DELIVERABLE REPORT

Grant Agreement number: 688303

Project acronym: LUCA

Project title: Laser and Ultrasound Co-Analyzer for thyroid nodules

Funding Scheme: H2020-ICT-28-2015

Deliverable reported: D1.8 Data Management Plan

Due date: 30.04.2016

Name, title and organisation of partner: Katharina Krischak, EIBIR Gemeinnützige GmbH zur Förderung der Erforschung der biomedizinischen Bildgebung (EIBIR)

Project website address: www.luca-project.eu

Content

1) Introduction

This Data Management Plan (DMP) has been prepared by taking into account the template of the "Guidelines on Data Management in Horizon 2020"[1]. It presents details on the procedures of creating 'primary data' (data not available from any other sources) and their management.

The elaboration of the DMP will allow LUCA partners to address all issues related to data. However, the DMP will be a living document throughout the project. This initial version will evolve during the project and will be updated according to the progress of project activities.

The consortium will comply with the requirements of Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

Type of data, storage, recruitment process, confidentiality, ownership, management of intellectual property and access: The Grant Agreement and the Consortium Agreement are to be referred to for these aspects, particularly Articles 18, 23a, 24, 25, 26, 27, 29.3, 30, 31, 36, and 39 and "Annex I – Description of Action" of the Grant Agreement. The Grant Agreement was signed on November 10, 2015 while the Consortium Agreement was executed on December 18, 2015. The procedures that will be implemented for data collection, storage, access, sharing, protection, retention and destruction will be according to the requirements of the national legislation of each partner and in line with EU standards.

An ethical ethics-compliant approach will be adopted and maintained throughout the fieldwork process. The responsible partners will assure that the EU standards regarding ethics and data management are fulfilled.

2) Administrative Details

Project Name: Laser and Ultrasound Co-Analyzer for thyroid nodules

Project Identifier: LUCA

Funder: European Commission (Horizon 2020)

Grant number: 688303

Principal investigator and Project Data Contact: Prof. Turgut Durduran

Related policies: The members of the consortium agreed to follow the research data management policy of the partner institutions generating the data.

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

3) Data Management

This section will be subject to change as the project progresses, and reflects the current status within the consortium about the primary data that will be generated. The sub-sections below provide detailed information on the data sets, standards and metadata, and the respective data sharing and archiving and preservation procedures for the data sets collected at each partner institution:

a. Data sets collected at ICFO

Four types of data will be collected at ICFO:

1. "Component data": Design drawings (subsystems and LUCA system); (opto-) electronics board and component designs and specifications.
2. "Sub-system data": research laboratory data (test results of components), sub-systems and the LUCA system; research application data (dynamic range, sensitivity, repeatability, accuracy and other parameters defined in **WP4**).
3. "Evaluation data": Evaluation data which are the results from the end-user tests in clinics.
4. "Exploratory data": Exploratory data generated mainly within **WP5** by ICFO Knowledge & Technology Transfer unit and the Medical Optics group together (market, IP etc. analysis reports).

i. Data set descriptions

***What data** will be **generated or collected**?*	"Component data": ICFO group will be mainly in charge of the components related to diffuse correlation spectroscopy (DCS) sub-system. As such, we will generate design drawings, specifications and such for (a) source/laser, (b) detector, single photon counting avalanche photodiode, and (c) correlator unit. "Sub-system data": ICFO group will generate test results associated with components – electrical, optical, physical – and the DCS subsystem in its integrated form as a stand-alone system. DCS subsystem will be tested for its dynamic range (in intensity and in correlation decay times), sensitivity to small changes in scattered motion, repeatability over time and accuracy. Finally, the integrated LUCA system will be tested and we will focus on the DCS subsystem in its integrated form in the full LUCA platform. "Evaluation data": ICFO group will be involved in the evaluation of the data measured in the clinics by the end-users. ICFO group will be in charge of pre-processing, fitting, presentation and interpretation of the DCS data. "Exploratory data": ICFO Knowledge & Technology Transfer unit (ICFO-KTT) will work mainly with ICFO Medical Optics group but also with others to carry a market analysis, freedom-to-operate analysis and others. This data will be generated and managed at ICFO. We note that all these actions are collaborative and we expect significant overlaps and data sharing between partners.

*What is its **origin**?*	"Component data" and "Sub-system data" will be internal to the group and to the project. The measurements will be carried out at ICFO by ICFO. "Evaluation data" will be generated at IDIBAPS in close collaboration with IDIBAPS. "Exploratory data" will be generated at ICFO-KTT using external databases, studies and sources.
*What are its **nature, format and scale**?*	A wide range of data formats and scales will be generated. 1. Drawings and designs will use industry standard software and will, primarily, be confidential in nature. We will, as much as possible, generate publicly accessible versions for dissemination purposes. These will be stored in forward compatible, time-tested formats. Specifics will arise by M18. 2. Research application data on the testing of LUCA will follow non-standard formats common to each laboratory, in this case ICFO, doing the testing and will be stored in binary and text files. They will be associated with an electronic notebook which will include links to analysis scripts (Matlab, R, Excell, custom-software). The processed data will be saved in a report format and will be publicly available once cleared in terms of IP and exploitation issues by the appropriate committee in LUCA project as foreseen by the description of action. 3. Clinical data will be stored in electronic report forms, in formats that are to be designed and specified in LUCA tasks appropriate to the agreed rules on the system. The raw data will be associated with appropriate electronic notebooks , it will be anonymized as described in the ethical procedures, and parts pertaining to the identifiable patient information will be destroyed according to the ethical procedures and approvals that are due M24. This is a task of IDIBAPS. The processed data will be publicly available in summary as well as for individual subjects and shared through the LUCA web-site. Details will depend on the final system and the outputs that are tasks to be completed by M24. 4. Market analysis data will be confidential and will be shared within the consortium as reports and numbers. A summary will be published as part of the appropriate project deliverables. 5. Supporting data used in academic peer reviewed publications will be made available, after publication, via a recognised suitable data sharing repository (e.g. zenodo or national repository if available). This policy will be followed unless a partner or IEC can show that disseminating this data will compromise IP or other commercial advantage as detailed below. The project will use the metadata standards and requirements of the repository used for sharing the data. At the ICFO group, long-term access is ensured by the following measures: 1. Forward compatible, time-tested formats such as text files (comma-separated values, open-source formats such as R data-tables), and/or open-source binary formats (such as open document spreadsheets, open document text) and/or custom made binary formats (with definition files stored in standard text formats) will be utilized with associated descriptive documentation.

	2. All data will be stored in a secure hard-drive that is backed up every night by an incremental back-up script (rsbackup) to an external drive. Both drives are regularly replicated and upgraded at roughly three year intervals. 3. All desktop computers used by the ICFO personnel involved in the project is centrally managed by ICFO information technology (ICFO-IT) department which utilizes secure folders on the ICFO servers that are backed up automatically and internally by ICFO-IT. 4. All instrument control computers are kept outside internet and are "ghosted" after every major upgrade. "Ghost" copies are kept by ICFO-IT in open-source formats. 5. All electronic designs are stored, managed and accessed through the ICFO electronics workshop and are assigned unique identifiers. We note that ICFO Medical Optics group has a proven track-record in long-term data storage and access going back to the PI's earlier work from late 90s.
__To whom__ could it be __useful__?	"Component data": In the short-term, this type of data is only useful for the internal LUCA partners. In the medium-term, it will be useful for our other projects and some of these components are expected to become products. Some information may be used in scientific publications and presentations as described below. "Sub-system data" and "evaluation data" are useful both internally for our developments and upgrades but also for scientific publications. The data will be useful to the end-user community and the biophontonics community and will also be of interest to endocrinologists, the biomedical optics community, the ultrasonics community, radiologists, and biomedical engineers. "Exploratory data" is mainly useful internally and, in the medium-term, may be useful for industrial partners for exploitation purposes, e.g. for fund-raising. It will also be useful for future grant applications where higher TRL levels are foreseen.
Do __similar data sets exist__? Are there possibilities for __integration and reuse__?	This is a unique device and a data-set. There are possibilities to combine processed data for review papers on optics + ultrasound combinations in biomedicine as well as for reviews on applications of diffuse optics in cancer.

ii. Standards and metadata

How will the __data be collected/generated__?	"Component data" and "sub-system data" will be generated by laboratory tests using test equipment and using design software. "Evaluation data" will be generated mainly from ex vivo phantom measurements and by data acquired from the subjects. "Exploratory data" will be generated by studies of external databases, interviews with end-users and others. Details are described in the specific work-packages.
Which community	The lack of community data standards is one of the points that we explicitly

data standards or methodologies *(if any) will be used at this stage?*	discuss and attempt to contribute in LUCA project. Here, we mean the community of biomedical optics researchers using diffuse optical methods. Standards of a second community, the ICFO community, will be used. As mentioned above, there are standard methods internal to ICFO Medical Optics group, those handled by ICFO-IT, those handled by ICFO electronics workshop and those handled by ICFO-KTT.
How will the data be ***organised during the project?***	"Component data" and "sub-system data" generated by ICFO will follow a convention where the acronym of each component – stored at a shared bill-of-materials document -- , the date, the time will be used to uniquely identify the data set. Each data set will be associated with an electronic notebook kept in an open-source data format as described above. All software and main texts will be kept in a subversion repository managed by the ICFO-IT for version control. "Evaluation data" will follow the conventions defined jointly by IDIBAPS, HEMO and ECM who are the main drivers of the clinical studies and the final software suites. ICFO Group will follow their naming conventions.
Metadata *should be created to describe the data and aid discovery.* ***How will you capture this information?***	This will be captured in electronic notebooks, in header files in open-source format (described above) and in case-report files. The exact details are being defined as the systems mature.
Where will it be recorded?	All internal data will be kept according to the different units at ICFO and their standard practices. We will work collectively with the other LUCA partners to arrange the external data in standard formats. As explained above, every data-set is associated with an electronic notebook, appropriate header file and comments. These will be recorded in the storage system(s) described above.

iii. Data Sharing

Where and how *will the data be made available and* ***how can they be accessed****? Will you share data via a data repository, handle data requests directly or use another mechanism?*	Internal to the project, the ICFO data will be shared using generic cloud-storage (mainly Dropbox) wherever appropriate, e.g. when the shared data is not very sensitive or is incomprehensible for an intruder. Otherwise, it will be shared by encrypted files (PGP encryption) using ICFO's own cloud system that is managed by ICFO-IT. Brief reports, spreadsheets and such will be shared by the TEAMWORK framework set by EIBIR. Externally, we will use the project web-site as the main gateway for sharing data. We will post, after IP clearance, appropriate data sets alongside publications on journal web-sites.
To whom *will the data be made available?*	Bulk of the data will be widely accessible for end-users, however, there may be some data, such as market studies, IP portfolios that will be shared with entities and people related to the exploitation activities.
What are the	We will use the LUCA web-site for all dissemination. The processed data will

technical mechanisms for dissemination *and necessary* ***software*** *or other tools for enabling* ***re-use*** *of the data?*	be presented in a way that it is cross-platform and software independent to the best of our abilities. If some software or dataset we generate becomes of value for the general biomedical optics community, we will consider developing a unique web-site for this purpose.
Are any ***restrictions on data sharing*** *required and* ***why****?*	There will be restrictions based on the need for securing publications prior to public release and for exploitation purposes. These are defined in the project DOA.
What ***strategies*** *will you apply* ***to overcome or limit restrictions****?*	We will utilize procedures such as embargo until publication, anonymising and simplification.
Where (i.e. in which repository) *will the data be deposited?*	As mentioned above, there are no community defined standards for the biomedical diffuse optics community. Therefore, we will utilize the project website, possibly dedicated websites for specific outputs and journal websites.

iv. Archiving and preservation (including storage and backup)

What procedures will be put in place for ***long-term preservation of the data****?*	As described above and repeated below, there are set of procedures for ICFO generated data. At the ICFO group, Long-term access is ensured by the following measures: 1. Forward compatible, time-tested formats such as text files (comma-separated values, open-source formats such as R data-tables), and/or open-source binary formats (such as open document spreadsheets, open document text) and/or custom made binary formats (with definition files stored in standard text formats) will be utilized with associated descriptive documentation. 2. All data will be stored in a secure hard-drive that is backed up every night by an incremental back-up script (rsbackup) to an external drive. Both drives are regularly replicated and upgraded at roughly three year intervals. 3. All desktop computers used by the ICFO personnel involved in the project is centrally managed by ICFO information technology (ICFO-IT) department which utilizes secure folders on the ICFO servers that are backed up automatically and internally by ICFO-IT. 4. All instrument control computers are kept outside internet and are "ghosted" after every major upgrade. "Ghost" copies are kept by ICFO-IT in open-source formats. 5. All electronic designs are stored, managed and accessed through the ICFO electronics workshop and are assigned unique identifiers. We note that ICFO Medical Optics group has a proven track-record in long-

	term data storage and access going back to the PI's earlier work from late 90's.
How long will the data be preserved *and what will its* ***approximated end volume*** *be?*	Apart from the certain aspects of the clinical datasets, which will be managed by IDIBAPS, there are no limitations on the preservation of the data. We will follow academic standards and aim for a ten year preservation of the data. As mentioned above, PI is able to access, re-use and re-analyse data from late 90s. The approximate end-volume of this data will be less than one terabyte.
Are ***additional resources and/or is specialist expertise*** *needed?*	No. We are all experts in the management of datasets of this size. Internally, ICFO-IT manages the general policies, makes suggestions on good-practices and ensures security against intrusions.
Will there be any ***additional costs*** *for archiving?*	The costs are budgeted within the project and internally.

b. Data sets collected at POLIMI

Three types of data will be collected by POLIMI:

1. "Component data": specification and designs of laser sources, detectors and timing electronics, including the electronic boards for operating them.
2. "Sub-system data": research laboratory data (test results of components), sub-systems and the LUCA system; research application data (dynamic range, sensitivity, repeatability, accuracy and other parameters defined in **WP4**).
3. "Evaluation data": Evaluation data that are the results from the end-user tests in clinics.

i. Data set descriptions

What data *will be* ***generated or collected****?*	"Component data": POLIMI will be in charge of the components related to time-resolved spectroscopy (TRS) sub-system. As such, POLIMI will generate specifications and design drawings for (a) laser sources, (b) detectors, namely SPAD (Single-Photon Avalanche Diodes) or SiPMs (Silicon PhotoMultipliers), and (c) timing electronics (TDC, Time-to-Digital Converter). "Sub-system data": POLIMI will generate test results associated with components – electrical, optical, physical – and the TRS subsystem in its integrated form as a stand-alone system. TRS subsystem will be tested for performances assessment. Finally, the integrated LUCA system will be tested and we will focus on the TRS subsystem in its integrated form in the full LUCA platform. "Evaluation data": POLIMI will be involved in the evaluation of the data measured in the clinics by the end-users, in particular for pre-processing, fitting, presentation and interpretation of the TRS data. We note that all these actions are collaborative and we expect significant overlaps and data sharing between partners.

*What is its **origin**?*	"Component data" and "Sub-system data" will be generated within the group and the project. The measurements will be carried out at POLIMI by POLIMI. "Evaluation data" will be generated at IDIBAPS.
*What are its **nature, format and scale**?*	A wide range of data formats and scales will be generated. 1. Drawings and designs will use industry standard software and will, primarily, be confidential in nature. We will, as much as possible, generate publicly accessible versions for dissemination purposes. These will be stored in forward compatible, time-tested formats. Specifics will arise by M18. 2. Research application data on the testing of LUCA will follow non-standard formats common to each laboratory, in this case POLIMI, doing the testing and will be stored in binary and text files. Matlab/Excell script will be provided for the reading of these files. The processed data will be saved in a report format and will be publicly available once cleared in terms of IP and exploitation issues by the appropriate committee in LUCA project as foreseen by the description of action. 3. Clinical data will be stored in electronic report forms, in formats that are to be designed and specified in LUCA tasks appropriate to the agreed rules on the system. The raw data will be anonymized as described in the ethical procedures, and parts pertaining to the identifiable patient information will be destroyed according to the ethical procedures and approvals that are due M24. This is a task of IDIBAPS. The processed data will be publicly available in summary as well as for individual subject s and shared through the LUCA web-site. Details will depend on the final system and the outputs that are tasks to be completed by M24. 4. Supporting data used in academic peer reviewed publications will be made available, after publication, via a recognised suitable data sharing repository (e.g. zenodo or national repository if available). This policy will be followed unless a partner or IEC can show that disseminating this data will compromise IP or other commercial advantage as detailed below. The project will use the metadata standards and requirements of the repository used for sharing the data. At POLIMI, Long-term access is ensured by the following measures: 1. Forward compatible, time-tested formats such as text files (comma-separated values, and/or open-source binary formats (such as open document spreadsheets, open document text) and/or custom made binary formats (with definition files stored in standard text formats) will be utilized with associated descriptive documentation. 2. All data will be 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 and DEIB 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. 3. All instrument control computers are kept outside internet.

	4. All electronic designs are stored, managed and accessed through the POLIMI electronics workshops and are assigned unique identifiers. We note that POLIMI group has a proven track-record in long-term data storage and access going back to 80's.
***To whom** could it be **useful**? Does it underpin a scientific publication?*	"Component data": In the short-term, this type of data is only useful for the internal LUCA partners. In the medium-term, it will be useful for our other projects and some of these components are expected to become products. Some information may be used in scientific publications and presentations as described below. "Sub-system data" and "evaluation data" are useful both internally for our developments and upgrades but also for scientific publications. We submit articles to target journals for the end-user community (i.e. .Journal of Clinical Endocrinology and Nutrition, European Journal of Endocrinology, Clinical Endocrinology and Thyroid in Endocrinology field, and Radiology, European Journal of Radiology and American Journal of Radiology in Radiology field) and for the biophotonics community (e.g. Biophotonics, Applied Optics, Biomedical Optics Express, Journal of Biomedical Optics, Nature Photonics). This is a multidisciplinary project and we expect that the range of journals will expand as the project progresses and may include endocrinology, biomedical optics, ultrasonics, radiology, biomedical engineering and others.
*Do **similar data sets exist**? Are there possibilities for **integration and reuse**?*	This is a unique device and a data-set. There are possibilities to combine processed data for review papers on optics+ultrasound combinations in biomedicine as well as for reviews on applications of diffuse optics in cancer.

ii. Standards and metadata

*How will the **data be collected/generated**?*	"Component data" will be generated by laboratory tests using test equipment and using design software. "Subsystem data" will be generated mainly from ex vivo phantom measurements and by data acquired from the subjects.
*Which community **data standards or methodologies** (if any) will be used at this stage?*	The lack of community data standards is one of the points that we explicitly discuss and attempt to contribute in LUCA project. Here, we mean the community of biomedical optics researchers using diffuse optical methods. POLIMI have already experienced other EU multidisciplinary projects during which exchange of data with different formats was crucial. Standard Matlab scripts were prepared in order to read data from the POLIMI format and convert them into other formats.
*How will the data be **organised during the project?***	"Component data" generated by POLIMI will be stored in folders and files within a root folder (whose name is the projects's one, "LUCA") that will contain all the information concerning the project. Each component will have a dedicated folder and the various releases of the component data will have a progressive numbering.

	"Sub-system data" generated by POLIMI will follow the standard convention applied by the Biomedical Optics Group, where the files are stored in a folders with the name of the project, and organized in subfolders indicating the different experiments/WP activities. The name of the files is composed of three parts: a three letter identifier to indicate the experiments/activity, a letter indicating the nature of the file (e.g. "m" in-vivo experimental measurement, "p" phantom measurement, "s" instrument response function measurement) and a progressive number. In the header of the file all the other information useful for the univocal identification of the data set are stored. An extensive description of the experiment and each file details are also written in the logbook of the laboratory involved. "Evaluation data" will follow the conventions defined jointly by IDIBAPS, HEMO and ECM who are the main drivers of the clinical studies and the final software suites. POLIMI Group will follow their naming conventions.
Metadata** should be created to describe the data and aid discovery. **How will you capture this information?	Metadata will be captured in text files describing how the data are stored in files and folders, how and when the data have been collected, the importance of the data, etc.
Where will it be recorded?	All internal data will be kept according to the different units at POLIMI and their standard practices. We will work collectively with the other LUCA partners to arrange the external data in standard formats. These will be recorded in the storage system(s) described above. Additionally, the data will be stored also in laptop and desktop computers routinely used in laboratory activities.

iii. Data Sharing

***Where and how** will the data be made available and **how can they be accessed**? Will you share data via a data repository, handle data requests directly or use another mechanism?*	Internal to the project, the POLIMI data will be shared using cloud-storage systems (such as OneDrive) via encrypted files. Brief reports, spreadsheets and such will be shared by the TEAMWORK framework set by EIBIR. Externally, we will use the project web-site as the main gateway for sharing data. We will post, after IP clearance, appropriate data sets alongside publications on journal web-sites.
***To whom** will the data be made available?*	Data describing the details of the developed components will be restricted only to the partner of the consortium working on connected topics. General data describing the performance of the developed components and how to exploit them will be widely accessible.

*What are the **technical mechanisms for dissemination** and necessary **software** or other tools for enabling **re-use** of the data?*	We will use the LUCA web-site for dissemination. The processed data will be presented in a way that it is cross-platform and software independent to the best of our abilities. If some software or dataset that we generate becomes of value for the general biomedical optics community, we will consider developing a unique web-site for that purposes.
*Are any **restrictions on data sharing** required and **why**?*	There will be restrictions based on the need for securing publications prior to public release and for exploitation purposes. These are defined in the project DOA. Furthermore, any patient data that could be used to identify the patients will be properly anonymized prior to sharing and the link between the patient ID and the dataset will be permanently destroyed after an appropriate time based on the ethical protocols and procedures that are approved. This is IDIBAP's responsibility and the POLIMI group will receive data that is already anonymized according to these principles.
*What **strategies** will you apply **to overcome or limit restrictions**?*	We will utilize procedures such as embargo until publication.
***Where (i.e. in which repository)** will the data be deposited?*	As mentioned above, there are no well-established community defined standards for the biomedical diffuse optics community. Therefore, we will utilize project web-site, possibly dedicated web-sites for specific outputs and journal web-sites.

iv. Archiving and preservation (including storage and backup)

*What procedures will be put in place for **long-term preservation of the data**?*	As mainly described above and repeated below: Long-term access is ensured by the following measures: 1. Forward compatible, time-tested formats such as text files (comma-separated values, and/or open-source binary formats (such as open document spreadsheets, open document text) and/or custom made binary formats (with definition files stored in standard text formats) will be utilized with associated descriptive documentation. 2. All data will be 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 and DEIB 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. 3. All instrument control computers are kept outside internet. 4. All electronic designs are stored, managed and accessed through the

	POLIMI electronics workshops and are assigned unique identifiers. We note that POLIMI group has a proven track-record in long-term data storage and access going back to 80s.
How long will the data be preserved *and what will its* ***approximated end volume*** *be?*	Apart from the certain aspects of the clinical datasets – which will be managed by IDIBAPS, there are no limitations on the preservation of the data. We will follow academic standards and aim for a ten year preservation of the data. As mentioned above, POLIMI group is able to access, re-use and re-analyse data from early 90s. The approximate end-volume of this data will be less than one terabyte.
Are ***additional resources and/or is specialist expertise*** *needed?*	No. We are all experts in the management of datasets of this size. Internally, POLIMI IT managers make suggestions on good practices and ensure security against intrusions.
Will there be any ***additional costs*** *for archiving?*	The costs are budgeted within the project and internally.

c. Data sets collected at IDIBAPS

Two types of data will be collected at IDIBAPS:

1. "Clinical data": Clinical data that is related to healthy volunteers and patients included as participants in **WP5**.
2. "Evaluation data": Evaluation data that are the results from the end-user tests in clinics.

i. Data set descriptions

What data *will be* ***generated or collected****?*	"Clinical data": IDIBAPS will be involved in the recruitment of healthy volunteers and patients included in the pilot study as participants in **WP5**. Data will be related to medical history, physical examination, laboratory and ultrasound parameters. "Evaluation data": Evaluation data that are the results from the end-user tests in clinics. We note that all these actions are collaborative and we expect significant overlaps and data sharing between partners.
What is its ***origin****?*	All data will be generated within the project. Some will reflect the confidential know-how of an individual partner; others will be generated in collaboration. "Clinical data" will be generated at IDIBAPS. Data storage will be performed maintaining the anonymity of volunteers and following current legislation. No biological samples related to the study will be stored. Once analyzed samples collected will be destroyed according to the existing protocols in the CDB (Centre de Diagnòstic Biomèdic) of Hospital Clinic of Barcelona. The encoding list will be destroyed once all the participants are measured with LUCA device

	and data is analyzed, to be sure no extra information is required. "Evaluation data" will be generated at IDIBAPS.
What are its ***nature, format and scale****?*	A wide range of data formats and scales will be generated. 1. Research application data on the testing of LUCA will follow non-standard formats common to each laboratory doing the testing and will be stored in binary and text files. They will be associated with an electronic notebook which will include links to analysis scripts (Matlab, R, Excell, custom-software). The processed data will be saved in a report format and will be publicly available once cleared in terms of IP and exploitation issues by the appropriate committee. 2. Clinical data: regarding to personal data, the standard regulatory guidelines will be followed at the national and international level: Spanish law and Directive 95/46/EC of the European Union, on protection of personal data. The only sensitive data that will be collected and/or processed are related to health and ethnicity. A database will be created with the variables of interest of the participants, both volunteers and patients. This database is only available to a member of the Hospital Clínic (Dr. Mireia Mora). The variables collected to register and treat patients' vital information will be included in another database associated to the code number of the participant. These variables include: name, date of birth and medical record number. This database will only be available to the members of the Hospital Clinic, since it is responsible for the clinical patients in routine clinical practice. The other members of the project will not have the data of the participants, only the code number assigned coding and the study variables for their analysis. It is not expected that the immediate results of this research project carry out important ethical implications. 3. Evaluation data will be stored in electronic report forms, in formats that are to be designed and specified in LUCA tasks appropriate to the agreed rules on the system. The raw data will be associated with appropriate electronic notebooks , it will be anonymized as described in the ethical procedures, and parts pertaining to the identifiable patient information will be destroyed according to the ethical procedures and approvals that are due M24. The processed data will be publicly available in summary as well as for individual subject s and shared through the LUCA web-site. Details will depend on the final system and the outputs that are tasks to be completed by M24. 4. Conformity data will be generated and stored according to the industry standards and will be mainly public. It will be shared as a report. 5. Market analysis data will be confidential and will be shared within the consortium as reports and numbers. A summary will be published as part of the appropriate project deliverables. 6. Supporting data used in academic peer reviewed publications will be made available, after publication, via a recognised suitable data sharing repository (e.g. zenodo or national repository if available). This policy will be followed unless a partner or IEC can show that disseminating this data will compromise IP or other commercial advantage as detailed below. The project will use the metadata standards and requirements of the repository used for sharing the data.

***To whom** could it be **useful**? Does it underpin a scientific publication?*	"Clinical data" (anonymized) and "evaluation data" are useful both internally for our developments and upgrades but also for scientific publications. The data will be interesting to the end-user community, the biophontonics community, endocrinotlogists, the biomedical optics community, the ultrasonics community, radiologists, and biomedical engineers. "Exploratory data" is mainly useful internally and, in the medium-term, may be useful for industrial partners for exploitation purposes, e.g. for fund-raising. It will also be useful for future grant applications where higher TRL levels are foreseen.
*Do **similar data sets exist**? Are there possibilities for **integration and reuse**?*	This is a unique device and a data-set. There are possibilities to combine processed data for review papers on optics + ultrasound combinations in biomedicine as well as for reviews on applications of diffuse optics in cancer.

ii. Standards and metadata

*How will the **data be collected/generated**?*	"Clinical data" will be collected from healthy volunteers and patients that will agree to participate. Healthy participants will be selected among those who have participated in previous work on thyroid with diffuse optics. They will be asked if they want to participate again in this project, completely voluntary. Patients will be selected from those who are followed by the endocrinology department of the Hospital Clinic of Barcelona and because of the condition will be surgically treated with total thyroidectomy. Data will be generated from to medical history, physical examination, laboratory and ultrasound parameters that will be obtained from the clinical practice. "Evaluation data" will come from subjects measurements with the LUCA device. "Exploratory data" will be generated by studies of external databases, interviews with end-users and others. Details are described in the specific work-packages.
*Which community **data standards or methodologies** (if any) will be used at this stage?*	Data storage, and where applicable sharing, will be performed maintaining the anonymity of volunteers and following current legislation. No biological samples related to the study will be stored. Once analyzed samples collected will be destroyed according to the existing protocols in the CDB (Centre de Diagnòstic Biomèdic) of Hospital Clinic of Barcelona. Data pertaining to this study, both clinical, laboratory and imaging, are not included in the conventional medical story, they will be included in a separate file in a protected place. Medical images, such as ultrasounds and MRIs will be stored in a storage system for images called PACS that allows you to store and transfer images in DICOM format.
*How will the data be **organised during the project?***	"Clinical data" will follow the standard procedure in accordance with the guidelines outlined in the Declaration of Helsinki and complies with the national legislation currently in effect in Spain, specifically, the Law of Biomedical Research (*Ley de Investigación Biomédica*) enacted in 2007. This law regulates the ethical evaluation of research projects in Spain that involve human subjects, and it designates and authorizes the local Clinical Research

	Ethics Committees for the review of all types of research projects involving humans, as well as when handling personal data. In this sense, the LUCA study in Spain will fulfill all national and European ethical requirements. Participants will be codified using "LUCA" followed by the "CO" for controls and "CA" for cases and followed by number established by the order of evaluation, for example, LUCA_CO_1, LUCA_CO_2, LUCA_CA_1… All the information obtained and written in the clinical protocol will be introduced in the database using both categorical and numeric variables as suitable. Excell and SPSS database will be used with restricted access. "Evaluation data" will follow the conventions defined jointly by IDIBAPS, HEMO and ECM who are the main drivers of the clinical studies and the final software suites.
Metadata** should be created to describe the data and aid discovery. **How will you capture this information?	This will be captured in electronic notebooks, in header files in open-source format (described above) and in case-report files. The exact details are being defined as the systems mature.
Where will it be recorded?	All internal data will be kept according to the different units at IDIBAPS and their standard practices. We will work collectively with the other LUCA partners to arrange the external data in standard formats. As explained above, every data-set is associated with an electronic notebook, appropriate header file and comments. These will be recorded in the storage system(s) described above.

iii. Data Sharing

***Where and how** will the data be made available and **how can they be accessed**? Will you share data via a data repository, handle data requests directly or use another mechanism?*	Data storage, and where applicable sharing, will be performed maintaining the anonymity of volunteers and following current legislation. No biological samples related to the study will be stored. Once analyzed samples collected will be destroyed according to the existing protocols in the CDB (Centre de Diagnòstic Biomèdic) of Hospital Clinic of Barcelona. The encoding list will be destroyed once all the participants are measured with LUCA device and data is analyzed, to be sure no extra information is required. At the latest, this will take place upon the completion of the project. The realization of this project will involve the voluntary participation of unpaid volunteers. Any use of data or samples follows local regulations, and international, especially: Declaration of Helsinki (World Medical Association), as amended in 2000, European Convention on Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine (Oviedo, April 1997). The partners involved in these aspects are committed to reporting all aspects of the studies to the project committees. This includes written informed consent documentation, part of the protocol for human research studies. Supporting data used in academic peer reviewed publications will be made available, after publication, via a recognised suitable data sharing repository (e.g. zenodo or national repository if available). This policy will be followed

	unless a partner or IEC can show that disseminating this data will compromise IP or other commercial advantage. The project will use the metadata standards and requirements of the repository used for sharing the data. Brief reports, spreadsheets and such will be shared via the project internal collaboration platform Teamwork. Externally, we will use the project website as the main gateway for sharing data approved for dissemination. We will post, after IP clearance, appropriate data sets alongside publications on journal web-sites.
***To whom** will the data be made available?*	We aim to make bulk of the data widely accessible; however, there may be some data, such as market studies, IP portfolios that will be shared with entities and people related to the exploitation activities. Clinical data of subjects will be internal in IDIBAPS and will not be shared.
*What are the **technical mechanisms for dissemination** and necessary **software** or other tools for enabling **re-use** of the data?*	We will use the LUCA website for all dissemination. The processed data will be presented in a way that it is cross-platform and software independent to the best of our abilities. If some software or dataset that we generate becomes of value for the general biomedical optics community, we will consider developing a unique web-site for that purposes.
*Are any **restrictions on data sharing** required and **why**?*	There will be restrictions based on the need for securing publications prior to public release and for exploitation purposes. These are defined in the project DOA. Furthermore, any patient data that could be used to identify the patients will be properly anonymized prior to sharing and the link between the patient ID and the dataset will be permanently destroyed after an appropriate time based on the ethical protocols and procedures that are approved. This is IDIBAPS responsibility and the ICFO group will receive data that is already anonymized according to these principles.
*What **strategies** will you apply **to overcome or limit restrictions**?*	We will utilize procedures such as embargo until publication, anonymising and simplification.
***Where (i.e. in which repository)** will the data be deposited?*	As mentioned above we will utilize project web-site, possibly dedicated web-sites for specific outputs and journal web-sites. Within the LUCA consortium we will use the project management platform Teamwork where data files can be up- and downloaded in folders organised by WP and/or specific topics with a version management and the possibilities to restrict the access and add tags.

iv. Archiving and preservation (including storage and backup)

*What procedures will be put in place for **long-term preservation of***	Data storage will be performed maintaining the anonymity of volunteers and following current legislation. No biological samples related to the study will be stored. Once analyzed samples collected will be destroyed according to the existing protocols in the CDB (Centre de Diagnòstic Biomèdic) of Hospital Clinic

***the data**?*	of Barcelona. The encoding list will be destroyed once all the participants are measured with LUCA device and data is analyzed, to be sure no extra information is required. At the latest, this will take place upon the completion of the project. Any use of data or samples follows local regulations, and international, especially: Declaration of Helsinki (World Medical Association), as amended in 2000, European Convention on Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine (Oviedo, April 1997). The partners involved in these aspects are committed to reporting all aspects of the studies to the project committees. This includes written informed consent documentation, part of the protocol for human research studies. Regarding to personal data, the standard regulatory guidelines will be followed at the national and international level: Spanish law and Directive 95/46/EC of the European Union, on protection of personal data. The only sensitive data that will be collected and/or processed are related to health and ethnicity. A database will be created with the variables of interest of the participants, both volunteers and patients. This database is only available to a member of the Hospital Clínic (Dr. Mireia Mora). The variables collected to register and treat patients' vital information will be included in another database associated to the code number of the participant. These variables include: name, date of birth and medical record number. This database will only be available to the members of the Hospital Clinic, since it is responsible for the clinical patients in routine clinical practice. The other members of the project will not have the data of the participants, only the code number assigned coding and the study variables for their analysis. It is not expected that the immediate results of this research project have important ethical implications. Data pertaining to this study, both clinical, laboratory and imaging, are not included in the conventional medical story, they will be included in a separate file in a protected place. Medical images, such as ultrasounds and MRIs will be stored in a storage system for images called PACS that allows you to store and transfer images in DICOM format.
***How long will the data be preserved** and what will its **approximated end volume** be?*	According to the Biomedical Investigation Law, there is no need to preserve the data. However, we aim for at least five year preservation of the data. The approximate end-volume of this data will be less than one terabyte.
*Are **additional resources and/or is specialist expertise** needed?*	No. We are all experts in the management of datasets of this size.
*Will there be any **additional costs** for archiving?*	The costs are budgeted within the project and internally.

d. Data sets collected at HEMO

Four types of data will be collected at HemoPhotonics:

1. "Component data": Design drawings (subsystems and LUCA system); Firmware and Software for micro-controllers etc.; (opto-) electronics boards; component designs and specifications.
2. "Sub-system data": laboratory evaluation data (test results of components) for sub-systems and the LUCA system; device application data (dynamic range, sensitivity, repeatability, accuracy and other parameters defined in **WP4**); compliance testing and documentation.
3. "Evaluation data": Evaluation data that are the results from the end-user tests in clinics.
4. "Exploratory data": Exploratory data generated mainly within exploitation plan (market reports; market & IP strategy, IP analysis reports etc.).

i. Data set descriptions

***What data** will be **generated or collected**?*	"Component data": HemoPhotonics will mainly provide or contribute to components related to diffuse correlation spectroscopy (DCS) sub-system, develop internal control electronics, specific firmware, as well as operation and control software. We will therefore generate schematic and design drawings, software and firmware codes, application documentation, specifications etc. "Sub-system data": HemoPhotonics will generate or contribute to test results associated with components – electrical, optical, physical – and the DCS subsystem in its integrated form as a stand-alone system. Furthermore HemoPhotonics will perform and document functional and compliance tests on the sub-system level as well as for the integrated LUCA system. "Evaluation data": HemoPhotonics will be involved to some aspects of evaluation of the data measured in the clinics by the end-users. In particular, HemoPhotonics will generate evaluation code for optical data in collaboration with ICFO and POLIMI for the LUCA device implementation based on clinical evaluations. Furthermore, end-user feedback e.g. on usability of the LUCA device in clinical settings will be collected. "Exploratory data": In collaboration mainly with ICFO and the industrial partners, HemoPhotonics will provide contributions to the exploitation aspects of LUCA like market analysis, exploitation strategy, freedom-to-operate analysis etc.
*What is its **origin**?*	"Component data" will be generated by HemoPhotonics. "Sub-system data" will be generated by HemoPhotonics, ICFO, POLIMI, VERMON, ECM "Evaluation data" will be generated at IDIBAPS in collaboration with ICFO. Specific evaluation code to be developed for implementation in the LUCA system will be generated by HemoPhotonics. "Exploratory data" will be mainly generated at ICFO-KTT using external databases, studies and sources.
*What are its **nature, format***	A wide range of data formats and scales will be generated.

***and scale**?*	1. Drawings and designs will use industry standard software and will, primarily, be confidential in nature. As much as possible, publicly accessible versions will be generated for dissemination purposes. These will be stored in forward compatible, time-tested formats. Specifics will arise by M18. 2. Software and firmware code will be developed in standard development suites for C++, VHDL on a dedicated computer system. Codes will be confidential. 3. Application data on the testing of LUCA will follow non-standard formats in binary and text files and evaluated with internal scripts based on common software tools (Excel, Matlab, etc.) on dedicated computer system. The processed data will be saved in a report format and will be publicly available once cleared in terms of IP and exploitation issues by the appropriate committee in LUCA project as foreseen by the description of action. 4. Exploitation strategy, market analysis, freedom-to-operate analysis etc. data will be confidential and will be shared within the consortium as reports and numbers. A summary will be published as part of the appropriate project deliverables. Long-term access is ensured by the following measures: 1. All data will be stored in a secure hard-drive that is backed up bi-weekly to an external drive. Both drives will be regularly replicated and upgraded at roughly three year intervals. 2. All developed intermediate and released firmware and software code will be stored under proper consecutive version assignments. 3. All mechanical and electronic design files will be stored, managed with assignment of unique identifiers.
***To whom** could it be **useful**? Does it underpin a scientific publication?*	"Component data": In the short-term, this type of data is only useful for the internal LUCA partners. In the medium-term, it will be useful for our other projects and when some of these components might become products. "Sub-system data" and "evaluation data" are useful internally for our developments and upgrades. They may support occasionally scientific publications. "Exploratory data" is mainly useful internally and, in the medium-term for product exploitation as well as e.g. fund raising purposes addressing higher technology readiness levels.
*Do **similar data sets exist**? Are there possibilities for **integration and reuse**?*	This is a unique device and a data-set.

ii. Standards and metadata

*How will the **data be collected/genera***	"Component data" and "sub-system data" will be generated by laboratory tests using test equipment, using design and development software.

***ted**?*	"Evaluation data" will be generated mainly from ex vivo phantom measurements and by data acquired from the subjects. "Exploratory data" will be generated by studies of external databases, interviews with end-users and others. Details are described in the specific work-packages.
*Which community **data standards or methodologies** (if any) will be used at this stage?*	Community data standards in this area of research do presently not exist but the LUCA project attempts to contribute to future standardization.
*How will the data be **organised during the project?***	"Component data" and "sub-system data" generated by HemoPhotonics will follow a convention where the acronym of each component – stored at a shared bill-of-materials document -- , the date, the time will be used to uniquely identify the data set. All software and main texts will be kept in a subversion repository. "Evaluation data" will follow the conventions defined jointly by IDIBAPS, HemoPhotonics and ECM who are the main drivers of the clinical studies and the final software suites.
Metadata** should be created to describe the data and aid discovery. **How will you capture this information?	This will be captured in header files in open-source format. The exact details are being defined as the systems mature.
Where will it be recorded?	Every data-set is associated with an electronic notebook, appropriate header file and comments and will be recorded in the storage system described above.

iii. Data Sharing

***Where and how** will the data be made available and **how can they be accessed**? Will you share data via a data repository, handle data requests directly or use another*	Internal to the project, the HemoPhotonics data will be shared using generic cloud-storage (mainly Dropbox) wherever appropriate, e.g. when the shared data is not sensitive or incomprehensible for outsiders. Brief reports, spreadsheets and such will be shared by the Teamwork framework set by EIBIR. Externally, we will use the project web-site as the main gateway for sharing data. We will post, after IP clearance, appropriate data sets alongside publications on journal web-sites.

mechanism?	
***To whom** will the data be made available?*	Apart of dissemination related activities of WP6, most of HemoPhotonics generated data is restricted to internal use.
*What are the **technical mechanisms for dissemination** and necessary **software** or other tools for enabling **re-use** of the data?*	We will use the LUCA web-site for all dissemination. The processed data will be presented in a way that it is cross-platform and software independent to the best of our abilities.
*Are any **restrictions on data sharing** required and **why**?*	For most of the data generated by HemoPhotonics, restrictions on device technology (hardware and software) are required to allow a successful exploitation of the developments in future products.
*What **strategies** will you apply **to overcome or limit**.*	Where appropriate, IP protection measure will be implemented.
***Where (i.e. in which repository)** will the data be deposited?*	Where appropriate, we will use the project website to make data available. Supplementary data will be accessible in publications available on journal websites and the project website.

iv. Archiving and preservation (including storage and backup)

*What procedures will be put in place for **long-term preservation of the data**?*	As described above, to ensure long-term access, HemoPhotonics will implement the following measures: 1. Forward compatible, time-tested formats such as text files (comma-separated values, open-source formats), and/or open-source binary formats (such as open document spreadsheets, open document text) and/or custom made binary formats (with definition files stored in standard text formats) will be utilized with associated descriptive documentation. 2. Codes are based on standard languages with long-term availability (e.g. C++, VHDL). 3. All data will be stored in a secure hard-drive that is backed up bi-weekly to an external drive. Both drives are regularly replicated and upgraded at roughly three year intervals.

	4. All mechanical and electronic designs are stored and assigned unique identifiers.
How long will the data be preserved *and what will its* ***approximated end volume*** *be?*	We aim for a ten year preservation of the data. The approximate end-volume of this data will be less than one terabyte.
Are ***additional resources and/or is specialist expertise*** *needed?*	No.
Will there be any ***additional costs*** *for archiving?*	The costs are budgeted within the project and internally.

e. Data sets collected at VERMON

Three types of data will be collected at VERMON:

1. "Component data": Design drawings of the probe; components, mechanical parts and specifications.
2. "Sub-system data": research laboratory data (test results of components, images), research application data (US probe performance, mechanical and safety validation and other specification validation as defined in the **WP3**).
3. "Exploitation data": Market and competition assessment data, cost models, pre-product datasheet. Patent list (competition, FTO and patents resulting from LUCA) in relation with **WP7** activities.

i. Data set descriptions

What data *will be* ***generated or collected****?*	"Component data": VERMON will be mainly in charge of the components related to the multimodal probe. As such, we will generate design drawings, specifications, process definition. "Sub-system data": VERMON will generate test results associated with the probe compliance to specifications. Tests data will deal with mechanical assessment, process validation, US component performance and safety compliance. "Exploitation data": VERMON will contribute to the data collection necessary to setup a thorough exploitation plan. These include market data forecasts, potential end-user identification and manufacturing costs. Patent datasets will be created to assess competition and FTO as well as to monitor IP protection of

	the LUCA project results. We note that all these actions are collaborative and we expect significant overlaps and data sharing between partners.
*What is its **origin**?*	"Component data" and "Sub-system data" will be generated with VERMON's internal design and test tools. "Exploitation data" will be essentially derived from market studies, patent database extraction and more generally from the web.
*What are its **nature, format and scale**?*	A wide range of data formats and scales will be generated. 1. Drawings and designs will use industry standard software and will, primarily, be confidential in nature. We will, as much as possible, generate publicly accessible versions for dissemination purposes. These will be stored in forward compatible, time-tested formats. 2. Research application data on the testing of LUCA probe will follow formats dependent from the test workbenches. Usually, the measurement data will be stored in Matlab or Excel formats. The data files have small sizes (few tens of Ko). 3. Market analysis data will be confidential and will be shared within the consortium as reports and numbers. A summary will be published as part of the appropriate project deliverables. These data will be stored in data servers of VERMON. The IS infrastructure is based on redundant hard-drive with a weekly and monthly backup.
***To whom** could it be **useful**? Does it underpin a scientific publication?*	"Component data": In the short-term, this type of data is only useful for the proper interaction between LUCA partners and to keep the internal knowledge within VERMON. In the long-term, if further developments and designs occur, this type of data will be shared on a business-to-business basis. "Sub-system data" are useful both internally for our developments and upgrades but also for assessing the performance indicators of the LUCA solution. Generic performance data can be public for dissemination purposes towards possible end-users and customers. "Exploitation data" is company confidential by default. For proper coordination of exploitation in the LUCA consortium some data subsets or aggregated data can be shared.
*Do **similar data sets exist**? Are there possibilities for **integration and reuse**?*	This is a unique device and a data-set. There are possibilities to combine processed data for review papers on optics + ultrasound combinations in biomedicine as well as for reviews on applications of diffuse optics in cancer.

ii. Standards and metadata

*How will the **data be collected/genera***	Data will be generated from different sources from internal tools and testbenches to data accessible by the web.

***ted**?*	
*Which community **data standards or methodologies** (if any) will be used at this stage?*	Not applicable for VERMON
*How will the data be **organised during the project?***	VERMON has an internal methodology to keep track of the data generated by each project/product which is based on several data management tools : • A project management tools keeps track of project development. This internally developed database records the project responsibilities and all the project step validation. • Designs and measurement files are stored in a dedicated server following a common folder infrastructure. Each probe has a root folder with subfolders related to "Specifications", "Design History Files (DHF)" and "Preliminary study". The DHF folder has a standard organisation related to each development step of the project and history of each processed probe with Quality check sheets. Most of the documents have dedicated templates, giving a formal and easy check of their version and level of approval. • Mechanical designs files are managed by our design tool (TopSolid, Missler Software) giving access to each parts and sub-parts with a versioning and user-rights management.
Metadata** should be created to describe the data and aid discovery. **How will you capture this information?	Not applicable object for VERMON
Where will it be recorded?	All internal data will be kept according to the IS infrastructure in VERMON and with its standard practices. We will work collectively with the other LUCA partners to arrange the external data in standard formats.

iii. Data Sharing

***Where and how** will the data be made available and **how can they be accessed**? Will you share data via a data repository, handle data*	Internal data is stored internally in VERMON with no access from outside the company network. Externally, we will use the project web-site as the main gateway for sharing data. We will post, after IP clearance, appropriate data sets alongside publications on journal web-sites.

requests directly or use another mechanism?	
***To whom** will the data be made available?*	We aim to make bulk of the data widely accessible; however, there may be some data, such as market studies, IP portfolios that will be shared with entities and people related to the exploitation activities. Specific data will be shared among the LUCA consortium to ensure the proper advancement of the project. Different levels of sharing may be considered: only one person, several people belonging to one partner, a group of partners (WP group, topic group,…) or to the whole consortium.
*What are the **technical mechanisms for dissemination** and necessary **software** or other tools for enabling **re-use** of the data?*	We will use the LUCA web-site for all dissemination. The processed data will be presented in a way that it is cross-platform and software independent to the best of our abilities. If some software or dataset that we generate becomes of value for the general biomedical optics community, we will consider developing a unique web-site for that purposes.
*Are any **restrictions on data sharing** required and **why**?*	There will be restrictions based on the need for securing publications prior to public release and for exploitation purposes. These are defined in the project DOA.
*What **strategies** will you apply **to overcome or limit restrictions**?*	Data which have been approved for public release, after confidentiality and IP clearance, either on the project website or dissemination documents will be, by purpose, without limitations. Possible access restrictions to scientific publication may be dictated by the publication editors. Whenever possible, we will target editors which offer free access.
***Where (i.e. in which repository)** will the data be deposited?*	As mentioned above we will utilize project web-site, possibly dedicated web-sites for specific outputs and journal web-sites. Within the LUCA consortium we will use the project management platform where data files can be uploaded/downloaded in folders organised by WP and/or specific topics with a version management and the possibilities to restrict the access and add tags.

iv. Archiving and preservation (including storage and backup)

*What procedures will be put in place for **long-term preservation of the data**?*	VERMON internal infrastructure has been designed for long-term data storage and retrieval. We use dedicated internal servers for each tool. These servers are mirrored with a RAID infrastructure located in a separate room with regular storage backup (daily/weekly/monthly). This IS infrastructure cannot be accessed from outside of VERMON's network.
How long will	Internally to VERMON, project archives as old as 15 years ago can actually be

the data be preserved *and what will its* ***approximated end volume*** *be?*	retrieved in a fast and thorough way.
Are ***additional resources and/or is specialist expertise*** *needed?*	VERMON has two people dedicated to IS management.
Will there be any ***additional costs*** *for archiving?*	The costs are budgeted within the project and internally.

f. Data sets collected at ECM

Four types of data will be collected at ECM:

1. "Component data": Ultrasound beamformer specifications, electronic boards schematics and design, processing software specification and source code, FPGA firmware specifications and source code, mechanical drawings.
2. "Sub-system data": Ultrasound probe integration test reports, ultrasound image evaluation test reports, integration test reports of Luca demonstrator, integration report of the communication protocol between ultrasound and optical components.
3. "Evaluation data": Evaluation data which are the results from the end-user tests in clinics.
4. "Exploratory data": Market and competition analysis reports, cost structure, commercial product datasheet, business plan.

i. Data set descriptions

What data *will be* ***generated or collected****?*	"Component data": ECM will provide data related to the ultrasound beamformer hardware, firmware and software. Generated data will be made of mechanical drawings, electronic schematics, software and firmware source codes, specification documents. "Sub-system data": ECM will generate test results associated with ultrasound system performance including probe integration, image quality assessment, interaction with the optical components, functional and compliance test reports at the sub-system level and for the integrated LUCA system. "Evaluation data": ECM will be involved in the evaluation of the data measured in the clinics by the end-users. ECM will be in charge of generation of the ultrasound image and display of the optical measurements results. End-user feedback on the LUCA device performance in clinical settings will be collected. "Exploratory data": In collaboration mainly with ICFO and the industrial partners, ECM will provide contributions to the exploitation aspects of LUCA like market

	analysis, exploitation strategy, freedom-to-operate analysis etc.
*What is its **origin**?*	"Component data" will be generated by ECM. "Sub-system data" will be generated by ECM, HemoPhotonics, ICFO, POLIMI, VERMON. "Evaluation data" will be generated at IDIBAPS in collaboration with ICFO. ECM will be involved in supporting the clinical investigators with the ultrasound subsystem performance. "Exploratory data" will be mainly generated from market analysis reports, potential customers need analysis using external databases, studies and available reports.
*What are its **nature, format and scale**?*	A wide range of data formats and scales will be generated. 1. Drawings and designs will use industry standard software and will, primarily, be confidential in nature. As much as possible, publicly accessible versions will be generated for dissemination purposes. These will be stored in forward compatible, time-tested formats. Specifics will arise by M18. 2. Software and firmware code will be developed in standard development suites for C++, VHDL on a dedicated computer system. Codes will be confidential. 3. Application data on the testing of LUCA will follow non-standard formats in binary and text files and evaluated with internal scripts based on common software tools (Excel, Matlab, etc.) on dedicated computer system. The processed data will be saved in a report format and will be publicly available once cleared in terms of IP and exploitation issues by the appropriate committee in LUCA project as foreseen by the description of action. 4. Exploitation strategy, market analysis, freedom-to-operate analysis etc. data will be confidential and will be shared within the consortium as reports and numbers. A summary will be published as part of the appropriate project deliverables. Long-term access is ensured by the following measures: 1. All data will be stored in ECM data server secured by a redundant hard drive system that is totally backed up once a week and incrementally backed up on a daily basis. 2. All developed firmware and software code will be stored under proper consecutive version assignments. 3. All mechanical and electronic design files will be stored, managed with assignment of unique identifiers according to ECM Quality system requirements.
***To whom** could it be **useful**? Does it underpin a scientific publication?*	"Component data": In the short-term, this type of data is only useful for the internal LUCA partners. In the medium-term, it will be useful for our other projects and when some of these components might become products. "Sub-system data" and "Evaluation data" are useful internally for our developments and upgrades. They may support occasionally scientific publications. "Exploratory data" is company confidential by default. For proper coordination of exploitation in the LUCA consortium some data subsets or aggregated data can

	be shared.
*Do **similar data sets exist**? Are there possibilities for **integration and reuse**?*	This is a unique device and a data-set. There are possibilities to combine processed data for review papers on optics + ultrasound combinations in biomedicine as well as for reviews on applications of diffuse optics in cancer

ii. Standards and metadata

*How will the **data be collected/generated**?*	"Component data" and "sub-system data" will be generated by laboratory tests using test equipment, using design and development software. "Evaluation data" will be generated mainly from ex vivo phantom measurements and by data acquired from the subjects. "Exploratory data" will be generated by studies of external databases, interviews with end-users and others. Details are described in the specific work-packages.
*Which community **data standards or methodologies** (if any) will be used at this stage?*	Not applicable for ECM.
*How will the data be **organised during the project?***	"Component data" and "sub-system data" generated by ECM will be managed according to the existing quality procedure related to documentation control under the requirements of ISO 13485 standard. "Evaluation data" will follow the conventions defined jointly by IDIBAPS, HemoPhotonics and ECM who are the main drivers of the clinical studies and the final software suites.
Metadata** should be created to describe the data and aid discovery. **How will you capture this information?	Not applicable for ECM.
Where will it be recorded?	Every data-set will be recorded in the ECM storage system described above.

iii. Data Sharing

Where and how *will the data be made available and* ***how can they be accessed****? Will you share data via a data repository, handle data requests directly or use another mechanism?*	Data are stored internally in ECM servers with no access from outside the company network. Externally, we will use the project web-site as the main gateway for sharing data.
To whom *will the data be made available?*	We aim to make bulk of the data widely accessible; however, there may be some data, such as market studies, IP portfolios that will be shared with entities and people related to the exploitation activities. Specific data will be shared among the LUCA consortium to ensure the proper advancement of the project. Different levels of sharing may be considered: only one person, several people belonging to one partner, a group of partners (WP group, topic group) or to the whole consortium.
What are the ***technical mechanisms for dissemination*** *and necessary* ***software*** *or other tools for enabling* ***re-use*** *of the data?*	We will use the LUCA web-site for all dissemination. The processed data will be presented in a way that it is cross-platform and software independent to the best of our abilities. If some software or dataset that we generate becomes of value for the general biomedical optics community, we will consider developing a unique web-site for that purposes.
Are any ***restrictions on data sharing*** *required and* ***why****?*	There will be restrictions based on the need for securing publications prior to public release and for exploitation purposes. These are defined in the project DOA.
What ***strategies*** *will you apply* ***to overcome or limit restrictions****?*	We will use procedures as embargo until publication in order to implement IP protection measures.
Where (i.e. in which repository) *will the data be deposited?*	As mentioned above we will utilize project web-site, possibly dedicated web-sites for specific outputs and journal web-sites. Within the LUCA consortium we will use the project management platform where data files can be uploaded/downloaded in folders organised by WP and/or specific topics with a version management and the possibilities to restrict the access and add tags.

iv. Archiving and preservation (including storage and backup)

*What procedures will be put in place for **long-term preservation of the data**?*	As described above, to ensure long-term access, ECM will implement the following measures: 1. Forward compatible, time-tested formats such as text files (comma-separated values, open-source formats), and/or open-source binary formats (such as open document spreadsheets, open document text) and/or custom made binary formats (with definition files stored in standard text formats) will be utilized with associated descriptive documentation. 2. Codes are based on standard languages with long-term availability (e.g. C, C#, VHDL). 3. All data will be stored in ECM data server secured by a redundant hard drive system that is totally backed up once a week and incrementally backed up on a daily basis. 4. All mechanical and electronic designs are stored and assigned unique identifiers according to the ECM Quality procedure related to Documentation Control, under the requirements of ISO 13485 standard.
***How long will the data be preserved** and what will its **approximated end volume** be?*	We aim for a ten year preservation of the data. The approximate end-volume of this data will be less than one terabyte.
*Are **additional resources and/or is specialist expertise** needed?*	No.
*Will there be any **additional costs** for archiving?*	The costs are budgeted within the project and internally.

g. Data sets collected at UoB

One type of data will be collected at UoB:

1. "Simulated data": Data produced using numerical models for evaluation using phantoms.

i. Data set descriptions

What data** will be **generated or	"Simulated data": UoB group will be mainly in charge of the computational tools that predict physical systems. Only data from these computational

***collected**?*	models will be generated. We note that all these actions are collaborative and we expect significant overlaps and data sharing between partners.
*What is its **origin**?*	"Simulated data" will be internal to the group and to the project.
*What are its **nature, format and scale**?*	A wide range of data formats and scales will be generated. 1. Research application data on the testing of LUCA will follow non-standard formats common to each laboratory, in this case UoB, doing the modelling and will be stored in binary and text files. They will be associated with an electronic notebook which will include links to analysis scripts (Matlab, Excell, custom-software). The processed data will be saved in a report format and will be publicly available once cleared in terms of IP and exploitation issues by the appropriate committee in LUCA project as foreseen by the description of action. 2. Supporting data used in academic peer reviewed publications will be made available, after publication, via a recognised suitable data sharing repository. This policy will be followed unless a partner or IEC can show that disseminating this data will compromise IP or other commercial advantage as detailed below. The project will use the metadata standards and requirements of the repository used for sharing the data. At the UoB, Long-term access is ensured by the following measures: 1. Forward compatible, time-tested formats such as text files (comma-separated values, open-source formats such as R data-tables), and/or open-source binary formats (such as open document spread sheets, open document text) and/or custom made binary formats (with definition files stored in standard text formats) will be utilized with associated descriptive documentation. 2. All data will be stored in a secure hard-drive that is backed up every night by an incremental back-up script (rsbackup) to an external drive. Both drives are regularly replicated and upgraded at roughly three year intervals. 3. All desktop computers used by the UoB personnel involved in the project is centrally managed by UoB information technology department which utilizes secure folders on the servers that are backed up automatically and.
***To whom** could it be **useful**? Does it underpin a scientific publication?*	"Simulated data": In the short-term, this type of data is only useful for the internal LUCA partners. In the medium-term, it will be useful for our other projects. Some information may be used in scientific publications and presentations as described below. The data will be interesting to the end-user community and the biophontonics community. We submit articles to target journals for these communities (e.g. Biophotonics, Applied Optics, Biomedical Optics Express, Journal of Biomedical Optics, Nature Photonics).
*Do **similar data sets exist**? Are*	There are possibilities to combine simulated data for review papers on optics + ultrasound combinations in biomedicine as well as for reviews on applications

*there possibilities for **integration and reuse**?*	of diffuse optics in cancer.

ii. Standards and metadata

*How will the **data be collected/generated**?*	"Simulated data" will be generated using computational models. Details are described in the specific work-packages.
*Which community **data standards or methodologies** (if any) will be used at this stage?*	The lack of community data standards is one of the points that we explicitly discuss and attempt to contribute in LUCA project. Here, we mean the community of biomedical optics researchers using diffuse optical methods.
*How will the data be **organised during the project?***	"Simulated data" will follow the conventions defined jointly by IDIBAPS, HEMO and ECM who are the main drivers of the clinical studies and the final software suites. UoB Group will follow their naming conventions.
Metadata** should be created to describe the data and aid discovery. **How will you capture this information?	This will be captured in electronic notebooks, in header files in open-source format (described above) and in case-report files. The exact details are being defined as the systems mature.
Where will it be recorded?	All internal data will be kept according to the different units at UoB and their standard practices. We will work collectively with the other LUCA partners to arrange the external data in standard formats. As explained above, every data-set is associated with an electronic notebook, appropriate header file and comments. These will be recorded in the storage system(s) described above.

iii. Data Sharing

***Where and how** will the data be made available and **how can they be accessed**? Will you share data via a data repository, handle data requests directly*	Internal to the project, the UoB data will be shared using generic cloud-storage (mainly Dropbox) wherever appropriate, e.g. when the shared data is not very sensitive or is incomprehensible for an intruder. Otherwise, it will be shared by encrypted files (PGP encryption) using UoB's own cloud system that is managed by its IT department. Brief reports, spreadsheets and such will be shared by the TEAMWORK framework set by EIBIR. Externally, we will use the project web-site as the main gateway for sharing data. We will post, after IP clearance, appropriate data sets alongside publications on journal web-sites.

or use another mechanism?	
***To whom** will the data be made available?*	Bulk of the data will be widely accessible, however, there may be some data, such as market studies, IP portfolios that will be shared with entities and people related to the exploitation activities.
*What are the **technical mechanisms for dissemination** and necessary **software** or other tools for enabling **re-use** of the data?*	We will use the LUCA web-site for all dissemination. The processed data will be presented in a way that it is cross-platform and software independent to the best of our abilities. If some software or dataset that we generate becomes of value for the general biomedical optics community, we will consider developing a unique web-site for that purposes.
*Are any **restrictions on data sharing** required and **why**?*	There will be restrictions based on the need for securing publications prior to public release and for exploitation purposes. These are defined in the project DOA. Furthermore, any patient data that could be used to identify the patients will be properly anonymized prior to sharing and the link between the patient ID and the dataset will be permanently destroyed after an appropriate time based on the ethical protocols and procedures that are approved. This is IDIBAP's responsibility and the UoB group will receive data that is already anonymized according to these principles.
*What **strategies** will you apply **to overcome or limit restrictions**?*	We will utilize procedures such as embargo until publication, anonymising and simplification.
***Where (i.e. in which repository)** will the data be deposited?*	As mentioned above, there are no community defined standards for the biomedical diffuse optics community. Therefore, we will utilize project web-site, possibly dedicated web-sites for specific outputs and journal web-sites.

iv. Archiving and preservation (including storage and backup)

*What procedures will be put in place for **long-term preservation of the data**?*	At the UoB group, Long-term access is ensured by the following measures: 1. Forward compatible, time-tested formats such as text files (comma-separated values, open-source formats such as R data-tables), and/or open-source binary formats (such as open document spreadsheets, open document text) and/or custom made binary formats (with definition files stored in standard text formats) will be utilized with associated descriptive documentation. 2. All data will be stored in a secure hard-drive that is backed up every night by an incremental back-up script. Both drives are regularly replicated and upgraded at roughly three year intervals.

	3. All desktop computers used by the UoB personnel involved in the project is centrally managed by UoB IT department.
How long will the data be preserved *and what will its* ***approximated end volume*** *be?*	Apart from the certain aspects of the clinical datasets – which will be managed by IDIBAPS, there are no limitations on the preservation of the data. We will follow academic standards and aim for a ten year preservation of the data. The approximate end-volume of this data will be less than one terabyte.
Are ***additional resources and/or is specialist expertise*** *needed?*	No. We are all experts in the management of datasets of this size. Internally, UoB-IT manages the general policies, makes suggestions on good-practices and ensures security against intrusions.
Will there be any ***additional costs*** *for archiving?*	The costs are budgeted within the project and internally.