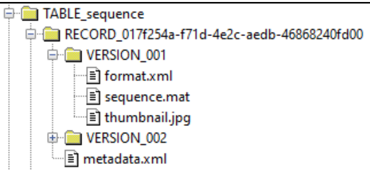
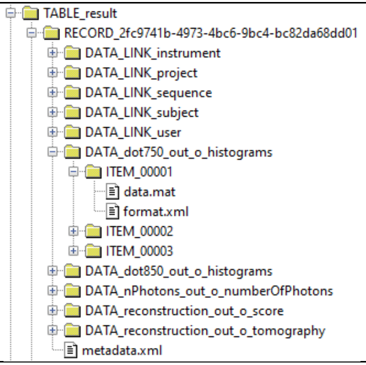
They are all organized as records (i.e. a container for data), which are stored in the eight tables (i.e. a container for records) of a database. In order to avoid a dependency to a third-party software, to make the data easy to read from an external program and to ease the data recovery as well as the custom use of the database, a dedicated solution has been developed. The database is written on the hard drive as a hierarchy of folders and files. It has the following structure:

 databaseCOT DBROOT_SOLUS TABLE_device TABLE_instrument TABLE_processing TABLE_project TABLE_result TABLE_sequence TABLE_subject TABLE_user RECORD_0ce66f17-b3a6-4d51-9433-fa98ff3267d9 RECORD_1c9b0c99-2f46-45bc-a1fe-4b408db94755 RECORD_2af696d1-1f3e-49c8-81c2-d85b4a872c3b RECORD_2bc8fb48-7d25-4ad4-8ec2-ddc4e522f28f	**Box A:** (top folder) installation folder **Box B:** (subfolder of A) root folder for the database named `SOLUS' (prefix DBROOT_) **Box C:** (subfolders of B) set of tables required for the software (prefix TABLE_) **Box D:** (subfolders of C) set of records for the selected table (prefix RECORD_)

Each table stores one of the eight type of data. The folder of a table has the prefix *TABLE_*. It contains the records and the file *metadata.xml* that describes the configuration.

The name of the folder of a record is *RECORD_key*. The key is a unique identifier that is automatically generated at the record creation (refer to java UUID for more information about the identifier). There are three types of records depending on data to save and policy for history:

	Single	**Version**	**Bag of data**
Number of data	One	One	On purpose
History	No	Yes	No
Tables	User, project, device, processing	Instrument, sequence, subject	Result
Common files	- **metadata.xml:** short description of the record - **format.xml:** definition the object that is saved - **misc:** file(s) or folder(s) containing the data (examples below: definition.xml, sequence.mat, data.mat)		
Thumbnail	One	One per version	One
Specific folders	None	**VERSION_###** contains the data for version number ###	**DATA_LINK_*** contains a pointer to another record of the database **DATA_x_out_y** contains the data of output named y that are generated by the item named x **ITEM_#####** is a subfolder of DATA_*. It contains the data for iteration #####

Structure	 TABLE_device RECORD_2bab0f43-fe63-40bb-b19d-705f515dfb2c definition.xml format.xml metadata.xml	 TABLE_sequence RECORD_017f254a-f71d-4e2c-aedb-46868240fd00 VERSION_001 format.xml sequence.mat thumbnail.jpg VERSION_002 metadata.xml	 TABLE_result RECORD_2fc9741b-4973-4bc6-9bc4-bc82da68dd01 DATA_LINK_instrument DATA_LINK_project DATA_LINK_sequence DATA_LINK_subject DATA_LINK_user DATA_dot750_out_o_histograms ITEM_00001 data.mat format.xml ITEM_00002 ITEM_00003 DATA_dot850_out_o_histograms DATA_nPhotons_out_o_numberOfPhotons DATA_reconstruction_out_o_score DATA_reconstruction_out_o_tomography metadata.xml

The open data will be organized closely to this internal software structure. However, the exact content of the open database as well as the release of software tools to ease the reading of these data is still subject to discussion.

3.2. Making data openly accessible:

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so**
- **Specify how the data will be made available**
- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?**
- **Specify where the data and associated metadata, documentation and code are deposited**
- **Specify how access will be provided in case there are any restrictions**

Data will be made "as open as possible, as closed as necessary". In this respect, all data described above will be made open apart from:

- algorithms for data analysis, which could be considered for IP protection
- personal data subject to privacy protection, accordingly to GDPR Regulation (EU) 2016/679, and ethical provisions.
- SW elastography images.

In particular, clinical data retrieved at Ospedale San Raffaele (OSR) will be directly saved without any personal data. As defined in Deliverable D5.1(“Definition of the clinical protocol”) the following data will be saved for scientific purposes:

- age
- menopausal status
- pathology results
- results from color Doppler
- BI-RADS score

Due to the fact that the possibility to share SW images as open-access is still ongoing, more details will be given in the final version of the DMP, due by month 48.

The software will automatically assign a numerical ID to each patient for data analysis, as described in section 3.1.

All specifications required to access the data will be inserted in the data repository. The segmentation of US images, and in general the extraction of optical properties for suspect lesions/inhomogeneities require advanced analysis tools, generally pertaining to the methods of inverse problems in diffuse optics. If already published or not involved in IP protections, the algorithms will be described in detail to permit replications. Inclusion of software tools for data processing will be considered if not causing significant overburden distracting important energies from the fulfilment of the project aims.

A three-phase process for data storage is foreseen. Initially, data will be collected by the SOLUS prototype and stored locally on the instruments, while other information will be gathered by clinicians and recorded on paper (as described in Deliverable D5.1). In the second phase, all collected data will be stored at

POLIMI data warehouse, apart from protected clinical information which will be retained at OSR. This will permit construction of the database and initial tests on analysis. In the third phase, when data acquisition is complete, data will be uploaded on an open repository. At present, the choice is for Zenodo, because of perfect match with requirements, and increased interest in the International community. Still final decision will be made close to the actual deposition (not earlier than month 36) to take into account the updated status of public repositories.

3.3. Making data interoperable:

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.**
- **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?**

The realm of clinical optical data is at present not covered by standards or specific vocabularies. The numerosity of the clinical study limits its potential use mainly to researchers and operators within the field. The definition of metadata and in particular the fields in the XLM will match the vocabularies most often covered by scientific publications in diffuse optics.

3.4. Increase data re-use (through clarifying licenses):

- **Specify how the data will be licenced to permit the widest reuse possible**
- **Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed**
- **Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why**
- **Describe data quality assurance processes**
- **Specify the length of time for which the data will remain re-usable**

A 6-12 months embargo after acceptance of relevant publications will be considered.

Data will be made available and reusable through open data repositories for periods of 10 years.

4. ALLOCATION OF RESOURCES

Explain the allocation of resources, addressing the following issues:

- **Estimate the costs for making your data FAIR. Describe how you intend to cover these costs**
- **Clearly identify responsibilities for data management in your project**
- **Describe costs and potential value of long term preservation**

Since data deposit in a local data warehouse and an external repository will not start earlier than 2 years from now, cost estimate will be performed at due time since policies and costs are rapidly changing under great internal and external pressure on data preservation and sharing. In general terms, it is highly probable that no extra-costs will be incurred for the storage of data since the overall dimension of data will be handled by standard POLIMI data facilities and fit in the free allowances of Zenodo repository.

Concerning the POLIMI warehouse, in September 2018 the Institution proposed to each Department the possibility of a support for data storage related to European projects. This data storage will be directly managed by the central ICT staff of POLIMI, providing high-level services in terms of backup, robustness, protection and restricted access. A request has been already sent to get access to that space. The feasibility of that opportunity will be definitely updated in the Final version of DMP due by month 48.

Dr Andrea Farina is responsible for the coordination of the overall data management.

5. DATA SECURITY

Address data recovery as well as secure storage and transfer of sensitive data

The second phase of data storage will be performed internally at a data warehouse of POLIMI and at OSR for protected clinical information. No access external to the consortium will be possible.

The actual data repository in force for the research group at POLIMI is located in secure hard-drives provided by a redundant system (RAID 5) that is backed up every week by an incremental back-up script (rsbackup) to other external servers. The data servers are located in the basement of the Physics Department of Politecnico di Milano in a restricted access area. The data servers have an access controlled by passwords, and they are part of a VLAN without access from outside the POLIMI institution. The VLAN at which not only the data servers are connected but all the PCs used for this project is part of an institutional network protected by a firewall. We note that the POLIMI group has a proven track-record in long-term data storage and access going back to the 80s.

In case POLIMI benefits of access to the institutional repository, the data stored will inherit the high-level security of the overall institutional network. More detail about the security level will be specified in the Final version of the DMP (M48).

In the final phase, the public repository will be chosen to grant requirements of long-term secure storage. The most probable choice - Zenodo - already fulfils all requirements.

Sensitive data - mostly personal data of the clinical study - will not be shared and will be stored only at OSR to comply with the privacy policies foreseen in the clinical protocol.

6. ETHICAL ASPECTS

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

The Clinical protocol (Deliverable 5.1 - Definition of the clinical protocol - produced at M3) and the ethical requirements in terms of protection of personal data (Deliverable D5.2 - Approval of clinical protocol by ethical committee - due at month 36) set specific requirements for anonymization of data and protection of personal data of patients. These requirements will be strictly followed and will prevent sharing of information.

All data stored at POLIMI data warehouse and deployed at public repository will be completely anonymized.

The patient information and consent will follow the guidelines set forth in ISO 14155 for patient information and informed consent, and will imply also sharing of data excluding sensitive data.

7. OTHER

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

At present, the main local procedures for data management are related to the requirements of sensitive data protection described in the clinical protocol (Deliverable D5.1) and operated by OSR. No other prescriptive procedures are identified so far.

8. CONCLUSIONS

In summary, we have shown that two independent datasets will be generated: i) Phantom data and ii) Clinical data. These two datasets will be organized into an XML-based database whose folders are saved on the hard drive of the clinical machine. Every measurement will be associated to an ID number for anonymization. A copy of this database will be placed in a local repository hosted at the Physics Department of Politecnico di Milano with restricted access to the consortium members. A request of space with restricted permission on the recent institutional repository has been sent, whose result will be discussed in the last version of the DMP, due by month 48.

The two datasets will be, eventually, uploaded for open-access on two independent public repositories on Zenodo, except for:

- algorithms for data analysis, which could be considered for IP protection.
- personal data subject to privacy protection, accordingly to GDPR Regulation (EU) 2016/679, and ethical provisions.
- SW elastography images.

The policy on SW elastography images is still under discussion, thus updates on that will be given at month 48 with the final version of the document.