ePROCESSOR

D1.2

Data Management Plan (DMP)
v1.0

Document information

Work package:	**WP1: Coordination and Management**
Contract number:	956702
Project website:	www.eprocessor.eu
Author(s)	Nehir Sonmez (BSC)
Contributors	BSC: Santiago Marco-Sola, Lluc Alvarez, Julianna Anguelova UNIBI: Jens Hagemeyer THALES: Nicolas Ventroux EXA: Yannis Papaefstathiou
Reviewer(s)	EXAPSYS
Dissemination Level	PU
Nature	ORDP
Key words	data sets, open access, FAIR data
Contractual deadline:	30/9/2021

EuroHPC Joint Undertaking

This project has received funding from the European High-Performance Computing Joint Undertaking Joint Undertaking (JU) under grant agreement No 956702. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Spain, Sweden, Greece, Italy, France, and Germany.

Change Log

Version	Author(s)	Comments and Description of change
0.1	Nehir Sonmez	First version
0.2	Yannis Papaefstathiou (EXAPSYS)	Reviewed version
1.0	Nehir Sonmez	Final version for submission

Contents

1. Executive Summary

This deliverable presents the data management plan (DMP) of the eProcessor project, which describes the data management life-cycle for all datasets to be collected, processed and/or generated along the lifetime of the project. The DMP complies with the EC's objective of making research data findable, accessible, interoperable, and reusable (FAIR). Concretely, this deliverable describes, among others:

- Which datasets will be generated, collected and processed, considering both (i) the development and execution of the eProcessor application use-cases and (ii) the research activities towards the development of the eProcessor ASIC, as well as the emulation/simulation and software development technology.
- Which methodology and standards are to be applied to these eProcessor datasets.
- How these datasets will be managed and stored during the lifetime of the project, and after the end of it.
- How these datasets will be made (openly) accessible.

To summarize, the datasets managed or created in the eProcessor project are:

Dataset	Description
HPC, AI, ML, DL and Bioinformatics applications benchmarking	Publicly available benchmarks like NAS, Wavefront Alignment (WFA), Smith-Waterman, FM-Index, European Distributed Deep Learning Library (EDDLL) and European Computer Vision Library (ECVL) will be used to evaluate the eProcessor architecture. We envision using other publicly available high performance data analytics workloads for AI/ML/DL. We will additionally use the Smart Mirror (from the smart home use case), and the aerial surveillance use-case application as primary workloads.
eProcessor architecture performance modeling	The performance of the eProcessor architecture will be evaluated using microkernels with open data sets and synthetic data. The test cases and the model infrastructure will be distributed as open source.
eProcessor architecture source code	In eProcessor, we will provide IP and results of open source nature. The consortium might also choose to derive hardware/software components from the open source community.

2. Datasets Summary

In the eProcessor project, performance numbers from the High Performance Computing (HPC), HPDA (high performance data analytics, including Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL), and Bioinformatics) applications will be generated with the goal to compare the performance of key applications with (i) the simulation infrastructure, (ii) the emulation platform, (iii) and with the resulting eProcessor chip. This is expected to be a small set of data, in the order of several hundred MBs.

Performance data will be collected such as execution time, and energy consumption, and other metrics will be derived, such as speedup, IPC and GFlops/Watt. There is no existing

data that will be used, and the data will be generated from the performance of benchmarks and applications. Those data will be tracked, in order to compare the expected improvements at each iteration. This is also expected to be a small (but valuable) dataset.

The resulting data will be useful for researchers working on similar approaches for building and accelerating software/hardware co-designed Exascale systems, and in particular for those targeting HPC, HPDA and Bioinformatics applications. Using a common set of workloads will allow comparisons with other systems.

eProcessor will generate four main types of datasets:

1. The source code for a performance model used to evaluate the architecture using small microkernels with synthetic input data.
2. The RTL (Register-Transfer Level) source code for the eProcessor (including the on-chip accelerators, NoC, IOMMU, L2 caches) that can be mapped to the hardware emulator (FPGA-based system).
3. The source code of the software components and tools that will form the software stack for the eProcessor.
4. Datasets generated and used for the evaluation of the eProcessor software and hardware stack, including the full use-case applications and the microbenchmarks.

The eProcessor project will also manage the personal data from the partners of the consortium, as stated in D1.1 under GDPR. Therefore, in this document, we will not make any references to this type of data.

3. FAIR Data

3.1 Making data findable, including provisions for metadata

The performance data provided by the performance simulation infrastructure and the eProcessor emulator will be small, in the order of a few Gigabytes. It will be organized by application and system configuration, making the results easy to navigate and find. The results will be in a standard format based on the output of the applications, and thus it will be relatively small for any standard identification mechanisms.

The input data for the applications, especially the synthetic applications, will be either auto-generated or from publicly available sources when possible. There are no plans to persist the data, other than the method used to generate or obtain the input data. Thus, no unique identifiers (i.e., Digital Object Identifiers) are required to locate those data.

Any source code developed in the eProcessor project will be made available, when possible, in open source repositories provided by BSC, with a standard organization and relevant documentation, while any changes made to the source files of GEM5 will be propagated to the GEM5 repository so as to be distributed as open-source in the GEM5 community.

3.2 Making Data Openly Accessible

All input and output data generated in this project may follow the Open Access policy. However, when there is a huge amount of data used as part of the application data set, only the generation and/or source of the data sets for the applications may be provided and accessible by the community through the repository to demonstrate the validity of the eProcessor implementation.

The source code for any components licensed as open source will be included in a public Git repository[1], while it will be linked to repositories of the relevant software tools (e.g. GEM5, Linux, etc.) when applicable. When needed, new Git projects and repositories will be created for the various parts of the eProcessor project. Furthermore, Git submodules will be used to link the integrated version with the corresponding Git project of each of the eProcessor components (software and/or hardware), when applicable.

As for the use-case applications, the dataset used for the aerial border surveillance use-case will be made openly accessible. The reference data sets and corresponding data provided in the Smart Mirror do not contain individual-related information. All data sets, e.g. gesture, object, face or voice recognition data sets are anonymized, no individual-related information is stored. Furthermore, all data have been collected with the informed consent of the users.

For the bioinformatics application use cases, the project utilizes simulated and real datasets for development and performance evaluation purposes. These datasets are representative of the application use cases targeted in the project. These datasets contain sequencing data (i.e., DNA sequencing and reference tags) that are synthetically generated and derived from actual sequencing experiments of the order of a few million sequences.
In the case of simulated datasets, all the data are generated using publicly available and open-source simulators. Datasets and simulation parameters used will be made public to support the transparency and reproducibility of the experiments. In the case of authentic sequencing datasets, including genome reference assemblies and sequencing datasets, all the data will be obtained from public sequence archival resources, particularly the European Nucleotide Archive (ENA) and the NCBI Sequence Read Archive (SRA). All the datasets supporting the developments and findings of this study will be made openly available, correctly referencing the accession numbers of the datasets and studies related to them. All non-simulated datasets are derived from third-party public resources, and they contain no personal or traceable information, and pose no ethical concerns.

For any models, codes, or other information developed by either eProcessor partners, or 3rd parties, an agreement with the owner will be required.

3.3 Making data interoperable

No specific data format will be provided to the datasets needed to evaluate the performance of the eProcessor due to their small size.

[1] The eProcessor Gitlab is currently in private mode in https://gitlab.bsc.es/eprocessor. It will be made public when all mentioned data is available and ready in the corresponding format.

The data generated when the eProcessor will be running our use-case applications, will allow to compare the eProcessor performance with that reported by other sources, to better determine the advances on the programming and runtime support for heterogeneous systems, as well as the novel eProcessor accelerators. This information will be included in scientific documents to properly determine the advances of the eProcessor architecture.

The simulation data will be published with enough detail in order to allow other scientists to compare their results with the ones generated in this project, as well as research communities, such as the GEM5 simulator community.

Reference data sets of the Smart Mirror use case (video or audio data, like gesture, object, face or voice data sets) are mostly self-explaining and therefore can be intuitively used by other researchers.

The dataset used for the aerial border surveillance use-case will be publicly available in a standard and open format for better interoperability.

3.4 Increase data re-use (through clarifying licences)

During the project, data reuse options for the data sets will be dealt with on a case by case basis, aiming, when possible, to keep them public, accessible, free of charge and reusable under request, under a standard reuse non-contagious license, such as the BSD license (important for industrial exploitation), in line with the obligations set out in the Grant Agreement. This will enable the widest reuse or inheriting the license types from the different data sources as explained in the data summary description, and taking care of each partners' business constraints or legal limitations on them.

4. Other research outputs

The main research outputs for eProcessor are in the forms of journal and conference publications. The performance data of the eProcessor will be clearly explained and depicted to enable reproducible results. The accompanying source codes will also be made publicly available, as appropriate.

5. Allocation of Resources

There is no additional cost of making our performance data FAIR, as it does not need any 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 eProcessor project. After three years, we consider that the data will not have any value, as results will be superseded by new datasets obtained from future developments. Data will still be present in project publications and public repositories, when appropriate.

6. 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 applied. Each partner is in charge of backing up its data and all partners have strongly secured IT systems protecting their local storage servers.

With respect to the Smart Mirror use-case in eProcessor, data security in terms of data privacy is very crucial, which means that all data are processed locally and not sent to the cloud, and no data (sound or image) are recorded at any time.

7. Ethical Aspects

No ethical or legal issues are directly derived from the generated data. For the aerial border surveillance application, reference data sets do not contain individual-related data, since only anonymized sea or terrestrial vehicles, such as boats or cars, are targeted by the algorithm. There are then no ethical or legal issues regarding those data.

The Bioinformatics data are derived from third-party public resources, and they contain no personal or traceable information, thus pose no ethical concerns either.

The smart home use case, represented by the smart mirror, is driven by the interaction of users with their living environment. For this purpose, the usage of cameras and microphones is inevitable. After all, the smart home should adapt to the individual needs of its residents. Within the eProcessor project, data security is very crucial, which means that all data are processed locally and not sent to the cloud, and no data (sound or image) are recorded at any time. In order to be able to distinguish between people, for example, for the smart mirror, feature vectors of the trained users are stored, which do not allow any conclusions to be drawn about their appearance. Nothing is stored from unknown and untrained persons. Other saved user data are limited to the information displayed for a known person. For training DNNs, the datasets used are mainly open-source, and do not contain individual-related data, (e.g. generated or stored personal data). The only exception is gesture recognition, for which a non-public dataset was created. For this dataset, all participating persons have already given their written consent.

8. Acronyms and Abbreviations

BSD - Berkeley Software Distribution
DNN - Deep Neural Network
FAIR - findability, accessibility, interoperability, reusability
FPGA - Field-programmable gate array
GDPR - General Data Protection Regulation
GPL - General Public License