

DATA MANAGEMENT PLAN

Grant Agreement number	24DIT03
Project short name	A3SmartML
Project full title	Automated, adaptive and uncertainty-aware smart measurements using machine learning
Data management plan	1 st ☒ 2 nd ☐
Confidentiality status	PU - Public, fully open

TABLE OF CONTENTS

1	Data management plan.....	2
1.1	Data summary.....	2
1.2	Findable, Accessible, Interoperable and Re-usable (FAIR) Data.....	3
1.2.1	Making data findable, including provisions for metadata.....	3
1.2.2	Making data accessible	3
1.2.3	Making data interoperable.....	5
1.2.4	Increase data re-use	6
1.3	Other research outputs	7
1.4	Allocation of resources	8
1.5	Data security	8
1.6	Ethics	9
1.7	Other issues.....	9
2	Open science: research data management.....	9

1 Data management plan

1.1 Data summary

Questions	Answers
1 Will you re-use any existing data and what will you re-use them for? State the reasons if re-use of any existing data has been considered but discarded.	This project will re-use: - Internal data from the participants. - Publicly available data. The data will be used for: - Enhancing and augmenting datasets used within the project. - Training of ML/DL models generated within the project. - Testing and Benchmarking estimation procedures developed within the project. - Testing and validating uncertainty evaluation approaches developed within the project. - Testing and developing smart measurement strategies.
2 What types and formats of data will the project generate or re-use?	The project will generate measurement data in several formats, such as: - Structured data in *.mat container format (sHSI) - Structured data in *.npy container format (PCM, sHSI) - Structured data in *.hdf5 container format (LIMM) - Image data as *.png (PCM) - Gwyddion volume data (HS-AFM) - Numerical data in CSV and JSON format Software will be generated and re-used. Mainly written in python (*.py) and Matlab (*.m).
3 What is the purpose of the data generation or re-use and its relation to the objectives of the project?	The data generated or re-used will be subject to the developed of smart measurement strategies. The data will be generated by measurements and simulations. They will be used for method development, validation and for presentation at conferences and in peer-reviewed publications. The data generated and re-used is in close relation to the objectives, since objectives 1-3 perform method development on applications with corresponding data sets, e.g. - Objective 1 requires the development and training of ML/DL models on data generated by the partners - Objective 2 requires the validation of uncertainty evaluation methods based on ground-truth data generated by the partners - Objective 3 develops adaptive measurement protocols for smart measurement strategies, based on generated data by the partners Objective 4 is mainly concerned with the generation (measurement and simulation) of the data itself.
4 What is the expected size of the data that you intend to generate or re-use?	The overall size of the data is expected to be > 1TB.
5 What is the origin/provenance of the data, either generated or re-used?	The data generated will originate from actual measurements performed by the partners and simulations developed within the project. The data re-used originates from open-source software libraries, e.g. - https://github.com/CambridgeCIA/LION - https://github.com/manuelmarschall/CompressedFTIR
6 To whom might your data be useful ('data utility'), outside your project?	The data might be useful to: - Stakeholders from industry: optical industry and metrology, scanning-based instrument manufacturers, large-scale user facilities, material

	processing and manufacturing companies, semiconductor industry and quality control. - Other research groups working on smart measurement strategies and ML/DL models for the solution of inverse problems. This especially includes NMI - NMIs and DIs.
--	---

1.2 Findable, Accessible, Interoperable and Re-usable (FAIR) Data

1.2.1 Making data findable, including provisions for metadata

Questions	Answers
7 Will data be identified by a persistent identifier?	Yes, if applicable, DOI or commit/tag on Git repositories will be used.
8 Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.	The metadata created for all of the project's deposited datasets will be open under a Creative Commons Public Domain Dedication (CC 0) or equivalent (to the extent legitimate interests or constraints are safeguarded), in line with the FAIR principles (in particular machine-actionable) and provide information at least about the following: datasets (description, date of deposit, author(s), venue and embargo); the European Partnership on Metrology funding; grant project name (Automated, Adaptive and Uncertainty-Aware Smart Measurements using Machine Learning), acronym (A3SmartML) and number (24DIT03); licensing terms; persistent identifiers for the dataset, the authors involved in the action, and, if possible, for their organisations and the grant. Where applicable, the metadata will include persistent identifiers for related publications and other research outputs. The project additionally develops in A4.1.6 of Annex 1 to the Grant Agreement a data harmonisation approach for the data sets generated in the project. This includes appropriate labels and metadata tables. The metadata tables will be added to ensure that the data adheres to the digital SI and FAIR+T principles. The metadata tables describe each data set (including sample info, acquisition parameters, units, etc.).
9 Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use?	Yes, keywords will be provided in the metadata. Keywords may include: - Smart Measurements - Linear inverse problems - Machine Learning or Deep Learning - Automated Measurements - Adaptive Measurement protocols
10 Will metadata be offered in such a way that it can be harvested and indexed?	Yes, the metadata will be added to ensure that the data adheres to the FAIR+T principles. Additionally, data will be stored on the trusted repository Zenodo. The metadata are indexed in a searchable resource. Metadata are licensed under CC0, except for email addresses. All metadata are exported via Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and can be harvested.

1.2.2 Making data accessible

Questions	Answers
Repository:	
11 Will the data be deposited in a trusted repository?	The data and associated metadata, documentation and code will be deposited in the trusted open access repository Zenodo (https://zenodo.org). Additionally, either to allow for active development of software or to ensure long-term data storage, the following repositories will be used: - PTB's Gitlab server (https://gitlab1.ptb.de)

Questions	Answers
	- PTB's Open Access Repository (https://oar.ptb.de) - Github (https://github.com)
12 Have you explored appropriate arrangements with the identified repository where your data will be deposited?	No, the data will be uploaded via a standard procedure and require no special arrangements.
13 Does the repository ensure that the data are assigned an identifier? Will the repository resolve the identifier to a digital object?	Yes, Zenodo will assign an identifier (DOI) to each of the project's deposited datasets. The repository will resolve the identifier to a digital object.
Data:	
14 Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.	All the data that are needed to validate the results presented in scientific publications will be made openly available as the default unless there is a specific reason not to publish the data. <u>Datasets which cannot be shared – voluntary restrictions</u> Other data may be made available on a case-by-case basis if it is relevant for third parties. The following data will not be made publicly available: - Data obtained with the permission of third parties, but the third parties have not agreed to make the data publicly available. - Data that compromises the protection of a participant(s) intellectual property. - Pre-existing internal data of a participant, but the participant has not agreed to make the data publicly available. The level of data made available will also be considered, for example, raw and pre-processed data will not be provided unless there is a clear reason for doing so. It is planned to combine software and generated data with pre-existing data. However, no pre-existing data will be made public unless the corresponding partner has agreed to make the data publicly available. Therefore, published datasets for training ML/DL models and published data to validate uncertainty evaluation methods and scientific publications may differ from the internally used data for the method development. Hence, there should not be any legal / contractual reasons for not sharing the data.
15 If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.	The data used in scientific publications will be made available for re-use as soon as is reasonably possible. Most data produced is open-source and no IP protection is envisioned. In the case of sensitive data, an embargo period of 12 months will be applied whilst a patent application is pending.
16 Will the data be accessible through a free and standardised access protocol?	Yes, Zenodo provides well described conditions for free and standardised access (see http://about.zenodo.org/policies/).
17 If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?	There are no restrictions on the use of the published data, but users will be required to acknowledge the project and the source of the data in any resulting publications, according to the latest version of the CC-BY license.

Questions	Answers
18 How will the identity of the person accessing the data be ascertained?	There is no need to ascertain the identity of persons accessing the data.
19 Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?	No there is no need for a data access committee (DAC) in this project. The participants responsible for their data will ensure that the data contained in published data sets are approved for open sharing. The appointed corresponding author, with responsibility for the data, will decide alone or, if appropriate, jointly with the collaborating partners, about granting access to the data.
Metadata:	
20 Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?	In Zenodo, metadata are licensed under CC0, except for email addresses. All metadata are exported via OAI-PMH and can be harvested. The metadata will contain information to enable the user to access and process the data.
21 How long will the data remain available and findable? Will metadata be guaranteed to remain available after data are no longer available?	The data will remain available and findable for the lifetime of the Zenodo repository, which is expected to be a minimum of 20 years. If data are withdrawn from Zenodo, the DOI and the URL of the original object are retained. In case of closure of the Zenodo repository, best efforts will be made by Zenodo to integrate all content into suitable alternative institutional and/or subject based repositories.
22 Will documentation or reference about any software be needed to access or read the data and will this be included? Will it be possible to include the relevant software (e.g. in open source code)?	The software that will be developed by the consortium will be needed to read the data. This software will be shared and documented.

1.2.3 Making data interoperable

Questions	Answers
23 What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?	The datasets will use the trusted repository's basic metadata schema for administrative data, which is compliant with the recommended standards used by DataCite (https://search.datacite.org/) and OpenAIRE (https://www.basesearch.net/). For individual datasets, the specific vocabularies and methodologies of the GUM (procedure) will be used for all applications (subject-independent), if appropriate. All numeric values include appropriate units according to Digital SI guidelines.
24 In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated	Mapping will not be required as the terminology used will be chosen to be compatible with the existing literature.

ontologies or vocabularies to allow their re-use, refinement or extension?	
25 Will your data include qualified references ¹ to other data (e.g. other data from your project, or datasets from previous research)?	Yes, the project's datasets that will be deposited in the chosen repository (e.g. Zenodo) might include qualified references to other datasets from the same project or from previous research.

1.2.4 Increase data re-use

Questions	Answers
26 How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?	A Readme file in markdown format or plain text format will be provided together with the data, in order to enable data analysis and to facilitate data re-use.
27 Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement?	The data will either be licensed under the latest available version of the Creative Commons Attribution International Public License (CC BY) or a license with equivalent rights as set out in the Grant Agreement. Users will be required to acknowledge the consortium and the source of the data in any resulting publications. Alternatively, the Creative Commons Public Domain Dedication License (CC 0) will be used.
28 Will the data produced in the project be useable by third parties, in particular after the end of the project?	Any data published in open-access journals will be usable by third parties after the datasets have been deposited in Zenodo. The data that do not relate to peer-reviewed publications will be made available for re-use on a case-by-case basis.
29 Will the provenance of the data be thoroughly documented using the appropriate standards?	Yes, the provenance and context of the data will be thoroughly documented in accordance with relevant community and domain-specific standards. The data will be accompanied by detailed information on how they were captured, processed, analysed, and validated. Additional contextual and methodological information necessary for interpretation and reuse will also be provided.
30 Describe all relevant data quality assurance processes.	Data quality will be assured through several quality assurance procedures: - Repeated and comparison measurements. - Performing unit-tests and regression testing for developed software - Use of controlled vocabularies and standard terminology. - Metrological characterisation of the measurement set-ups. - Verification / validation of the data collected (if appropriate). - Peer-review of publications based on the data. The Mathmet quality assurance tools (QAT) will be considered https://www.euramet.org/european-metrology-networks/mathmet/activities/quality-assurance-tools
31 Further to the FAIR principles, DMPs should also	<u>Allocation of resources</u>

¹ A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: <https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/>)

Questions	Answers
address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.	<u>The estimated costs for making the (data and) other research outputs FAIR are 7 570 € (personnel costs) (see question 34).</u> The costs for making the data and research outputs FAIR are contained in the project budget. This includes, in particular for the data generated through measurements and simulations, the personnel costs dedicated to the task 4.1. of Annex 1 and the participants involved in the development of deliverable D6. The costs for making other research outputs FAIR are included in the project's budget and will be claimed if compliant with the Grant Agreement's conditions. Where feasible, long-term preservation will be ensured by depositing the other research outputs in repositories. <i>Security of other research outputs</i> The participating NMIs maintain ISO/IEC 17025-compliant laboratories; industrial partners follow comparable measurement quality procedures. The participants will store other research outputs on their organisations' networks, which are protected by firewall, backups etc. Other research outputs will also be stored in the project's SharePoint environment, with password-protected login. Deposition in public repositories will provide additional security as they have multiple replicas in a distributed file system which is backed up on a nightly basis. This project will not generate sensitive other research outputs. The other research outputs will be safely stored in open access repositories. <i>Ethical aspects</i> There are issues that could impact on the sharing of other research outputs. - Information relating to other research outputs acquired from third parties, e.g., manufacturers, will not be shared without their explicit consent. - Information relating to other research outputs collected by the consortium at commercial sites will not be shared without the site owner's explicit consent. The project will not share other research outputs with identifiable personal information. Sensitive information relating to the other research outputs will be collected, separated as soon as possible, and kept secure.

1.3 Other research outputs

Questions	Answers
32 In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).	The software, which includes solvers for linear inverse problems using statistical prior knowledge and interpretable regularization techniques, developed in the project will be released under a GNU-GPL license. The developed measurement protocols, which include the adaptive scanning procedures for hyperspectral imaging, will be published as a good practice guide on the project's website and on a trusted platform (e.g. Zenodo). The same holds also for all other project deliverables, which include a description of the experimental prototypes that implement smart measurements for scanning hyperspectral imaging and photocurrent mapping of semiconductors.
33 Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide	As far as possible, the FAIR data approaches specified in questions 7-30 above will be applied to the management of this project's other research outputs. This commitment will be met by releasing, e.g., the new software that will be developed in the project under open access license and by placing the reports on the developed experiment prototypes in a trusted repository.

Questions	Answers
sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.	

1.4 Allocation of resources

Questions	Answers
34 What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?	The estimated costs for making the data and other research outputs Findable, Accessible, Interoperable and Re-usable (FAIR) are 7 570 € (personnel costs, in particular for the leading and supporting participants of A4.1.6 and A4.1.7 of the Grant Agreement Annex 1). These costs have been kept to a minimum by using a free repository (Zenodo) and by making only relevant data and other outputs FAIR.
35 How will these be covered? Note that costs related to research data/output management are eligible as part of the European partnership on metrology grant (if compliant with the Grant Agreement conditions).	The costs for making the data FAIR are included in the project's budget and will be claimed if compliant with the Grant Agreement's conditions. Long term preservation will be ensured by depositing the data within repositories (Zenodo, PTB's Open Access Repository). There are no costs associated with the long-term preservation of the data in these repositories.
36 Who will be responsible for data management in your project?	Each participant will be responsible for the management of its data. Additionally, the coordinator, supported by the WP leads will ensure that the management tools and guidelines outlined above are enforced on a case-by-case basis using the following instruments: - The EURAMET publication database, which automatically informs the coordinator upon acceptance/rejection of an upload - The Zenodo community https://zenodo.org/communities/a3smartml/ which collects all publications and published data sets - Regular communication at project meetings
37 How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?	Long term preservation will be ensured by depositing the data within repositories (Zenodo, PTB's Open Access Repository). There are no costs associated with the long-term preservation of the data in these repositories. The data will serve as benchmark cases for the developed methods and tools developed in the project. It will enable stakeholders to test and understand the methods developed in the project. Hence, the data will facilitate the uptake of the project's results. The participants responsible for their data will decide on a case-by-case basis on what data will be kept for how long, taking into account the Grant Agreement and the guidelines outlined above.

1.5 Data security

Questions	Answers
38 What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	<i>Data recovery and secure storage</i> All participants are either accredited to, or work in compliance with, the ISO 17025 standard on the "General requirements for the competence of testing and calibration laboratories". The participants will store data on their organisations' networks, which are protected by firewall, backups etc. Data will also be stored in the project's SharePoint environment, with password protected login. Deposition in the Zenodo public repository and code repositories, such as Gitlab or Github will provide additional security as it has multiple replicas in a distributed file system which is backed up on a nightly basis.

	<i>Transfer of sensitive data</i> This project will not generate sensitive data.
39 Will the data be safely stored in trusted repositories for long term preservation and curation?	Yes, the data will be safely stored in the Zenodo open access repository or in PTB's Open Access Repository. Zenodo and the underlying Invenio Framework for digital repositories were designed according to the Open Archival Information Systems (OAIS) reference model. PTB's Open Access Repository is stored on two physically and geographically separated servers that are regularly backed up. PTB is working towards German Initiative for Network Information (DINI) certification.

1.6 Ethics

Questions	Answers
40 Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics report(s) and the ethics section in the Annex 1.	This project has been given "Ethics clearance" from the Partnership Ethics Review 2024. There are possible issues that could impact on data sharing. - Data acquired from third parties, e.g., manufacturers, will not be shared without their explicit consent. - Data collected by the consortium at commercial sites will not be shared without the site owner's explicit consent.
41 Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	There are no market and customer surveys planned in the project. Nevertheless, if any sensitive data are collected, they will be separated as soon as possible and kept secure.

1.7 Other issues

Questions	Answers
42 Do you, or will you, make use of other national / funder / sectorial / departmental procedures for data management? If yes, which ones (please list and briefly describe them)?	Data management will be compliant with: - The research data policy of the European Partnership on Metrology. - European laws about data security and the protection of privacy (e.g. GDPR). - Institutional guidelines. - Scientific community guidelines.

2 Open science: research data management

Statement	Put an X in the box to confirm	Or, list any exceptions to this
All participants have adhered to the requirements of the project's GA and CA with respect to open science: research data management (GA Article 17 and its Annex 5) for this reporting period	☒	

