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 8.2: First release of the Data Management Plan

Leader partner:	Beneficiary 1, POLIMI
Author(s):	Antonio Pifferi
Work Package:	8
Estimated delivery:	M6
Actual delivery:	28 April 2017
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 Data Management Plan (DMP) was generated using the DMP online tool provided by the “Digital Curation Centre” (www.dmponline.dcc.ac.uk) following the template for EU H2020 projects.

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 inserted 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 lesion optical properties
- exploit database of breast optical properties.

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	
	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	
	US segmentation	region of interest	XML	
	recovered lesion properties	μ_a, μ_s' values	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.

Naming conventions will be specified in a more advanced version of the DMP foreseen at month 24 of the SOLUS Project, and still before the actual data collection (starting after month 24).

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.

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 as foreseen in the clinical protocol (Deliverable D5.1) and ethical provisions.

Final decisions on these two aspects and specific identification of closed data, or data subject to specific embargo related to IP policies will be taken in the updated DMP at month 24. Related access policies will be defined at due time.

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 Ospedale San Raffaele. 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 taken close to the actual deposition (not earlier than m36) 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**

Licensing policies will be defined later (around M24) when the general dissemination, IP protection and exploitation policies are more clearly drawn.

Typically, a 6-12 months embargo after acceptance of relevant publications can 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.

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 perfomed internally at a data warehouse of POLIMI and at Ospedale San Raffaele 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 stored 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 POLIMI group has a proven track-record in long-term data storage and access going back to the 80s.

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 Ospedale San Raffaele 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 M36) 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 some part 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 Ospedale San

Raffaele. No other prescriptive procedures are identified so far. However, since local policies are rapidly evolving to cope with the increased demand for Open Data and Data Management, this section will be updated in a future release (M24) to describe the actual situation.