A Model-driven development framework for highly Parallel and EneRgy-Efficient computation supporting multi-criteria optimisation

D8.3 Data Management Plan (DMP)

Version 0.2

Documentation Information

Contract Number	871669
Project Website	www.ampere-euproject.eu
Contractual Deadline	30.06.2020
Dissemination Level	PU
Nature	R
Author	Olivera Vujatovic (BSC) and Sara Royuela (BSC)
Contributors	Eduardo Quiñones (BSC), Nadia Tonello (BSC) and Dirk Ziegenbein (BOSCH)
Reviewer	Tommaso Cucinotta (SSSA) and Alessandro Biondi (SSSA)
Keywords	data management, datasets, accessibility

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871669.

Change Log

Version	Description Change
V0.1	First draft by Olivera Vujatovic and Sara Royuela
V0.2	Final version after deliverable review

Table of Contents

1. Executive Summary

This deliverable presents the Data Management Plan (DMP) of the AMPERE project, which describes the data management life-cycle for all datasets to be collected, processed and/or generated along the lifetime of the project.

Concretely, this deliverable describes, among others:

- Which datasets will be generated, collected and processed, considering both, the development and execution of the AMPERE application use-cases and the research activities towards the development of the AMPERE technology.
- Which methodology and standards will be applied to datasets.
- How datasets will be stored and handled during the lifetime of the project, and after the end of it.
- How the datasets will be made (openly) accessible.

2. Datasets

AMPERE is developing a novel software architecture to help the development of complex Cyber-Physical Systems of Systems (CPSoS). To do so, AMPERE will develop a new generation of programming environments and toolboxes for low energy and highly parallel computing, capable of implementing correct-by-construction CPSoS, in which the constraints captured by the system model are efficiently transformed into the parallel programming models supported by the underlying parallel architecture, and so providing the level of performance required.

The capabilities developed in the AMPERE project will be demonstrated through the envisaged use case applications, which are highlighted next (further details on the use-cases will be possible to find in deliverable D1.1 of AMPERE project):

- *Intelligent Predictive Cruise Control (PCC)*: this use-case, provided by BOSCH and framed in the automotive domain, consists in extending an Adaptive Cruise Control (ACC) system for cars with data from the *electronic horizon* (e.g., topographical data like curvature, inclines, or speed limits) as well as the *front video camera* to improve fuel efficiency. An overview of this use case is provided in Figure 1.

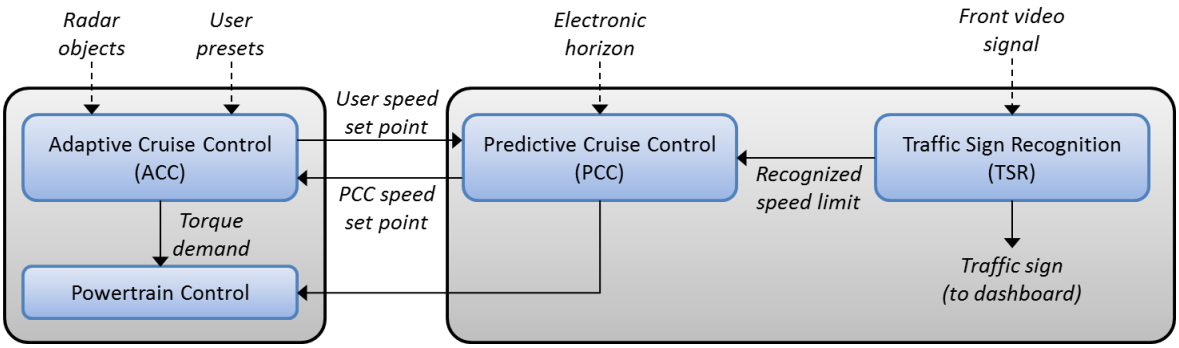

Figure 1. Predictive Cruise Control (PCC) use case overview.

- *Obstacle Detection and Avoidance System (ODAS)*: this use-case, provided by Thales Italy and framed in the railway domain, consists in extending an Advanced Driver-Assistance Systems (ADAS) functionality (i.e., obstacle detection and collision avoidance) based on data fusion from the tram vehicle sensors and artificial intelligence (AI) analytics. An overview of this use-case is provided in Figure 2.

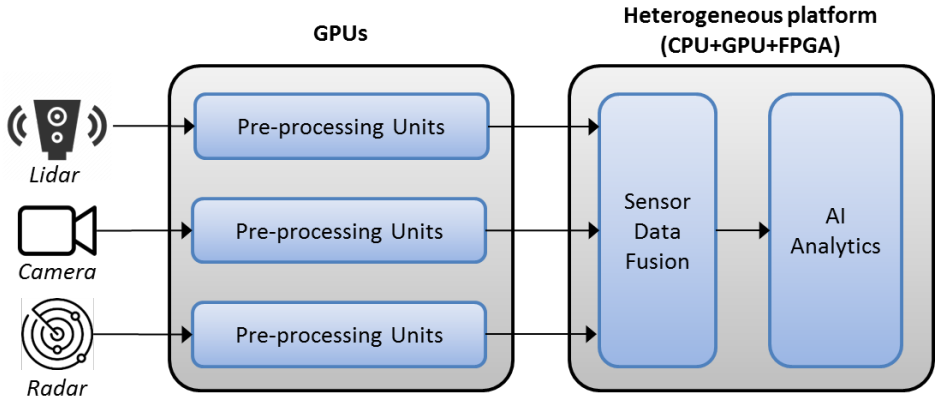

Figure 2. Obstacle Detection and Avoidance System (ODAS) use case overview.

The use-case applications introduced above will provide two kinds of data:

1. *Sensor outputs*: this data corresponds to the measurements performed via different sensors (i.e., cameras, lidars, and radars) and is not generated in real time. The use-cases will not be tested in real conditions, but rather a simulation environment will be used. Instead of capturing data from the environment in real-time, the simulation of the environment will be accomplished by using data previously generated and adequately anonymized. In case the partner responsible for the use-case decides to execute the experiments with data-sets generated in real-time, they will be conducted in restricted (and private) areas. In both cases, the responsible partner will handle the data in compliance with the EU General Data Protection and Regulation (GDPR).
2. *Specific application outputs*: this data corresponds to the information generated by the components in charge of each use case (i.e., decisions about the optimal speed in the PPC use-case, and decisions about obstacles causing collision in the ODAS-use case).

AMPERE will generate/utilize the following types of datasets:

1. *Datasets collected from sensors*. This data will be generated through simulation or will be previously recorded and appropriately anonymized to be used without any ethical concern in the AMPERE project.
2. *Datasets generated to evaluate performance, including the fulfillment of non-functional requirements, of the AMPERE software ecosystem*. The collection of this kind of data has the objective of comparing the evolution of the developments in AMPERE with the capabilities in current production systems. Public data will never include road/rail track areas videos, as they might have an impact on privacy. Performance data will be collected as average and maximum observed execution times, energy consumption values, and other derived metrics such as speed-up, worst-case response time, and GFlops/Watt, among others. Overall, this data will be generated from the execution of application benchmarks and application uses-cases. The resulting data will be relevant for researchers working on similar approaches.
3. *Datasets generated from the execution of the data analytics methods implemented by the AMPERE application use-cases*. This information will depend on the application use-case (either PCC or ODAS). Since the technologies developed are conceived to inform people (tram drivers and car drivers) about potential problems or optimal speed, but do not imply any actual decision (e.g., braking the car or the tram vehicle), no ethical constrains will arise, and the data can be distributed under GDPR.

The AMPERE project will also manage the personal data from the partners of the consortium under GDPR [1]. These data (e-mail addresses and contact details) will be used only for the project communication purposes.

3. FAIR (Findable, Accessible, Interoperable and Re-usable) data

The AMPERE project will apply the FAIR Data Principles [2] in order to provide guidance for scientific data management and stewardship, and also with the objective of promoting maximum use of the research data produced in the frame of the project. FAIR is a set of guidance principles of great importance to stakeholders because the help in accomplishing the next requirements regarding data:

- *Findable*: data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.
- *Accessible*: metadata and data are understandable to humans and machines, and data is deposited in a trusted repository.
- *Interoperable*: Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- *Reusable*: data and collections have a clear usage licenses and provide accurate information on provenance.

The next subsections describe how the AMPERE project will address the fulfillment of these principles.

3.1. Making data Findable (including provisions for metadata)

Given the huge amount of data expected to be generated/utilized by the AMPERE application use-cases, only those results that may be relevant and helpful for a more comprehensive and thorough understanding of the AMPERE software architecture will be accessible to the community through the project publications and the project data repository.

Concretely, AMPERE aims at applying an open-data approach to the following types of datasets, upon which a unique *Digital Object Identifier (DOI)* will be assigned:

1. The source code of those software components and tools licensed as open-source for a complete list of components in the AMPERE. Note that this applies to the basic elements of the AMPERE software architecture, while the developed source code, for any of the two use-case applications, will not be open source.
2. The datasets generated from the execution of the two use-cases and related to the performance evaluation of them. Performance data may include average and maximum observed execution times, energy consumptions, speed-up, worst-case response time, and GFlops/Watt, among others. This information will be included in reports, whitepapers, deliverables, research papers, etc.

Overall, these datasets have a great value to be utilized for evaluating the developed applications and thus the native, utilized AMPERE software architecture.

The performance data, as gathered from the evaluation of AMPERE for evaluation purposes, will be included within publications and scientific papers describing the features and innovations of AMPERE.

3.2. Making data openly accessible

The open-data identified in Section 3.1, with all the specified limitations, will be made accessible as follows:

1. The source code of the AMPERE software components licensed as open-source will be included in the Git repository suitably created to contain a complete and integrated version of the AMPERE software development ecosystem. Some components of AMPERE have already their own versioning system not integrated in Git (e.g., AMALTHEA [3]). To keep track of the versions used in these cases, the git repository will include links to the official releases, or will create a Git tag with

a copy of the corresponding source code. Finally, for software components already using Git for version control, the AMPERE Git repository will include a fork or a sub-module of the original repository.

2. The historicized datasets used for the evaluation of the use-cases will also be stored in the Git repository of the AMPERE project. The data obtained through sensors that could potentially include limited access information will be anonymized before it is used for the AMPERE use-cases. AMPERE might limit or even deny the access to real-time data on restricted areas. In this case, pictures and videos from employees and/or identifiable car vehicles/motorcycles might appear. In this case, the data access might be restricted/denied if somehow impairing the GDPR for managing personal data, as explained in https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN.

To facilitate the access to this data, the public project website will include documentation describing how to access the AMPERE datasets and how to download and use it in full or in specific parts.

3.3. Making data interoperable

The use of metadata standards to access the data is still under discussion between the consortium members. Among others, the Metadata Standards Directory [4] provided by the Research Data Alliance is being considered.

No specific data format will be provided to the datasets needed to evaluate the performance of the AMPERE project due to the small size. This information will be included in scientific documents to properly determine the advances on the AMPERE technology capabilities. Anyhow, all the data will be saved as produced from the corresponding sensor/device, equally those resulting from the application elaboration. However, those data are stored and described (through formerly cited metadata) utilizing largely known formats and contents, and thus making them easily reusable.

3.4. Increase data Re-use (through clarifying licenses)

The performance evaluation and the historicized open-data datasets generated by the applications will be licensed under Creative Commons [5] to allow for the widest possible reusage, since this license allows both commercial and non-commercial use of the data without any restriction. There will be no embargo on the data.

However, their usage will be constrained to mere scientific and research investigations

4. Allocation of resources

There is no additional cost for making the AMPERE datasets:

- The source code of the open-source software components and tools that will form the AMPERE ecosystem will be included in repositories by each owner. The Git repository including the integrated version of the AMPERE will be covered with BSC resources if needed.
- The performance evaluation datasets will be maintained at BSC facilities and included in publications.

5. Data security

The datasets collected or generated by the AMPERE project does not require to apply any data security policies. Any data previously collected, including any personal or private data that could be considered sensitive to be protected (e.g., video and images of citizens like pedestrians, tram passengers, and

motorcycles' drivers, as well as recognized automobiles through identifiable car plates) will be anonymized adequately by the contributing party before it is used in the project.

6. Ethical aspects

Since the AMPERE project will not collect real-time data, and previously collected data will be anonymized before it is used in the project, there are no ethical aspects to be addressed. In case the partners responsible of the use-case decide to conduct experiments with data-sets generated in real-time, they will be executed on restricted (and private) areas of the respective companies. In this case, each partner will guarantee the privacy on the company employees.

7. Acronyms and Abbreviations

- ADAS – Advanced Driving Assistant System
- AI – Artificial Intelligence
- CPSoS – Cyber Physical System of Systems
- CPU – Central Processing Unit
- DMP – Data Management Plan
- DOI – Digital Object Identifier
- FPGA – Field Programmable Gate Array
- GPU – Graphics Processing Unit
- GDPR – General Data Protection and Regulation
- GFLOPS – Giga FLoating Point Operations Per Second
- NGAP – Next Generation Autonomous Positioning
- ODAS – Obstacle Detection and Avoidance System
- ORDP – Open Research Data Pilot
- PCC – Predictive Cruise Control

8. References

[1] P. Voigt and A. Von dem Bussche, "The EU general data protection regulation (GDPR)," A Practical Guide, 1st Ed., Cham: Springer International Publishing, 2017.

[2] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. a. S. L. B. d. S. a. B. P. E. Boiten and others, "The FAIR Guiding Principles for scientific data management and stewardship (vol 15, 160018, 2016)," *Scientific data,* vol. 6, 2019.

[3] C. Wolff, L. Krawczyk, R. Höttger, C. Brink, U. Lauschner, D. Fruhner, E. Kamsties and B. Igel, "AMALTHEA—Tailoring tools to projects in automotive software development," *8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications,* vol. 2, pp. 515-520, 2015.

[4] A. Ball, S. Chen, J. Greenberg, C. Perez, K. Jeffery and R. Koskela, "Building a disciplinary metadata standards directory," *International Journal of Digital Curation,* vol. 9, no. 1, pp. 142-151, 2014.

[5] Creative Commons Corporation, "Creative Commons," 2020. [Online]. Available: https://creativecommons.org/.