D9.2 – Data Management Plan (DMP)

Project Information

Grant Agreement Number	780839
Project Full Title	Multiplex phOtonic sensor for pLasmonic-based Online detection of contaminants in milK
Project Acronym	MOLOKO
Funding scheme	IA
Start date of the project	January 1st, 2018
Duration	42 months
Project Coordinator	Stefano TOFFANIN (CNR)
Project Website	http://www.moloko-project.eu

Deliverable Information

Deliverable n°	9.2
Deliverable title	Data Management Plan (DMP)
WP no.	9
WP Leader	BEWARRANT
Contributing Partners	ALL
Nature	ORDP: Open Research Data Pilot
Authors	Giovanna Guidicelli, Francesco Mercuri (CNR)
Contributors	All Consortium Partners
Reviewers	Stefano Toffanin (CNR)
Contractual Deadline	30/04/2018
Delivery date to EC	15/05/2018

Dissemination Level

PU	Public	✓
PP	Restricted to other programme participants (incl. Commission Services)	
RE	Restricted to a group specified by the consortium (incl. Commission Services)	
CO	Confidential, only for the members of the consortium (incl. Commission Services)	

Document Log

Version	Date	Author	Description of Change
V1.0	07/05/2018	Giovanna Guidicelli	Collection of the data description from each partner
V2.0	10/04/2018	Giovanna Guidicelli	Completion of the draft of the deliverable
V3.0	14/05/2018	Stefano Toffanin	Revision of the draft by Coordinator
V4.0	15/05/2018	All partners	Final revision, introduction of further information into the Data Tables

Table of Contents

The aim of the MOLOKO's Data Management Plan (DMP) is to *identify the project's research data and to describe how to make them findable, accessible, interoperable and re-usable (FAIR*). Following the H2020 Data Management Plan Template v1.0 available on the Participant Portal, all partners involved in research's activities (all except BeWarrant that is not supposed to experimentally or compile data) were asked to provide detailed information about the data generated during the entire project as reported in *1. Data Summary*.

In the Deliverable, we report then an initial analysis on how we intend to manage the amount of data produced in the project them. The first check-point of the whole architecture of the DMP is the release of the first scientific/technological publications that will be published within MOLOKO project: indeed, the data reported in the article will be made available and interoperable to the larger typologies of stakeholders. In order to avoid issues related to IP rights and their access, as a first step in the strategy of development of DMP *only data related to publications available to the public will be released.* In MOLOKO Project DMP is intended to be a living document in which information can be made available on a finer level of accuracy and details through updates as the implementation of the project progresses and when significant changes occur.

Further update on Data Management will be provided in the first and second reporting period. Indeed, the DMP is intended to be a living document in which information can be made available on a finer level of granularity through updates as the implementation of the project progresses and when significant changes occur.

1. Data Summary

Hereafter we report the tables where the information about the contents and the collection/generation of the data provided by each MOLOKO's Partners are reported. At this early stage of development of the project, the reported information will be processed in order to determine the general specifications and structures of the metadata that will be generated within the DMP.

Beneficiary 1- CNR	
What is the purpose of the data collection/generation and its relation to the objectives of the project?	CNR generates data related to the light-emitting component (OLET) of the optoplasmonic chip present in the MOLOKO sensor. In particular, the technical requirements and the geometrical, electrical and optical characteristics of the OLETs will be of necessary importance for the design, development and optimization of the optoplasmonic chip and its integration in the whole MOLOKO sensor. Data on (i) the design and (ii) the optoelectronic characterization of the whole optoplasmonic chip will be generated for its optimization and integration in the final MOLOKO sensor.
What types and formats of data will the project generate/collect?	Mainly documents (doc, docx, ppt, pptx, etc.), illustrations (png, jpeg, etc.), drawings (DWG, etc.) and raw data (opj, xls, txt, etc.).

Will you re-use any existing data and how?	Existing data are currently under evaluation to check the compatibility of some components to the requirements of the MOLOKO sensor. Most probably, existing data will be used as hints for the development of components with new specifications.
What is the origin of the data?	Data origin from the optoelectronic characterization instrumentation used to test both the OLET component and the optoplasmonic chip at CNR. Data will be also produced at CNR to design the MOLOKO components.
What is the expected size of the data?	Hundreds of megabytes/few gigabytes.
To whom might it be useful ('data utility')?	The data are of fundamental importance for some partners of the MOLOKO consortium and can be useful for possible scientific production (publications, patents, etc.).

Beneficiary 2 - PLASMORE	
What is the purpose of the data collection/generation and its relation to the objectives of the project?	Data are related to the optimization and characterization of the optical detection scheme of the prospected sensor and of its functionalization for detection purposes.
What types and formats of data will the project generate/collect?	Characterization of the plasmonic surfaces during the preparation process is producing different kind of images from SEM and AFM measurements. Optical experiments yield plot of spectral response and series of numeric values (columns) of reflectance versus wavelength. Sensitivity tests and detection results consist in image series of the active surface and/or sensorgrams, i.e. evolution of signal for each analyte as a function of time, describing the intensity and dynamic evolution of analyte capturing events.
Will you re-use any existing data and how?	Each data series represents a test or a confirm of the chip response and can be used to drive sensor optimization.
What is the origin of the data?	Several different characterization instruments (like Microscopes or photo-spectrometers) for the optoplasmonic chip preparation and test process;

	signal response of the produced prototypes for the detection output.
What is the expected size of the data?	Several Gbytes for images; a few kbytes for each simple batch of numeric results (sensorgram).
To whom might it be useful ('data utility')?	For project partners for internal use and process optimization; for sensor producer, provided the respect of intellectual property.

Beneficiary 3 - CSEM	
What is the purpose of the data collection/generation and its relation to the objectives of the project?	Models, simulation results, tables and illustrations for the optical simulations of the optoplasmonic chip in order to model the effects of different grating geometries, materials, wavelengths and geometrical setups, and to optimize the sensitivity. Most probably CSEM will generate measurement data from the testing of the different MOLOKO prototypes. They will comprise measurement data from the testing of the sensing of different analytes in milk. This allows to test the functionality of the sensors and to optimize it.
What types and formats of data will the project generate/collect?	Simulation setups (CSEM internal format), result tables (e.g. Excel) and illustrations (jpg, png). The database format for the sensor measurements is not yet defined. The data will most probably contain information on the date, time, location, and the measurement results of a certain test run.
Will you re-use any existing data and how?	Most probably, the data will not be reused, since they build on a setup and test conditions which are specific to the MOLOKO sensor and project.
What is the origin of the data?	The results from optical simulations and optimizations being run at CSEM. Measurement data will result from test measurements during the development and validation phase of the MOLOKO sensing device.
What is the expected size of the data?	Several hundreds of megabytes.
To whom might it be useful ('data utility')?	The result tables and certain illustrations may be useful for certain technical implementation tasks developed by project partners, but likely will be of

	no use for external users since the data will be highly specific to the MOLOKO project.

Beneficiary 4 - ISS	
What is the purpose of the data collection/generation and its relation to the objectives of the project?	A. Food Health Technology Assessment. B. Integration within the BEST platform.
What types and formats of data will the project generate/collect?	A. Indicators, recommendations, technical reports, good practices, standards (numerical, text, checklists). B. Experimental (numerical).
Will you re-use any existing data and how?	A. Data from literature, technical reports, economics; previously-existing data from partners on milk dairy chain management (comparison, reference). B. data previously collected by BEST platform (comparison, reference).
What is the origin of the data?	A. MOLOKO Partners, Literature databases, Standards and Laws databases. B. Alert Project (Sistema Integrato di biosensori e sensori (BEST) per il monitoraggio della salubrità e qualità e per la tracciabilità della filiera del latte bovino, INDUSTRIA 2015) and demonstration of the integration of the MOLOKO sensor within the BEST platform in real-setting conditions.
What is the expected size of the data?	A. <2 TB B. <100 GB
To whom might it be useful ('data utility')?	A&B: Consortium partners; scientific community; analytical instrumentation and diary industry.

Beneficiary 5 – PARMALAT	
What is the purpose of the data collection/generation and its relation to the objectives of the project?	Validation of analysis system (MOLOKO sensor) and methodology.

What types and formats of data will the project generate/collect?	Numerical, recorded on Excel or Access database.
Will you re-use any existing data and how?	Historical dataset, obtained with traditional/standard method, within our supply chain; Expected concentration values (numbers) about interest analytes reported in literature.
What is the origin of the data?	Data collected from the MOLOKO sensor and data collected from the analytical methods currently in use (IR of Milkoscan and Fossomatic, Elisa Kit, Delvo Test).
What is the expected size of the data?	~ 100 Mb
To whom might it be useful ('data utility')?	Partners within the Consortium, Dairy supply chain (farms), Dairy industry.

Beneficiary 6 - FRAUNHOFER	
What is the purpose of the data collection/generation and its relation to the objectives of the project?	• giving data background for reports • sharing data between the partners • giving comprehensive data background for combined publications The main source of data are measurements and simulations related to the optimization and characterization of photodetector to be integrated within the optoplasmonic chip and engineering and implementation of the microfluidic module in the functional sensor.
What types and formats of data will the project generate/collect?	• CAD-Data of microfluidic structures (.stl, .step) • measurement data in raw format (.csv, .xlsx) or report data format (.docx, .pdf) • script data to process measurement data (.py) • microscopy images and pictures raw and processed data (.tiff, .png, .eps, .jpg) • movie data raw and compressed (.mov., .avi))
Will you re-use any existing data and how?	We will use existing data for reference purposes only to compare new developed systems to gold standards.
What is the origin of the data?	The origin of existing and/or new data are our own developed, in-house characterization methods and setups.

What is the expected size of the data?	• mov's and tiff's are large sizes (each .mov about 500MB and each .tiff of about 20MB) • single CAD-assembly folders can have up to 2GB • the other files are of irrelevant small size
To whom might it be useful ('data utility')?	• CAD-data is of interest in terms of interface definition between the partners and a coherent overall system design • experimental data for cross-checking of method and result • raw data (.mov and .tiff) are sometimes reasonable to share, if other partners have certain software or know-how of analysing the data, which the original partner has not.

Beneficiary 7 - VTT	
What is the purpose of the data collection/generation and its relation to the objectives of the project?	Data generated in project by VTT are related to realization and characterization of novel antibodies, their nucleotide and amino acid sequences. Moreover, data about the binding properties of the antibodies will be produced.
What types and formats of data will the project generate/collect?	Reports about protocols for realization of Novel recombinant antibody sequence and their characterization (e.g. sensorgrams collected from SPR analysis).
Will you re-use any existing data and how?	Apart the use of already collected data as complementary information in IPR protection activity and later for publications, previously-collected data on specific antibodies collected by standard techniques will be used for comparison.
What is the origin of the data?	Mainly the data are generated in laboratory as procedures/protocols (text) and/or as collection of numerical values (columns of values) by the instrumentations for characterizing the antibodies.
What is the expected size of the data?	Few kbytes for each simple batch of numeric results (sensorgram).
To whom might it be useful ('data utility')?	MOLOKO consortium partners and possible end-users (i.e. providers of analytical/biodiagnostics instruments).

Beneficiary 9 - WR	
What is the purpose of the data collection/generation and its relation to the objectives of the project?	During the development and evaluation/validation of the MOLOKO device and the diagnostic methods, we will obtain data about the performance of the antibodies (binding properties) and the assays (sensitivity and specificity) and about the presence and levels of contaminants and quality parameters in milk.
What types and formats of data will the project generate/collect?	Analytical data about the performances of the antibodies and assays and about concentrations of the parameters to be measured.
Will you re-use any existing data and how?	Of some antibodies and assays (e.g. streptomycin and casein) the performance data are known in the Biacore biosensor format.
What is the origin of the data?	Concentrations of parameters which are obtained during the project.
What is the expected size of the data?	Few kbytes for each simple batch of numeric results (sensorgram).
To whom might it be useful ('data utility')?	MOLOKO consortium partners and possible end-users (i.e. providers of analytical/biodiagnostics instruments, farmers).

Beneficiary 10 - NEBIH	
What is the purpose of the data collection/generation and its relation to the objectives of the project?	The data collected would be directly linked to the laboratory performance of the MOLOKO sensor. Together with data obtained by other laboratory methods on the same samples, the validation of the MOLOKO equipment would be possible.
What types and formats of data will the project generate/collect?	The data will be collected via laboratory measurement processes, and then several performance criteria will be calculated (generated) from those data. The format of the raw data: • Sample Id (string) • Measurement Id (string)

	• Measured parameter (string) • Laboratory measurement method (string) • Measured value (floating point number) • Measurement unit (string) • Chromatograms (picture files)
Will you re-use any existing data and how?	Re-using of existing data is not planned, since the validation procedure requires using the same identical set of samples. Existing data may be used for comparison.
What is the origin of the data?	The data will be collected/generated in laboratory, from the measurements of the samples collected and (in some cases) spiked with the analyte to be measured. We will use an external software (Interval) for the validation process.
What is the expected size of the data?	Data are in the order of magnitude of 500 MB – 1 GB.
To whom might it be useful ('data utility')?	Data would be used for validation purposes in the project itself. Besides, the data could be useful for future MOLOKO users, as background data on performance criteria. In addition, data will be useful for the laboratory accreditation procedure as well.

Beneficiary 11 - QCL	
What is the purpose of the data collection/generation and its relation to the objectives of the project?	A. To assist in developing and enacting strategies for the commercial exploitation of the project results. B. Patent data will be collected for an IPR freedom-to-operate task. C. Collation of Information on products, prices and applications for potentially competitive technologies.
What types and formats of data will the project generate/collect?	A. Benchmarks, market sizes, market reports, surveys and direct communication (text & numerical). B. Data collected from EU patent service website, MOLOKO partners (text). C. Data collated from manufacture and supplier information (text & numerical).

Will you re-use any existing data and how?	A. Data from literature, MOLOKO partners, market databases and reports. B. EU patent service website. C. Published information.
What is the origin of the data?	A. MOLOKO partners, literature databases, publically accessible internet services. B. EU patent service website, MOLOKO partners. C. Manufacturers and Suppliers of analytical and biodiagnostic instrumentations.
What is the expected size of the data?	A. <10 GB B. < 10 GB C. <10GB
To whom might it be useful ('data utility')?	A: MOLOKO partners, companies, organisations and/or investors with commercial interest in the project results. B: MOLOKO partners. C: MOLOKO partners, companies, organisations and/or investors with commercial interest in the project results and sensor implementation.

Beneficiary 12 - MILKLINE	
What is the purpose of the data collection/generation and its relation to the objectives of the project?	The data collection and generation aims to detect abnormal milk from a quality and safety point of view by means of the MOLOKO detection methodology. Under the current legislation, abnormal milk cannot be commercialize and represents a loss for the farmer and a risk for the consumer.
What types and formats of data will the project generate/collect?	Statistics (numerical), indicators, recommendations, technical reports, good practices, standards (numerical, text, checklists).
Will you re-use any existing data and how?	Data from literature and data generated from standard methodologies available in the market as comparison with data collected by the MOLOKO sensor.
What is the origin of the data?	Data are generated by direct measurements of analytes presence (Y/N) and concentration

	(quantitative) in milk samples both in tanks and in in-line (when possible) reals setting conditions.
What is the expected size of the data?	~ 100 Mb.
To whom might it be useful ('data utility')?	To the end-user who can be the farmer, the final customer and the milk process industry on one side, to the research on the other side. Consortium Partners and scientific community.

2. FAIR data

2. 1. Making data findable, including provisions for metadata

Data description

The description of procedures to generate data is associated to a **dataset** (i.e. collection of data).

At this stage of development of the project, the specific typology and total number of variables in a single dataset table (see Data Summary) cannot be defined *a-priori*.

The procedures for the identification of data are defined as follows:

- ☐ Each dataset is initially assigned to a unique ID, automatically generated through a Universally Unique Identifier (UUID) application.
- ☐ Each dataset is also associated to a Digital Object Identifier (DOI). The service is provided by the DOI (www.doi.org) community through a request to a local Registration Agency (RA).

The use of a DOI guarantees, at the same time, unique identification of the single dataset and the possibility of automatic data web retrieval.

After this step, the dataset is univocally associated to an identifier.

The implementation of the data description depends on the typology of datum considered.

In most cases, a text description is appropriate. In this case, data are described by compiling a form (data description template), available to all users.

Metadata

One or more metadata files are generated for each dataset. The metadata are identified by the same unique ID of the related dataset, with a different suffix/extension.

Each metadata file is uploaded in a standardized format, depending on the dataset considered.

Appropriate templates will be available for download to all MOLOKO's partners in the Collaborative Platform.

The metadata files are linked to the data descriptor, and directly accessible through a web link.

The link is realized through the definition of an URL address, related to the metadata ID, which is also associated to a web resource on the cloud correlated to the MOLOLKO Collaborative Platform.

2.2. Making data openly accessible

By default and as a first case-study of data management, only data related to publications will be made openly available In general, the General Assembly will decide on a case-by-case basis which data can be released in order to avoid issues related to IP rights protection or access.

All data and metadata files will be uploaded onto a cloud storage and sharing facility specifically dedicated to MOLOKO project.

The MOLOKO private cloud is provided by means of OneDrive utility in Microsoft Office365 (see D10.4). The unique ID allows the retrieval of data and metadata files to registered users.

The main features of the MOLOKO cloud storage and sharing facility are the following:

- ☐ Access through the MOLOKO webpage
- ☐ Restricted access to registered users only if needed
- ☐ 1 TB of storage
- ☐ Sharing of documents/metafiles by/with intra- and extra-consortium users
- ☐ Document status and drafting can be checked online.

In addition to local storage, public metadata and datasets will be made available to users once publications are finalized through the MOLOKO'S website and through the OpenAIRE sharing web platform. In particular, relevant MOLOKO's metadata and dataset will be uploaded by the involved researchers to the **ZENODO** (https://zenodo.org/) platform, compiling project-related information.

This will enable automatic data extraction from the OpenAIRE platform, thus ensuring accessibility through a standard platform for Open Data access.

2.3. Making data interoperable

All data will be made available in standard/open formats compliant with commercial/open software in order to allow as much as possible data exchange between researchers and institutions.

Standard vocabulary for metadata description will be used, in case this will not be possible a mapping of more common ontologies (i.e. diagnostics, optoelectronics, plasmonics, assay, calibration, electronics, module, validation, demonstration…) will be provided. In this case, specific technical contribution from specialists in semantics and logics will be considered.

2.4. Increase data re-use (through clarifying licences)

The data will be made available according to Open Licenses such as Creative Commons.

The data will be available for re-use upon decision of the General Assembly, in order to avoid issues related to IP rights protection or access. Once the data are made openly available they will remain open.

Within the strategy of development of DMP, the dataset that will be firstly available are those reported in publications originated from the consortium, thus intrinsically made for being reused. However, specific agreements with the Editors of scientific/technological journals will be considered and provided.

The data quality is assured by each partner, that bears the responsibility of them. The tools necessary for describing and identifying the dataset and for preparing the metafiles will be provided by the General Assembly in strict collaboration with the Coordinator and the Exploitation and Innovation Managers.

3. Allocation of resources

The costs for making data FAIR include the costs of the cloud facility and of personnel involved in collecting and managing data:

- ☐ Set up of the data space in the Project Collaborative Platform
- ☐ Implementation of the UUID generator
- ☐ DOI registration request
- ☐ Preparation of templates for:
  - Data descriptor (general, text format, pdf output)
  - Metadata: text template, spreadsheet template
- ☐ Data collection
- ☐ Generation of the data descriptor
- ☐ Generation of the metadata
- ☐ Upload to the private cloud server
- ☐ If public, upload data to ZENODO/OpenAIRE

CNR as project's coordinator will be responsible for the data management.

The overall cost of DMP according to the reported cost-items is considered and covered by the *Open Access* cost the project budget (see Annex 2 in Grant Agreement).

4. Data security

Data will be stored and shared in the private Collaborative Platform with restricted access (username + password) to authorized users. As an initial step, only the Consortium Partners will have access to the cloud storage where dataset and metadata are filed. Other options about managing store data according to the inputs from the General Assembly.