This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777533.

PROviding Computing solutions for ExaScale ChallengeS

D1.2	Revised Data Management Plan		
Project :	PROCESS H2020 - 777533	Start / Duration:	01 November 2017 36 Months
Dissemination[1]:	PU	Nature[2]:	R
Due Date:	30 April 2019	Work Package:	WP 1
Filename[3]	D1.2_QDMPlan_v1.0.docx		

ABSTRACT

This deliverable includes an update of the Data Management Plan (DMP) of the project, focusing on refining the outlooks of the service pilots and the use cases they serve.

[1] PU = Public; CO = Confidential, only for members of the Consortium (including the EC services).

[2] R = Report; R+O = Report plus Other. Note: all "O" deliverables must be accompanied by a deliverable report.

[3] eg DX.Y_name to the deliverable_v0xx. v1 corresponds to the final release submitted to the EC.

Deliverable Contributors:	**Name**	**Organization**	**Role / Title**
Deliverable Leader[4]	Maximilian Höb	LMU	Coordinator
Contributing Authors[5]	Matti Heikkurinen	LMU	Writer
	Jan Schmidt	LMU	Contributing author
Reviewer(s)[6]	Ruben Riestra	INMARK	WP9 leader
Final review and approval	Maximilian Höb	LMU	Coordinator

Document History

Release	Date	Reasons for Change	Status[7]	Distribution
0.8	2019-04-03	Initial version	Draft	WP1
0.9	2019-04-12	Reviewing	In Review	Consortium
1.0	2019-04-30	Final version	Released	

[4] Person from the lead beneficiary that is responsible for the deliverable.

[5] Person(s) from contributing partners for the deliverable.

[6] Typically person(s) with appropriate expertise to assess the deliverable quality.

[7] Status = “Draft”; “In Review”; “Released”.

Table of Contents

Executive Summary

This updated Data Management Plan (DMP) presents an overview of the evolution of the goals, background and constraints the project needs to take into account when generating and publishing reusable datasets. The initial assessment of the role of the use cases being focused on providing tools aimed at making orders of magnitude improvements in the efficiency and convenience of extracting value from existing data assets has not changed, thus most of the new material in this updated DMP is stemming from the experiences gained in the work that is directly related to the use cases. For convenience, the detailed background analysis of the use cases presented in the deliverable D1.1 has been included in this document as an annex.

List of Tables

1 Updated Data Management Plan

1.1 Introduction

As stated in the Deliverable D1.1, the requirements of the PROCESS DMP stem from the EC Grant Agreement. These requirements need to be applied in the contexts of the five service pilots included in PROCESS, each of them with unique opportunities and constraints related to the reuse of the data generated.

These foundations of the DMP have not changed during the first 18 project months, but for convenience the initial analysis of the situation from the deliverable D1.1 is included in the Annex 1 of this document.

1.2 Potential Data assets

The deliverable D1.1 identified two groups of sources for reusable datasets:

- The work focused on the use cases and supporting the communities around it
- The work on the general purpose exascale data solution that supports the use cases.

The first group was assumed to contain the majority of the assets that are used, and the experiences of the first 18 months tend to support this initial assessment. It is possible that the validation step of the PROCESS solution will generate datasets that could be used as basis for more general benchmarks of extreme data applications, but this remains to be confirmed during the second half of the project.

1.3 Common approaches to data management

The initial assessment framework determining whether the project should publish a dataset or not is still valid:

1. Would publishing the dataset raise potential privacy issues (e.g. allowing de-anonymising subjects)?
2. Does the license of the original dataset allow publishing a derived set (IPR)?
3. Does publishing dataset lead to potential savings (in terms of time, computational resources etc.) when compared to re-generating the derived set?
4. Does the project need to keep a derived dataset already for its internal use (e.g. for testing, benchmarking, validation)?
   a) Would these datasets be needed by third parties to fully validate correct behaviour of PROCESS tools?
5. Can a suitable (i.e. a repository that would facilitate discovery of the data asset by its intended users), managed repository that would be willing to host the dataset be found?
6. If not, can the long-term commitments needed for formally publishing a dataset be met by the partners?

Applying this framework has produced following observations:

- There are potential derived datasets stemming from use cases (UC) #1 and #4. However, at least in the case of UC#4, it is likely that publishing the algorithm would be more efficient way of allowing reuse. In case of UC#1, the project is using derived dataset internally. However, these data assets are deemed to be very specific to the project and not of interest to third parties.
- The dataset used as a starting point for the UC#3 has certain licensing issues that need to be taken into account. However, the complementary datasets identified do not carry this limitation.
- There are some promising – albeit early stage – discussions that indicate that the resource requirements of long-term preservation of datasets can possibly be met through collaborative arrangements with other projects.

1.4 Use-case specific data management aspects

The following sections will present an update of the potential reusable datasets generated in the context of each of the use cases. The details of the data processing workflows and requirements are presented in the deliverable D4.1.

1.4.1 UC#1

Background datasets

The ongoing UC#1 activities are focused on the Camylyon17 and Camelyon16 datasets, with the other background datasets kept as candidates for further testing and validation at the end of the project. The project may also gain access to datasets collected and used by the ExaMode project[8], in which case the use of other already published datasets will have a considerably lower priority.

Table 1 Updated background dataset summary of use case 1 – data sets under active study highlighted

Dataset name	Estimated size	Description	Format	Annotations
Camelyon17	>3TB	1000 WSI, 100 patients	BIGTIFF	XML file
Camelyon16	>1TB	400 WSI	BIGTIFF	XML file + Binary Mask
TUPAC16	>3TB	WSI	BIGTIFF	CSV file
TCGA	>3TB	WSI	BIGTIFF	TXT file
PubMed Central	~5 million images	Low resolution	Multiple formats	NLP of image captions
SKIPOGH	>30TB	WSI	BIGTIFF	
ExaMode	Tens of TB	*TBD*	*TBD*	*TBD*

Data generated

As described in the original DMP, the UC#1 will generate two types of data assets for the project internal use:

- Derived datasets based on one of the published ones (Camelyon17, Camelyon16 ...).
- Actual neural networks trained by the datasets.

Publishing approach

As in the original DMP – the project will focus on documenting the processes used to develop derived datasets.

1.4.2 UC#2

Background datasets

The work in the UC#2 will rely on accessing the data from the LOFAR Long Term Archive (LTA)[9] and produce tools that allow more efficient use of the LTA contents.

Publishing datasets retrieved from LTA is not deemed necessary at this stage, as any actual analysis performed by third party users would need to access the official archive as an

[8] https://cordis.europa.eu/project/rcn/220703/factsheet/en

[9] https://lta.lofar.eu/

authoritative source of data. Furthermore, it is possible to validate the UC#2-related service pilot by using files filled with random values.

1.4.3 UC#3

Background datasets

The work in the UC#3 was based on using the UNISDR community as a pilot community for advanced PROCESS tools. The original dataset consists of about 2TB of data and is openly accessible via http://unisdr.mnm-team.org. This resource is described in more detail in Annex 2 of this deliverable.

As the new, community-based process used by UNISDR is still evolving, the project is at the moment considering enhancing the original datasets with other assets. The primary candidate for testing this approach is based on the datasets produced by the CliMex project[10]. The project has generated fifty climate simulation models for the time period of 1950 to 2100 covering Central Europe and North-Eastern North America. This data is used as an input for hydrological simulations to identify extreme flooding scenarios associated to climate change. The CliMex project will aim at publishing ~200TB dataset during 2019, with a suitable open license (details of the exact license still under review). The integration work to make these two datasets available through a single interface is ongoing.

In parallel to this technical work, the project is investigating ways to benefit from the synergies with the LEXIS project[11], that has two pilot activities dealing with disaster risk modelling (Earthquake and Tsunami, Weather and Climate).

1.4.4 UC#4

Background datasets

The background datasets are the private transaction records kept by LSY that are used as basis for generating statistically similar simulated datasets for testing the ancillary pricing mechanism.

Data generated

The simulated data needs to be evaluated based on the checklist presented in the section "Common approaches to data management". It is likely that the value of large datasets is relatively low, i.e. it wouldn’t add value to publishing of the generation algorithm.

1.4.5 UC#5

The used datasets base on pre-processed data from the Copernicus Sentinel data.

Data generated

Tools, documentation and parameter files for the pre-processed Copernicus data used in PROMET[12] and output generated with PROMET.

Publishing approach

There are three potential channels that could be interested in the data assets generated in the UC#5 context

1. Users of the PROMET software
2. Broader agronomy research community interested in easy access to satellite data

[10] https://www.climex-project.org/

[11] https://lexis-project.eu/

[12] Mauser, W. ; Bach, H.: “PROMET - large scale distributed hydrological modelling to study the impact of climate change on the water flows of mountain watersheds.”, DOI : 10.1016/j.jhydrol.2009.07.046

3. Providers of generalised Copernicus access services

Evaluating which of these channels are the best ones for promotion of third-party reuse is still ongoing.

2 Annex 1: introduction and background of the PROCESS DMP

2.1 Introduction

The requirements for the Data Management Plan (DMP) are laid out in grant agreement (GA) and supporting documentation provided by the EC. The GA states that:

> *"Regarding the digital research data generated in the action ('data'), the beneficiaries must:*
>
> > *(a) deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate — free of charge for any user — the following:*
> >
> > > *(i) the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible;*
> > >
> > > *(ii) other data, including associated metadata, as specified and within the deadlines laid down in the data management plan"*

The prerequisites for complying with these requirements include:

- Identifying the data generated that could form the basis of becoming a reusable data asset
- Identifying and securing access to optimal repositories for long-term preservation of the data
- Review and refine the metadata so that it provides information that is relevant and understandable also when separate from the PROCESS project context. This is important already for the project internal use as the data assets of the PROCESS project are multi-disciplinary in nature.
- Map the data with publications made by the project
- Having necessary *due diligence* processes in place to ensure publication of data will not - directly or indirectly - raise any additional compliance issues.

Fulfilling these requirements in a multifaceted project such as PROCESS requires a two-stage approach: creating an initial data management plan describing potential reusable data asset that project can generate during its lifetime, and common principles and approaches used to choose optimal approaches for making them reusable in the longer term. The initial data management plan will be refined during the project lifetime as the nature of the data assets generated will become clearer. However, it should be noted that due to the interdisciplinary nature of the project, it is likely that the project will generate several data management plans to match the specific requirements and community conventions of each of the disciplines involved. Maximising the potential for reuse will also depend on successful identification of the potential secondary user communities, as this is a prerequisite for successful reviewing and refining of metadata specifications and identification of the optimal repositories that are to be used for storing the data generated during the PROCESS lifetime.

One of the common characteristics of the PROCESS use cases is that they do not do primary data collection themselves. Instead, they will generate data, either based on the publicly available datasets or (especially in case of the UC#4, see next section) provide simulation results based on statistical distributions of actual, private datasets that are used as background of the project work.

2.2 Background

PROCESS is a project delivering a comprehensive set of mature services prototypes and tools specially developed to enable extreme scale data processing in both scientific research and advanced industry settings. These service prototypes are validated by the representatives of the communities around the five use cases:

- UC#1: Exascale learning on medical image data
- UC#2: Square Kilometre Array/LOFAR
- UC#3: Supporting innovation based on global disaster risk data/UNISDR
- UC#4: Ancillary pricing for airline revenue management
- UC#5: Agricultural analysis based on Copernicus data

From the data management perspective, each of these five use cases presents challenges that are complementary to each other, and with different potential for direct generation of exploitable data assets. This mapping is presented in the table below:

Table 2 Key challenges of use cases

Use case	Key challenge	Type of reusable data asset
UC#1	Machine learning using massive, public datasets; exploitation requires high degree of privacy	More challenging datasets based on the published ones (e.g. with noise of artefacts simulating mistakes made during the scanning of a tissue slide, rotation of regions of interest etc.)
UC#2	Extreme volume of data (LOFAR reduced data set 5-7PB per year, SKA centrally processed data rate: 160Gbps)	For the most part the data assets will remain in LOFAR LTA (long term archive), however a disk copy of test observation could be useful for software testing and validation
UC#3	Usability of extreme scale tools to support emerging big data user communities: The UNISDR Global Assessment Report (GAR) datasets have been made publicly available for non-commercial use since early 2017. The process to be used for the 2019 will be fundamentally different, with considerably larger group of experts with heterogeneous and evolving data curation practices involved in the data production and curation.	The 2015 and 2017 GAR datasets and the results of the CIMA showcase.

UC#4	Very large datasets, extreme responsiveness requirements, high financial risks/potential rewards; exploitation requires demonstrating high degree of security and auditability of the PROCESS solutions.	Tools, documentation and parameter files for generating simulated transaction datasets
UC#5	Support wide range of uses of a very large dataset of satellite images (growing at the rate of 7.5PB per month)	Tools, documentation and parameter files for accessing Copernicus data, possibly specialised derived sets of data (e.g. time series of specific location)

All these use cases have distinct communities, practices and documentation/metadata conventions, thus any component that can be used as part of all five demonstrators can be considered a proven, generalizable data management component with very high exploitation and uptake potential.

3 Annex 2: description of the original UNISDR dataset

The original UC#3 service pilot dataset consists of:

- Hazard data, consisting of groups of simulated scenarios for each of the natural hazards earthquake, tsunami, riverine flood, cyclonic wind, storm surge and so on) used in the analysis. Each set of simulated scenarios must comply with the certain key requirements, such as being mutually exclusive, collectively exhaustive and having an annual frequency of occurrence associated with them.
- Exposure data, describing each exposed asset with a set of attributes such as their geographical location, structural characteristics, construction material type, economic value (among others).
- Vulnerability data, characterising the exposed asset with a set of attributes describing their relevant characteristics that determine how sensitive they are to different hazards at different intensities.

The overall structure of the dataset is relatively complex. For example, in the most comprehensive hazard category (flooding) there are scenarios generated for 163 countries, stored in a ”directory per country” structure. There are almost 6,6 million flood scenarios that form the basis of the risk assessment process. The other hazard scenarios can be handled with smaller granularity, since e.g. the slight variation in the location of the epicentre of an earthquake has considerably less impact on the likely outcomes compared to flooding scenarios. Altogether, there are five different hazard types that are assessed based on probabilistic modelling, with scenario generation handled by fairly autonomous expert teams.