Towards EXtreme scale Technologies and Accelerators for euROhpc hw/Sw Supercomputing Applications for exascale

WP8: Project Management and Exploitation

D8.3: Data Management Plan

http://textarossa.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme, EuroHPC JU, grant agreement No 956831

TEXTAROSSA

Towards EXtreme scale Technologies and Accelerators for euROhpc hw/Sw Supercomputing Applications for exascale

Grant Agreement No.: 956831

Deliverable: D8.3: Data Management Plan

Project Start Date: 01/04/2021 **Duration**: 36 months

Coordinator: *AGENZIA NAZIONALE PER LE NUOVE TECNOLOGIE, L'ENERGIA E LO SVILUPPO ECONOMICO SOSTENIBILE - ENEA , Italy.*

Deliverable No	D8.3
WP No:	WP8
WP Leader:	ENEA
Due date:	M6 (September 30, 2021)
Delivery date:	20/10/2021

Dissemination Level:

PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

This project has received funding from the European Union's Horizon 2020 research and innovation programme, EuroHPC JU, grant agreement No 956831

DOCUMENT SUMMARY INFORMATION

Project title:	Towards EXtreme scale Technologies and Accelerators for euROhpc hw/Sw Supercomputing Applications for exascale
Short project name:	TEXTAROSSA
Project No:	956831
Call Identifier:	H2020-JTI-EuroHPC-2019-1
Unit:	EuroHPC
Type of Action:	EuroHPC - Research and Innovation Action (RIA)
Start date of the project:	01/04/2021
Duration of the project:	36 months
Project website:	textarossa.eu

WP8 Project Management and Exploitation

Deliverable number:	D8.3					
Deliverable title:	Data Management Plan					
Due date:	M6					
Actual submission date:						
Editor:	Xavier Martorell (BSC)					
Authors:	Xavier Martorell (BSC), Claire Chen (BULL)					
Work package:	WP8					
Dissemination Level:	Public					
No. pages:	14					
Authorized (date):	20/10/2021					
Responsible person:	Xavier Martorell (BSC)					
Status:	Plan	Draft	Working	Final	Submitted	Approved

Revision history:

Version	Date	Author	Comment
0.1	2021-07-25	Xavier Martorell (BSC)	Preliminary outline
0.2	2021-09-04	Claire Chen (BULL)	Information on confidential data
0.3	2021-09-15	Xavier Martorell (BSC)	Information on Open Data
1.0	2021-10-05	Xavier Martorell (BSC)	Including comments from reviewers
1.1	2021-10-08	Xavier Martorell (BSC)	Minor edits / final version

Quality Control:

Checking process	Who	Date
Checked by internal reviewer	Alessandro Lonardo and	2021-09-29
	Project Technical Commettee	
Checked by Task Leader		
Checked by WP Leader	Massimo Celino	
Checked by Project Coordinator	Massimo Celino	

COPYRIGHT

ACKNOWLEDGEMENTS

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement no 956831. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Italy, Germany, France, Spain, Poland.

Please see http://textarossa.eu for more information on the TEXTAROSSA project.

The partners in the project are AGENZIA NAZIONALE PER LE NUOVE TECNOLOGIE, L'ENERGIA E LO SVILUPPO ECONOMICO SOSTENIBILE (ENEA), FRAUNHOFER GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG E.V. (FHG), CONSORZIO INTERUNIVERSITARIO NAZIONALE PER L'INFORMATICA (CINI), INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), BULL SAS (BULL), E4 COMPUTER ENGINEERING SPA (E4), BARCELONA SUPERCOMPUTING CENTER-CENTRO NACIONAL DE SUPERCOMPUTACION (BSC), INSTYTUT CHEMII BIOORGANICZNEJ POLSKIEJ AKADEMII NAUK (PSNC), ISTITUTO NAZIONALE DI FISICA NUCLEARE (INFN), CONSIGLIO NAZIONALE DELLE RICERCHE (CNR), IN QUATTRO SRL (in4). Linked third parties of CINI are POLITECNICO DI MILANO (CINI-POLIMI), Università di Torino (CINI-UNITO) and Università di Pisa (CINI-UNIPI); linked third party of INRIA is Université de Bordeaux; in-kind third party of ENEA is Consorzio CINECA (CINECA); in-kind third party of BSC is Universitat Politècnica de Catalunya (UPC).

The content of this document is the result of extensive discussions within the TEXTAROSSA © Consortium as a whole.

DISCLAIMER

Table of contents

List of Acronyms

Acronym	Definition
FAIR	Findable, Accessible, Interoperable, Reusable
FPGA	Field-Programmable Gate Array
GFISCO	Go FAIR International Support & Coordination Office
IDV	Integrated Development Vehicle
MPI	Message Passing Interface
OmpSs	OpenMP Superscalar
OpenACC	Open Accelerator
OpenMP	Open Multiprocessing

Executive Summary

This document explains the Data Management Plan (DMP) for the Textarossa project. It is developed with the objective of making research data findable, accessible, interoperable, and reusable, following the FAIR approach promoted by the EU.

The DMP describes the types of data we are considering to manage in the project, and how these data are managed, taking into consideration the possibility to hold confidential information, the availability of the data after the project terminates, and potential standardization of the project IPs.

The DMP will be updated, if needed, during the development of the project.

1 Introduction

The Textarossa project will be managing various types of data. This document summarizes their characteristics and way of usage. We will end up providing most of the data, and specifically those that are not confidential by any partner following the FAIR (Findable, Accessible, Interoperable, and Reusable) scheme, as described in Section 2.

1.1 Open Data

Open data includes data that is used as input for Textarossa. Being initial description of products in development (OmpSs, StarPU…) or input data for applications, and the results obtained from their executions. Results are mostly performance results, as the project is centered on Computer Science aspects, not on the particular scientific results obtained from the applications.

1.2 Confidential Data

Data is considered confidential if some partner(s) define it to be this way. At the time of writing of this deliverable we have identified as confidential the specification/description of the blade that Atos will provide to the project to implement liquid cooling. This particular data, and any other that have the needs to stay consortium-confidential, will be managed in a secured way, as stated in section 4.

2 FAIR Data

Achieving FAIR (Findable, Accessible, Interoperable, and Reusable) [1] data follows these 4 steps:

2.1 Making data findable, including provisions for metadata

The project will use data repositories based on a filesystem-like set of directories, and accessible through the Internet by a web browser. Data will be classified given its high-level characteristics as:

- Documentation, specifications, sorted by product/development
- Application input, output, and results, sorted by platform/application

Results obtained from the experiments with Textarossa benchmarks and applications may include performance, energy consumption, resource occupation, etc. of the applications running on the Textarossa Integrated Development Vehicles (IDVs). Textarossa will have two main IDVs developed in the context of the project: IDV-A by ATOS, focusing on the cooling technology by reusing one existing GPU blade with X86/64 host, and IDV-E by E4, featuring ARM and accelerator(s) (e.g. FPGA).

Data indexing, to make data findable, will be implemented by means of the proper use of directory structures, and the search engines provided by the repository platform. The amount of data managed by the Textarossa project is relatively small, and we do not foresee the need to implement a database-like type of access.

2.2 Making data openly accessible

Public datasets will be open for public access from public repositories such as Zenodo (https://www.zenodo.org/) or the European Data Portal (https://data.europa.eu) the project will also favor publications of research & development in Open Access forums, Conferences and Journals.

Confidential information, mostly in the form of product specifications/descriptions will be kept in the Intranet side of the project web.

2.3 Making data interoperable

Application input, output and results will be organized in such a way that it can be compared with former results obtained from previous versions of the products/developments. This way, Textarossa will favor the assessment of the advances that we will be achieving in the project according the identified Key Performance Indicators (time-to-solution, energy-to-solution, etc).

2.4 Increase data reuse (through clarifying licenses)

Application input, output and results will be structured in a way that can be easily reusable, keeping them public, and free of charge, possibly through a GPL-like license, or inheriting the license types from the different data sources (specifically for input data), and taking care of each partners' business constrains or legal limitations they may have.

3 Allocation of resources

There is no additional cost for making the Textarossa project data FAIR, as it does not need any special treatment.

Data will be kept for three years after the Textarossa project has finished. After three years, we consider that the data will not have value anymore, as it may become superseded by new datasets obtained from future developments. Textarossa generated data will still be present in project publications, and public repositories.

4 Data security

Open data will not require to apply additional security policies in the project, given that the data used and produced do not include any private or personal information that could be considered as sensitive. If necessary, the Textarossa project will ensure that data is kept safely through regular backups. In case the data repositories already provide this feature, it will not be reimplemented.

Confidential data will be kept protected inside the project Intranet, with careful access policies, such as strong password requirements. Regular backups will also be used for keeping the information safe.

5 Ethical aspects

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

6 Standardization

Within the Textarossa project, we will develop several hardware/software components that will be considered for standardization. Among them, at the beginning of the project we have:

- Hardware IP for task scheduling: The hardware task scheduler will provide runtime acceleration via hardware management of task creation and task dependences. In case it includes the possibility for the programmer to interact with it, it has some opportunity to be included as a non-mandatory part of the OpenMP Standard [2].

Regarding standardization, the approach is twofold. On the one hand, Textarossa developments will be designed and developed to be compliant with the related standards (OpenMP, MPI…); and on the other hand, in case Textarossa proposes new features, we will try to promote them to be incorporated in such standards. In this regard, BSC is currently participating in the OpenMP, OpenACC, and MPI standardization bodies and will act as the contact with such bodies.

7 Conclusions

This deliverable presents the Data Management Plan for the Textarossa Project. We have identified data in the form of:

- Specifications/descriptions of products
- Input, output data for applications
- Results obtained from the experiments with applications
- Publications in Open Access Forums, Conferences and Journals

Among the information listed, we consider that all of if will be in Open Access from the Textarossa website. Only a few specifications/descriptions related to products being used in the project, like the Atos racks with liquid cooling, will be considered as consortium confidential, and they will be protected in the project repository with secured access for partners only.

Data published by the project consortium will be kept available online for 3 years, after the finalization of the project.

Textarossa will favour the standardization of hardware/software components, like the hardware IP component that we are developing for task scheduling.

8 References

[1] GO FAIR International Support & Coordination Office (GFISCO), Fair Principles. https://www.go-fair.org/fair-principles/

[2] OpenMP ARB, OpenMP Specification 5.1. https://www.openmp.org/specifications