D6.2 "DATA MANAGEMENT PLAN"

Version 1.7

Document Information

Contract Number	780681
Project Website	https://legato-project.eu/
Contractual Deadline	31/05/2018
Dissemination Level	PU
Nature	ORDP: Open Research Data Pilot
Author	Sergi Madonar (BSC)
Contributors	Xavier Martorell (BSC), Raúl de la Cruz (BSC), Miquel Pericas (CHALMERS), Nils Kucza (UNIBI), Gunnar Billung-Meyer (CHR)
Reviewers	Valerio Schiavoni (UNINE), Christof Fetzer (TUD)

The LEGaTO project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No 780681

Change Log

Version	Description of Change
V 1.0	Initial draft for internal review
V1.1	Contributions from OmpSs and Smart City use case added
V1.2	Contributions from Smart Home use case added
V1.3	Contributions from Secure IoT gateway added
V1.4	Review of the document
V1.5	Revision of the document for Interim Review comments
V1.6	Internal review of the document
V1.7	Final version including changes requested in the Periodic Review.

Comments

This deliverable shall be updated if new data is generated or used in the context of the project or different processes have to be applied to manage the data.

Index

1. Executive Summary 4
2. Data Summary 4
3. FAIR Data 6
   3.1. Making data findable, including provisions for metadata 6
   3.2. Making data openly accessible 7
   3.3. Making data interoperable 7
   3.4. Increase data re-use (through clarifying licences) 7
4. Allocation of resources 7
5. Data security 8
6. Ethical aspects 8
7. Further support in developing your DMP 8

1. Executive Summary

LEGaTO partners elaborate the Data Management Plan to comply with EC's objective of making research data findable, accessible, interoperable, and reusable (FAIR). The DMP will be updated during the lifespan of the project. The DMP describes the life cycle for all data sets that will be collected, processed, or generated by the project. It is a document outlining how research data will be handled during a research project, and even after the project is completed, describing what data will be collected, processed, or generated, and following which methodology and standards, whether and how this data will be shared and/or made open, and how it will be curated and preserved.

The datasets managed or created in project are:

Dataset		Description
OmpSs apps benchmarking		To quantify the evolution and improvement, data from the performance of OmpSs applications will be generated.
Smart City Use Case	Network status	Real-time measurements of the city network with the objective to validate the capacity of different energy-efficient architectures.
	Wind fields	Wind speed and direction are gathered from sensors in the city.
	Pollution	Concentration values for NO_2 and O_3 ($\mu g/m_3$) obtained from sensors allocated along the city.
Smart Home Use Case		Interaction between sensors and actuators. Some information will be processed and generated in this interaction.
Secure IoT Gateway		Latency and throughput measurements of the different hardware configurations.

2. Data Summary

In LEGaTo project, **performance numbers from OmpSs applications** will be generated with the goal to compare the positive evolution of the developments in OmpSs, and their performance in the various applications ported. It is expected to be a small set of data.

Performance data will be collected as execution time, and energy consumption, and other metrics will be derived, as speedup, and GFlops/Watt. There will not be any previous data used and the data will be generated from the performance of benchmarks and applications.

The result data will be useful for researchers working on similar approaches for programming heterogeneous systems, as it will allow comparisons with their systems.

Within the **Smart City application**, several experiments will be conducted in order to validate the feasibility of energy-efficient architectures (GPUs, FPGAs and DFEs) and their task-based programming paradigm to enhance the urban-scale air quality model of the use case.

These experiments will use both synthetic and input data collected in campaigns by air-quality stakeholders from high density populated areas at Barcelona city. On the other hand, the experiments will produce high-resolution nowcasting maps of wind fields and concentration values

for certain pollutants that will be compared with real data acquired in campaigns to validate the model.

Urban-scale pollutant dispersion models require of two critical inputs: i) high-resolution (tens of meters) near-surface wind fields resulting from air flowing through urban-scale morphologies (buildings) and, ii) sensors and emission inventories used to characterize the pollutant sources (mainly derived from vehicle combustion).

The expected data provided to the community through this use case includes two groups of data sources:

1. Real-time measurements from the monitoring network on a half-hourly basis.
2. High-resolution (10 m) nowcasting maps at any point in the target area on a half-hourly basis; offering the following attributes for each group (real-time and simulated):
    a) Wind fields: wind speed (ms^{-1}) and wind direction (degrees)
    b) Concentration values for NO_2 and O_3 ($\mu g/m3$)

The Smart City use case will use existing data resulting from the H2020 GrowSmarter project. The data used will be: i) the computational mesh domain representing the urban area of interest, ii) and the output data generated at that moment for comparison and validation of the correctness and accuracy of the new implementation for the LEGaTO project.

The used data is provided by a wide spectrum of sources:

- **The computational mesh domain of Barcelona city:** is generated with a BSC in-house meshing software using merged and filtered data such as topography, LiDaR and cadastre coming from national entities offering them as Open Data.
- **The air-quality data campaigns for validation:** are acquired through agreements with governmental agencies devoted to air quality and environmental assessment.
- **The real-time sensing data:** is obtained from monitoring nodes with wind/air quality sensors deployed within the H2020 GrowSmarter project in a neighborhood area of Barcelona city.
- **Initial boundary conditions for the wind field:** are generated after interpolating on-line meteorological data from weather forecast agencies.

The size of the data involved in the Smart City use case may vary depending on the area extension to be simulated by means of the Computational Fluid Dynamics code (Alya). In a real case of a whole city like Barcelona, meshes can easily reach hundreds of millions of elements, which leads to input data files of 5 to 10 GBytes. This size includes computational mesh definition (topology) and initial boundary conditions of the meteorological and particle data (wind field and pollutants emissions). Finally, the output data generated may require up to 10GBytes of space for one month of operational simulation.

The data produced by the air quality model will be an invaluable tool for public administrations and health agencies that are tasked to monitor the evolution of air quality over time and, eventually, also to make use of model forecasts in order to react promptly to possible air pollution episodes affecting vulnerable groups of citizens. In the context of the H2020 GrowSmarter project, the Barcelona city council may use this information to provide prompt warnings to citizens. The other two lighthouse cities of the GrowSmarter project (Koln and Stockholm) have already shown interest in this solution and may incorporate this solution in the following years.

Other target customers of our solutions are public and private entities interested on assessing transport of particulate matter from emission areas such as cargo ports, quarries, cement and chemical factories near (< 5 kilometers) urban areas. Collaterally, all these studies could be used to

elaborate sustainable urban mobility plans and to design future urban plans. Environmental measures at urban-scale is a topic of growing interest not only for public entities but also for private companies that manipulates large quantities of data (e.g. Google: https://environment.google/projects/airview/, PlumeLabs: https://plumelabs.com/en/).

In the **Smart Home application,** the interaction of sensors, information processing, and actuators is examined. For a satisfying users' experience in those environments a high computational power is required. Therefore, a capable local system or cloud service is inevitable. Edge computing or local processing instead of cloud services can mitigate these issues. The data bandwidth required for pure cloud computing is also not negligible. For example, a full HD video stream needs a bandwidth of around 3,5 Mbit/s. In a smart environment many streams and multiple algorithms working on those streams are used and increase required communication bandwidth. Sensor-near signal pre-possessing reduces this data volume.

Within the LEGaTO project, two different aspects of the smart home are investigated individually:

- **Using the LEGaTO tools to optimize frameworks** used within the smart home environment (e.g. face recognition, speech recognition, object recognition). Porting these frameworks to an embedded device for local or edge computing.
- **Anomaly detection and behavior prediction** in smart home environment. Based on raw sensor data, messages send by the unified middleware (robotic service bus – RSB), and development of a situation memory for machine learning or classical statistical methods.

The used frameworks are open source. Each evaluated and optimized framework will be provided in a dedicated repository for open access. For benchmarking and profiling dummy reference data sets are provided in each respective repository. Due to the data privacy law a release of a large data set of pictures and audio recording is not possible. A small sample collection and pretrained graphs are published instead. For speech recognition open source resources like OpenSLR (LibriSpeech) or Common Voice are used as language models.

For the behavior prediction and anomaly detection a flexible and addressable situation memory as a semantic knowledge-base is developed in the LEGaTO project. Therefore, individual-related sensor data are collected. These data sets can be used internally, but not open accessible due to the data privacy law. The evaluation of the model for this situation memory just began, therefore a specification is not finalized. The dataset will include message send over the unified middleware as well as sensor data of smart home devices like thermometers, heater regulators, and window tilt sensors.

For the development of the **Secure IoT gateway** a good performance of the VPN processing is important for some scenarios. Therefore, the throughput and latency of different hardware configurations will be evaluated. The incurring data is non-personal.

3. FAIR Data

3.1. Making data findable, including provisions for metadata

The performance data obtained by the **OmpSs benchmarking** will be of a reduced size (around GigaBytes), so it is not expected to need any standard identification mechanism.

The information obtained within the **Smart City use case**, due to the characteristics of the simulated data, there are no plans to use persistent and unique identifiers such as Digital Object Identifiers to locate the data produced in the Smart City use case.

For the **Smart Home use case**, due to the characteristics of the reference data sets, there are currently no plans to integrate digital object identifiers.

The performance data collected for the **Secure IoT gateway** is of a very limited size and therefore has no need for a standard identification mechanism.

3.2. Making data openly accessible

All input and output data generated in this project may follow the Open Access policy. However, given the huge amount of data that some models could produce, only some specific results will be accessible to the community through the project publications and the repository to demonstrate the validity of the LEGaTO implementation.

On the other hand, the urban-scale air quality model is being developed at the BSC using the in-house Alya multi-physics code. Therefore, to allow access to the final source code, an agreement with the BSC will be required.

The reference data sets and corresponding provided in the **Smart Home use case** do not contain individual-related information and public accessible via the repository together with test use case of the frameworks used in LEGaTO.

3.3. Making data interoperable

For the OmpSs performance data, it will be possible to be compared with performance data obtained from other sources, to better determine the advances on the programming and runtime support for heterogeneous systems.

The simulation data will be published with enough detail in order to allow other scientists to compare their air quality models' results with the ones generated in this project.

Reference data sets of the Smart Home use case (video or audio data) are self-explaining and therefore can be intuitively used by other researchers.

The performance data of the Secure IoT gateway is self-explaining and therefore can be intuitively used.

3.4. Increase data re-use (through clarifying licences)

During the project data re-use options for the data sets will be dealt on a case by case basis, aiming, when possible, to keep them public, accessible, free of charge and re-usable under request, under a GPL-like licence such as ODC Open Database License (ODbL) or inheriting the licence types from the different data sources as explained in the data summary description, and taking care of each partners' business constrains or legal limitations on them.

4. Allocation of resources

There is no additional cost of making our performance data FAIR, as it does not need a special treatment.

Performance data will be under the responsibility of BSC, the coordinator of the project.

Data will be kept for three years after the LEGaTO project. After three years we consider that the data will not have value anymore, as results will be superseeded by new datasets obtained from future developments. Anyway, data will still be present in project publications and public repositories when appropriate.

5. Data security

There is no need for applying data security policies in the project given that the data used and produced do not include any personal or private data that could be considered as sensitive. Regular backups for keeping the information safe will be used.

6. Ethical aspects

No ethical or legal issues are directly devised for the generated data.

However, the use of operational air quality models in urban areas could have ethical implications for public administrations. An operational model may foster the opportunity for public administrations to shorten inequality with respect to air quality breathed by citizens of lower classes in urban areas. It is usual in developing countries that the impoverished communities live in areas with higher pollution levels due to its inherent poverty and their lack to access better located neighborhoods. This air quality model could be used, if they wish, as a tool by city councils and public welfare departments to elaborate urban plans.

In the **Smart Home use case**, reference data sets do not contain individual-related data because it is not generated or stored personal data. Therefore, no ethical or legal issues are directly devised.

7. Further support in developing your DMP

No other procedures for data management for the performance data.