



Human-centric algebraic machine learning

Data Management Plan - Revised

Deliverable D8.11

WP8 - Dissemination, Exploitation and
Collaboration



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952091.

Consortium



ALGEBRAIC AI



Chantalimaud
Foundation



EPROSIMA
The Middleware Experts



German
Research Center
for Artificial
Intelligence



TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

uc3m



Project

Title: ALMA: Human Centric Algebraic Machine Learning
 Acronym: ALMA
 Coordinator: Proyectos y Sistemas de Mantenimiento SL (eProsima)
 Grant agreement: 952091
 Call: H2020-EIC-FETPROACT-2019
 Program: Horizon 2020
 Start: 01 September 2020
 Duration: 54 months
 Website: <https://alma-ai.eu>
 E-mail: alma@eprosima.com
 Consortium: **Proyectos y Sistemas de Mantenimiento SL (eProsima) (EPROS)**, Spain
Champalimaud Foundation (CF), Portugal
ALGEBRAIC AI (ALG), Spain
German Research Center for Artificial Intelligence (DFKI), Germany
Inria (Inria), France
TU Kaiserslautern (TUK), Germany
FIWARE Foundation e.V. (FF), Germany
Technical Research Centre of Finland (VTT), Finland
Carlos III University (UC3M), Spain

Deliverable

Number: **D8.11**
 Title: **Data Management Plan - Revised**
 Month: 54
 Work Package: WP8 - Dissemination, Exploitation and Collaboration
 Work Package leader: Proyectos y Sistemas de Mantenimiento SL (eProsima)
 Deliverable leader: Proyectos y Sistemas de Mantenimiento SL (eProsima)
 Deliverable type: ORPD: Open Research Data Pilot
 Dissemination level: Public (PU)
 Date of submission: 28/02/2025
 Version: v2.1
 Status: Finished

Version history

Version	Date	Responsible	Author/Reviewer	Comments
v1.0	16/02/2021	EPROS	Raúl Sánchez-Mateos Lizano	Initial draft
v1.1	17/02/2021	EPROS	Javier Paris Uhryn	First review (RV1)
v1.2	17/02/2021	EPROS	Raúl Sánchez-Mateos Lizano	Updated Informed Consent Form
v1.3	30/06/2022	EPROS	Javier Paris Uhryn	Apply PO and EU Reviewers suggestions
v1.4	01/07/2022	EPROS	Juan López Fernández	Second review (RV2)
v2.0	27/01/2025	EPROS	Denisa Alexandru	Third review (RV3)
v2.1	14/02/2025	EPROS	Raúl Sánchez-Mateos Lizano	Final review

Executive summary

Algebraic Machine Learning (AML) is a relatively new machine learning technique based on algebraic representations of data. Unlike statistical learning, AML algorithms are robust regarding the statistical properties of the data and are parameter-free. The aim of the EU-funded ALMA project is to leverage AML properties to develop a new generation of interactive, human-centric machine learning systems. These systems are expected to reduce bias and prevent discrimination, remember what they know when they are taught something new, facilitate trust and reliability and integrate complex ethical constraints into human-artificial intelligence systems. Furthermore, they are expected to promote distributed, collaborative learning.

This report corresponds to *Deliverable D8.11 - Data Management Plan - Revised* of the ALMA project. This deliverable covers the specification of what data is open, details what data the project has generated, whether and how it has been exploited or made accessible for verification and re-use, and how it is curated and preserved. The ALMA project produced data from experiments and benchmarking of the AML engine on different scenarios, along with data relative to the different use cases. The performance of the different system tools also produced some reportable data. This document has been updated to reflect the most recent state of the project.

Contents

Executive Summary	4
Contents	5
Acronyms	6
1 Introduction	7
2 Data Cycle	8
3 Open access policy of ALMA	8
4 Data summary	9
4.1 Data description	9
5 FAIR data	10
5.1 Making data findable	10
5.2 Making data openly accessible	11
5.3 Making data interoperable	13
5.4 Increase data re-use	14
6 Allocation of resources	14
7 Data security	15
8 Ethical aspects	15
8.1 Ethical self assessment	15
8.2 Collection of Personal Sensitive Data	16
8.3 Informed consent	16
9 Other issues	16
10 Datasets	16
A Informed Consent Form Template	23
References	26

Acronyms

AI	Artificial Intelligence.
AML	Algebraic Machine Learning.
CA	Consortium Agreement.
CDR	Common Data Representation.
CERN	European Organization for Nuclear Research.
DDS	Data Distribution Service.
DL	Deep Learning.
DMP	Data Management Plan.
DOI	Digital Object Identifier.
DPO	Data Protection Officer.
DR	Disciplinary Repository.
EU	European Union.
FAIR	Findable, Accessible, Interoperable and Reusable.
GA	Grant Agreement.
GDPR	General Data Protection Regulation.
HPC	High Performance Data Center.
IDL	Interactive Data Language.
IPR	Intellectual Property Rights.
IR	Institutional Repository.
ML	Machine Learning.
OpenAIRE	Open Access Infrastructure for Research in Europe.
ORDP	Open Research Data Pilot.

1 Introduction

The ALMA research and innovation project aims to change the Artificial Intelligence (AI) paradigm by creating a new Machine Learning (ML) approach known as AML. AML is a new ML technique based on the algebraic representation of data. The goal of the EU-funded ALMA project is to leverage the properties of AML to develop a new generation of interactive, human-centered machine learning systems.

During the course of the ALMA project, a wide variety of data has been generated from R&D activities; mathematical publications describing the operation and capabilities of the AML, reports describing the ALMA system as a whole, datasets developed for the input of ML models, results and models generated through these activities, and theoretical and numerical analyses of the ALMA prototypes along with their validation.

ALMA complies with Article 29 of the Grant Agreement (GA) for the dissemination of results and open access to them. According to this article, partners shall disseminate their results by publishing them through open means, including scientific publications of any format. However, the obligation to protect results in Article 27, the confidentiality obligations in Article 36, the security obligations in Article 37 or the obligations to protect personal data in Article 39 are maintained.

Since AML technology is state-of-the-art, it is important that the 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 Open Research Data Pilot (ORDP) in Horizon 2020, ALMA has made its research data Findable, Accessible, Interoperable and Reusable (FAIR) and has followed the ORDP principles, i.e., "*as open as possible, as closed as necessary*", encouraging data management as an essential part of research best practice. 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 Intellectual Property Rights (IPR), privacy concerns, security, as well as data management and preservation issues*", as stated in the Guidelines on FAIR data management in Horizon 2020 published by the European Commission [1].

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

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

- Name, reference and versioning of the data.
- 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 [1] and is written with the The FAIR Guiding Principles for scientific data management and stewardship [2] and the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679) [3] as references.

2 Data Cycle

The amount of data that the ALMA project has collected and generated is significant. Furthermore, since the project data are highly diverse, and in order to improve their re-use throughout the project (widely expanded in time) it has been agreed to follow the data life cycle model described in [4] and shown in Figure 1. In the first phase of the project a first version of the DMP was presented, stipulating what to do with the data collected and generated. These data have been subsequently integrated, prepared and analyzed for use. During this phase the data were described, named and identified as explained in this document. In a third phase, the data were presented to the ALMA consortium to assess their impact, determine which datasets will be published and made Open Access. Since ALMA 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 was defined how long the data will be kept for later re-use, both by ALMA consortium members and third party researchers. Data may be stored in a cloud infrastructure hosted by any of the ALMA consortium members, an Institutional Repository (IR), a Disciplinary Repository (DR), or a High Performance Data Center (HPC) provided by public or private organizations.

The activity of assuring data quality through the process of data description and evaluation is carried out during the entire data life cycle. This implies that, as stated above, the present document was updated at any stage in the course of the ALMA project.

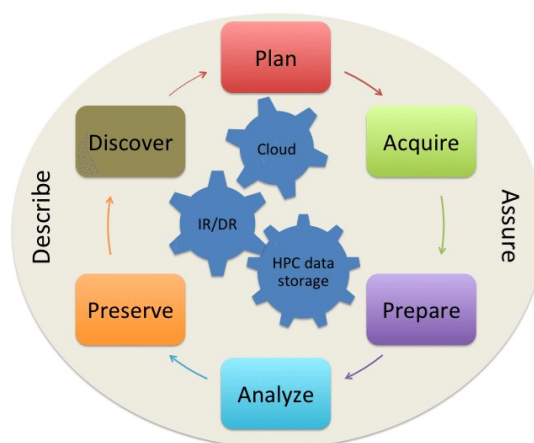


Figure 1: The Big Data Life Cycle Model [4].

3 Open access policy of ALMA

According to Article 29.1 of the GA, *"unless it goes against their legitimate interests, each beneficiary must - as soon as possible - 'disseminate' its results by disclosing them to the public by appropriate means (other than those resulting from protecting or exploiting the results), including in scientific publications (in any medium)."*

Thus, and as stipulated in this same document, all project deliverables and generated data were made publicly available, except for the ethical reports, whose distribution is private and internal to ALMA consortium members. The research data have been uploaded to a public domain repository, and/or published in scientific journals, thus making it openly accessible. Furthermore, the project deliverables

can be found on the official ALMA website¹.

However, for each of the datasets and knowledge generated, the ALMA consortium analyzed the potential conflicts of openly publishing the results versus the commercialization and application of IPR protection, thus deciding what information were made publicly available and when it remained private. Since ALMA is based on the use of a state-of-the-art ML (AML) technology, 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 29.3, which states that *"As an exception, the beneficiaries do not have to ensure open access to specific parts of their research data under Point (a)(i) and (iii), if the achievement of the action's main objective (as described in Annex 1) would be jeopardised by making those specific parts of the research data openly accessible. In this case, the data management plan must contain the reasons for not giving access."*

4 Data summary

This section describes each of the datasets collected, processed or generated in the R&D process of the ALMA project.

According to the European Union (EU) guidance proposed for the DMP, the data collected, processed and generated within the ALMA project have been characterized to meet the FAIR data specifications.

4.1 Data description

First, since the goal of the project is the creation of an ecosystem that allows the use and development of the AML technology, a large amount of data was necessary for the evaluation of the AML computational engine as well as the systems belonging to the ecosystem that surrounds it. These datasets were generated by ALMA consortium members or collected from public data repositories, which provide datasets for the use, evaluation, training and refinement of ML technologies. The ALMA proprietary datasets were added and specified in this document in a table that contains the following aspects.

- *ID* of the dataset that uniquely identifies it.
- *Name* of the dataset.
- *Lead partner* that is responsible of the dataset, its maintenance, and the data in it. This includes ensuring compliance with the data management and ethics requirements of the project.
- *WP* referencing the Work Package which this dataset has been used or generated.

All generated datasets are characterized with the following parameters:

- Unique identifier, whose composition is detailed in section A.
- Extended name.
- Data types.
- Data format.

¹<https://alma-ai.eu/>

- Reuse of existing data.
- Mechanisms for data generation.
- Expected size of the data.
- Utility of the data.
- Potential reuse of the data.
- Data dissemination.

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

5 FAIR data

This section develops the methodology that has been followed in order to ensure that data management in the context of the ALMA 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 do 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

ALMA 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 ALMA project. Thus, through the use of precise, detailed and accurate metadata, ALMA project datasets will be easily found by other researchers. The metadata used in ALMA (data type, quality, availability, author, content description, versioning, etc.) were deposited in standardized formats in the ALMA research data repositories, which are publicly accessible. Regarding bibliographic metadata, these comply with Article 29.2 of the ALMA Grant Agreement 952091. These have a standard format and include the following:

- *"the terms "European Union (EU)" and "Horizon 2020" ²;*
- *the name of the action, acronym and grant number;*
- *the publication date, and length of embargo period if applicable, and*
- *a persistent identifier." Name*

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

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

ALMA_XX_DS_ID_YY_ZZ

where

²All publications of the ALMA project indicate the EU funding of the project by including the following statement: "ALMA project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952091."

- **ALMA**: name of the project.
- **XX**: date of creation according to ISO 8601³ format (YYYYMMDD).
- **DS**: dataset name. The name must not contain blank spaces and the PascalCase⁴ 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 have been defined to facilitate the search of ALMA datasets. Furthermore, these metadata incorporates keywords that briefly describe and define the data. Some keywords frequently used to describe the project's research data are: *algebra*, *machine learning*, *human-computer interaction*, *artificial intelligence*, *optimization*, *robotics*, *Data Distribution Service (DDS)*.

5.2 Making data openly accessible

ALMA project datasets were initially held in a shared repository, accessible to the ALMA consortium members. The purpose is that the data have been treated, processed and evaluated prior to publication. The final or stable version of the datasets available in the ALMA private repository were evaluated by the consortium using the guidelines defined in 3. Mathematical publications were published in arXiv⁵,

³<https://www.iso.org/iso-8601-date-and-time-format.html>

⁴<https://techterms.com/definition/pascalcase>

⁵<https://arxiv.org/>

a free distribution service and open access archive for academic articles. Materials on this site are not peer-reviewed by arXiv.

Moreover, in case IPR or patent protection actions were applied, some the project results have been published in ZENODO⁶, a general-purpose open access repository developed under the Open Access Infrastructure for Research in Europe (OpenAIRE) program and operated by European Organization for Nuclear Research (CERN).

The project prioritized transparency and accessibility by ensuring that all peer-reviewed scientific publications were freely accessible to the public. This was achieved through publishing in Open Access scientific journals or conferences, which do not charge readers for access to the published materials. In cases where a journal did not provide an Open Access option, the project took additional steps to maintain accessibility by depositing an author's version of the article in a publicly accessible repository. This approach maximized the availability of the project data and findings to any interested party, regardless of institutional or financial barriers.

The data have been 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. Some formats for the metadata types are the following:

- PDF format for documents and publications.
- Preferable PNG 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 published on the ALMA 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.
- Interactive Data Language (IDL) files for the generation of Common Data Representation (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 AML. In case the data structure requires it, JSON format can also be used for the datasets.

Furthermore, binary data were generated and it was published as a result of the experiments to be carried out within the ALMA project.

Regarding the distribution means to make the data open, two ways of distributing the datasets are distinguished: internal distribution, accessible only to ALMA consortium members, and public distribution,

⁶<https://zenodo.org>

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 have been used:

- Proprietary data centers of the institutions collaborating in the project.
- Google Drive file sharing and collaborative work service managed by *Proyectos y Sistemas de Mantenimiento SL (eProxima)*⁷.
- Private section of the ALMA project official website.

For open access to the data the following platforms have been used:

- arXiv for the publication of mathematical articles.
- ZENODO⁸ for the publication of data and metadata of the results and progress of the research, development, deployment and testing processes.
- Github⁹ to host all source code defined as public and open to the user. This platform has also been used for the development of a file sharing system where the public data of the project are included, whether they are scientific publications, generated datasets, or conferences and presentations of the ALMA project in different events.
- YouTube to host the presentations and speeches at the events defined in the Grant Agreement 952091, thus granting higher exposure to the project.
- ALMA project web page.

To regulate access to private data, the institution hosting the data granted 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 ALMA project. These have been carefully selected to ensure the interoperability of the data. All data generated in the context of the ALMA project conforms 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 has been used for the description of data and metadata. ALMA adopts the DataCite¹⁰ Metadata Schema for metadata standardization. Furthermore, the keywords defined in Section 5.1 have been 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 ALMA (ethics, human-computer interaction, mathematical, computational, etc.), a standardized vocabulary was created to achieve interoperability of the data across the project's disciplines.

⁷<https://www.eproxima.com/>

⁸<https://zenodo.org/>

⁹<https://github.com/>

¹⁰<https://datacite.org/>

5.4 Increase data re-use

ALMA data have been made Open Access as soon as the data were reviewed by other members of the ALMA consortium and a stable version of the data have been submitted. However, an embargo period has been applied by the publishing journal if the data were subject to IPR protection or were awaiting patent protection [5].

As defined in the project's GA (Article 31.3 and 31.4), the data must be available to be re-used up to at least one year after the completion of the project. Each collaborator responsible for its results 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 ALMA project.

Data quality assurance measures shall be dictated and conducted by the institution that generates the data as the data owner. These measures 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 the costs assigned to FAIR data compliance in the ALMA project, how these costs were 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 was 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 ALMA. This 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 arXiv: free of charge.
- Cost of publication in ZENODO: free of charge.
- Cost of Open Source licenses: free of charge.
- Cost of Creative Commons copyright licenses: free of charge.

According to Article 8.1 of the Consortium Agreement (CA) 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, ownership of the results was joined to both collaborators. Partners must guarantee the right of access to the data to other beneficiaries, to exploit their own results, and to other affiliated entities up to one year after the termination of the ALMA project.

During the course of the project, research results, data and metadata have been shared by each collaborator in the private Google Drive data repository managed by *Proyectos y Sistemas de Mantenimiento*

SL (eProxima). Moreover, each collaborator has a backup of their results in the proprietary IR, DR or HPC of each institution.

Concerning the deliverables specified in the GA of the ALMA project, their publication, maintenance and re-use *Proyectos y Sistemas de Mantenimiento SL (eProxima)*, guaranteeing their availability for at least one year after the project's completion.

7 Data security

The ALMA project guarantees secure access to the data of all ALMA 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 *Proyectos y Sistemas de Mantenimiento SL(eProxima)*.
- Confidential data is encrypted upon transmission between consortium members. The passwords for access to such data are 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 ALMA project, special attention was 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 ALMA project. When any of the datasets contained confidential personal data or sensitive information, these data were subject to evaluation by both the organization owning the data and the ALMA 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 raised ethical issues in the development of *WP7 - Use cases* as the AML algorithms learns through interaction with humans. Specifically, Task 7.1 used public and anonymized medical images that contains personal data. These images come from previous Deep Learning (DL) studies performed by other research groups, and the organizations that exploit these images applied to the *Champalimaud Foundation* ethics committee for their use - strict ethical approval procedures were followed for these specific cases.

As discussed above, the research involves humans as the AML algorithms learns to interact with volunteers. The volunteers are mainly the researchers involved in the project or related departments. Proper ethical procedures working together with the ethical board of the involved organizations were established before starting any experiments with human subjects.

The ethics of the project comply with Article 34 of the GA and the applicable international, EU and national law (in particular, the EU export control Regulation No 428/2009) and do not involve any potential misuse of the research results.

8.2 Collection of Personal Sensitive Data

The ALMA partners avoided collecting personal sensitive data, and none of the partners has requested the participation of children or people unable to give their consent, specifically:

- Task 7.1: there are public sets of data already anonymized, thus the ALMA project does not deal with sensitive personal data.
- Task 7.3 (living scenarios): only university personnel was recruited to perform the experiments (written consent). The *Carlos III University*, as host institution, confirms that it has appointed a Data Protection Officer (DPO) for this task.

8.3 Informed consent

When it was necessary to collect data from individual participants for interviews, workshops or any other activity of the ALMA project, the participant was required to read and sign an Informed Consent Form. An example of this form can be found in Annex A of this document.

9 Other issues

There are no other issues to report within the ALMA project.

10 Datasets

This section provides a detailed definition of each of the datasets leveraged within the context of the ALMA project.

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
DDROUTER	DDS-Router	EPROS	WP6	T6.1	DDS Router implementation, validation, testing, and documentation. See eProsima DDS-Router GitHub repository ¹¹ .	Code and Text	C++ code, Python code, CMake, Shell, PowerShell, reStructured-Text(RST), HTML, CSS	New	Manual generation	Source code and user manual documentation of the DDS-Router Tool. It serves as the base for WAN communication in AML-IP	Every contribution is peer-reviewed by a member of eProsima	Open	None
AMLIP	AML-IP	EPROS	WP6	T6.1, T6.2, T6.3, T6.4	AML-IP implementation, validation, testing and documentation. See eProsima AML-IP GitHub repository ¹² .	Code and Text	C++ code, Python code, CMake, SWIG, Dockerfile, reStructured-Text(RST), HTML, CSS, Shell	New	Manual generation	Source code and user manual documentation of the AML-IP framework in charge of data exchange between Algebraic Machine Learning (AML) nodes through local or remote networks.	Every contribution is peer-reviewed by a member of eProsima.	Open	None
FDSDM	FastDDS-Monitor	EPROS	WP6	T6.1, T6.3, T6.4	Fast-DDS Monitor implementation, validation, testing, and documentation. See eProsima Fast-DDS Monitor GitHub repository ¹³ .	Code and Text	C++ code, QML, CMake, QMake, reStructured-Text(RST), HTML, CSS	Existing Data	Manual generation	Source code and user manual documentation of the Fast-DDS Monitor in charge of monitoring the state of the AML-IP networks	Every contribution is peer-reviewed by a member of eProsima	Open	None

¹¹<https://github.com/eProsima/DDS-Router>

¹²<https://github.com/eProsima/AML-IP>

¹³<https://github.com/eProsima/Fast-DDS-monitor>

AMLDDDEV	AML-Dashboard	EPROS	WP6	T6.4	AML-Dashboard implementation. See eProxima AML-Dashboard GitHub repository ¹⁴ .	Code	Python code, JavaScript code, Dockerfile, HTML	New	Manual generation	Source code of the AML-Dashboard, a user-friendly graphical interface that allows interacting with AML-IP and AML	Every contribution is peer-reviewed by a member of eProxima	Open	None
AMLUSER	AML-Dashboard	INRIA	WP4	T4.3	AML-Dashboard allowing user-defined input data to train personalized AML models that can be applied in place. It also allows comparing AML with any TensorFlow model trained on the same data.	Code	Python code, JavaScript code, Dockerfile, HTML	New	Manual generation	Source code of the AML-Dashboard, a user-friendly graphical interface that allows interacting with AML-IP and AML	Every contribution is peer-reviewed by a member of INRIA.		
AMLTOOLS	AML-Tools	EPROS	WP6	T6.4	Open-source documentation for AML-Toolkit that allow users to develop, train, and deploy Algebraic Machine Learning (AML) models. See eProxima AML-Tools GitHub repository ¹⁵ .	Code and Text	Python code, reStructuredText(RST), HTML, CSS	New	Manual generation	User manual documentation of the AML-Toolkit.	Every contribution is peer-reviewed by a member of eProxima	Open	None

¹⁴<https://github.com/eProxima-Private/AML-Dashboard>

¹⁵<https://github.com/eProxima-Private/AML-Tools>

WP7FOLDING	ALMA WP7 Folding	UC3M	WP7	T7.3	Dataset accompanying the deliverables "D7.5 - Design of the benchmark tested for robotized ironing and garment folding - Initial" and "D7.6 - Design of the benchmark tested for robotized ironing and garment folding - Final" as part of the "ALMA: Human Centric Algebraic Machine Learning" project, CORDIS - EU research results Grant agreement ID: 952091.	Compressed file composed by: 10 010 png images, labels.txt file as comma separated values.	Image files in png format	New	Automatic generation created with cloth2d.py (dataset generator) at https://github.com/roboticslab-uc3m/gymnasium-alma	-	Every contribution is peer-reviewed by a member of eProxima	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
DDSROUTER	GitHub, ReadTheDocs, and eProxima social media channels will be used to disseminate DDS-Router progress and updates	GitHub and ReadTheDocs metadata and keywords to identify repositories	GitHub and ReadTheDocs metadata	DDS-Router source code and documentation will be openly accessible on GitHub and ReadTheDocs	None (software dependencies will be on the installation manual)
AMLIP	GitHub, ReadTheDocs, and eProxima social media channels will be used to disseminate AML-IP progress and updates	GitHub and ReadTheDocs metadata and keywords to identify repositories	GitHub and ReadTheDocs metadata	AML-IP source code and documentation will be openly accessible on GitHub and ReadTheDocs	None (software dependencies will be on the installation manual)
FDDSM	GitHub, ReadTheDocs, and eProxima social media channels will be used to disseminate FastDDS Monitor progress and updates	GitHub and ReadTheDocs metadata and keywords to identify repositories	GitHub and ReadTheDocs metadata	FastDDS Monitor source code and documentation will be openly accessible on GitHub and ReadTheDocs	None (software dependencies will be on the installation manual)
AMLDDDEVEL	GitHub, ReadTheDocs, and eProxima social media channels will be used to disseminate AML-Dashboard progress and updates	GitHub and ReadTheDocs metadata and keywords to identify repositories	GitHub and ReadTheDocs metadata	AML-Dashboard source code and documentation will be openly accessible on GitHub and ReadTheDocs	None (software dependencies will be on the installation manual)
AMLUSER	GitHub	GitHub	GitHub	AML-Dashboard source code and documentation will be openly accessible on GitHub and ReadTheDocs	Dockerfile containing AML engine
AMLTOOLS	GitHub, ReadTheDocs, and eProxima social media channels will be used to disseminate AML-Toolkit progress and updates	GitHub and ReadTheDocs metadata and keywords to identify repositories	GitHub and ReadTheDocs metadata	AML-Tools source code and documentation will be openly accessible on GitHub and ReadTheDocs	None (software dependencies will be on the installation manual)
WP7FOLDING	Zenodo, GitHub, ReadTheDocs, and RoboticsLab-UC3M social media channels will be used to benchmark disseminate progress and updates.	Zenodo, GitHub and ReadTheDocs metadata and keywords to identify repositories	Zenodo, and ReadTheDocs metadata	https://cordis.europa.eu/project/id/952091/results	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
DDROUTER	GitHub server	GitHub server	-	Access Control	12.74 MBs	Raúl Sánchez-Mateos	Free
AMLIP	GitHub server	GitHub server	-	Access Control	11.48 MBs	Raúl Sánchez-Mateos	Free
FDDSM	GitHub server	GitHub server	-	Access Control	28.16 MBs	Raúl Sánchez-Mateos	Free
AMLDDDEVEL	GitHub server	GitHub server	-	Access Control	67.46 MBs	Raúl Sánchez-Mateos	Free
AMLUSER	GitHub server	GitHub server	-	Access Control	<100 MBs	Janin Koch	Free
AMLTOOLS	GitHub server	GitHub server	-	Access Control	34.38 MBs	Raúl Sánchez-Mateos	Free
WP7FOLDING	https://zenodo.org/records/14864393	DOI 10.5281/zenodo.14864392	-	Access Control	<40 MBs	-	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
DDROUTER	eProsimas open source software such as eProsimas Fast DDS, DDS Pipe, and Dev Utils	Yes	Internally by incorporating findings to other eProsimas products (such as AML-IP) or externally by any other user / developer / company that wants to re-use the source code	Open	-	Apache License 2.0
AMLIP	eProsimas open source software such as eProsimas Fast DDS or DDS Router	Yes	Internally by incorporating findings to other eProsimas products or externally by any other user / developer / company that wants to re-use the source code	Open	-	Apache License 2.0
FDDSM	eProsimas open source software such as eProsimas Fast DDS.	Yes	Internally by incorporating findings to other eProsimas products (such as AML-IP) or externally by any other user / developer / company that wants to re-use the source code	Open	-	GPLv3
AMLDDEVEL	Marcelle Toolkit	Yes	Internally by incorporating findings to other eProsimas products or externally by any other user / developer / company that wants to re-use the source code	Open	-	Apache License 2.0
AMLUSER	Marcelle Toolkit (owned by INRIA)	Yes	Externally by any other user / developer / company that wants to re-use the source code	Open	-	Apache License 2.0
AMLTTOOLS	-	Yes	Externally by researchers, developers, and industry professionals who want to implement AML models in their applications	Open	-	Apache License 2.0
WP7FOLDING	See https://github.com/roboticslab-uc3m/gymnasium-agma readme files	Yes	Intended for supervised learning, 10 000 png images with ground truth labels for training and cross-validation, and 10 png images with ground truth labels for additional testing and/or visualization.	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

Algebraic Machine Learning (AML) is a relatively new machine learning technique based on algebraic representations of data. Unlike statistical learning, AML algorithms are robust regarding the statistical properties of the data and are parameter-free. The aim of the EU-funded ALMA project is to leverage AML properties to develop a new generation of interactive, human-centric machine learning systems. These systems are expected to reduce bias and prevent discrimination, remember what they know when they are taught something new, facilitate trust and reliability and integrate complex ethical constraints into human-artificial intelligence systems. Furthermore, they are expected to promote distributed, collaborative learning.

Title: ALMA: Human Centric Algebraic Machine Learning
 Acronym: ALMA
 Coordinator: Proyectos y Sistemas de Mantenimiento SL (eProsima)
 Grant agreement: 952091
 Call: H2020-EIC-FETPROACT-2019
 Program: Horizon 2020
 Start: 01 September 2020
 Duration: 48 months
 Website: <https://alma-ai.eu>
 E-mail: alma@eprosima.com
 Consortium: **Proyectos y Sistemas de Mantenimiento SL (eProsima)**
 (EPROS), Spain
 Champalimaud Foundation (CF), Portugal
 ALGEBRAIC AI (ALG), Spain
 German Research Center for Artificial Intelligence
 (DFKI), Germany
 Inria (Inria), France
 TU Kaiserslautern (TUK), Germany
 FIWARE Foundation e.V. (FF), Germany
 Technical Research Centre of Finland (VTT), Finland
 Carlos III University (UC3M), Spain



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952091.



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 2020 research and innovation programme under grant agreement No 952091.



Detailed Consent

I hereby confirm my participation in Project Title pilot testing to be held in the context of the ALMA project.

I have been informed by the head of research about the purpose of ALMA 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 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 the 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 2020 research and innovation programme under grant agreement No 952091.

References

- [1] 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).
- [2] 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>.
- [3] 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>.
- [4] 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.
- [5] European Commission. *Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020*. Mar. 2017. URL: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf (visited on Feb. 11, 2021).