D7.2 – Data Management Plan

Version 1.0

Document Information

Contract Number	780622
Project Website	https://class-project.eu/
Contractual Deadline	30th June 2018
Dissemination Level	CO
Nature	R
Author(s)	Luca Chiantore (MOD), Eduardo Quiñones (BSC)
Contributor(s)	Roberto Cavicchioli (UNIMORE), Guadalupe Moreno (BSC), Isabel García (BSC)
Reviewer(s)	Guadalupe Moreno (BSC), Isabel García (BSC)

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 (BSC)	Initial Draft
0.2	Roberto Cavicchioli (UNIMORE)	Contributions Use-Case Data / internal review
0.3	Guadalupe Moreno (BSC) / Isabel García	Internal review
0.4	Eduardo Quiñones (BSC)	Contribution SA / internal review
0.5	Guadalupe Moreno (BSC) / Isabel García	Final internal review
1.0	Eduardo Quiñones (BSC)	Final version ready for EC evaluation

Table of contents

1 Executive Summary

This deliverable presents the data management plan (DMP) of the CLASS 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 CLASS application use-cases and the research activities towards the development of the CLASS technology.
- Which methodology and standards will be applied to CLASS 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 Data Summary

CLASS is developing a novel software architecture (CLASS SA for short) to help big data developers to efficiently distribute big-data workloads 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 will be 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 will generate four main types of datasets:

1. The source code of the software components and tools that will 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 will be collected as average and maximum observed execution time, energy consumption and other metrics derived such as speedup, worst-case response time, GFlops/Watt, etc. This data will be generated from the execution of application benchmarks and application use-cases. The result 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 will be applied to guarantee the privacy of Modena citizens). In particular, this information will be:

    - Traffic and pedestrian flows,
    - available parking slots,
    - pollution level and
    - any event that can impact on traffic conditions.

    This information will be compiled at strategic points of the MASA within specific periods of times (see CLASS Deliverable D1.1 for further information) and used to implement the advanced software services of the CLASS application use-cases.

4. Datasets generated from the execution of the data analytics methods implemented by the CLASS application use-cases. This information will depend on the application use-case. In particular:

   - *Digital Traffic Sign Application*. It will offer the opportunity to change traffic conditions dynamically, according to real-time traffic information collected by the sensor infrastructure. In case of accidents, traffic signals will advise the "best path to follow", reducing the induced traffic impact and improving the driver experience. It will dynamically create "green routes" for emergency vehicles (e.g., ambulances, fire-fighters and police vehicles) as well, adjusting the frequency of the traffic lights to reduce the time of intervention.
     Hence, information about: (1) the impact of dynamically changing the traffic conditions based on real-time information; and (2) "green routes" for emergency vehicles will be collected.

   - *Smart Parking Application.* It will provide information about the available parking spots in the monitored area. Detection of free parking lots will be based on the City IOT sensors and the smart car prototypes, equipped with sensors and Vehicle-to-Infrastructure capabilities. The application will elaborate a consistent real time representation of the available parking slots in the area on a map.
     Hence, information about the usage of parking resources in the MASA will be collected.

   - *Obstacle Detection Application*. it will alert drivers to general objects and vulnerable road users that may cross the driving path. City cameras and sensors data will be elaborated in real-time in order to detect critical situations that may endanger the safety of the driver and of VRU. The identification of potentially hazardous situations will be enforced at the different levels of the compute continuum, with a different precision and latency. Edge-side embedded computing platforms will pre-elaborate potential risks using low-latency/low-precision algorithms for screening the frames to send to cluster and data center levels. Advanced sensor fusion capabilities enforced at higher levels will then be able to cover in detail the hazardous situations not previously identified by the edge-side sensors.
     Hence, information about potentially hazardous situations to alert drivers about vulnerable road users that may cross the driving path.

Moreover, regarding the data collected and anonymized from MASA and vehicle sensors (point number 2 or previous list), different methods of data aggregation will be applied to further guarantee the privacy of Modena citizens and fulfil Italian regulation (CLASS Deliverable D1.1 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 will also manage the personal data from the partners of the consortium as stated in D8.2 under GDPR. Therefore, in this document we will not make references to this type of data.

3 FAIR data

3.1 Making data Findable (including provisions for metadata)

Given the huge amount of data expected to be generated by the CLASS application use-cases, only those results that may be relevant for the development and execution of smart systems (with special interest of automotive and smart city sectors) will be accessible to the community through the project publications and the project data repository.

Concretely, CLASS aims to apply an open-data approach 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-sources (see CLASS Deliverable D2.1 for a complete list of components in the CLASS SA).
2. The historicized datasets from MASA and vehicle sensors. The City of Modena is still evaluating if the first-level aggregation datasets from the MASA and vehicle sensors will follow an open-source approach. The disaggregated datasets will not be offered as open-data.
3. The dataset generated from the execution of the three use-case.

Overall, these datasets have great value to be reused for statistical analysis of traffic flow, crowd behaviour, pollution control, emergency vehicle routing and so on.

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.

3.2 Making data openly accessible

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

1. 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[1], OpenWhisk[2]. Moreover, a new Git project will be created, including a complete integrated version of the CLASS software development ecosystem (see CLASS Deliverable D2.1). For such a purpose, Git *submodules* will be used to link the integrated version with the corresponding Git projects of each CLASS software component.

2. The historicized datasets from MASA and vehicle sensors, and datasets generated from the execution of the use-cases will be conserved in the City of Modena facilities and will be made openly available. The consortium is still evaluating the most suitable open-data infrastructure required. Platforms such as Zenodo[3] are also being evaluated. In case the consortium decides to offer the first-level aggregation datasets as open-data, the same infrastructure will be used.

   The Modena City Council will keep the data for at least three years after the end of the project. 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.

[1] https://github.com/bsc-wdc/compss
[2] https://github.com/apache/incubator-openwhisk
[3] zenodo.org

To facilitate the access to this data, the public project website will include documentation describing how to access the CLASS 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.

CLASS is also evaluating the impact of using the data models offered by the FIWARE open initiative to facilitate the readability and interoperability of the datasets generated by the MASA and vehicle sensors, and the CLASS use-cases. Concretely, the first analyses have shown interest on the *Alert, Parking, Transportation* and *Environment* data models (see CLASS Deliverable D1.1 for further information).

No specific data format will be provided to the datasets needed to evaluate the performance of the CLASS SA due to the small size. This information will be included in scientific documents to properly determine the advances on the CLASS technology capabilities.

3.4 Increase data De-use (through clarifying licenses)

The performance evaluation, historicized and application's generated datasets open-data will be licensed under *Creative Commons* to let the widest reuse possible of it, since this licence allows both commercial and non-commercial use of the data without any restriction. There will be no embargo on the data.

4 Allocation of resources

There is no additional cost for making the CLASS datasets identified in Section 2 FAIR:

- The source code of the open-source software components and tools that will form the CLASS SA will be included in GitHub by each owner (see CLASS Deliverable D2.1 for the corresponding owners). The GitHub including the integrated version of the CLASS SA will be covered with BSC resources if needed.
- The performance evaluation datasets will be maintained at BSC facilities and included in publications.
- The rest of the open-data will be stored within the City of Modena facilities for at most three years after the end of the project. The City of Modena will have the full responsibility of storing the metadata, maintaining it available and sharable. The infrastructure and personnel funds granted from the European Community will cover the storage, hardware and staff time to manage the servers on which the data will be stored.

5 Data security

The datasets collected or generated by the CLASS project does not require to apply any data security policies, that the data do not include any personal or private data that could be considered sensitive to be protected. Regular backups for keeping the information safe will be used.

The non-aggregated datasets collected by MASA and vehicle sensors will be stored in a secure *Storage Area Network* (SAN) belonging to the City of Modena in which a private *Virtual*

[4] http://rd-alliance.github.io/metadata-directory/

Protected Network (VPN) access to the City servers is allowed by authorised personnel. If the consortium decides not to offer first-aggregated datasets openly, they will be stored in the same SAN.

6 Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

The Ethical aspects treated along the project about the collection of data from the citizens of the City of Modena (including video recording) needed to develop and execute the CLASS use-case and GDPR are developed and described in CLASS Deliverables D8.1 and D8.2, respectively.