This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952110

MARILIA deliverable: Data Management Plan, DMP
Deliverable number: D1.4

MARILIA			
MARA-BASED INDUSTRIAL LOW-COST IDENTIFICATION ASSAYS			
Project nr:	952110	Call reference:	H2020-EIC-FETPROACT-2019
Start date:	September 1st, 2020.	Duration:	24 months

Deliverable identification			
Leading beneficiary:	AIT	Planned delivery date:	M6
Related WP:	WP1	Actual delivery date:	M6
Dissemination level:	Public		

Contributors	
Beneficiary name	Contributor(s)' name(s)
AUSTRIAN INSTITUTE OF TECHNOLOGY (AIT)	Yasaman Ahmadi, Natalie Mutter, David Kutak
RUDER BOSKOVIC INSTITUTE (RBI)	Ivo Piantanida, Ivo Crnolatac
DAY ONE SRL (D1)	Paolo de Stefanis, Gianluca Giordani
IREN SPA (IR)	Nicola Bazzuro, Guido Scarafia, Micaela Tiso
FACULTY OF SCIENCE UNIVERSITY OF ZAGREB (UZ)	Branimir Bertoša, Zoe Jelić Matošević

Deliverable Reviewers		
Version	Reviewer	Date
1.0	Ivan Barišić	15.02.2021.
2.0	All	24.02.2021.

Table of content

1. Context and objectives 3
2. Description of the performed tasks and obtained results 3
2.1 General DM procedures & infrastructure 3
2.1.1 AIT AUSTRIAN INSTITUTE OF TECHNOLOGY GMBH (AIT) 3
2.1.2 RUDER BOSKOVIC INSTITUTE (RBI) 5
2.1.3 DAY ONE SRL (D1) 6
2.1.4 IREN SPA (IR) 7
2.1.5 FACULTY OF SCIENCE UNIVERSITY OF ZAGREB (UZ) 7
2.2 Data set descriptions 9
2.2.1 MARILIA-AIT-001 9
2.2.2 MARILIA-AIT-002 10
2.2.3 MARILIA-AIT-003 11
2.2.4 MARILIA-AIT-004 12
2.2.5 MARILIA-AIT-005 13
2.2.6 MARILIA-RBI-001 14
2.2.7 MARILIA-D1-001 15
2.2.8 MARILIA-IR-001 17
2.2.9 MARILIA-UZ-001 20
2.3 Data exchange within MARILIA consortium 22
3. Conclusion 24

1. Context and objectives

This document represents the first version of the MARILIA project data management plan (DMP). The aim of this first version of DMP is mainly to describe the general data management (DM) procedures the consortium members have agreed on and to give an overview of the procedures and infrastructure in place for data management at the beneficiaries' sites. It provides for a preliminary survey, which will be updated (as required) over the course of the project to give an actual overview of the data generated in the MARILIA project.

The aims of this first version of DMP are:

- Ensuring that all participating parties have proper procedures for data storage, data exchange, data safety and security and that they are committed to adhere to those procedures.

- Facilitating the decision-making process on which data should be made accessible to the public. The MARILIA consortium partners are committed to follow a research data policy that is "as open as possible and as closed as necessary". However, since generation and subsequent exploitation of IP is one important objective of MARILIA project, some data to be generated within the project cannot be published until the protection of the IPR is guaranteed. The DMP will document the dissemination level of each data set and, in case that a data set cannot be made public, the reasons for keeping it confidential.

- To describe the data management life cycle for the data sets that are collected, processed or generated by the project, in order to facilitate their short- and long-term findability, accessibility, interoperability, and re-usability, both within the consortium and, for those data that can be made public, outside the consortium.

2. Description of the performed tasks and obtained results

Since several DM activities are usually independent of a specific data set but rather depend on the general data handling procedures established at a beneficiary's site, we have decided to split the DM information accordingly. Therefore, two templates were sent to each beneficiary: one for the general data management procedures of the beneficiary and one to be filled out specifically for each data set. Some questions were repeated in the data set specific template, to allow for deviating DM procedures for special data sets, if necessary (e.g. different security levels or storage procedures).

The following sections present the information collected from all beneficiaries.

2.1 General DM procedures & infrastructure

2.1.1 AIT AUSTRIAN INSTITUTE OF TECHNOLOGY GMBH (AIT)

Data manager:

Name: Ivan Barišić

Telephone number: +43 664 88390643

E-mail: Ivan.Barisic@ait.ac.at

Available / necessary resources

Technical resources and operations team

The common network and system infrastructure and solutions are centrally administered by a highly specialized IT operations team. Project specific infrastructure and solutions are administered directly by the project team or the organizational unit. The central part of the IT operations is based on defined service levels and a documented scope. The infrastructure and solutions are built for a 24/7 operational availability. The support coverage is defined in the service levels.

Data is stored on storage systems from renowned manufacturers with support contracts that provide short time recovery of hardware and software system failures. For the purposes of data redundancy and availability, RAID disk architecture as well as redundant power supplies are standard. All storage and backup systems are on the premises of AIT.

For collaboration and data exchange with external partners, various managed file sharing services (e.g. FTP, SharePoint, OneDrive for Business) are available, where data recovery is partly possible, depending of the chosen service. Generally, there exists no data classification in AIT, there is no distinction made between project data and business data. Therefore, the backup process for these data types is the same. Data management and backup is documented in operational guidelines. The common backup is executed in a

cascading scheme – daily incremental, weekly full, monthly full, and yearly full. The monthly and yearly backups are kept in dedicates safes, which are geographically dislocated from the data centre.

Access rights are managed centrally and are documented in a ticketing system.

Data management and backup of specific project-related data can be handled with higher restrictions or with a differentiated support, if necessary.

Data access and security

General data security measures

All accounts and resources are protected by complex passwords that must be changed at least once every three months. Systems are configured to regularly update operating system software, server applications, client software and malware protection software.

Windows-based clients are registered company-wide with the System Center Configuration Manager (SCCM), where detailed client information is listed and software updates are rolled out. Non-windows based clients have to be equipped with the newest available software patches by the person which is in charge of the system. Systems are protected with firewalls and accounts follow the least-privilege principle. By default, all laptops and smartphones are encrypted and securely managed.

Access right management

Authentication is based on a central directory service. Authorization is done based on group membership specific to the research project. Group membership to read or alter data is granted by the project manager.

Additional requests are recorded comprehensibly in the internal ticketing-system.

Remote access is possible through a secure VPN access solution and two factor authentication. Project specific: If required, it is possible to monitor the changes of data (along with the reason of change provided by the user) by adding version control to the file repositories. Access rights are managed within the lifecycle of the project.

Risk management

AIT focuses on two aspects of risk management initiatives – user awareness and preventive technology.

With regard to user awareness several guidelines are defined to support users in handling data. In addition, users are frequently informed about new developments and threats.

With regard to preventive technology there are several security measures in place to ensure data protection from unauthorized access.

Secure Access and Transfer

Every connection to a data-processing system from a remote location is done by certificate-based authentication and strong encryption methods (e.g. for the “Online Document Sharing Service”). Secure and safe data transfer is managed through SSL based protocols (e.g. HTTPs, FTPs, SFTP, SCP) or a virtual private network.

Storage and backup (short and mid-term)

Data storage

The available storage space on the file storage is separated by location and department. Permission for access is claimed comprehensibly over the internal ticketing system. The permission management is organized with active directory groups by the central IT.

Specific project initiatives can be handled with higher restrictions or with a differentiated support.

System related backups

Backup of virtual machines, backup of local client data, backup of Linux systems and full image client backups are designed for disaster recovery, not for mid-term preservation. On request it is possible to preserve a system for mid or long term.

General backup procedure

Company data is stored for 10 years after the project end, because, according to the internal AIT-quality management specifications, project data must be kept for this period.

- File-Service: on a regular base, data is secured on LTO-tapes:
  - Daily: differential backup, where the LTO tapes are overwritten weekly.
  - Every Friday: full backup, where the LTO tapes are overwritten each month. Exception: No overwrite of the last full backup is made in a month. This LTO tape is stored securely in a data safe (security class EN 1047-1) with restricted access.
- Exchange-Service: daily full backup to disk.

- VM: daily backup to disk of the whole centrally managed virtual infrastructure, with an available restore period of the last 7 days. Longer-term backups have to be requested separately.
- SharePoint: daily differential backup and weekly full backups to a file share which is kept for 30 days.
- FTP: no backup needed because it is only used for data exchange.
- OneDrive for Business: managed, externally hosted (EU) cloud storage solution for each user for data exchange and project activities with a guarantee of high availability. Therefore, there exists no centrally managed backup strategy.

Data backup and recovery of the central infrastructure is the responsibility of the central IT. Decentralized initiatives can be handled with higher restrictions or with a differentiated support.

Archiving and preservation (long term)

According to the internal AIT quality management specifications, project data must be stored for 10 years after the end of the project (see above: general backup procedure). For this period, AIT can guarantee the availability and the restricted access to stored data for eligible persons. Besides the standard backup procedure, AIT has no further dedicated central data archiving system. Because of that, AIT cannot guarantee the unchangeability of stored data. If necessary, data archiving needs to be executed through decentralized initiatives.

2.1.2 RUDER BOSKOVIC INSTITUTE (RBI)

Data manager

Name: Ivo Piantanida

Telephone number: +385 1 45 71 326

E-mail: pianta@irb.hr

Available / necessary resources

1. Laboratory for Biomolecular Interactions and Spectroscopy (LBIS) has internal database hardware (4TB) linked over RBI intranet and protected by user-defined password; used for experimental data archiving. The responsible person is Laboratory Head (Ivo Piantanida), who authorizes access.

2. Also, Ruđer Bošković Institute has “Full-text Institutional Repository” (FULIR, http://fulir.irb.hr/). The following types of documents can be deposited in FULIR: full-texts of papers published in scientific journals or conference proceedings, PhD and MA thesis, book chapters, monographs, reports, manuals, PPT or poster presentations from conferences and other meetings, and also audio, video and AV materials connected to RBI work. The repository is implemented, maintained and developed by RBI Library. The library also provides support with self-archiving to RBI staff.

The costs for making data FAIR are only related to GOLD OPEN access publication of results and are covered by corresponding research project.

All mentioned resources are available for the long-term preservation (over 10 years).

Data access and security

1. Experimental data: will be saved at LBIS’s internal database hardware (4TB) linked over the RBI intranet. The access is protected by user-defined password. The responsible person is Laboratory Head (Ivo Piantanida), who authorizes access.

2. Open access data for open public information: Ruđer Bošković Institute “Full-text Institutional Repository” (FULIR, http://fulir.irb.hr/), managed by the RBI Library staff protocols.

Storage and backup (short and mid-term)

1. Experimental data: monthly backup is provided by saving data on the external disk drive, which is stored with the Laboratory Head (Ivo Piantanida), who authorizes access. That ensures mid-term data storage (up to 5 years). In case of an accident, the external drives with backups are stored separately from the main storage unit of Lab (intranet connected) and can be accessed by I. Piantanida or other senior members of the team (permanently employed scientists involved in MARILIA project: I. Crnolatac, D. Saftić).

Archiving and preservation (long term)

All the resources as mentioned are available for the long-term preservation (over 10 years).

2.1.3 DAY ONE SRL (D1)

Data manager

Name: Paolo De Stefanis

Telephone number: +39 347 79 27 523

E-Mail: paolo@day-one.biz>

Available / necessary resources

Data management in the project follows the internal procedures of the company.

In general, the company collects marketing data, including certain information about individuals.

These, for instance, include customers, suppliers, business contacts, employees and other people that the organization approaches to understand and predict customers and user's behaviour with respect to new technology.

The internal policy describes how this personal data must be collected, handled and stored to meet the company's data protection standards – and to comply with the law.

In particular, the internal procedure foresees that the data management approach:

- Complies with data protection law and follows good practice
- Protects the rights of customers, staff and partners
- Is transparent about how it stores and processes individuals' data
- Protects the company from the risks of a data breach

In this sense, data are usually collected through one-to-one interviews or focus groups, to which individuals participate after having read and agreed with an informed consent.

Data are anonymised and used for statistical purposes, generally to indicate the product characteristics that users would love to see in a new system.

Data are stored in Day One's cloud space, which is protected through state-of-the-art double-security facilities. Data are only accessible to the DPO and to the staff involved in the specific project.

Data access and security

The company appoints the following responsible:

1. **Data owners**, who take care of:
   - ensuring that data is governed in accordance with the policy
   - managing corporate data in their area of responsibility, including data provided to or by contractors or third parties
   - the establishment of validation rules for data entry and data correction in their area of responsibility
   - identifying and documenting authorities for access to data and levels of access
   - authorising downloads and uploads of corporate data
   - authorising appropriate access to corporate data, including to restricted data
2. **System administrators**, who are responsible for:
   - providing and removing access to data users as specified by data owners
   - ensuring that data systems are operating efficiently
   - monitoring the transfer of data
   - ensuring that appropriate safeguards exist to protect data and that appropriate disaster recovery and business continuity procedures are in place
   - ensuring that data users' devices are able to access the system.
3. **Data users**:
   - are responsible for accessing, entering, maintaining and using data in accordance with rules set by data owners
   - are responsible for ensuring that all access to data through their user account is relevant and appropriate to the work being undertaken
   - are responsible for ensuring that subsequent use and distribution of data accessed through their user account is valid and appropriate
   - must not disclose corporate data to unauthorised persons without the consent of the relevant data owner
   - must not disclose their password to anyone.

Storage and backup (short and mid-term)

D1 has the unlimited data storage at its disposal, and all data are being backed up on Google Cloud automatically. Data manager, Paolo de Stefanis is the person responsible for data backup and recovery.

Archiving and preservation (long term)

Data are stored in the Google Drive archive of the company, which is protected through state-of-the-art security features (double-access security).

2.1.4 IREN SPA (IR)

Data manager

Name: Nicola Bazzurro

Telephone number: +393355695217

E-mail: nicola.bazzurro@gruppoiren.it

Available / necessary resources

The common network and system infrastructure and solutions are centrally administered by a highly specialized IT operations team. The project specific infrastructure and solutions are administered directly by the project team or the organizational unit.

The infrastructure and solutions are built for a 24/7 operational availability. For collaboration and data exchange with colleagues or external partners, various managed file sharing services (e.g. Microsoft Teams, Box, OneDrive for Business) are available, where data recovery is partly possible, depending of the chosen service.

Data access and security

All the raw data are saved within accounts that are protected by complex password. A remote access is possible through a secure VPN access solution.

An exclusive project folder will be created for the project. The project folder will only be accessible to the approved personnel and project team members who need access to complete their tasks. Permissions to other files are set by the data manager. Only the people with read, write and delete access are permitted to add data.

Authorization is done based on group membership specific to the research project. Group membership to read or alter data is granted by the project manager.

Storage and backup (short and mid-term)

MARILIA project folders will be stored in shared drive and backed up daily. For non-electronic data, the data will be scanned and converted to electronic PDF files.

Data backup and recovery of the central infrastructure is the responsibility of the central IT.

Archiving and preservation (long term)

An exclusive project folder will be created for MARILIA project. The MARILIA project folder will only be accessible to approved personnel and project team members who need access to complete their tasks. The access control is set as (1) No Access, (2) Access with Read only, (3) Access with Read, Write and Delete.

2.1.5 FACULTY OF SCIENCE UNIVERSITY OF ZAGREB (UZ)

Data manager

1. Name: Branimir Bertoša

Telephone number: +385 915026969

E-mail: bbertosa@chem.pmf.hr

2. Name: Zoe Jelić Matošević

Telephone number: +385919313217

E-mail: zoejm@chem.pmf.hr

Available / necessary resources

All the data will be saved at the work stations that will be purchased within the project, and the additional backup of the data will be made using large capacity external hard disks. The access will be password protected and provided to all team members.

The costs for making data FAIR related to GOLD OPEN access publication of results and are covered by the corresponding research project. The previously mentioned data managers will decide on time resolution of the simulations that will be kept in long-term storage, and summaries of key analyses and raw analysis data will also be kept.

All the resources as mentioned are available for the long-term preservation (over 10 years).

Data access and security

All the data will be saved at the work stations that will be used to produce the simulations, the backup will be made using large capacity external hard disks. The access to the data will be password protected and provided to all team members, as well as to the consortium members from RBI and AIT.

Storage and backup (short and mid-term)

Computer simulations will be stored on the workstations at which they will be produced. Backup of the data on a monthly basis will be provided by saving data on high capacity external disk drive, stored with the Group Head (Branimir Bertoša), who authorizes access. That ensures mid-term data storage (up to 5 years). In case of an accident, the external drives with backups are stored separately from the workstations.

Archiving and preservation (long term)

All the resources as mentioned are available for the long-term preservation (over 10 years).

2.2 Data set descriptions

2.2.1 MARILIA-AIT-001

Data set reference / name / creator:

(Internal) reference: MARILIA-AIT-001

Persistent and unique identifier: Not available yet.

Name: Electrophoresis image data.

Created by: AIT: Natalie Mutter.

Data set description:

Data format: Electronic: tif, jpeg, png.

Origin of the data: Lab experiments.

Hardware used for data generation: Imaging systems, ChemiDoc Touch (BioRad), UVP GelDoc 310 (Biospectrum).

Software used for data generation: Image Lab Software (BioRad), VisionWorks LS Software.

Typical file size: Kilobytes.

Approximate amount of data: Megabytes.

Short description of the data: The images recorded show the result of DNA electrophoresis or protein electrophoresis experiments. Detailed descriptions will be found in the electronic labbook.

Purpose of data generation: Evaluation of the cloning, expression and purification of recombinant proteins required for the development of the pathogen detection assay (WP2).

Standards and metadata / Data interoperability:

In our electronic labbook (eLabFTW) each image will be liked to a specific experiment where a detailed description exists of the experimental procedure.

Data sharing and data re-usability:

Dissemination level: CO.

Embargo period: Until publication/patent.

Repository/repositories used or planned for upload: Published data will be made available via Pubmed.

Further details on data sharing: The data will be accessible and shared within the AIT competence unit Molecular Diagnostics.

Explanation why CO data cannot be made public: The data cannot be shared due to intellectual property and commercial issues.

Data access and security:

As described in the general part.

Storage and backup (short and mid-term, during the project):

As described in the general part.

Archiving and preservation (long term, after the project):

As described in the general part.

Ethical aspects:

None

2.2.2 MARILIA-AIT-002

Data set reference / name / creator:

(Internal) reference: MARILIA-AIT-002

Persistent and unique identifier: Not available yet.

Name: Cloning, expression and purification protocols.

Created by: AIT: Natalie Mutter, Yasaman Ahmadi.

Data set description:

Data format: Electronic: pdf, doxc, xlsx.

Origin of the data: Lab experiments.

Hardware used for data generation: Nanodrop 2000C, Epoch Microplate spectrophotometer (Biotek).

Software used for data generation: Word, Excel, Adobe, eLabFTW, NanoDrop 2000/2000c software, Gen5 data analysis software.

Typical file size: Kilobytes.

Approximate amount of data: Megabytes.

Short description of the data: Data from the preparation and amplification of plasmid DNA and expression and purification of proteins. Quality control data of the plasmids or proteins. Detailed descriptions will be found in the electronic labbook.

Purpose of data generation: Preparation of recombinant proteins required for the development of the pathogen detection assay (WP2).

Standards and metadata / Data interoperability:

In our electronic labbook (eLabFTW) all data and protocols will be liked to a specific experiment where a detailed description exists of the experimental procedure.

Data sharing and data re-usability:

Dissemination level: CO.

Embargo period: Until publication/patent.

Repository/repositories used or planned for upload: Published data will be made available via Pubmed .

Further details on data sharing: The data will be accessible and shared within the AIT competence unit Molecular Diagnostics.

Explanation why CO data cannot be made public: The data cannot be shared due to intellectual property and commercial issues.

Data access and security:

As described in the general part.

Storage and backup (short and mid-term, during the project):

As described in the general part.

Archiving and preservation (long term, after the project):

As described in the general part.

Ethical aspects:

None.

2.2.3 MARILIA-AIT-003

Data set reference / name / creator:

(Internal) reference: MARILIA-AIT-003

Persistent and unique identifier: Not available yet.

Name: DNA and oligonucleotide sequence data.

Created by: AIT- Ivan Barisic, Yasaman Ahmadi, Natalie Mutter.

Data set description:

Data format: Electronic: xlsx, txt, fasta, ab1.

Software used for data generation: Excel, BioEdit, ApE plasmid editor.

Hardware used for data generation: PCs.

Typical file size: Kilobytes.

Approximate amount of data: Megabytes.

Short description of the data: DNA sequence data is a letter code comprising A (adenine), C (cytosine), G (guanine) and T (thymine) corresponding to a nucleotide. The sequence data either will be used to synthetize DNA or will be obtained from sequencing experiments to control cloned DNA. Some sequences will be published within scientific publications.

Purpose of data generation: Preparation of recombinant proteins required for the development of the pathogen detection assay (WP2).

Standards and metadata / Data interoperability:

Oligonucleotide sequences will be saved together with the additional information's such as their corresponding name, plasmid, restrictions sites.

Data sharing and data re-usability:

Dissemination level: CO.

Embargo period: Until publication/patent.

Repository/repositories used or planned for upload: Published data will be made available via Pubmed.

Further details on data sharing: The data will be accessible and shared within the AIT competence unit Molecular Diagnostics.

Explanation why CO data cannot be made public: The data cannot be shared due to intellectual property and commercial issues.

Data access and security:

As described in the general part.

Storage and backup (short and mid-term, during the project):

As described in the general part.

Archiving and preservation (long term, after the project):

As described in the general part.

Ethical aspects:

None.

2.2.4 MARILIA-AIT-004

Data set reference / name / creator:

(Internal) reference: MARILIA-AIT-004

Persistent and unique identifier: Not available yet.

Name: Source code for software.

Created by: AIT – David Kuťák, Lucas da Cunha Melo, Haichao Miao.

Data set description:

Data format: Electronic - various source code files (.js, .ts, .html, .py, …).

Hardware used for data generation: Personal computers.

Software used for data generation: IDEs (integrated development environments).

Typical file size (for electronic data): Kilobytes to Megabytes.

Approximate amount of data: Megabytes to Gigabytes.

Short description of the data: During the project, several files with source code will be created in order to develop new software tools. Selected parts of the software and developed ideas will be made available via scientific publications.

Purpose of data generation: Development of a software for molecular visualization and modelling.

Re-use of existing data: Software is built on top of existing publicly available frameworks.

Standards and metadata / Data interoperability:

Documentation of the source code will be either created directly in the corresponding source file or separately in an additional document (metadata). Manuals of the software will be stored in the repository system. If possible, standardized input and output formats will be used to allow interoperability with other software applications.

Data sharing and data re-usability:

Dissemination level: CO.

Explanation why these data cannot be made public: The data cannot be shared due to intellectual property and possible commercial and publication issues.

Embargo period: None.

Repository/repositories used or planned for upload: The source code will be stored in a distributed revision control system that will be hosted on the GitHub servers.

Further details on data sharing: Several core repositories will be used for storing the different branches of the software codebase and its individual parts. In addition, each contributor will have the possibility to keep own versions of the software in their own repository.

Data access and security:

In addition to the description in the general part, access to the data will be given on a per-user basis. The repository will be hosted on the GitHub servers and accessible only via secure authentication to approved users.

Storage and backup (short and mid-term, during the project):

The data will be stored on the GitHub servers as well as cloned on the computers of the individual software contributors. The available storage space should be sufficient.

Archiving and preservation (long term, after the project):

As described in the general part.

Ethical aspects:

None.

2.2.5 MARILIA-AIT-005

Data set reference / name / creator:

(Internal) reference: MARILIA-AIT-005

Persistent and unique identifier: Not available yet.

Name: Bacteria-binding protein (BBP) related data

Created by: AIT- Yasaman Ahmadi.

Data set description:

Data format: XLSX, docx.

Origin of the data: Online database tool (CAMP, APD, DBAASP, BACTIBASE, etc.) and lab experiment.

Hardware used for data generation: Enspire plate reader.

Software used for data generation: Excel, docx.

Short description of the data: List of suitable BBPs, found in different databases (CAMP, APD, DBAASP, BACTIBASE, etc.) and details for each BBP (including target organism, PDB file, purification method, etc) are listed. The data will be used to select BBP. Protocols and data will be generated to measure and record the catalytic activity of horseradish peroxidase (HRP) enzymes and to analyse the affinity of expressed BBP-HRP conjugates to different bacteria.

Purpose of data generation: To make a list of available and suitable BBPs to use it for conjugation to split or full HRP enzyme, and to evaluate enzymatic activity and affinity of expressed BBP-HRP conjugate to find the best binders.

Standards and metadata / Data interoperability:

No existing standards for reference.

Data sharing and data re-usability:

Dissemination level: CO. The data cannot be shared due to intellectual property and commercial issues.

Embargo period: Until publication/patent.

Potential Users: Data is useful for consortium partners.

Repository/repositories used or planned for upload: The data will be saved in the online labbook and internal MARILIA folder and will be linked to the specific experiment where a detailed description exist of the experimental procedure.

Further details on data sharing: The data will be accessible and shared within the AIT business unit Molecular Diagnostic.

Licences for data re-use: N/A.

Data access and security:

As described in the general part.

Storage and backup (short and mid-term, during the project):

As described in the general part.

Archiving and preservation (long term, after the project)

As described in the general part.

Ethical aspects:

None.

2.2.6 MARILIA-RBI-001

Data set reference / name / creator:

(Internal) reference: MARILIA-RBI-PO

Persistent and unique identifier: Not available yet.

Name: Synthetic procedures for protein-oligonucleotide conjugates (POC).

Created by: RBI: I. Piantanida, Ž. Ban, I. Crnolatac, D. Saftić.

Data set description:

Data format: Electronic: Word Documents, Origin files, spectrophotometric data, mass spectrometry data, microcalorimetry data.

Software used for data generation: Microsoft Word, Origin software package, FS5 Edinburgh Inst. Fluorimeter, Cary WinUV, Cary Eclipse, Jasco Spectra Manager, Agilent MassHunter, TA Instruments NanoAnalyzer, TA nanoDSC.

Hardware used for data generation: Spectrophotometers, LC-MS, Microcalorimeters, gel-electrophoresis.

Short description of the data: Synthetic procedures used for POC preparation. LC-MS analysis for structure elucidation, Spectrophotometric, gel-electrophoresis and Microcalorimetric measurements for characterisation.

Purpose of data generation: Preparation and characterisation of POCs.

Re-use of existing data: No.

Standards and metadata / Data interoperability:

Word document (docx) files will be used for reporting and storage of data where possible, due to its interoperability. The synthetic procedures will be described and saved in docx files, gel-electrophoresis images will be incorporated into docx. The synthetic procedures metadata will be stored in laboratory notebooks. The Instrument-generated raw spectrophotometric data will be imported in Origin and the resulting diagrams transferred into docx. The LC-MS data will be saved as portable document format (pdf) files. The microcalorimetric instrument generated raw data will be analysed, and the resulting diagrams will be transferred in docx. The raw metadata from all instruments will be collected and stored on the corresponding instruments hard disks in a separate folder with a project identifier (MARILIA).

Data sharing and data re-usability:

Dissemination level: CO = Confidential, only for members of the consortium (including the Commission Services), as the experimental details are the potential material for patent protection.

Embargo period: Until work is published, or patent filed, foreseen 2 years after the end of the project.

Potential Users: Consortium partners.

Repository/repositories used or planned for upload: Once the results are published, Ruđer Bošković Institute has “Full-text Institutional Repository” (FULIR, http://fulir.irb.hr/).

Further details on data sharing: The data will be accessible to and shared between the members of the consortium.

Licences for data re-use: Licences for data re-use are issued by the project coordinator, AIT.

Data access and security:

As described in the general part.

Storage and backup (short and mid-term, during the project):

As described in the general part.

Archiving and preservation (long term, after the project):

All the resources as mentioned are available for the long-term preservation (over 10 years).

Ethical aspects:

None.

2.2.7 MARILIA-D1-001

Data set reference / name / creator:

(Internal) reference: MARILIA-D1-DS01

Persistent and unique identifier: N/A

Name: House of Quality Data

The data set includes the replies to questionnaires made with stakeholders for the assessment of the product required features. Such replies are both qualitative and quantitative and are collected in a form which is then used to fill out the so called “House of Quality”, as described in deliverable D1.2.

Created by: D1: Paolo De Stefanis.

Data set description:

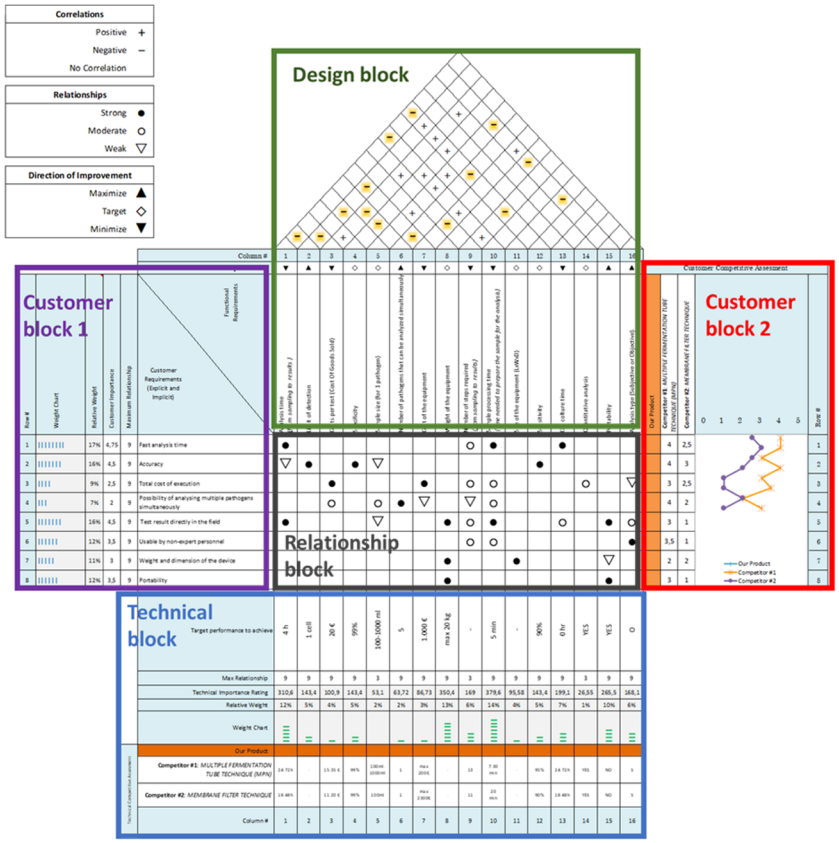
Data format: The data are first collected in a text document (.doc) and subsequently processed through a spreadsheet (.xls), which is designed to replicate the House of Quality.

Origin of the data: Interviews with stakeholders.

Hardware used for data generation: No hardware used.

Software used for data generation: Microsoft Word and Microsoft Excel.

Short description of the data: Data are collected in a spreadsheet organised as described by the following figure:



Purpose of data generation: The purpose is to help the product development team to understand the key product features required by the intended customers, prioritise them, and compare the target technical specifications with the best alternatives on the market.

Re-use of existing data: No.

Standards and metadata / Data interoperability:

Data are commented in a document, which gives a more detailed interpretation of the values included in the House of Quality, for supporting the product development team taking decisions on the features to include. No keywords nor metadata are included or are to be used. Data collected are in principle confidential, so their purpose is not to be shared with the research or industrial community.

Data sharing and data re-usability:

Dissemination level: CO = Confidential, only for members of the consortium (including the Commission Services). Data are to be kept confidential as they represent the key features to be included in the product to be commercially exploited and describe the innovation that will distinguish the product from the competitors. By sharing such data, the exploitation potential of the Marilia solution could be affected.

Embargo period: 5 years from the project end date.

Potential Users: Such data are useful for the consortium partners to take decisions on the product features to implement.

Repository/repositories used or planned for upload: Data are stored in the project internal repository. The data are available in .doc and .xls format and are freely accessible by the consortium partners.

Licences for data re-use: N/A

Data access and security:

Data includes the features of state-of-the-art devices, which are not sensitive, and prospected features to be included in the product, which are per se sensitive. There is no risk that the data is procured illegally and manipulated. Data are stored in D1's cloud repository, which is protected through state-of-the-art data security infrastructure, and through the project repository. Data are accessible to the project partners and D1's personnel. Username and password must be provided to access the files.

Storage and backup (short and mid-term, during the project):

As described in the general part.

Archiving and preservation (long term, after the project):

As described in the general part.

Ethical aspects:

None.

2.2.8 MARILIA-IR-001

Data set reference / name / creator:

(Internal) reference: MARI-IREN-DS01

Persistent and unique identifier: N/A

Name: Microbial quantification data.

Created by: IR: Nicola Bazzurro.

IREN will provide Requirements for evaluating the performance of quantification methods for nucleic acid target sequences — quantitative (qPCR) and digital (dPCR) real-time PCR, providing generic requirements for evaluating the performance, and ensuring the quality of methods used for the quantification of specific nucleic acid sequences (targets). The standard is applicable to the quantification of DNA (deoxyribonucleic acid) target sequences using either dPCR or qPCR amplification technologies.

Iren also provides reference to the microbiological analysis. Bacteria presence is detected using selective media. Filtering membranes (0.45 µm filter) are used to filter the water. Subsequently, the filters are placed on selective media plates, incubated and then colonies are counted. The following parameters are screened: total colony count at 37°C (most-probable number (MPN)/100ml), *E. coli* (MPN/100ml). Commercially available detection units such as the Colilert and Enterolert are also used. More precisely, *E. coli* are detected using Colilert®, a commercially available enzyme-substrate liquid-broth medium (IDEXX Laboratories, Inc., Westbrook, Maine) that allows the simultaneous detection of total coliforms and *E. coli*. It is available in the MPN or the presence/absence (PA) format. The MPN method is facilitated by using a specially designed disposable incubation tray called the Quanti-Tray®. Both methods can be done in the field or in laboratory.

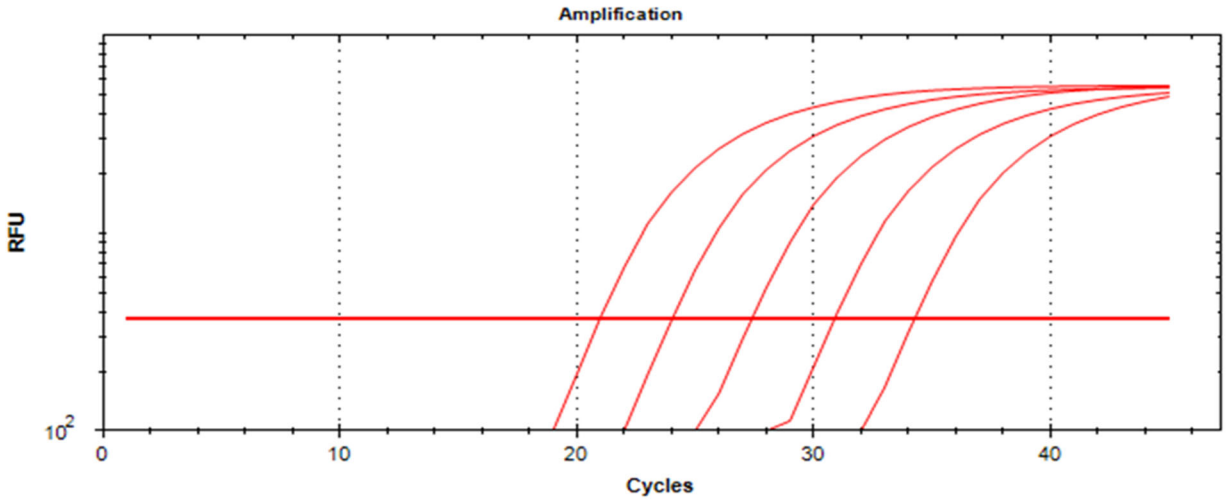
Results as MPN are calculated following ISO regulation qPCR real-time absolute quantification curves.

Each sample will be subjected to DNA extraction and quantification in terms of amount of *Escherichia coli* genomic copies per liter by means of real-time qPCR.

This technique allows to monitor DNA amplification in real time through monitoring of fluorescence.

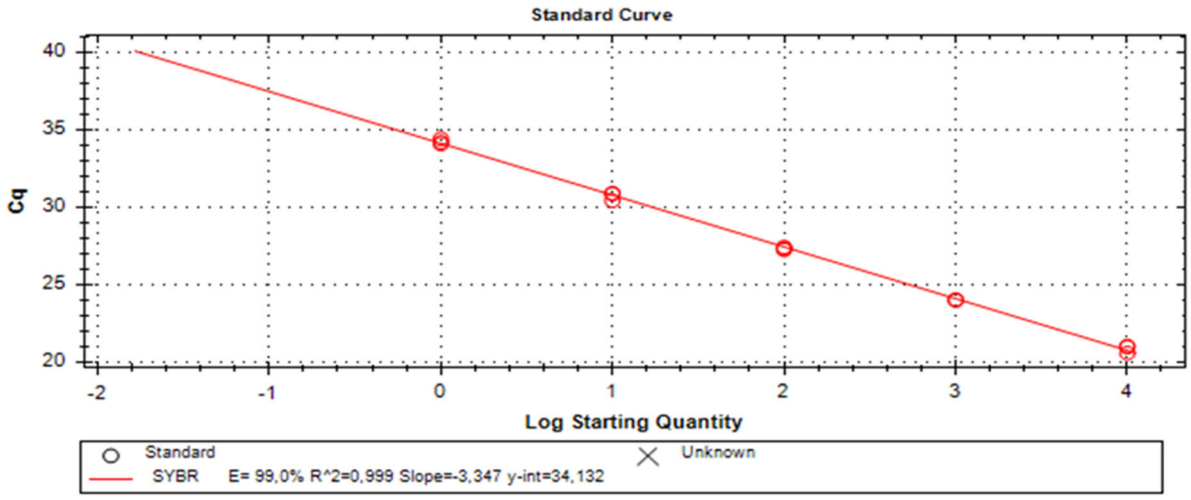
Fluorescence signal is measured after each cycle and the intensity of the fluorescent signal reflects the momentary amount of DNA amplicons in the sample at that specific time.

In the figure below cycles of PCR represented as function of fluorescence signal (RFU).



The point at which the fluorescence intensity increases above the detectable level corresponds proportionally to the initial number of template DNA molecules in the sample. This point is called the quantification cycle (Cq; different manufactures of qPCR instruments use their own terminology, but since 2009, the term Cq is used exclusively) and allows determination of the absolute quantity of target DNA in the sample according to a calibration curve constructed of serially diluted standard samples (usually decimal dilutions) with known concentrations or copy numbers.

The figure below shows a standard curve used to determine the unknown concentration in terms of genome copy (log starting quantity) as a function of Cq.



Data set description:

Data format: pdf, pcrd, xls, jpeg, txt, laboratory notebook.

Origin of the data: Lab experiments aimed at validating *E. coli* detection method, public archives, standard samples: *Escherichia coli* (Migula) Castellani and Chalmers (ATCC® 10798), database describing all the samples used for the validation.

Hardware used for data generation: Realtime qPCR output files-CFX 96 wells Biorad, can be read with the dedicated app, namely MAESTRO, available from Biorad.

Software used for data generation: Excel, power point, word, MAESTRO (BIORAD), photoshop, anteprima.

Short description of the data: Each column of the database file (created with excel) will contain name of the sample numerically assigned, source of water used for each test, number of bacterial colonies; date of creation of the sample; date of DNA extraction; date of analysis by qPCR; amount of E. coli copy detected. Each row will contain data for one given sample.

Purpose of data generation: Evaluation of water sample with reference methods (WP4 T4.2).

Re-use of existing data: Re-use of any existing data is not envisaged at this time.

Standards and metadata / Data interoperability:

The analytical data deriving from the chemical and biological analyses conducted in the laboratory are managed through the LIMS software LABVANTAGE 6 which allows end users to manage the laboratory flow, supporting them in the following activities: Sample log, Sample reception, Management and approval of results, Sample Review, Sample Approval, Reporting Tools (CoA, Test Reports, Spreadsheets, Trends, etc).

LABVANTAGE 6 also allows the configuration of the following master data: Test Methods, Parameters, Measure Units, Products, Sample Templates, Formulations, Batch templates, Specification Limits, Instruments, Reagents. LABVANTAGE 6 consists of the following main pages (as default): LIMS, Lab Admin, System Admin.

The functionality of data extraction from LABVANTAGE to Excel consists in the generation of a report in xls format having in the first column the list of all the analyses for the chosen samples.

In the following columns an example of the results for each sample is reported:

Sample_ID			210044863	210044864	210044865
Point Code			P01 grezza	P02 grezza	P03 rete
Point description					
Note					
Customer					
Delivery Data			23/01/2020	23/01/2020	23/01/2020
Sample Data			23/01/2020	23/01/2020	23/01/2020
Sample procedure					

The name of the generated excel file has the following format: DD-MM-YYYY_HH-MI.xls.

The report is generated after having searched the samples according to specific queries (by instrument, by method or other), and having selected the samples to be extracted (there is possibility to select a group of continuous or discontinuous samples, and heterogeneous between them - this is the case with different analyses). In the event that no type of extraction is selected, the procedure extracts the parameters present in the selected samples according to the indication of one of the different types of grouping provided in the parameter list.

The features of Excel extraction are described below.

There is only one model and corresponds to the template described (except for the order of the extracted parameters). The extraction of the results will take place after they have been entered in the LIMS.

The LIMS calculates the results (e.g.: ionic charges) thanks to the implementation of Calculation primitives.

In the event that it is realized that a reanalysis is necessary, with consequent modification of a result, after the extraction has taken place and the comparison with historical data has been carried out, the result must be modified in the LIMS and the sample must be re-extracted.

After selecting the samples to be extracted, the user can choose the type of extraction from those configured on the setup page or choose from one of the fields shown in the parameter list.

Data sharing and data re-usability:

Dissemination level: CO = Confidential, only for members of the consortium (including the Commission Services). Data cannot be shared as Public for intellectual property and commercial reasons.

Embargo period: N/A.

Potential Users: Consortium partners.

Repository/repositories used or planned for upload: The produced data will be shared between the consortium partners in a manner described in Section 2.3 of this document.

Licences for data re-use: Not applicable as we do not envisage to re-use data.

Data access and security:

As described in the general part.

Storage and backup (short and mid-term, during the project):

As described in the general part.

Archiving and preservation (long term, after the project):

As described in the general part.

Ethical aspects:

None.

2.2.9 MARILIA-UZ-001

Data set reference / name / creator:

(Internal) reference: MARILIA-UZ-DS01

Persistent and unique identifier: Not available yet.

Name: MD-simulations.

Created by: UZ: Branimir Bertoša, Antun Barišić, Zoe Jelić Matošević, Sanja Škulj.

Data set description:

Data format: PDB files (Protein Data Bank format), trajectory files (for example amber PARM7 files, NetCDF files, xtc files - commonly used formats for computer simulation of biochemical systems), text files (input files for conducting simulations).

Origin of the data: Molecular dynamics simulations run with software that are commonly used in academic community (for example, Amber software and/or Gromacs).

Hardware used for data generation: Computer workstations and server Isabella at SRCE (https://www.srce.unizg.hr/isabella/). The exact specifications of machines are not yet available – machines are currently being purchased.

Software used for data generation: MODELLER (for PDB models), Amber, Gromacs.

Short description of the data: Dataset MARILIA-UZ-001 will contain all molecular dynamics simulations obtained within the project as well as the relevant input files and analyses. Each simulation obtained by Amber software (in the NetCDF file format) is accompanied by the topology (Amber PARM7 format) and coordinate files (Amber coord file format), as well as input files for tleap (needed to generate topology and coordinate files from PDB file) and input files for molecular dynamics (needed to run the simulations). More detail about these formats can be found here: https://ambermd.org/FileFormats.php. All simulations performed by Gromacs software require topologies (in the top and tpr file format), parameter files (in the itp files) and coordinate files (usually in gro file format) as input files. All simulation details are written in mdp file format. As output trajectory files (usually xtc files) and other files (usually log, edr and cpt file formats) will be generated. More detail about these formats can be found here: https://manual.gromacs.org/documentation/current/reference-manual/topologies/topology-file-formats.html. Similar is with the other commonly used programs for computer simulations of biochemical systems. Files related to analysis of the simulations are deemed to be metadata and will also be provided as figures in .jpeg or .pdf format as well as raw data in regular .txt format.

Purpose of data generation: The purpose of generating molecular dynamics simulations is to gain insight about the systems studied in the project on a molecular level. The generated simulations can be used to predict the behavior of altered and engineered proteins and thus push the experimental work in the right direction, as well as to enable a deeper and more detailed interpretation of the experimental data.

Re-use of existing data: Published crystal structures from the Protein Data Bank (up to this point the structure under PDB ID 1H5A has been used) are used as a starting point for generating new models which will be used as starting structures for MD simulations. Sequences available at https://www.uniprot.org/ will likely be used for modelling proteins for which no crystal structure in the Protein Data Bank is available.

Standards and metadata / Data interoperability:

All the data contained in this dataset is in standard formats defined by the widely used program package Amber and the Protein Data Bank. It is possible that some of the data generated by analyses might be in formats that are not as widely used, but these will be followed by the appropriate annotation.

Data sharing and data re-usability:

Dissemination level: CO = Confidential, only for members of the consortium (including the Commission Services). Data are potential material for patent protection.

Embargo period: Until work is published, or patent filed, foreseen within 2 years.

Potential Users: This data can be useful to AIT as well as RBI as structural insights that will be gained from the simulations can help in resolving issues arising during protein purification or modification, planning of synthesis and choosing the strategies for protein purification or modification, synthesis, fusion protein generation etc. After being made publicly available, the data can be useful to other researchers interested in modelling the horseradish peroxidase enzyme.

Repository/repositories used or planned for upload: One of the workstations purchased at UZ from MARILIA project funds will be set up as a server on which data from MARILIA-UZ-001 will be stored. Consortium members will be provided with password-protected accounts so as to be able to connect via secure shell (ssh) and secure copy (scp) protocols, to view or download the data.

Licences for data re-use: N/A

Data access and security:

The data from MARILIA-UZ-001 will be made accessible as previously described over a secure shell protocol for consortium members via password-protected accounts. The data in MARILIA-UZ-001 does not contain any sensitive personal information so we think that the described system utilising password-protected access to a data storage server is sufficient, both in terms of accessibility and security.

Storage and backup (short and mid-term, during the project):

Monthly backup is provided by saving data on external disk drives, stored with Branimir Bertoša, who authorizes access. That ensures mid-term data storage (up to 5 years). In case of an accident, the external drives with backups are stored separately from main storage units (intranet connected) and can be accessed by Branimir Bertoša or other members of the team (Zoe Jelić Matošević, Antun Barišić, Sanja Škulj).

Archiving and preservation (long term, after the project):

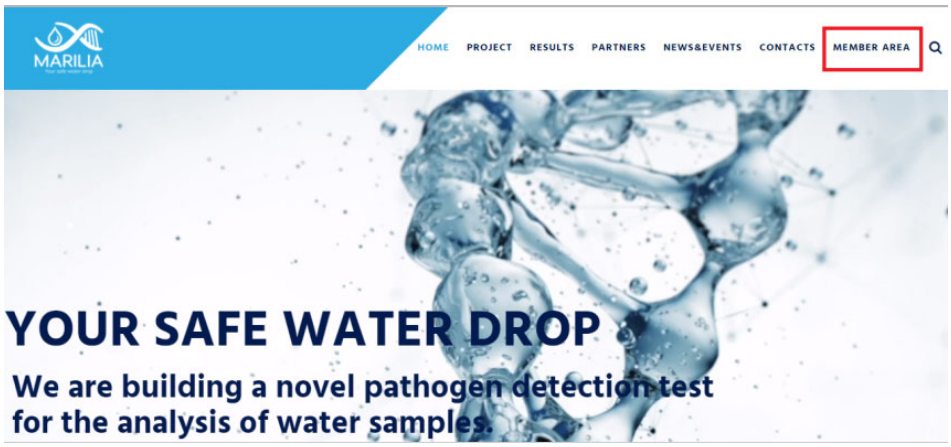
All the resources as mentioned are available for the long-term preservation (over 10 years).

Ethical aspects:

None.

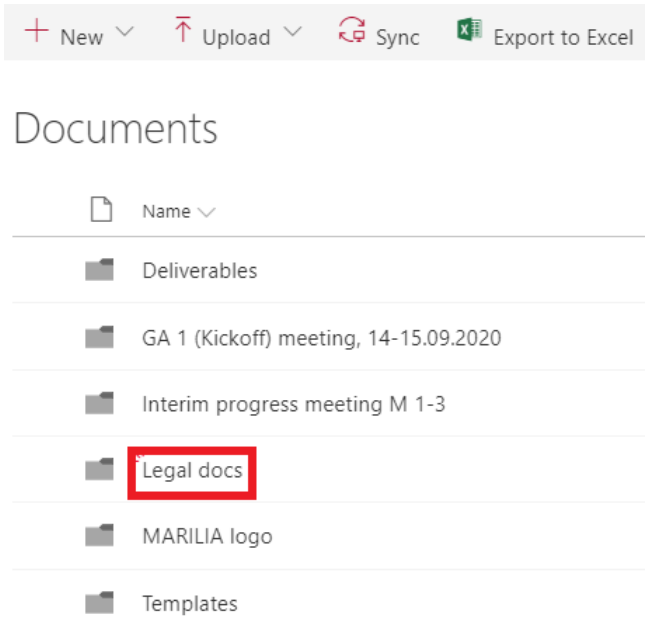
2.3 Data exchange within MARILIA consortium

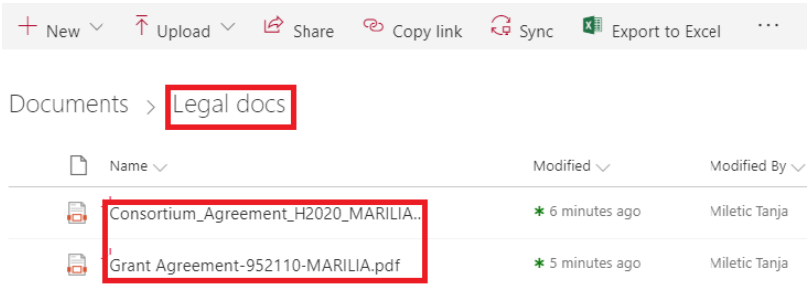
For the secure data exchange within the consortium, AIT has set up a secure exchange system accessible via a public web-address (https://portal.ait.ac.at/sites/marilia) which is also linked from the MARILIA website (https://www.mariliaproject.eu). The website includes a link to the project internal communication and management platform ("Member Area") developed in Microsoft SharePoint. This system has a very flexible design to tailor different sections to the specific needs of the project and is accessible only to MARILIA team members. User access rights to the exchange system are managed by the IT department of AIT, and access rights are granted by the Coordinator.



Included in the exchange system is a document centre where different documents can be centrally hosted and shared. This allows project members to access current versions of reference documents and guidelines.

All relevant project documents shall be stored on the SharePoint in the pertinent folder. For example: "Legal documents" folder will contain Consortium Agreement and Grant Agreement; "Templates" folder will contain project relevant templates to be used by the beneficiaries, such as Deliverable template; For every project meeting a separate folder is created, which will contain all the relevant documents which pertain to that meeting (agenda, minutes of the meeting, attendance lists).





In accordance with MARILIA project Grant Agreement Annex 1, Part B, the partners have pledged to adhere to the following guidelines:

➢ **Data curation and preservation**

- Uniform standards for data curation and preservation will be established and followed by all project partners.
- Each project partner will establish a central network drive for storing project-relevant data that is accessible only by project team members (password protection). Each folder on the network drive will allow applying specific read and writing access limitations, thus reflecting the category of data confidentiality.
- The data on the project network drives will undergo daily incremental and weekly full backups. The rooms for storing the backup tapes will be physically separated from the locations of the network drives.
- Data that originally incurs in non-electronic format will be scanned and saved electronically according to the standards cited above.
- Archiving periods will be followed according to the respective legislation for the specific data types with a minimal period of five years following the end of the project.

➢ **Data standards**

- Design data, raw data of measurements and metadata will be saved and archived in general accessible formats not requiring proprietary software.
- Project wide standards for data cataloguing and annotation will be agreed on in the DMP to ensure an efficient and unambiguous data retrieval.
- Standards for data annotation and metadata for publicly available data must be used according to the regulations of the chosen data repositories and will be determined during the project.

➢ **Data confidentiality**

Three different levels of confidentiality will be defined due to IP issues and marketing strategies:

- Level 1. Classified data
  These data cannot be accessible to the public at any time. This data comprises personal and medical data. However, it is not foreseen to collect such data within the MARILIA project.
- Level 2. IPR relevant data
  These data are confidential during the run time of MARILIA but may be made accessible to the public at a later stage. Potentially IPR relevant data that may be commercially or industrially exploited will be regulated by the consortium agreement that will be set up before the beginning of MARILIA. After the IPR are secured, the consortium will make the data accessible to the public (free of charge) provided there are no other legitimate reasons preventing this.
- Level 3. Public data
  This concerns data that can be made accessible to the public, e.g. there is no conflict with the obligation to protect results, the confidentiality obligations, the security obligations or the obligations to protect personal data.

The above stated conventions and guidelines underpin the consortiums commitment for making the data FAIR, in accordance with the FAIR[1] guiding principles for data resources, both for the use outside the consortium as well as within the consortium:

- Making data findable, including provisions for metadata
- Making data openly accessible
- Making data interoperable
- Increase data re-use (through clarifying licences)

3. Conclusion

The first steps in building and maintaining a consistent and well-documented data management have been undertaken, resulting in this first version of a data management plan. As the EC acknowledged in their "Guidelines on Data Management in Horizon 2020", a DMP is not a fixed document, but evolves during the lifespan of the project. Additional data sets therefore, might arise over the course of the project requiring the DMP to be updated accordingly during the project lifetime.

This first version of the DMP serves mainly as a survey to collect the information about which data management procedures are currently implemented by each beneficiary, and in which manner; also, the survey will aim to identify where some room for improvement might exist. As part of this action, we have also put efforts into raising an awareness between the consortium partners about the importance of good data management including the proper documentation.

The information collected from the consortium members for this initial version of the DMP also revealed an aspect of which we haven't been aware to its full extent during the writing of the MARILIA proposal: Although MARILIA has declared itself to be part of the "open data pilot" and the consortium is still committed to give the general public access to its research data, some data will have to be kept confidential until the related IPR is secured. Thus, at the current time, hardly any research data within MARILIA are labelled as PU ("public"). As soon as IPR is secured, we will change the status of data sets from CO ("confidential, only for members of the consortium and the involved EC services") to PU, provided there are no other reasons for confidentiality (as outlined in the MARILIA grant agreement). At this time, we will also choose the appropriate repositories and document them in the DMP.

The DMP itself is a public document and as such will be made accessible to the public via MARILIA project website.

[1] http://www.nature.com/articles/sdata201618