European Innovation Council

Funded by the European Union

HORIZON-EIC-2021-PATHFINDEROPEN-01

EIC Pathfinder Open 2021

VerSiLiB

Versatile Amplification Method for Single-Molecule Detection in Liquid Biopsy

Start date: 01/04/2022, Duration 48 months

Deliverable D7.3
Data Management Plan

Due date: 30/09/2022

Submission date: 26/09/2022

Responsible work package leader: Sanna Aikio, WP7, VTT

Deliverable type: Report

Version: V1.0

Dissemination level: Public

This project has received funding from the European Union's European Innovation Council (EIC) under Horizon Europe program. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council. Neither the European Union nor the granting authority can be held responsible for them.

Authors

Sanna Aikio, VTT
Sanna Uusitalo, VTT
Giuseppe Spoto, UNICT
Patrizio Giacomini, IRE
Jakub Dostalek, AIT & FZU
Stefan Fossati, AIT
Jiri Perutka, PRO
Prateek Singh, FAD

Revision log

Version	Date [DD/MM/YYYY]	Change
V0.1	16/09/2022	Final draft
V1.0	26/09/2022	Final

Reviewed by

Role	Name	Date
Work package leader	Sanna Aikio	16/09/2022
Coordinator	Sanna Aikio	26/09/2022

Keywords

Data summary, FAIR data, data security

Executive summary

This document contains the first version of the Data Management Plan describing project's data management practices. Plan will be updated annually and the final plan will be drawn up at the end of the project.

Contents

1. Data summary

Research data summaries by partner are provided in Sections 1.1-1.7 below. The summaries list the methods used to generate the data, covered subject areas, main formats and types of the datasets, estimated data volumes, and software used to analyse the datasets. The data analysed in connection with the project work will be produced by the project, but the possibility of reuse of any existing relevant open data will also be examined carefully.

Quality control measures will be taken to maintain the accuracy of data during the project. Potential re-utilization of any opened data will be ensured by careful documentation of datasets as well as description and publication of data collection methods, protocols, workflows and models.

1.1. VTT data summary

Research data will be generated with simulations using COMSOL Multiphysics simulation software (Comsol). The data will cover the main subject areas of the research: nanoparticle transport dynamics in the microfluidics. Main formats and types used for the datasets will be: mph. Estimated volume of data will be ca. 10 GB. Datasets will be processed and analysed using the COMSOL Multiphysics software and other numerical analysis software such as Matlab (Mathworks).

Research data will be generated with microscopes and collected using cameras. The data will cover the main subject areas of the research: experiments to study the nanoparticle transport dynamics in the microfluidics by magnetic actuation. Main formats and types used for the datasets will be: tiff, bmp. Estimated volume of data will be ca. 5 GB. Datasets will be processed and analysed using various analysis software.

Research data will be generated and collected using VerSiLiB reader. The data will cover the main subject areas of the research: development and integration of the magnetic cycler and the optical reader. Main formats and types used for the datasets will be: tiff, bmp, txt. Estimated volume of data will be ca. 10 GB. Datasets will be processed and analysed using numerical analysis software such as Matlab (Mathworks).

Research data will be generated with Veeco Dektak 150 profilometer and collected using computer system. The data will cover the main subject areas of the research: surface topographies and profiles of the master molds and replicated components. Main formats and types used for the datasets will be: surface profiles 2D data in data or csv file format. Estimated volume of data will be ca. ~0.2MB /dataset, typically in total some tens of MB in one project. Datasets will be processed and analysed using profilometer software (e.g. Vision, Gwyddion) or spreadsheet program (e.g. Excel or Origin).

Research data will be generated with Wyco NT3300 Optical profiler profilometer and collected using computer system. The data will cover the main subject areas of the research: surface topographies and profiles of the master molds and replicated components. Main formats and types used for the datasets will be: surface profiles 3D data (.opd file format). Estimated volume of data will be ca 1-2MB /dataset, typically < 1GB in one project. Datasets will be processed and analysed using profilometer software (e.g. Vision, Gwyddion).

Research data will be generated with Neoscope JCM-5000 JEOL Tapletop SEM and collected using computer system. The data will cover the main subject areas of the research: imaging and analysing physical dimensions of the master molds and replicated components. Main formats and types used for the datasets will be digital images in file formats such as .jpg, .tiff .bmp. Estimated volume of data will be ca. < 1 GB. Datasets will be processed and analysed using Microsoft Office tools Microsoft Office Picture Manager, Microsoft Word, Microsoft Power Point. And with graphics software like GIMP, Photoshop etc.

Research data will be generated with SmartScope ZIP 250 and collected using computer system. The data will cover the main subject areas of the research: optical imaging and analysing physical dimensions of the master molds and replicated components. Main formats and types used for the datasets will be digital images in file formats such as .jpg, .tiff .bmp. Estimated volume of data will be ca. < 1 GB. Datasets will be processed and analysed using Microsoft Office tools Microsoft Office Picture Manager, Microsoft Word, Microsoft Power Point. And with graphics software like GIMP, Photoshop etc.

1.2. UNICT data summary

Research data will be generated with fluorescence signals and collected using [a Leica DM IL LED inverted microscope equipped with a Leica DFC monochrome 7000 GT camera] system. The data will cover the main subject areas of the research: fluorescence detection from VerSiLiB prototype. Main formats and types used for the datasets will be: tiff. Estimated volume of data will be ca. 10 GB. Datasets will be processed and analysed using 'Fiji ImageJ 1.52' software.

Research data will be generated with UV-vis measurements and collected using VersaWave microvolume spectrophotometer system. The data will cover the main subject areas of the research: peptide nucleic acid probes characterization and magnetic microparticles functionalization strategy. Main formats and types used for the datasets will be: txt, csv. Estimated volume of data will be ca. 0.01 GB. Datasets will be processed and analysed using VersaWave software.

Research data will be generated with plasmonic biosensor and collected using Surface Plasmon Resonance imaging apparatus (SPRi) (GWC Technologies) system. The data will cover the main subject areas of the research: SPRi assay development for DNA target detection from biological samples with functionalized active surfaces. Main formats and types used for the datasets will be: tiff, csv. Estimated volume of data will be ca. 100 GB. Datasets will be processed and analysed using V++ (version 4.0, Digital Optics Limited) software.

1.3. AIT data summary

Within this project, following data will be generated at different stages and with a range of different applications. First, theoretical studies on the optical properties of metallic nanostructures will be generated using numerical simulations. The raw data comprises of files generated by FDTD-solutions simulation software by Lumerical (file type .fsp) and exported field data and spectra (file type .csv) and field distribution images (file typ .png).

Furthermore, experimentally prepared samples will be characterized using AFM from Molecular Imaging (file type .mi) and electron microscopy from Zeiss (file type .png). The optical characterization is carried out mostly on home-built systems, saving data in the ASCII format (file type .txt). Data analysis will be carried out in OriginLabs Origin (file type .opju).

The raw data will be processed, and the results summarized in form of deliverables (file type .docx) and presentations (file type .pptx) and finally published in the form of journal articles.

The data is organized in folders according to analysis type and each sample carries a unique identifier that allows to match the data obtained by different modalities in a logical manner. A text file including metadata about the data at hand will be included in the directory in human readable .txt format. Due to the large size of numerical simulation results, the total volume of data generated is estimated to be around 500GB.

The results of the theoretical studies are used to design and prepare nanophotonic surfaces, and the results of the experimental characterization are used to iteratively improve the designs in the numerical simulations.

Protocols, simulations and optical measurement results obtained in previous projects carried out by the project partners will be used to inform the design of novel nanophotonic architectures.

The data generated during this project will be essential to achieve the project goals of delivering high performing substrates for plasmon enhanced fluorescence. It will be further interesting for researchers and development engineers in the field of optical biosensing and fluorescence imaging.

1.4. IRE data summary

Research data will be generated by a variety of pieces of equipment, depending on the specific task and intended deliverable.

Genomics: the BRAF mutational status will be assessed in genomic tumor DNA (gDNA) and circulating tumor DNA (ctDNA) by Next Generation Sequencing (NGS). Targeted OncoMine and panCancer ThermoFisher panels will be run on the ThermoFisher Ion S5 sequencer. Alternatively, digital PCR (dPCR) data will be generated using custom BRAF assays on the QuantStudio 3D (ThermoFisher).

Proteomics: there are several ways CSPG4 expression will be measured: flow cytometry with specific antibodies on a ThermoFisher Attune acoustic flow cytometer. Sandwich ELISA on a semiautomated ELISA reader.

Antibodies: data will be generated by growing/cryopreserving hybridomas through standard tissue culture equipment. Antibodies will be harvested from supernatants. Complementarity Determining Regions will be assessed by NGS of cDNA using a variety of targeted and untargeted NGS approaches. Scatchard plots will be performed by home-made ultrafast equipment for quick ligand-to-ligand equilibrium relaxation.

General Biochemistry: nucleic acid and protein quality will be assessed by electrophoretic assays and by built-in assay matrics (NGS).

Human samples and data: gDNAs and ctDNAs will be obtained following informed consent and stored in our Institutional Biobank in compliance with EU regulation. Data will be elaborated by RedCAP on our GDPR-compliant institutional platform (pseudoanonymised). Data to be shared with the partners will be double-blinded. RedCAP generates a number of FHIR-compatible, interoperable outputs for data handling.

All activities: Datasets will be processed and analysed by descriptive statistics using standard statistical packages, e.g. by IBM-SPSS v.21.0.

1.5. FZU data summary

Within this project, following data will be generated at different stages and with a range of different applications. First, theoretical studies on the optical properties of metallic nanostructures will be generated using numerical simulations. The raw data comprises of files generated by FDTD-solutions simulation software by Lumerical (“.fsp”) and exported field data and spectra (“.csv”) and field distribution images (“.png”).

Furthermore, experimentally prepared samples will be characterized using AFM from Molecular Imaging (“.mi”) and electron microscopy from Zeiss (.png). The optical characterization is carried out mostly on home-built systems, saving data in the ASCII format (“.txt”). Data analysis will be carried out in OriginLabs Origin (“.opju”).

The raw data will be processed, and the results summarized in form of deliverables (.docx) and presentations (.pptx) and finally published in the form of journal articles.

The data is organized in folders according to analysis type and each sample carries a unique identifier that allows to match the data obtained by different modalities in a logical manner. A text file including metadata

about the data at hand will be included in the directory in human readable .txt format. Due to the large size of numerical simulation results, the total volume of data generated is estimated to be around 500GB.

The results of the theoretical studies are used to design and prepare nanophotonic surfaces, and the results of the experimental characterization are used to iteratively improve the designs in the numerical simulations.

Protocols, simulations and optical measurement results obtained in previous projects carried out by the project partners will be used to inform the design of novel nanophotonic architectures.

The data generated during this project will be essential to achieve the project goals of delivering high performing substrates for plasmon enhanced fluorescence. It will be further interesting for researchers and development engineers in the field of optical biosensing and fluorescence imaging.

1.6. PRO data summary

Research data will be generated and collected with an Illumina MiSeq instrument (2 × 300). The data will cover the main subject areas of the research: affinity modulation of nanobodies against human CSPG4 antigen. The main formats and types used for the datasets will be FASTQ. The estimated volume of data will be ca. 4 GB. Datasets will be processed and analyzed using Illumina software and custom-built scripts and algorithms.

Research data will be generated and collected with an ELISA reader (photometer). The data will cover the main subject areas of the research: affinity modulation of nanobodies against human CSPG4 antigen. The main formats and types used for the datasets will be CSV. The estimated volume of data will be ca. 100 MB. Datasets will be processed and analyzed using Microsoft Excel.

1.7. FAD data summary

Research data will be generated with 'Solidworks' 3D CAD design software. The data will cover the main subject areas of the research: 3D design of the microfluidic chips. Main formats and types used for the datasets will be: STEP. Estimated volume of data will be ca. 10 GB.

Research data will be generated with 'COMSOL Multiphysics' simulation software. The data will cover the main subject areas of the research: Flow in microfluidic systems. Main formats and types used for the datasets will be: mph. Estimated volume of data will be ca. 20 GB. Datasets will be processed and analysed using COMSOL Multiphysics software.

Research data will be generated with microscopes and collected using cameras. The data will cover the main subject areas of the research: experiments to validate the accuracy of the fabricated chips. Main formats and types used for the datasets will be: tiff. Estimated volume of data will be ca. 5 GB. Datasets will be processed and analyzed using imageJ software.

2. FAIR[1] data

2.1. Making data findable, including provisions for metadata

Discipline compliant metadata elements will be used describing the data to aid data discovery and potential re-use. List of metadata elements and metadata standards used are provided in a separate spreadsheet. Metadata including descriptions and keywords of opened data will be made available via FAIR compliant

[1] FAIR = Findable, Accessible, Interoperable, and Reusable

repository for searching and discovery after project closure. Persistent identifiers provided by the repository will be used in identifying and linking to datasets.

General overview										
ID	WP	Resource type	Title	Version	Date of creation	Owner	Creator	Contributors	Software used to create data	Origin and method
1.1	*X*	*Dataset*	*Dataset title*	*X.X*	*DD.MM.YYYY*	*Partner abbreviation*	*N.N*	*N.N. (DataCurator), N.N (DataCollector), N.N (ContactPerson)*	*Name of the software*	*Collected during the project with XX sofware and open source packages YY and ZZ*

Content description			Technical description			
Description and relation to the project objectives	Subjects (keywords)	List of variables	Software used to create the file	File format	Necessary software	File size
Short description	*XXX, YYY, ZZZ*	*TBD*	*Software name*	*.xxx*	*Description of compatible software to use the data*	*ca. XX MB*

Sharing and preservation							
Use / Users	Rights / Licence	Dissemination action	Communication action	Access	Restrictions	Repository for open data	Permanent identifier (e.g. DOI, URN)
TBD	*Creative Commons licence e.g. CC BY-SA*	*Add, if any*	*Add, if any*	*Open (permission based)*	*None (but some part of data might be proprietary)*	*e.g. Zenodo*	*Add, if any*

Figure 1 Spreadsheet (shown here piecewise) to collect information of the data sets. Spreadsheet contains sections for 'General information', 'Content description', 'Technical description' and 'Sharing and preservation'.

2.2. Making data openly accessible

Decisions concerning the sharing of datasets will be taken by the decision-making body of the consortium. Coordinator in collaboration with project participants will take all the appropriate measures to make relevant data openly available and usable for third parties for study, teaching and research purposes.

If, after project closure, permission to re-use the data is required, all requests for further use of data will be considered carefully and whenever possible approved by the person mandated with the task. Permission for data use will be granted providing there are no IPR or confidentiality issues involved or any direct overlap of research questions with the primary research. Permission will be provided by request using the appropriate procedure described in connection with other metadata.

Primary focus in data sharing will be on the data underlying prospective scientific publications ensuring the validation of results presented in publications. In addition to summary data, also operational or raw data will be opened, when benefits and possibilities for successful raw data re-use are recognized and there are no confidentiality or commercialization issues involved / identified.

Published and FAIR-compatible data will be archived in a public and trusted repository. Unless no discipline-specific archive platform is available, generic and certified repository services using standardized access protocols e.g., CSC's IDA, CERN's Zenodo or EUDAT's B2SHARE, will be used to enhance long-term accessibility and re-usability of the data. Metadata of the datasets will be opened under public and open copyright license, CC0.

Justification for possible case-specific embargo for published data will be decided by project consortium. Embargo will be sought, if necessary, in connection with possible IPR protection or any potential patent, utility mode etc. application based on project results.

No definite period or time limit is planned for access to data. However, the opened data will be deposited in a repository, which guarantees for foreseeable future the data integrity on bit level. No perpetual data curation policy to guarantee full long-term digital preservation of datasets is planned at this point.

2.3. Making data interoperable

Variables and value names will be constructed following general data processing conventions and standards common to the research subject. List of value names and used vocabulary will be provided in a separate list. Examples of vocabulary information to be managed within the project will be e.g. units of observation, list of variables with the name and label of each variable as well as its values and value labels, frequency distribution of each variable, information on the classifications used and meanings of abbreviations used.

2.4. Increasing data re-use

After the project completion ownership of datasets will belong to the grant beneficiaries that generated them. Creative Commons license CC-BY-SA or CC-BY or similar public copyright license will be used for any opened datasets, unless there are compelling reasons to select more restricted type of public license. Creative commons licenses will by default include also a disclaimer of liability for the re-use of opened data.

Data quality will be assured by following appropriate quality control and curation methods e.g., rigorous control of any incoming data by well-managed data profiling (formats, value distributions and data consistency and completeness will be assessed for any incoming data); logically defined data pipeline with centralized data management preventing duplicate data entering the system; capturing and documenting data conditions and scenarios with their dependencies and conditions; maintaining data integrity with checksums and triggers, if necessary; enhancing data and metadata lineage traceability for the pipeline, thus enabling more effective data governance. Research teams will regularly check the quality of not just the data, but also related software, algorithms, and workflows when and if changes are made in them.

3. Other research outputs

Any other project outputs, which will be needed for verifying or analysing the data - software, algorithms, workflows, protocols or models - will be opened alongside the corresponding data.

4. Allocation of resources

Making research data quality-controlled, FAIR-compatible and as open as possible should be considered by the consortium members while allocating resources to the project. Costs related to making data and other research outputs FAIR may include direct and indirect costs. Direct costs are included in the budget estimates, considering the eligibility rules and the usual accounting practises of each project participant.

The direct costs may include for example personnel costs and costs related to storage, archiving, re-use and security. The costs are estimated and included in the Grant Agreement budget, and actual costs reported in the financial statements in each periodic report.

Each consortium member is responsible for covering their costs during pre- and post-grant phases with own funding. During the project, consortium partners will be responsible for managing and curating datasets at their possession. At the project ending, each consortium member will take appropriate measures to ensure long-term preservation and sharing of opened datasets.

5. Data security

At the beginning of the research project, the project consortium will decide and agree on the tasks, roles, responsibilities and rights relating to data collection, dataset management and data use.

During the project, datasets will be available only to those project participants or consortium members, who have been accredited by and their data usage has been approved by Principal Investigator or authorized project consortium member. Project participants will be responsible for curating, preserving, disseminating and deleting in appropriate manner the datasets in their possession. Retention time for curated datasets will be the same as for other project results.

Data collected or acquired within the project will be stored in a secure IT environment behind a firewall at project consortium members' premises or in secure cloud environment provided by project consortium members' authorized and security cleared IT service providers. Access to it will need registration and authentication. Responsible project participant at VTT will check applications for the use of data. Where access is granted to research data, this will be provided through secured telecommunications channels. EU GDPR regulation will be followed in storage and transfer of sensitive or personal data.

Long-term and secure preservation of published research data will be ensured by using only certified and OpenAIRE guidelines compatible repositories.

6. Ethical aspects

Privacy of data subjects will be secured by following closely the General Data Protection Regulation (Regulation (EU) 2016/679 of the European Parliament and of the Council). The project consortium has appropriate technical and organizational measures in place to carry out data protection during the project.

Processes that handle personal data have been designed and built with the GDPR principles taken into account. Specifically, informed consent for data sharing and long-term preservation is always included in questionnaires dealing with any personal data. Processes provide safeguards to protect research data (e.g. using pseudonymization or full anonymization where appropriate), and use the highest-possible privacy settings by default. No person or organization involved will unintentionally be identifiable directly or indirectly in the datasets. Any indirect reference to sensitive personal information or e.g. lines of businesses, branches or industries will be removed and destroyed after the anonymized dataset has been checked and validated.

After curation, no person-related data is available publicly without explicit, informed consent, of the data subject and – if no full anonymization is required – publicly available data cannot in any circumstances be used to identify a subject without additional information stored securely in a separate place. Project members will always retain an unambiguous and individualized affirmations of consent from the data subjects and the subjects will always have the right to revoke their consent at any time.

During and after closure of the project the project members will clearly disclose any datasets, which have been collected during the project and declare the lawful basis and purpose for their processing. In addition, project members will state how long the data in their possession will be retained and unambiguously declare, if it is being shared with any third parties or outside of the EEA. Data subjects of the project will have the right to request a portable copy of the data collected in a common format, and the right to have their data erased under specified circumstances. VTT employs a data protection (privacy) officer (DPO), who is responsible for managing compliance with the GDPR.

During the Grant Agreement preparation, the Ethics Review was cleared. Research integrity and ethical principles related to data collection and use are covered in the ethics self-assessment in the Section 4 of the Grant Agreement Part B. Handling of data related to clinical samples is discussed in the deliverable D7.2 'Ethics Plan'.

7. Progress compared to the Description of Action

First version of the project 'Data Management Plan' was prepared. The work is fulfilled 100%.