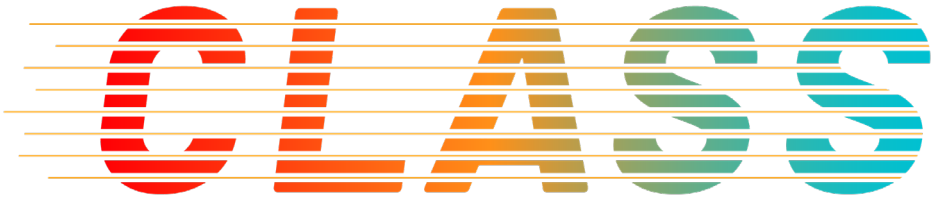

D7.2– Data Management Plan

Version 2.0

Document Information

Contract Number	780622
Project Website	https://class-project.eu/
Contractual Deadline	M6, 30th June 2018. Update M42, 30th June 2021.
Dissemination Level	PU
Nature	ORDP
Author(s)	MODENA, BSC
Contributor(s)	UNIMORE, BSC, MODENA
Reviewer(s)	UNIMORE
Keywords	datasets, open access, FAIR data

Notices: *The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No "780622".*

Change Log

Version	Author	Description of Change
0.1	Luca Chiantore (MOD)	Initial Draft
0.2	Roberto Cavicchioli (UNIMORE)	Contributions Use-Case Data / internal review
0.3	Guadalupe Moreno / Isabel García (BSC)	Internal review
0.4	Eduardo Quiñones (BSC)	Contribution SA / internal review
0.5	Guadalupe Moreno / Isabel García (BSC)	Final internal review
1.0	Eduardo Quiñones (BSC)	Final version ready for EC evaluation
1.1	Elli Kartsakli (BSC)	BSC updated version review
1.2	Luca Chiantore (MOD)	MOD updated version review
1.3	Roberto Cavicchioli (UNIMORE)	Internal review
2.0	BSC	Updated version D7.2

Table of contents

Executive Summary

This deliverable presents the data management plan (DMP) of the CLASS project, which describes the data management life-cycle for all datasets collected, processed and/or generated along the lifetime of the project. Concretely, this deliverable describes:

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

1 Data Summary

CLASS has developed a novel software architecture (CLASS SA for short) to help big-data developers to efficiently distribute big-data workflows along the compute continuum (from edge to cloud) in a complete and transparent way, while providing sound real-time guarantees. To do so, CLASS is adopting: (1) innovative distributed architectures from the high- performance domain; (2) timing analysis methods and energy efficient parallel architectures from the embedded domain; and (3) data analytics platforms and programming models from the big-data domain.

The capabilities of the CLASS SA have been demonstrated on a real smart-city use case, featuring a heavy sensor infrastructure to collect real-time data across a wide urban area, named *Modena Automotive Advanced Area* (or MASA for short), and prototype connected vehicles equipped with heterogeneous sensors/actuators and computing and connectivity capabilities.

CLASS has generate four main types of datasets:

1. **The source code** of the software components and tools that form the CLASS SA.
2. **Datasets generated to evaluate performance and real-time capabilities** of the CLASS SA, with the objective of comparing the evolution of the developments in CLASS SA applied to the smart city domain. Performance data have been collected as average and maximum observed execution time, energy consumption and other metrics derived such as speedup, execution time variability, etc. This data has been generated from the execution of application benchmarks and application use-cases, and the results have been analyzed in the final evaluation deliverables of the technical Work Packages (i.e, D1.6 [1], D2.8 [2], D3.6 [3], D4.6 [4] and D5.5 [5]). This data will be useful for researchers working on similar approaches in big-data.
3. **Datasets collected from the sensors located in the MASA and connected vehicles**, and processed by the CLASS SA (and upon which anonymization mechanisms are applied to guarantee the privacy of Modena citizens). In particular, this information includes the dynamics characteristics of traffic and pedestrian flows, including its bounding box, GPS position, speed and object orientation.
4. **Datasets generated from the execution of the data analytics** methods implemented by the CLASS application use-cases, in particular:
   - *Collision Detection Application*. It alerts drivers to general objects and vulnerable road users that may cross the driving path. City cameras and vehicle sensors data are processed and fused in real-time in order to detect critical situations that may endanger the safety of the driver and of Vulnerable Road Users (VRUs). This dataset contains information about potentially hazardous situations to alert drivers about vulnerable road users that may cross the driving path.

- The *air pollution application* uses the data coming from the distributed sensor infrastructure of the MASA to estimate the pollution emissions of current traffic conditions in real-time. Emissions are computed using the PHEMLight + COPERT V emission model, a simplified version of Passenger car and Heavy duty Emission Model (PHEM), which generates instantaneous fuel consumption and emission factors based on the vehicle engine power. The emissions are interpolated from emission curves containing the normalized engine's power output and vehicle data to obtain the emission and fuel consumption values. The emission and fuel consumption files contain data for the whole range of normalized power demands, combining several vehicles and emission behaviors into one average vehicle per PHEMlight emission class. The output obtained estimates vehicle fuel consumption and emissions of NOx, PM, CO, HC and NO[1] at a time resolution of 1Hz, for each vehicle and road segment.

Moreover, regarding the data collected and anonymized from MASA and vehicle sensors (point number 2 or previous list), different methods of data aggregation have been applied to further guarantee the privacy of Modena citizens and fulfil Italian regulation (CLASS Deliverable D1.1 [6] provides further details on the aggregation rules):

- *Disaggregated datasets.* The data collected and anonymized from MASA and vehicle sensors.
- *First-aggregation dataset.* Data generated by aggregating the information coming from the disaggregated datasets.
- *Historicized dataset.* A second data elaboration process is applied to the first aggregation dataset to obtain a coarse-grain granularity information level.

The CLASS project has also managed the personal data from the partners of the consortium as stated in D8.2 [7] under GDPR. Therefore, in this document we will not make references to this type of data.

2 Data availability and FAIR data

During the development of the CLASS SA and the execution of the CLASS use cases, a huge amount of data has been generated. However, due to a number of factors, the value and usability of these datasets by the public and interested stakeholders (such as automotive and smart city sectors) has been significantly reduced:

The data model employed in CLASS for the generation and aggregation of data generated in the use cases has been continuously evolving throughout the project's lifetime, thus impeding the collection of consistent output datasets. These changes have been driven by the need to achieve the real-time performance constraints during the execution of the use cases at MASA, as well as some difficulties encountered during the integration of all the CLASS SA components.

[1] NO: Nitrogen Oxide; CO: Carbon monOxid; PM: Particulate Matter; HC: Hydrogen and Carbon

During the lifetime of the project, it has not been feasible to execute the use case scenarios for a sufficiently long period of time, which would enable the collection of meaningful historical data. The main reasons behind this have been:

1. Difficulties in integration of the SA components (including those derived from the pandemic), resulting in higher instabilities of the CLASS SA prototype than expected.
2. Instabilities in the Modena/MASA infrastructure, which prohibited the stable execution of the use cases. In particular, the available computing infrastructure has been shared with other services, which often interfered significantly with the performance of the CLASS analytics.
3. Instability in the communications, especially between the fog nodes and the Modena data center, which impacted the execution of the real-time analytics and the collection of data.

Taking into account the above reasons, the CLASS consortium decided not to make the *datasets collected from the sensors located in the MASA and connected vehicles* and the *datasets generated from the execution of the CLASS use cases* publicly available, as originally planned.

In the remaining of this section, the efforts to ensure FAIR data are discussed.

2.1 Making data Findable (including provisions for metadata)

The knowledge generated by CLASS will be accessible to the community through the project publications and the project repository, with links and updates provided in the CLASS website.

As explained in the previous section, the datasets available throughout the lifetime of CLASS will not be made publically available. However, the possibility of generating meaningful datasets using the CLASS SA beyond the lifetime of the project is not discarded, considering that the CLASS SA is deployed and operational at the City of Modena. In that regard, it is worth mentioning the ongoing collaborations between members of the Traffair and CLASS projects regarding the *air pollutant application*, in which the data generated will be evaluated with the air quality models developed within Traffair.

The performance data from the evaluation of the CLASS SA for evaluation purposes will be included within publications and scientific papers describing the features and innovations of the CLASS SA.

2.2 Making data openly accessible

The open-data identified in Section 3.1 will be made accessible as follows:

The source-code of CLASS software components licensed as open-source will be included in a Git repository. In fact, most of the components are already Git projects,

e.g., COMPSs[2], OpenWhisk[3]. Moreover, a new Git project has been created[4], including a complete integrated version of the CLASS software development ecosystem (see CLASS Deliverable D2.7 [8], where the deployment of the CLASS SA is also described). For such a purpose, Git *submodules* are used to link the integrated version with the corresponding Git projects of each CLASS software component.

If any datasets are generated, platforms such as Zenodo[5] will be used. In this case, the Modena City Council has created a well-defined protocol in which the data will be kept for at least three years after its generation. After this period of time we consider that the data might not have value anymore, as results might be super seeded by new datasets obtained from future developments.

2.3 Making data interoperable

CLASS has also evaluated the possibility of aligning the datasets generated by the CLASS use-cases with the FIWARE open initiative, but this has not been possible to achieve within the lifetime of the project. However, FIWARE has been a point of reference at the time of designing the analytics data model, and the adaptation of the model to achieve compatibility with FIWARE is viable in the future if needed.

The definition of the CLASS data model has been guided by performance objectives, to optimize the data update, access and removal. This information has been included in Deliverable D1.6 [1].

2.4 Increase data Re-use (through clarifying licenses)

The performance evaluation, and any historicized and application's generated datasets will be licensed under *Creative Commons open-data license* to ensure the widest possible level of reuse, since this license allows both commercial and non-commercial use of the data without any restriction.

Acronyms and Abbreviations

- MASA – Modena Automotive Advanced Area
- SA – Software Architecture
- VRU – Vulnerable Road Users

[2] https://github.com/bsc-wdc/compss

[3] https://github.com/apache/incubator-openwhisk

[4] https://github.com/class-euproject

[5] https://zenodo.org/

References

[1] CLASS, "D1.6 - Use case evaluation," June 2021.

[2] CLASS, "D2.8 - Evaluation of the CLASS Software Architecture," June 2021.

[3] CLASS, "D3.6 - Validation of the CLASS egde computing subsystem," June 2021.

[4] CLASS, "D4.6 - Validation of the Cloud Data Analutics Service Management and Scalability Components," June 2021.

[5] CLASS, "D5.5 - Evaluation of teh data analytics platform," June 2021.

[6] CLASS, "D1.1 - Use case requirement specification and definition and first description of the sensing and datasets collected," June 2018.

[7] CLASS, "D8.2 - GEN- Requirement No.3," June 2018.

[8] CLASS, "D2.7 - Final Release of the CLASS Software Architecture," June 2021.