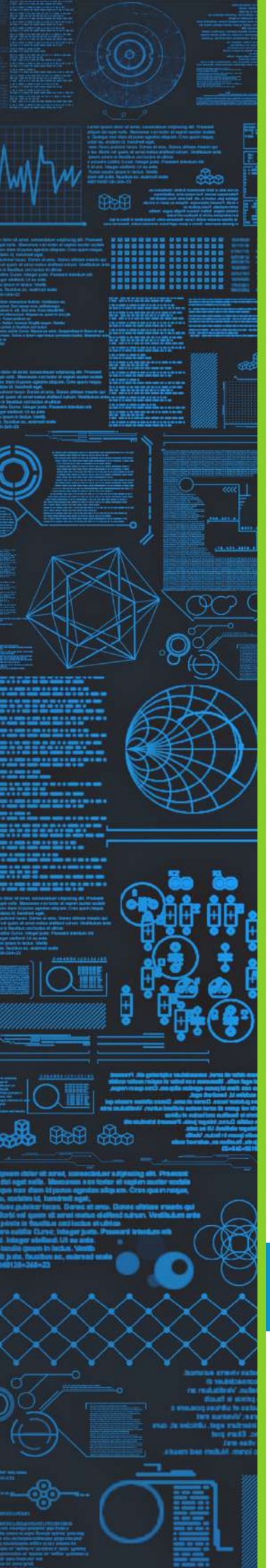

SustainML

Application Aware, Life-Cycle Oriented Model-Hardware Co-Design Framework for Sustainable, Energy Efficient ML Systems

Data Management Plan (M18)

Deliverable D7.2

WP7 - Dissemination, Exploitation, and Management

This project has received funding from the European Union's Horizon Europe research and innovation programme (HORIZON-CL4-2021-HUMAN-01) under grant agreement No 101070408

eProsima
The Middleware Experts

DFKI

Inria

upmem

RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau

IBM

Project

Title:	SustainML: Application Aware, Life-Cycle Oriented Model-Hardware Co-Design Framework for Sustainable, Energy Efficient ML Systems
Acronym:	SustainML
Coordinator:	eProsima
Grant agreement ID:	101070408
Call:	HORIZON-CL4-2021-HUMAN-01
Program:	Horizon Europe
Start:	01 October 2022
Duration:	36 months
Website:	https://sustainml.eu
E-mail:	sustainml@eprosima.com
Consortium:	**eProsima (EPROS)**, Spain **DFKI**, Germany **Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau (RPTU)**, Germany **University of Copenhagen (KU)**, Denmark **National Institute for Research in Digital Science and Technology (INRIA)**, France **IBM Research GmbH**, Switzerland **UPMEM**, France

Deliverable

Number:	**D7.2**
Title:	**Data Management Plan (M18)**
Month:	18
Work Package:	WP7 - Dissemination, Exploitation, and Management
Work Package leader:	eProsima
Deliverable leader:	eProsima
Deliverable type:	Report (R)
Dissemination level:	Public (PU)
Date of submission:	2024-03-31
Version:	v2.1
Status:	Finished

Version history

Version	Date	Responsible	Author/Reviewer	Comments
v1.0	16-03-2023	eProsima	Raúl Sánchez-Mateos Lizano	First draft
v1.1	24-03-2023	eProsima	Mario Domínguez López	Review
v1.2	30-03-2023	eProsima	Raúl Sánchez-Mateos Lizano	Applied suggested changes and final version
v2.0	18-03-2024	eProsima	Raúl Sánchez-Mateos Lizano	1. Review FAIR data policy and ethical issues of SustainML project 2. Update project datasets
v2.1	25-03-2024	eProsima	Jesús Poderoso Martín	Review

Executive summary

SustainML project aims to develop a design framework and an associated toolkit, the so-called SustainML, that will foster energy efficiency throughout the whole life-cycle of Machine Learning (ML) applications: from the design and exploration phase that includes exploratory iterations of training, testing and optimizing different system versions all the way to the final training of the production systems (which often involves huge amount of data, computation and epochs) and (where appropriate) continuous online re-training of the inference process during deployment. The framework will optimize the ML solutions based on the application tasks, across levels from hardware to model architecture. It will also collect both previously scattered efficiency-oriented research, as well as novel Green-AI methods. Artificial Intelligence (AI) developers from all experience levels can make use of the framework through its emphasis on human-centric interactive transparent design and functional knowledge cores, instead of the common blackbox and fully automated optimization approaches.

This report corresponds to *Deliverable D7.2 - Data Management Plan (M18)* of the SustainML project. This deliverable covers the specification of what data will be open, specifies the data that the project will generate, whether and how it will be exploited or made accessible for verification and re-use, and how it will be curated and preserved. This document will be updated to reflect the most recent state of the project as it progresses.

Contents

Acronyms

AI	Artificial Intelligence.
CDR	Common Data Representation.
CERN	European Organization for Nuclear Research.
DMP	Data Management Plan.
DOI	Digital Object Identifier.
DR	Disciplinary Repository.
EC	European Commission.
EU	European Union.
FAIR	Findable, Accessible, Interoperable and Reusable.
GA	Grant Agreement.
GDPR	General Data Protection Regulation.
HPDC	High Performance Data Center.
HW	Hardware.
IDL	Interface Description Language.
IPR	Intellectual Property Rights.
IR	Institutional Repository.
ML	Machine Learning.
OpenAIRE	Open Access Infrastructure for Research in Europe.
ORDP	Open Research Data Pilot.
SW	Software.

1 Introduction

The SustainML research and innovation project aims to develop a design framework and an associated toolkit, so-called SustainML, that will foster energy efficiency throughout the whole life-cycle of ML applications. The goal of the EU-funded SustainML project is to optimize the ML solutions based on the application tasks, across levels from hardware to model architecture.

During the course of the SustainML project, a wide variety of data will be generated from R&D activities; scientific publications describing the parameterization of ML tasks into quantifiable descriptions for their interpretation, catalogs for recycling and continuous learning to maximize the reusability of ML models avoiding repetitive training, catalogs of efficient ML practices, Hardware (HW) libraries for sustainable ML, source code of the implemented SustainML framework, and validation results of the methodology and processes used, among others.

SustainML will comply with Article 17 of the Grant Agreement (GA) for the communication, dissemination, open science and visibility of results. According to this article, partners shall disseminate their results by publishing them through open means, including scientific publications of any format. However, the confidentiality obligations in Article 13, the possible application of Intellectual Property Rights (IPR), access rights and rights of to protect results in Article 16, or the obligations to protect personal data in Article 15 are maintained.

Since SustainML is at an early stage and the underneath technology of SustainML framework is state-of-the-art, it is important that the eventual dissemination of the findings (data, publications, and results) is open for scrutiny by other researchers, potential future partners, and a comprehensive regulatory community.

As a project participating in the European Commission (EC)'s Open Research Data Pilot (ORDP) (OpenAIRE [1]), SustainML will make its research data Findable, Accessible, Interoperable and Reusable (FAIR) and follow the ORDP principles, i.e., "*as open as possible, as closed as necessary*", encouraging data management as an essential part of the research best practices. However, data sharing in the open domain may be restricted, taking into account "*the need to balance openness and protection of scientific information, commercialization and IPR, privacy concerns, security, as well as data management and preservation issues*", as stated in the guidelines on FAIR data management in Horizon Europe published by OpenAIRE [1].

The Data Management Plan (DMP) is, therefore, a document that describes the type of data that will be collected and generated throughout the project, the practices that will be carried out for handling the data, the policy to be applied to each type of data, the custody of the data, the access to the data by SustainML Consortium members and associated third party researchers, and SustainML's responsibility to preserve and reuse the data.

The DMP here presented is the first version of this deliverable (M6), which is intended to be a living document that will be reviewed and updated throughout the project, thus adjusting to the needs of the project. This document will be annually modified with each revision of the SustainML project, taking into account the intermediate updates that SustainML may require.

According to the FAIR principle for research data (Findable, Accessible, Interoperable and Re-usable) the SustainML DMP considers:

- Name, reference and versioning of the data.

[1] `https://www.openaire.eu/`

- Description of the data
- Metadata standards.
- Data management and sharing during and after project completion.
- Data hosting and preservation during and after project completion.

The following document makes use of the Horizon 2020 FAIR Data management Plan Template [2], following the guidelines on FAIR data management in Horizon Europe published by OpenAire [1], is written with the The FAIR Guiding Principles for scientific data management and stewardship [3], and the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679) [4] as references.

2 Data Cycle

The amount of data that the SustainML project intends to collect and generate is significant. Furthermore, since the project data is expected to be highly diverse, from Software (SW) to Hardware (HW) libraries, in order to guarantee the re-usal of the developed code and validation results throughout the project (widely expanded in time), it has been agreed to follow the data life cycle model described in [5] and shown in Figure 1.

In the first phase of the project, a first version of the DMP is presented, stipulating what to do with the data collected and generated. These data will be subsequently integrated, prepared and analyzed for use. During this phase the data are described, named and identified as explained in this document. In a third phase, the data are presented to the SustainML consortium to assess their impact, determine which datasets will be published and made Open Access. Since SustainML follows a policy of open access to the generated documents, this should be taken into account when assessing the publicity of the data (see next sections). Finally, for each dataset it is defined how long the data will be kept for later re-use, both by SustainML consortium members and third party researchers. Data may be stored in a cloud infrastructure hosted by any of the SustainML consortium members, an Institutional Repository (IR), a Disciplinary Repository (DR), or a High Performance Data Center (HPDC) provided by public or private organizations.

The activity of ensuring data quality through the process of data description and evaluation is accomplished during the entire data life cycle. This implies that, as stated before, the present document will be updated at any stage in the course of the SustainML project.

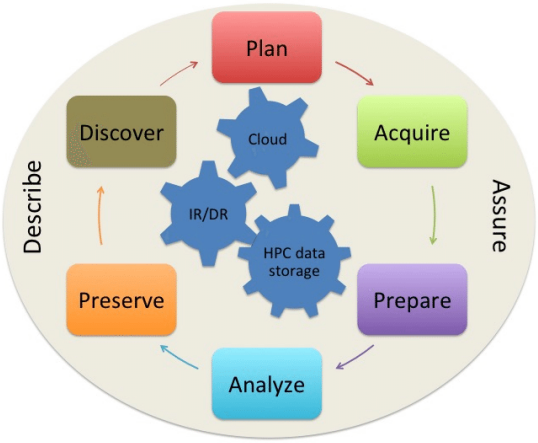

Figure 1: The Big Data Life Cycle Model [5].

3 Open access policy of SustainML

According to Article 17, *Dissemination of results* subsection, of the GA, "*The beneficiaries must disseminate their results as soon as feasible, in a publicly available format, subject to any restrictions due to the protection of intellectual property, security rules or legitimate interests.*"

Thus, and as stipulated in this same document, all project deliverables and generated data will be publicly available, except for the internal reports such as meeting notes or private shared documents, whose distribution will be private and internal to SustainML consortium members. The research data will be uploaded to a public domain repository, and/or published in scientific journals, thus making it openly accessible. Furthermore, the project deliverables can be found in the publications section of the SustainML official website [2].

However, for each of the datasets and knowledge generated, the SustainML consortium will analyze the potential conflicts of openly publishing the results versus the commercialization and application of IPR protection, thus deciding what information will be made publicly available and what information will remain private. Since SustainML is based on the use of a state-of-the-art ML task characterization techniques, the findings can be protected by the consortium by two means: *a*) preserving the data internally and only sharing these data outside the consortium after IPR protection measures have been undertaken, or *b*) applying for a patent protecting some output resulting from the project realization to obtain commercial benefits, in which case the results would be made public after the patent registration is granted.

Everything previously discussed is defined within the framework of Article 16 (Annex 5) of SustainML GA, which states that "*Beneficiaries which have received funding under the grant must adequately protect their results — for an appropriate period and with appropriate territorial coverage — if protection is possible and justified, taking into account all relevant considerations, including the prospects for commercial exploitation, the legitimate interests of the other beneficiaries and any other legitimate interests.*"

[2] `https://sustainml.eu`

4 Data summary

This section describes each of the datasets collected, processed or generated in the R&D process of the SustainML project. It is worth mentioning that the project is in an early development stage and the present document is the first version of the SustainML DMP, so the datasets are currently not completely defined. Future versions of this document will complete and enhance the information concerning these datasets. New datasets will be added to this section as the SustainML project progresses.

Moreover, according to the European Union (EU) guidance proposed for the DMP, the data collected, processed and generated within the SustainML project will be evaluated to fulfill the FAIR data specifications.

4.1 Data description

Since the objective of the project is the creation of a Machine Learning (ML) framework that allows the evaluation of the different ML algorithms previously studied, altogether with a set of proposed HW tools, a large amount of data will be needed for the creation of the behavioral patterns of the framework to be developed. These datasets may be generated by SustainML consortium members or collected from public data repositories, which provide datasets for the use, evaluation, training and refinement of ML techniques. Both, SustainML proprietary and already available datasets, will be added and specified in future releases of this document.

The datasets generated or intended to be generated, will be added in this document and characterized as follows:

- **Dataset summary**
  - *Dataset ID*: unique identifier of the dataset.
  - *Name*: name of the dataset.
  - *Lead Partner*: partner responsible of the dataset, its maintenance, and the data contained within it. This includes ensuring compliance with the data management and ethics requirements of the project.
  - *Work Package*: references the Work Package which this dataset has been used or generated.
  - *Work Package Task*: project task which this dataset has been used or generated.
  - *Description*: complete description of the dataset.
  - *Data types*: type of that it contains (C++/Python/Java source code, HW library, images, text, etc.)
  - *Data format*: format of the data.
  - *New / Existing data*: specifies if the data has been created within the SustainML project or already existed.
  - *Mechanisms for data generation*: specifies how the data was generated.
  - *Expected size of the data*: specifies the expected size of the data in Bytes.
  - *Utility of the data*: specifies how the data is going to be used within the project.
  - *Quality control procedures*: specifies the procedures to guarantee the quality of the data (peer reviewed, tests, validation procedures, etc.)
  - *Type of access (Open / Restricted)*: specifies the data protection issues.

  - *Ethical issues*: specify whether the data could raise any ethical issue.

- **Making data findable**
  - *Data dissemination*: specifies how the data will be available.
  - *Metadata standard*: specifies the metadata used to easily find the data.
  - *Type of associated metadata*: specifies the type of metadata used.
  - *How will the data be findable*: specifies the mechanisms used to make the data findable.
  - *Software required to use / read the data*: specifies if there is any software requirement for the user to read the data.
- **Storage of data**
  - *Data storage location (short-term)*: specifies where will be stored in the short term.
  - *Data storage location (long-term)*: specifies where will be stored in the long term.
  - *Storage media*: specifies the storage media to be used.
  - *Data security provisions*: specifies how data security is ensured, in case the data is not publicly accessible.
  - *Expected size of the data*: specifies the expected size of the data in Bytes.
  - *Person responsible*: specifies the person responsible for ensuring the availability of data in the short and long term.
  - *Cost*: specifies the cost of data storage.
- **Re-use data**
  - *Reuse of existing data*: if the data already existed, specifies how the data is going to be reused within the project.
  - *Potential reuse of data*: specifies how to the data is going to be reused.
  - *How will data be reused*: specifies how the data will be used within the framework of the project and the possibility of its use by the same or other entities within or outside the project at the end of the project.
  - *Type of access (Open / Restricted)*: specifies whether the data is freely accessible or access is restricted to the consortium.
  - *Type of IP / protection sought*: specifies whether the data requires intellectual property conditions to be applied to it.
  - *License*: specifies the license that applies to the data (Open Source license or more restrictive licenses).
- **Other Comments**: in this section, the person responsible for the data may specify other types of conditions for the publication and use of the data, if any.

A more detailed description for each of the generated datasets, including each of the parameters listed above, will be provided in the Section 10 of this document.

5 FAIR data

This section develops the methodology to be followed in order to ensure that data management in the context of the SustainML project meets the FAIR data criteria. It is worth mentioning that those datasets that are not public due to any of the points discussed in Section 3 will not meet these criteria. However, after these datasets are made available, i.e., public, this DMP will be updated to incorporate them and thus meet the FAIR criteria.

5.1 Making data findable

SustainML metadata are those research data that make it possible for third party researchers interested in the field to quickly and easily find the research material of the SustainML project. Thus, through the use of precise, detailed and accurate metadata, SustainML project datasets will be easily found by other researchers. The metadata used in SustainML (data type, quality, availability, author, content description, versioning, etc.) will be exposed in standardized formats in the SustainML research data repositories, which will be publicly accessible. Regarding bibliographic metadata, these will comply with Article 17 (Annex 5) of the SustainML Grant Agreement 101070408. These will have a standard format and will include the following:

- *"publication (author(s), title, date of publication, publication venue);*
- *Horizon Europe funding*[3]*;*
- *grant project name, acronym and number;*
- *licensing terms;*
- *persistent identifiers for the publication, the authors involved in the action and, if possible, for their organisations and the grant.*
- *Where applicable, (...) persistent identifiers for any research output or any other tools and instruments needed to validate the conclusions of the publication.*

The Digital Object Identifier (DOI) will be the persistent and unique identifier of the project's publications in open data repositories.

A naming convention will be used to classify and identify the research results. These names are composed according to the following rule:

$$SustainML\_XX\_DS\_ID\_YY\_ZZ$$

where

- **SustainML**: name of the project.
- **XX**: date of creation according to ISO 8601[4] format (YYYYMMDD).

[3] All publications of the SustainML project should indicate the EU funding of the project by including the following statement: "*This ¡insert document type¿ is supported by European Union's Horizon Europe research and innovation programme (HORIZON-CL4-2021-HUMAN-01) under grant agreement No 101070408, project SustainML (Application Aware, Life-Cycle Oriented Model-Hardware Co-Design Framework for Sustainable, Energy Efficient ML Systems).*"

[4] `https://www.iso.org/iso-8601-date-and-time-format.html`

- **DS**: dataset name. The name must not contain blank spaces and the PascalCase[5] format must be applied.
- **ID**: dataset identifier.
- **YY**: data type.
- **ZZ**: version number starting at `v1.0`. The rule applied to dataset versioning follows the same rule applied to software versioning, i.e. `vMajor.Minor`. A change to the Major version indicates a significant change in the dataset that may have repercussions on the intended use or context to which it applies. A Major version change should be made if:
  - major new data is added or removed;
  - the name of an attribute is changed;
  - the value of some data is changed due to temporal or spatial baseline changes;
  - new attributes are introduced;
  - the data generation model is changed;
  - the format of the data is changed;
  - a Major change is made to a dataset on which another dataset depends while introducing major changes in the latter.

  The changes in the Minor version are due to changes in the quality of the data and do not affect the scope of the initial dataset. A change in the minor version must be made if:
  - errors in the data are corrected;
  - the data generation model is rerun;
  - a Minor change is made to a dataset on which another dataset depends while introducing minor changes in the latter.

In order to make the datasets more easily findable by other third party researchers, metadata will be defined to facilitate the search of SustainML datasets. Furthermore, these metadata will incorporate keywords that briefly describe and define the data. However, at the current stage of the project it is not possible to precisely define the metadata to be generated. Nevertheless, some keywords that will be used frequently to describe the project's research data will be: *green-ai*, *machine-learning*, *human-computer-interaction*, *artificial intelligence*, *sustainml*, *energy-efficiency*, *machine-learning-frame*.

5.2 Making data openly accessible

SustainML project datasets will be initially held in a shared repository, accessible to all the SustainML consortium members. The purpose is that the data will be treated, processed and evaluated prior to publication. The final or stable version of the datasets available in the SustainML private repository will be evaluated by the consortium using the guidelines defined in Section 3. In case no IPR or patent protection actions are applied, these data will be published in ZENODO[6], a general-purpose open access repository developed under the Open Access Infrastructure for Research in Europe (OpenAIRE) program

[5] `https://techterms.com/definition/pascalcase`

[6] `https://zenodo.org`

and operated by European Organization for Nuclear Research (CERN). Furthermore, arXiv[7], a free distribution service and open access archive for academic articles, will be also considered for other type of publications as materials on this site are not peer-reviewed by arXiv.

The data will be stored and published in standard digital formats that do not require special software to handle them, apart from the basic software that a regular computer might already have. Although not all metadata types intended to be published have been clearly defined, it is possible to define some general formats such as:

- PDF format for documents and publications.
- Preferable PNG, SVG or GIF format for higher quality images, although these can also be published in other formats such as JPG, in case an image compression is required, or PDF format.
- Audio files will be published in MP3 (MPEG-1 Audio Layer III) format as lossy audio format, and WAV as uncompressed audio format.
- For video files there are several options depending on the use case:
  - WEBM format for videos to be published on the SustainML website or used in other activities requiring video streaming such as online lectures or conferences.
  - MP4 or AVI format for high quality videos stored in a database.
  - WMV format for compressed videos that require a smaller size, although this format is not compatible by default with Apple devices that do not have Windows Media Player software installed.
- Interface Description Language (IDL) files for the generation of CDR-compliant types.
- The source code files shall be published in the formats corresponding to the programming language. Some of these languages used are Python, C/C++ and CMake. Regardless of the use of these files, they can all be opened and read as plain text documents.
- Preferably CSV format for training, validation and test datasets used in SustainML framework. In case the data structure requires it, JSON format can also be used for the datasets.

It is also expected to generate binary data that will be published as a result of the experiments on the SustainML framework and conducted within the SustainML project. Due to the early stage of the project it is not possible to define the software required for processing these binary data. It will be specified in future versions of the SustainML DMP.

Regarding the distribution means to make the data open, two ways of distributing the datasets are distinguished: internal distribution, accessible only to SustainML consortium members, and public distribution, open to third party researchers interested in the subject and to a wider community. For the former, i.e., internal distribution to project partners, the following will be used:

- Proprietary data centers of the institutions collaborating in the project.
- Google Drive file sharing and collaborative work service managed by *eProsima*[8].
- Private section of the SustainML project official website.

For open access to the data the following platforms will be used:

[7] `https://arxiv.org`

[8] `https://www.eprosima.com`

- ZENODO[9] for the publication of data and metadata of the results and progress of the research, development, deployment and testing processes.
- arXiv for the publication of mathematical articles.
- Github[10] to host all source code defined as public and open to the user. This platform will also be used for the development of a file sharing system where the public data of the project will be included.
- YouTube to host the presentations and speeches during the events, conferences and workshops in which SustainML partners participate, thus granting higher exposure to the project.
- SustainML project web page.
- Other scientific journals yet to be specified.

To regulate access to private data, the institution hosting the data will grant access to each of the researchers and project members requesting access on an individual basis. In contrast, no registration or authentication of the requester is required to access public data.

5.3 Making data interoperable

The previous section defines the formats used for each of the data generated by the SustainML project. These have been carefully selected to ensure the interoperability of the data. All data generated in the context of the SustainML project will conform to the criteria defined in Section 5.2, thus making the data interoperable between platforms and facilitating data exchange and re-use between researchers, institutions, organizations, countries, etc.

For all data types, a standardized vocabulary will be used for the description of data and metadata. SustainML project offers to adopt the DataCite[11] Metadata Schema for metadata standardization. Furthermore, the keywords defined in Section 5.1 will be used for the arrangement, classification, and indexing of the data and metadata.

Since the standard vocabulary is highly dependent on the study field, and various fields of study are addressed in SustainML (artificial intelligence, human-computer interaction, hardware acceleration, sustainable computing, etc.), a standardized vocabulary will be created to achieve interoperability of the data across the project's disciplines. This vocabulary will be updated as the project progresses.

5.4 Increase data re-use

In order to achieve the widest re-usability of the SustainML public data, Creative Commons[12] licenses will be used to protect the authorship of the generated datasets. These public domain licenses will grant copyright permissions to the datasets, ensure data attribution and allow third parties to copy, distribute and use SustainML public results, without the need for the consortium to register the work. *Attribution-ShareAlike* and *Attribution-NonCommertial-ShareAlike* licenses will be considered for application to SustainML datasets. Both licenses allow third parties to distribute, modify, adapt and extend the data as long as they credit the author of the original data. Both licenses also stipulate that the

[9] https://zenodo.org/

[10] https://github.com

[11] https://datacite.org/

[12] https://creativecommons.org/

license for derived works must have identical terms as the original license. The difference between the two licenses is that the former allows the use of the derived works for commercial purposes while the latter does not.

SustainML data will be Open Access as soon as the data are reviewed by other members of the SustainML consortium and a stable version of the data can be submitted. However, an embargo period may be applied by the publishing journal if the data are subject to IPR protection or are awaiting patent protection. In either case, the SustainML consortium will ensure that the maximum embargo period is six months.

As defined in Article 16 Annex 5 (*Access rights for exploiting the results*) of SustainML project GA, request access to data will be available to be re-used up to at least one year after the completion of the project. Each collaborator is responsible for its results and must grant access rights to the data and metadata to other beneficiaries and other affiliated entities up to one year after the completion of the SustainML project.

Data quality assurance actions shall be dictated and conducted by the institution that generates the data as the data owner. These actions shall ensure that:

a) the datasets have no erroneous or obsolete records,

b) provide the necessary metadata for a complete description of the data,

c) follow the conventions stipulated in this document for compliance with the FAIR data policy, and

d) are regularly updated during their period of re-usability.

6 Allocation of resources

This section provides a forecast of the costs assigned to FAIR data compliance in the SustainML project, how these costs will be covered, the responsibility for data management and the long-term preservation of resources.

First, the costs derived from compliance with the FAIR data policy are as follows:

- Cost of publishing scientific articles in Gold Access journals. This will be borne by the owner organization of the article.
- Cost of the website operation: to be determined.
- Cost of the creation and maintenance of the Github repository that hosts the source code developed at SustainML. As all repositories will be public eventually, the service is free of charge.
- Cost of the creation and maintenance of the private repositories of each partner to keep their results. This cost is to be determined and will be in charge of each partner.
- Cost of publication in ZENODO: free of charge.
- Cost of publication in arXiv: free of charge.
- Cost of Open Source licenses: free of charge.
- Cost of Creative Commons copyright licenses: free of charge.

According to Article 16 Annex 5 (*Ownership of results*) of the SustainML GA each partner owns and is responsible for the management of the results it has produced, and shall operate on behalf of those results. In the event that two or more collaborators jointly generate results and is not possible to establish the contribution of each partner or separate them for maintaining their protection, ownership of the results will be joined to both collaborators. Partners must guarantee the right of access to the data to other beneficiaries, with the purpose of exploiting their own results, and to other affiliated entities up to one year after the termination of the SustainML project.

During the course of the project, research results, data and metadata will be shared by each collaborator in the private Google Drive data repository managed by *eProsima*. Moreover, each collaborator will have a backup of their results in the proprietary IR, DR or HPDC of each institution.

Concerning the deliverables specified in the GA of the SustainML project, their publication, maintenance and re-use will be also in charge of *eProsima*, guaranteeing their availability for at least one year after the project's completion.

7 Data security

The SustainML project guarantees secure access to the data of all SustainML consortium members by restricting and enabling access only to researchers/developers collaborating in the project. Therefore, guidelines with the main security mechanisms adopted are as follows:

- Access rights to the data repository shared by collaborators are managed individually for each researcher/developer/project collaborator. The access policy to the shared repository is managed exclusively by *eProsima*.
- Confidential data will be encrypted upon transmission between consortium members. The passwords for accessing to such data will be communicated directly, avoiding the use of e-mail.
- The security of the data centers of each of the collaborating organizations must be guaranteed by each organization.
- Data shall be stored in two separate locations to avoid data loss.
- A bi-weekly backup of the data generated should be performed to avoid data loss.

The security of the data preserved after the termination of the project must be guaranteed by the partners owning the data.

8 Ethical aspects

This section deals with the ethical and legal aspects related to the collection, handling and processing of personal data resulting from interaction with humans. In the context of the SustainML project, special attention will be paid to the mechanisms for storing, sharing, preserving and protecting the identity of individuals and organizations participating in the different research projects resulting from the SustainML project. When any of the datasets contains confidential personal data or sensitive information, these data will be subject to evaluation by both the organization owning the data and the SustainML consortium in order to prevent the publication of any personal data identifying the participants that could be misused by third parties.

8.1 Ethical self assessment

The project may raise ethical issues in the development of *WP4 - Interaction and User Studies* since the purpose of this work package is the creation of new interaction and visualization approaches that allow any user to interactively explore the alternatives of proficient ML models with intelligent agents. Users will be asked, through the SustainML framework, to express their objectives in terms of the problem they intend to solve with artificial intelligence, as well as to understand the different alternatives and possibilities that the framework proposes to them.

Specifically, the tasks of this work package that require previous interactions with the end users are the following:

- Task *T4.1 Developing ML project descriptions using human-computer partnership.* Users need the ability to express their needs, requirements and constraints for an ML model before even starting the ML project, which can be challenging for inexperienced and advanced ML developers. In order for the SustainML framework to assist developers with this process, it is first necessary to identify and understand their existing workflows to plan, search, and select ML models in their current work process. This will be done through interviews and observational studies of current techniques employed by ML professionals. This will result in the development of new interactive methods for any user to express, define and develop the most relevant descriptions using human-machine approaches.
- Task *T4.3 Exploring ML models using resource footprint estimation.* Developers highly depend on their personal and professional experience, as well as the existence of well-documented ML models and supporting toolkits to decide which ML model is best suited for their project. Therefore, it can be difficult to encourage developers to use the most practical and energy-efficient ML model. This task aims to design and develop new interactive visualizers in which developers are able to collaborate with intelligent systems to explore the most appropriate and energy-efficient ML models. Thus, direct collaboration with end users to test these developments is essential to execute this task.
- Task *T4.5 Evaluation.* The resulting system(s) will be incrementally tested with ML experts and novices, resulting in a formal evaluation of the different visualizations and interaction techniques in a single system.

As discussed above, the research involve humans. The volunteers will be mainly the ML developers, experienced ML researchers, and researchers involved in the project or related departments. Proper ethical procedures will be established in conjuction with the ethical board of the involved organizations, before starting any experiments with human subjects.

The ethics of the project comply with Article 14 of the GA and the applicable international, EU, and national laws, and do not involve any potential misuse of the research results. Moreover, the project will be developed in compliance with: EU 2021 Proposal for a Regulation laying down harmonized rules on artificial intelligence; Ethical guidelines provided by the Horizon 2020 Programme; Charter of

Fundamental Rights of the European Union; European Convention on Human Rights; National, EU and international legislation; The EU General Data Protection Regulation (GDPR).

8.2 Data collection and anonymization process

All the data collected for this project and its scope must fulfill a set of requirements in order to be suitable for its use within the SustainML project:

- In no case will data of minors or persons who cannot give their consent be collected or used.
- Any participant in the project must be informed of the nature of the project itself, and must fulfill an informed consent form facilitated in Annex A of this document.
- The data collected should in no case be sensitive or personal data.

All such data will be completely anonymized prior to use. This will not suppose a lost of information for the project, as the intention of this data will always be the design considerations for the SustainML framework-human interaction. Thus, the anonymized data will fulfill the same behaviour as non-anonymized data.

Anonymization processes will be applied depending on the data collected, the requirements of the specific experiment, and the consent of every participant in the experiment. The complete anonymization of the data will be understood as a process over the collected data where the final data cannot not be de-anonymized, i.e. it will not be possible to obtain personal information from the person behind a specific data.

8.3 Collection of Personal Sensitive Data

The SustainML partners will avoid collecting personal sensitive data, and none of the partners has planned to request the participation of children or people unable to give their consent.

8.4 Informed consent

In order to collect data from individual participants in interviews, workshops or any other activity of the SustainML project, the participant will be required to read and sign a *Informed Consent Form*. An example of this form can be found in Annex A of this document.

9 Other issues

There are no issues to report at this stage of the SustainML project.

10 Datasets

This section provides a detailed definition for each of the datasets used within the context of the SustainML project. Since the project is at a relatively early stage, not all datasets are yet defined. Future versions of this document will update this section with newly generated datasets and possible updates on the dataset already described.

ID	Name	Lead Partner	Work Package	Work Package Task	Description	Data Types	Data Format	New / Existing Data	Mechanisms for data generation	Utility of the Data	Quality Control Procedures	Type of access (Open / Restricted)	Ethical Issues
AFCIIDP	Automated Flow for Custom Instruction Integration in DPU Processors	RPTU	WP2	T2.3	Exploration of PiM processors Operators/Accelerators for ML/DNN layers efficiency optimization	Code	RTL (Verilog)	New	Manual generation	Allows energy efficiency measurements on various architectures experiments	Functions through kernels RTL simulation, effectiveness (energy) through standard power measurement EDA flow	Restricted	None
HCIMLE	ML-expert interviews	INRIA	WP4	T4.1, T4.2, T4.4	Interview data regarding the planning, selecting and testing of ML-approaches by experts	Text	.xls, .m4a, .txt	New	Audio recording and transcripts	Input for system development	-	Restricted according to GDPR	Personal information (will have approval from the ethics board of Inria Paris-Saclay)
HWKINN	Energy-efficient Hardware kernel implementations for NN architectures	RPTU	WP2	T2.1	Implementation Templates for different NN HW kernels	Code	SystemVerilog / RTL	New	Via a generation framework	Used for different HW architecture of NN Layers (DL/ML)	Via test and validation of the implemented kernels	Open	None
KGMLT	Knowledge Graph representing ML Tasks from description to algorithm	DFKI	WP1	T1.1 T1.2 T1.3	Basic structure covering the majority of ML Taks (types, description, pipeline, specification, algorithms, data types,...)	Metadata, Python Source Code	Turtle syntax for storing RDF data (.ttf)	New	Via a generation framework	Basic knowledge base to be used and build on in follow-up projects	based on high-ranked publications and expert knowledge	Open	None

ESMLFW	SustainML framework	EPROS	WP5	T5.1, T5.2, T5.3	SustainML framework implementation, validation and testing. This dataset will include the user manual documentation. See eProsima/SustainML GitHub repository[13].	Code and Text	Python code, C++ code, CMake, reStructured-Text (RST), HTML, CSS	New	Manual generation	This is the source code and user manual documentation of SustainML framework	Every contribution is peer-reviewed by a member of eProsima	Open	None
PASP	PyTorch Accelerator Simulator and Profiler	UPMEM	WP2	T2.3	Simulation-based accelerator profiling through PyTorch framework execution.	Code	Python code	New	Manual generation	Allows energy efficiency explorations on various accelerator architectures (CPU, GPU, PIM...)	Code review, simulated output benchmarking	Open	None
WP1SMLFW	WP1 SustainML framework nodes	EPROS	WP1, WP5	T1.1, T1.2, T1.3, T1.4	The design of the SustainML framework is based on a library that allows each partner to seamlessly integrate the code resulting from the research of each WP1. This dataset corresponds to the source code of the implementation of the WP1 nodes.	Code	Python code	New	Manual generation	Source code of WP1 integration into SustainML framework	Every contribution is peer-reviewed by a member of eProsima	Open	None
WP2SMLFW	WP2 SustainML framework nodes	EPROS	WP2, WP5	T2.1, T2.2, T2.3	The design of the SustainML framework is based on a library that allows each partner to seamlessly integrate the code resulting from the research of each WP2. This dataset corresponds to the source code of the implementation of the WP2 nodes.	Code	Python code	New	Manual generation	Source code of WP2 integration into SustainML framework	Every contribution is peer-reviewed by a member of eProsima	Open	None

[13] https://github.com/eProsima/SustainML

WP3SMLFW	WP3 SustainML framework nodes	EPROS	WP3, WP5	T3.1, T3.2, T3.3	The design of the SustainML framework is based on a library that allows each partner to seamlessly integrate the code resulting from the research of each WP3. This dataset corresponds to the source code of the implementation of the WP3 nodes.	Code	Python code	New	Manual generation	Source code of WP3 integration into SustainML framework	Every contribution is peer-reviewed by a member of eProsima	Open	None

Table 2: Datasets summary

Dataset ID	Data dissemination	Metadata standard	Type of associated metadata	How will the data be findable	Software required to use / read the data
AFCIIDP	GitHub	-	-	Ask UPMEM for permission	Standard EDA tools (RTL simulators / synthesizers / linters)
HCIMLE	No	-	-	-	-
HWKINN	Gitlab	-	-	-	-
KGMLT	GitHub	-	-	Ask DFKI for permission	Tool to open .ttf files (e.g. GraphDB, Blazegraph, Virtuso)
ESMLFW	GitHub, ReadTheDocs, and eProsima social media channels will be used to disseminate SustainML framework progress and updates	GitHub and ReadTheDocs metadata and keywords to identify repositories	GitHub and ReadTheDocs metadata	SustainML framework source code and documentation will be openly accessible on GitHub and ReadTheDocs	None (software dependencies will be on the installation manual)
PASP	GitHub	GitHub metadata and keywords to identify repositories	GitHub metadata	To be determined	PyTorch framework and PyTorch model
WP1SMLFW	GitHub, and eProsima social media channels will be used to disseminate SustainML framework progress and updates	GitHub metadata and keywords to identify repositories	GitHub metadata	SustainML framework source code and documentation will be openly accessible on GitHub and ReadTheDocs	None (software dependencies will be on the installation manual)
WP2SMLFW	GitHub, and eProsima social media channels will be used to disseminate SustainML framework progress and updates	GitHub metadata and keywords to identify repositories	GitHub metadata	SustainML framework source code and documentation will be openly accessible on GitHub and ReadTheDocs	None (software dependencies will be on the installation manual)
WP3SMLFW	GitHub, and eProsima social media channels will be used to disseminate SustainML framework progress and updates	GitHub metadata and keywords to identify repositories	GitHub metadata	SustainML framework source code and documentation will be openly accessible on GitHub and ReadTheDocs	None (software dependencies will be on the installation manual)

Table 3: Datasets description - Making Data Findable

Dataset ID	Data storage location (short-term)	Data storage location (long-term)	Storage media	Data security provisions	Expected size of the data	Person responsible	Cost
AFCIIDP	GitHub server	GitHub server	-	Access Control	MBs	UPMEM IT	-
HCIMLE	Inria with personal identifier	Inria without personal identifier	External hard-drive	Access Control / Inria Safe	$<$ 5 GB	Janin Koch, Wendy Mackay	-
HWKINN	Internal Gitlab server	External Gitlab server	-	Access Control	4-8 MB	RPTU-EIT	-
KGMLT	GitHub server	GitHub server	-	Access Control	$<$ 5 MB	DFKI-EI	-
ESMLFW	GitHub server	GitHub server	-	Access Control	$<$ 10 MB	Raúl Sánchez-Mateos	Free
PASP	GitHub server	GitHub server	-	-	MBs	UPMEM IT	-
WP1SMLFW	GitHub server	GitHub server	-	Access Control	Cannot be quantified at this stage of the project.	Raúl Sánchez-Mateos	Free
WP2SMLFW	GitHub server	GitHub server	-	Access Control	Cannot be quantified at this stage of the project.	Raúl Sánchez-Mateos	Free
WP3SMLFW	GitHub server	GitHub server	-	Access Control	Cannot be quantified at this stage of the project.	Raúl Sánchez-Mateos	Free

Table 4: Datasets description - Storage of Data

Dataset ID	Reuse of existing data	Potential reuse of data	How will data be reused	Type of access (Open / Restricted)	Type of IP / protection sought	License
AFCIIDP	Not planned	Implementation in next generation ML/DNN PiM processors	Internal projects	Restricted	-	-
HCIMLE	No	No	-	-	-	-
HWKINN	Not planned	Beneficial in different HW NN projects	In different national / international projects	Open	-	BSD
KGMLT	No	Follow-up projects	Internal / External follow-up projects	Open	-	-
ESMLFW	eProsima open source software as eProsima Fast DDS and eProsima DDS Router	Yes	Internally by incorporating findings to other eProsima products or externally by any other user / developer / company that wants to re-use the source code	Open	-	Apache License 2.0
PASP	Yes, PyTorch models and framework	-	Benchmarking	Open	-	Pytorch: BSD-3; HuggingFace transformers library: Apache 2.0; Model (or weights) of Mistral 7B: Apache 2.0; Weights of Mixtral 8x7B: Apache 2.0; Weights of Llama2-7B/13B/70B: Llama 2 license (https://llama.meta.com/llama-downloads/) kind of limited open-source license; Weights of GPT-J: Apache 2.0; Weights of GPT-XL: MIT
WP1SMLFW	Open source software that could be re-use out of the SustainML project scope	Yes	Internally by incorporating findings to other eProsima software or externally by any other user / developer / company that wants to re-use the source code	Open	-	Apache License 2.0
WP2SMLFW	Open source software that could be re-use out of the SustainML project scope	Yes	Internally by incorporating findings to other eProsima software or externally by any other user / developer / company that wants to re-use the source code	Open	-	Apache License 2.0
WP3SMLFW	Open source software that could be re-use out of the SustainML project scope	Yes	Internally by incorporating findings to other eProsima software or externally by any other user / developer / company that wants to re-use the source code	Open	-	Apache License 2.0

Table 5: Datasets description - Re-use data

A Informed Consent Form Template

Informed Consent Form for Participation [template]

Project Description

SustainML project aims to develop a design framework and an associated toolkit, so-called SustainML, that will foster energy efficiency throughout the whole life-cycle of ML applications: from the design and exploration phase that includes exploratory iterations of training, testing and optimizing different system versions through the final training of the production systems (which often involves huge amounts of data and computation and (where appropriate) continuous online re-training during deployment for the inference process. The framework will optimize the ML solutions based on the application tasks, across levels from hardware to model architecture. ML developers from all experience levels can make use of the framework through its emphasis on human-centric interactive transparent design and functional knowledge cores, instead of the common blackbox and fully automated optimization approaches.

Title:	SustainML: Application Aware, Life-Cycle Oriented Model-Hardware Co-Design Framework for Sustainable, Energy Efficient ML Systems
Acronym:	SustainML
Coordinator:	eProsima
Grant agreement ID:	101070408
Call:	HORIZON-CL4-2021-HUMAN-01
Program:	Horizon Europe
Start:	01 October 2022
Duration:	36 months
Website:	https://sustainml.eu
E-mail:	sustainml@eprosima.com
Consortium:	**eProsima (EPROS)**, Spain **DFKI**, Germany **Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau (RPTU)**, Germany **University of Copenhagen (KU)**, Denmark **National Institute for Research in Digital Science and Technology (INRIA)**, France **IBM Research GmbH**, Switzerland **UPMEM**, France

This project has received funding from the European Union's Horizon Europe research and innovation programme (HORIZON-CL4-2021-HUMAN-01) under grant agreement No 101070408.

Information Sheet

Purpose of the research and of data collection:

Research Title: _____

Coordinator: _____

Website: _____

Contact E-mail: _____

Activity Details: _____

_____

_____

Possible Risks: _____

_____

_____

Incentives: _____

_____

_____

This project has received funding from the European Union's Horizon Europe research and innovation programme (HORIZON-CL4-2021-HUMAN-01) under grant agreement No 101070408.

Detailed Consent

I hereby confirm my participation in __Project Title__ pilot testing to be held in the context of the SustainML project.

I have been informed by the head of research about the purpose of SustainML and of the __Project Title__ testing. I have been informed about the nature and scope of the data which will be recorded and stored and about the purpose for which it may be used.

I understand that my responses will be documented (via raw data, audio and video recordings, photographs, written reports) and will be stored until __Date of deletion (usually on year after the end of the project)__ .

I understand that I can request rectification or deletion of this data before this date whenever this data has not been completely anonymized.

My participation in the __Project Title__ pilot testing is voluntary and I understand that I can end my participation at any time. I understand that I do not have to give reasons for ending my participation if I do not want to.

☐ I have read the foregoing information; I have had the opportunity to ask questions about it and questions have been answered to my satisfaction. By signing this form, I acknowledge that I have understood and agreed to the above terms.

Participant Name: _____ Participant signature: _____

Researcher Name: _____ Researcher signature: _____

Place and Date: _____

This project has received funding from the European Union's Horizon Europe research and innovation programme (HORIZON-CL4-2021-HUMAN-01) under grant agreement No 101070408.

References

[1] OpenAIRE. *How to comply with Horizon Europe mandate for Research Data Management*. URL: https://www.openaire.eu/how-to-comply-with-horizon-europe-mandate-for-rdm (visited on Mar. 1, 2023).

[2] European Commission. *Open access and Data management*. URL: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm (visited on Feb. 11, 2021).

[3] Mark D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1 (Mar. 2016), p. 160018. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. URL: https://doi.org/10.1038/sdata.2016.18.

[4] European Union. *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. Apr. 2016. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

[5] Line Pouchard. "Revisiting the Data Lifecycle with Big Data Curation". In: *International Journal of Digital Curation* 10 (June 2015). DOI: 10.2218/ijdc.v10i2.342.