D7.2 Data Management Plan

Version 0.5

Document Information

Contract Number	825473
Project Website	https://elastic-project.eu/
Contractual Deadline	31st May 2019
Dissemination Level	Public
Nature	ORDP
Author(s)	GEST
Contributor(s)	THALIT, FLO
Reviewer(s)	Nadia Tonello (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 "825473". © 2018 ELASTIC Consortium Partners. All rights reserved.*

Change Log

Version	Author	Description of Change
V0.1	Vanessa Fernandez (BSC)	Initial Draft
V0.2	Gaetano ZUMBO (GEST)	Enhanced Draft, partially filled by GEST.
V0.3	Gaetano ZUMBO (GEST)	Further enhancement, due to THALIT contribution
V0.4	Jurgen Assfalg (FLO)	Contributions from FLO
V0.5	Nadia Tonello (BSC)	End-to-end document Revision
V1.0	Vanessa Fernandez (BSC)	Final version
		(Final Change Log entries reserved for releases to the EC)

Table of contents

1 Executive Summary

This deliverable presents the Data Management Plan (DMP) of the ELASTIC 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 ELASTIC application use-cases and the research activities towards the development of the ELASTIC 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

ELASTIC is developing a novel software architecture 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, ELASTIC 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 ELASTIC will be demonstrated through the envisaged use case applications (as reflected in deliverable *D1.1 - Use case requirements definition*):

- Predictive Maintenance (both rail track status and energy power consumption),
- NGAP (Next Generation Autonomous Positioning), ADAS (Advanced Driving Assistant System).
- Interaction between the public and the private transport in the City of Florence (FLO).

Note that, any use case application will provide two kinds of data:

- Sensor output (excluding external road/rail track areas videos that will not be made publicly available, as having potential privacy impact)
- Specific application outputs: in this case, consisting in either data on-the fly and/or data stored on duty and uploaded when tram returns to the depot.

ELASTIC will generate/utilize three main types of datasets:

1. Datasets generated to evaluate performance and real-time capabilities with the objective of comparing the evolution of the developments in ELASTIC applied to the above-cited use case applications as exploited within the Florence Tramway network. From any use case application, the public data will not include external road/rail track areas videos, as having possible privacy impact. 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.

2. Datasets collected from the sensors located either on board of tram vehicle or along the tramway. Those are described in *D1.2 - Initial sensing and datasets collected*. The same previously cited video limitations apply. Also, restrictions apply to data that external third parties (i.e. not belonging to the ELASTIC consortium) provide to FLO, and that FLO is authorized to store and process in the mobility supervisor but is not allowed to distribute and/or publish.

3. Datasets generated from the execution of the data analytics methods implemented by the ELASTIC application use-cases. This information will depend on the application use-case (either NGAP, ADAS, Predictive Maintenance, or Interaction between the public and the private transport in the City of Florence). As for the previous point, restrictions might apply to outputs of data analytics whenever these might reveal data provided to FLO by external third parties (i.e. not belonging to the ELASTIC consortium), and that FLO is authorized to store and process in the mobility supervisor but is not allowed to distribute and/or publish.

The ELASTIC project will also manage the personal data from the partners of the consortium as stated in D8.2 under GDPR (General Data Protection and Regulation). Therefore, in this document we will not make references to this type of data.

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

3.1 Making data Findable (including provisions for metadata)

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

Concretely, ELASTIC aims to apply 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-sources for a complete list of components in the ELASTIC. Note that this applies to the basic elements of the Elastic software architecture, while the developed source code, for any of the three use case applications, will not be open source.
2. The dataset generated from the execution of the three use-case.
3. The dataset collected by the sensors utilized by the three use cases.

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

Upon any dataset transfer/delivery, those will always be associated to a comprehensive description (metadata, comprising date/time stamp and *Digital Object Identifier*), with an associated legenda for explaining the utilized acronyms. Example: *Elect_Pow_<ID>*, where *Elect_Pow* means current Electric Power consumption value, as gathered by the dedicated sensor and relevant input to the Energy Consumption profile of the Predictive maintenance use case application, and <ID> will be an unique and growing identifier associated to the various collected power measurements.

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

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 ELASTIC software components licensed as open-source will be included in a Git repository. Some of the components are already Git projects, e.g., COMPSs[1]. Moreover, a new Git project will be created, including a complete integrated version of the ELASTIC software development ecosystem. For such a purpose, Git submodules will be used to link the integrated version with the corresponding Git projects of each ELASTIC software component.
2. The historicized datasets use-cases are stored on the tram vehicle data storage devices and (or subsequently) transferred onto the FLO/GEST control centers. Successively, comprehensive and coherent datasets (e.g. related to a complete tram route on a given date) are then made available via data transfer or Web accessible. The only data whose access is limited or denied are those impairing citizens' privacy and security, for instance picture and videos where either citizens and/or identifiable car vehicles/motorcycles are shown. In particular, the data access might be restricted/denied if somehow impairing the General Data Protection and Regulation (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 ELASTIC 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 Directory2 provided by the Research Data Alliance is being considered.

[1] https://github.com/bsc-wdc/compss
[2] http://rd-alliance.github.io/metadata-directory/

No specific data format will be provided to the datasets needed to evaluate the performance of the ELASTIC due to the small size. This information will be included in scientific documents to properly determine the advances on the ELASTIC 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, as outlined in deliverable D1.2, 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, 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.

However, their usage will be constrained to mere scientific and research investigations, and subject to privacy constraints (in case they contain info that might impair citizens' privacy and security).

4 Allocation of resources

There is no additional cost for making the ELASTIC datasets, as identified in Section *3 FAIR (*Findable, Accessible, Interoperable and Re-usable) data:

- The source code of the open-source software components and tools that will form the ELASTIC will be included in GitHub by each owner. The GitHub including the integrated version of the ELASTIC 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 largest part of datasets collected or generated by the ELASTIC project does not require to apply any data security policies. However, this excludes those data including any personal or private data that could be considered sensitive to be protected: for instance, video and images of citizens (e.g. pedestrians, tram passengers, motorcycles' drivers), and recognized automobiles (through identifiable car plates). Additional details in this respect are anticipated in next chapter *6 Ethical aspects*. Anyhow, regular backups for keeping the information safe will thus be used.

6 Ethical aspects

The Ethical aspects treated along the project about the collection of data from the citizens needed to develop and execute the ELASTIC use-case and GDPR are developed and described in ELASTIC Deliverables D8.1 and D8.2, respectively.

Acronyms and Abbreviations

ADAS: Advanced Driving Assistant System

DMP: Data Management Plan

DOI: Digital Object Identifier

GDPR: General Data Protection and Regulation

GFLOPS: Giga FLoating Point Operations Per Second

NGAP: Next Generation Autonomous Positioning

ORDP: Open Research Data Pilot