D2.1

Version	1.1
Author	REL
Dissemination	PU
Date	30-06-2017
Status	FINAL

D2.1 Data Management Plan

Project acronym	ELASTEST
Project title	ElasTest: an elastic platform for testing complex distributed large software systems
Project duration	01-01-2017 to 31-12-2019
Project type	H2020-ICT-2016-1. Software Technologies
Project reference	731535
Project website	http://elastest.eu/
Work package	WP2
WP leader	TUB
Deliverable nature	PUBLIC
Lead editor	Christos Roupas
Planned delivery date	30-06-2107
Actual delivery date	30-06-2017
Keywords	Open source software, cloud computing, software engineering, operating systems, computer languages, software design & development

Funded by the European Union

License

This is a public deliverable that is provided to the community under a **Creative Commons Attribution-ShareAlike 4.0 International** License:

http://creativecommons.org/licenses/by-sa/4.0/

You are free to:

Share — copy and redistribute the material in any medium or format.

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

For a full description of the license legal terms, please refer to:

http://creativecommons.org/licenses/by-sa/4.0/legalcode

Contributors

Name	Affiliation
Christos Roupas	RELATIONAL SA
Frances Pavlidis	RELATIONAL SA
Francisco Gortázar	URJC

Version history

Version	Date	Author(s)	Description of changes
V1	14/06/2017	C. Roupas	
V1.1	16/06/2017	F. Gortázar	Minor changes

Table of contents

5

Glossary

AGA	Annotated Model Grant Agreement
CNR	Centre Consiglio Nazionale delle Ricerche
DoA	Document of Acceptance
DMP	Data Management Plan
EC	European Commission
EDM	ElasTest Data Manager
H2020	Horizon 2020, EU funding Strategy for 2014 -2020
Q&A	Questions and Answers
REL	Relational SA
SuT	System under Test

Executive summary

The Data Management Plan (DMP) of the ElasTest will give an overview of the available research data arising from the project, the data accessibility, management and terms of use. The DMP reflects the current state of the discussions, plans and ambitions of the ElasTest partners, and will be updated as work progresses.

As specified in Section 2.2.1.2 on Part B of the DoA, this project requires an Open Data Management Plan and this task is in charge of creating it and monitoring its implementation. This plan defines how to share and spread data generated by the project. It does not only aim to contemplate the mechanisms for sharing ElasTest generated data, but also to enable third parties (i.e. other researchers not part of the project consortium) to provide further testing data that might be helpful for training and improving ElasTest Machine Learning and Cognitive Q&A algorithms.

The current document aims to present the information on how the actual DMP will be put together and what it will include in its next version to be delivered to the EC as of December 2017 when some data from the project will have become available.

Partner roles: REL coordinates this task generating the plan and monitoring its appropriate implementation by the rest of the partners on all data gathering and analysis activities.

Questions addressed:
- What data will be collected / generated in the course of the project?
- What data will be exploited? What data will be shared/made open?
- What standards will be used / how will metadata be generated?
- How will data be curated / preserved including after project completion?

1 Introduction

1.1 Why a Data Management Plan (DMP)?

It is a well-known phenomenon that the amount of data is increasing while the use and re-use of data to derive new scientific findings is more or less stable. This does not imply, that the data currently unused are useless - they can be of great value in future. The prerequisite for meaningful use, re-use or recombination of data is that they are well documented according to accepted and trusted standards. Those standards form a key pillar of science because they enable the recognition of suitable   data.
To ensure this, agreements on standards, quality level and sharing practices have to be negotiated. Strategies have to be fixed to preserve and store the data over a defined period of time in order to   ensure their availability and re-usability after the end of ElasTest.

1.2 What kind of data are considered in the DMP

The main purpose of a Data Management Plan (DMP) is to describe *Research Data* with the metadata attached to make them *discoverable*, *accessible*, *assessable*, *usable beyond the original purpose* and *exchangeable* between researchers.

According to the "Guidelines on Open Access to Scientific Publication and Research Data in Horizon 2020" (2015)*:*

"*Research data* refers to information, in particular facts or numbers, collected to be examined and considered and as a basis for reasoning, discussion, or calculation. In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images. The focus is on research data that is available in digital form."

1.3 What can be expected from ElasTest DMP?

 The ElasTest DMP will describe the lifecycle, review processes and data management policies on research data produced by ElasTest. The DMP reflects the current status of discussion within the consortium about the data that will be produced. It is not a fixed document, but evolves during the lifespan of the project.
The target audience of the DMP is all project members and research institutions using the data produced.

1.4 Data Management Plan Template

The structure of the DMP as of the next version (to be submitted in December 2017 will be) when data from the ElasTest project will be managed, will be according to the specifications set by the "Guidelines on FAIR Data Management in Horizon 2020" of the H2020 Programme.

Including the following sections:
1. Data Summary
2. FAIR Data
   a. Making data finable, including provision for meta data
   b. Making data openly accessible
   c. Making data interoperable
   d. Increase data reuse
3. Allocation of resources
4. Data security
5. Ethical aspects
6. Other

2 Data set

2.1 Description of data

The data to be collected during the life time of the Project will fall into one of two general categories:

- Data, including metadata needed to validate the results presented in scientific publications.
- Other data, including associated metadata which all participants of the project choose to make available in open access mode in addition to the data underlying publications.

Note: This of course should comply with the protection of results (Article 27), confidentiality obligations (Article 36), security obligations (Article 37) and protection of personal data (Article 39) of the H2020 AGA.

2.2 What data will we collect?

Several types of data will be collected and analysed during the research in the Project. This data consists of end-user information obtained through the interaction with the ElasTest system and through the research with end users during the pilot and in other phases of the Project.

Detailed data categories:

- Specification of software requirements, architectures and systems in different formats. They will be published as part of the open data plan because code is open source and component designs will be made available, and it is mandated by the DoA. Specification of tests in different formats. These come

from the different partners. They are useful for validation of architectural decisions.

- Machine readable representation of the software artefacts (e.g. code documentation, code comments, etc.) These come from the GitHub repositories. Each repository includes: code, tests and documentation. Information is used to recommend good practices when developing tests.

- Specification of Q&A provided by testers at different stages of the testing process. These come from CNR after anonymization and are used to validate the outcomes for the project.

- Runtime information of systems (e.g. logs, costs, energy, KPIs, etc.) These come from the EDM component. Useful to estimate costs, train machine-learning algorithms for early failure detection on tests, and build oracles based on the outcomes of those algorithms.

- Machine readable representation of testing knowledge (e.g. papers, books, guidelines, etc.)

- Anonymized information regarding tester and end-user satisfaction and subjective feelings on SuT and tests. These come from CNR after anonymization and are used to validate the outcomes for the project.

- Scientific publications.

2.3 How will the data be managed, collected or created?

For the Management of data produced from the project, an online repository offering specialised tools for managing data will be used. Access to data in the repository will be allowed depending on the specifications set by the partners.

The repository of choice for ElasTest is Zenodo by OpenAIRE and CERN. The characteristics of this repository which led to it being chosen, were its large capacity for data storage and management as it's maintained by the CERN as well as its guarantee that even in the event that the CERN no longer supported the repository then the data would be transferred to some other appropriate facility. Finally, this repository is equipped with tools for the management of the accessibility and licensing of the data, therefore allowing for a variety of different licenses and access levels.

2.4 What documentation and metadata will accompany the data?

The documentation that will accompany the data will be coherent with the Commission Recommendation of 23rd June 2009 number 498. Thus, data will contain information about:

- Contacts: Individual or organizational contact details of persons in charge of data management.

- Metadata update: information indicating the date of last version of data (or modifications occurred).

- Statistical presentation indicating: basic description of data, sector of

coverage, statistical unit for which data have been collected, statistical population, reference area, time coverage.

- Unit of measure.
- Reference period.
- Project funding information: European Union logo and information about Grant Agreement and the action/program that funds the project.
- Institutional mandate including procedures for data sharing and coordination between data producing agencies.
- Confidentiality: property of data indicating the extent to which their unauthorized disclosure could be prejudicial or harmful to the interest of the source or other relevant parties.
- Release policy including dissemination rules and purposes.
- Information about data collection (source, frequency and adjustments)
- Comments

2.5 How will we manage copyright and Intellectual Property Rights (IPR) issues?

The IPR ownership is defined by the Consortium Agreement and Grant Agreement related to the Project. Such access will be provided by accepting the terms and conditions of use, as appropriate. Materials generated under the Project will be disseminated in accordance with Consortium Agreement. Those that use the data (as opposed to any resulting manuscripts) shall cite it as follows: The data created by the ElasTest project, funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 643566. For reuse of this data, please, contact ElasTest Consortium.

2.6 How will the data be stored and backed up during the research?

The data will be managed by the consortium members, as well as, other scientists interested in the ElasTest project. The knowledge generated by the Project among partners, scientific community, target users and public at large during the Project are managed in two ways, depending on the data source:

1. The non-sensitive data will be organized into a repository that will contain all the knowledge produced by the Project partners. A restricted access is expected for the knowledge that will be used for exploitation purposes; open access for all the other knowledge. Specific attention must be paid to the creation of an open access to the data collected during the trials. To this end, row data will be organized in an exportable format to be used by the scientific community and practitioners for their own purposes. A registered access for data download will be the only request for their use, in order to understand which organization is interested in using them and for which particular scope.

2. To manage and store the sensitive data obtained, all partners from ElasTest must comply with relevant European and national regulations as well as with the standards of practice defined by relevant professional boards and institutions.

The EU legislation standards for human studies include the following regulations as for the protection of individuals:

   a. The Declaration of Helsinki in its latest version (recommendation for conduct of clinical research).
   b. Directive 95/46/EC on the protection of individuals with regard to processing of personal data and on the free movement of such data Council Directive 83/570/EEC amending Directives 65/65/EEC, 75/318/EEC and 75/319/EEC dealing with proprietary medicinal products.
   c. Directive 95/46/EC (amendment 2003) of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the protection of privacy, storing of personal data and on the free movement of such data.
   d. The charter of fundamental rights of the EU (2000/C 364/01).
   e. EU Good Clinical Practice Directive (2001/20/EC).

3 Register on numerical data sets generated or collected in ElasTest

The register has to be understood as a living document, which will be updated regularly during the project lifetime. The intention of the DMP is to describe numerical models or observation datasets collected or created by ElasTest during the lifetime of the project. The information listed below reflects the conception and design of the individual work packages at the beginning of the project. Because the operational phase of the project started in January 2017, there is no dataset generated or collected until delivery date of this DMP.

The data register will deliver information according to Annex 1 of the Horizon 2020 guidelines (2015) (in *italics)*:

1. **Data set reference and name:** *Identifier for the data set to be produced.*
2. **Data set description:** *Descriptions of the data that will be generated or collected, its origin (in case it is collected), nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.*
3. **Standards and metadata**: *Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created.*
4. **Data sharing**: *Description of how data will be shared, including access*

procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.).
In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, and commercial, privacy-related, security-related).

5. **Archiving and preservation (including storage and backup)**: *Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.*

4 Datasets collected within the Various WP's

A table such as that demonstrated below will be included for every WP in order to present all the relative information of the data addressed within the context of that WP.

WPX	
What types of data will the project generate/collect?	
Data set reference and name	
Data set description	
Standards and metadata	
Data Sharing	
Archiving and preservation (including storage and backup)	
Reported by	

Table 1. WP data sets

5 References (*Bibliography*)

[1] European Commission Participant Portal H2020 Online Manual, *Data Management*. Available from: http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm [3 February 2017].

[2] European Commission Participant Portal H2020 Online Manual, *Open access & Data*

management. Available from:
http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm [3 February 2017].

[3] European Commission H2020 Programme, *Guidelines on FAIR Data Management in Horizon 2020,* Version 3, 26 July 2016. Available from:
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf [25 January 2017]

[4] European Commission, *Guidelines on Data Management in Horizon 2020,* Version 2, 30 October 2015. Available from:
https://indico.cern.ch/event/444264/contributions/1950399/attachments/1197993/1741971/h2020-hi-oa-data-mgt_en_1.pdf [25 January 2017]

[5] European Commission, H2020 Programme, *AGA- Annotated Model Grant Agreement,* Version 4.0, 27 February 2017. Available from:
https://www.google.gr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&uact=8&ved=0ahUKEwjdwpKc_uLUAhXCmBoKHeC_BIYQFggwMAI&url=http%3A%2F%2Fec.europa.eu%2Fresearch%2Fparticipants%2Fdata%2Fref%2Fh2020%2Fmga%2Fgga%2Fh2020-mga-gga-multi_en.pdf&usg=AFQjCNHt2A5Vzeqnh0JCvAUotC9qoI-lFg
[20 April 2017]

[6] DMPONLINE. Available from: https://dmponline.dcc.ac.uk/ [3 February 2017]

[7] OpenAIRE. Available from: https://www.openaire.eu/ [6 March 2017]

[8] EUDAT, *EUDAT Services for Data Preservation.* Available from:
https://eudat.eu/data-preservation [6 March 2017]