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

D1.1	**Quality Plan and Initial Data Management Plan**		
Project :	**PROCESS H2020 - 777533**	**Start / Duration:**	**01 November 2017 36 Months**
Dissemination[1]:	**PU**	**Nature[2]:**	**R**
Due Date:	**30 April 2018**	**Work Package:**	**WP 1**
Filename[3]	**D1.1_QDMPlan_v1.0.docx**		

ABSTRACT

This deliverable summarises the processes the project will use to ensure its deliverables and other outputs are of high quality and suitable for their intended purposes. The document includes a short summary of the foundations of the quality processes outlined in the Annex 1 of the grant agreement and in the consortium agreement, presents the deliverable process and dissemination-related quality issues.

The deliverable includes also the first version of the Data Management Plan (DMP) of the project.

This version is a draft of D1.1 and is under review.

[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]	Nils gentschen Felde	LMU	Coordinator
Contributing Authors[5]	Nils gentschen Felde	LMU	Writer
	Maximillian Höb	LMU	Writer
Reviewer(s)[6]	Ruben Riestra	INM	WP9 leader
Final review and approval			

Document History

Release	Date	Reasons for Change	Status[7]	Distribution
0.8	2018-04-26	Initial version	Draft	
0.9	2018-04-28	Reviewing	In Review	
1.0	2018-04-30	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 report has been created to support all PROCESS team members in their daily work for PROCESS. The specific purpose of this document is to complement quality related aspects of the key contractual documents of the project:

- The EC Grant Agreement, including the Annex 1 "Description of Action" (DoA)
- PROCESS Consortium Agreement (GA)

From the practical point of view, these documents establish primarily the contents of the project deliverables and the schedule for their delivery to the Commission (Annex 1), and the structure and the decision-making processes of the PROCESS project organisation. The quality plan complements these two documents by describing in a more detailed manner the processes that are utilised to ensure that the project outputs are of the highest possible quality. The quality in this context refers to to the outputs being accurate and fit for the purposes stated explicitly in the DoA or implied in the informal communications (such as presentations), and complying certain technical and formal characteristics (e.g. using document template). The quality assessment will thus provide foundations for successful dissemination and exploitation activities.

This document contains the descriptions of the quality management processes and responsibilities of the different project organisations that provide additional guidance on the implementation of the clauses in the DoA and GA. These include:

- Release procedure of deliverables and other project outputs carrying the PROCESS brand
- Communication practices and tools
- Software quality approach
- Management of IPR-issues

The Data Management Plan (DMP) will present the overall goals, background and constraints the project needs to take into account when generating and publishing reusable datasets. The multifaceted nature of the project and its stakeholders is discussed in some detail, before presenting the potential data assets the project will generate or be responsible for. The use cases of the project are largely based on providing tools aimed at making orders of magnitude improvements in the efficiency and convenience of extracting value from existing data assets. For this reason, the data management plan includes relative detailed description of the heuristics for determining whether the project should publish an actual derived datasets based on already published data sources or just publish the algorithm and tools for replicating the steps.

The DMP will also include a short summary of the potential sources of reusable dataset emerging from each of the use cases.

List of Tables

1 Quality Plan

1.1 Summary of the quality-related processes in the DoA

The project management structure is based on the following three components:

- Strategic decisions: General Assembly (GA) with all partners represented
- Innovation management, with innovation manager monitoring any issues arising from work packages WP2 (Motivation, Future Requirements, and Emerging Opportunities) and WP9 (Dissemination, Engagement and Exploitation)
- Day-to-day coordination: Work package leaders coordinating the work within the work packages and exchanging information with the project coordinator and other WP leaders through the project executive committee. WP leaders will have autonomy in deciding on the approaches to be used within the work package, as long as they can support other WP leaders and the project coordinator in their tasks.

All of these groups will use the project Wiki (Confluence tool discussed in the next chapter) for circulating agendas and to record the minutes of the project meetings. The meetings of these three management structures have the following cycle:

- GA meetings: three times per year at the minimum,
- Day-to-day coordination: weekly Project Executive Committee (PEC) teleconferences,
- Innovation management: status review in the PEC meetings, in-depth analysis on demand and ad hoc contacts between WP2 and WP9 leaders when a need arises.

1.2 Summary of the quality-related processed in the Consortium Agreement

The PROCESS Consortium Agreement is based on the DESCA model, and as such it does not have a direct bearing to the project quality management. It includes certain provisions related to timing of the meeting invitations and finalisation of agendas - both for the ordinary and the extraordinary meetings. Any partner may call for an extraordinary meeting.

1.3 Tools used by the project

1.3.1 Collaborative online working spaces - Confluence and GitLab

The project uses a Confluence collaboration tool installed at LRZ as an online collaboration space. Confluence is a Wiki-style system that supports collaborative editing of pages and their relationships with each other. It provides fine-grained access control for the content and advanced collaboration mechanisms that allow users to subscribe to notifications (for example page edits), assign tasks and leave comments on the pages.

Confluence is used as the main hub collecting links to all project internal tools, documentation and outputs. The deliverables are edited primarily using the confluence system to minimise the barriers for contributions and the additional workload needed for manual integration of contributions from multiple sources.

The software developed by the project is stored in the GitLab installation provided by LRZ as a service for LMU (gitlab.lrz.de). This mature, widely-used version management system supports well-defined software development processes and facilitates uptake by the other developer communities using GitLab by providing familiar interface for the published software. The details of the software development process are described in some detail in the section *"Software Quality Assurance"* in this document.

1.3.2 Communication tools

The communication tools range from mailing lists (one for each work package, WP leaders' list and list containing all members of any of the PROCESS list) to conference call systems. The mailing lists are listed in the Confluence system and can be added if need arises.

The primary conference calling system used by PROCESS is Gotomeeting (https://www.gotomeeting.com/), which was deemed to provide the best balance of ease-of-use, support of multiple platforms, features (e.g. screen sharing and chat functionalities that are not available on traditional conference call systems) while complying with the IT security policies of all of the partners.

1.3.3 Meeting practices

The basic schedule of the meetings is already defined in the Description of Action (DoA), Annex 1 of the Grant Agreement. The minutes of the meetings are stored on the confluence system, linked to a common "Meetings" page. This provides a common repository and a way to track any corrections made to the meeting minutes as changes are logged (timestamp and username).

1.4 Deliverable process

The deliverable process was discussed in the kick-off meeting and resulted in a simple, straightforward approach documented on the Confluence page (Quality Assurance process). The process agreed is as follows:

- Every official piece of paper (e.g. deliverable) will be sent to the Executive Board two weeks before official deadline (eb@process-project.eu)
- Executive Board has one week to suggest changes/improvements
- No reaction means silent consent
- One week left in order to include changes or improve on document for authors
- Final version has to be circled after being published to all project members (all@process-project.eu)

This basic mechanism can be refined during the project lifetime if needed (e.g. to accommodate urgent requests from parties project collaborates with). The main quality-related issue is the tacit approval of deliverables - in a multifaceted project consisting of platform development and highly autonomous use case pilots, waiting for explicit approval from all the members would increase risks of delays while not necessarily improving the coverage of the review process.

1.5 Software quality assurance

The foundations of the software quality assurance are the guidelines developed and best practices documented by NLeSC in their internal software development guide (accessible at https://nlesc.gitbooks.io/guide/content/). This approach is a natural choice since NLeSC is responsible for the WP8 (Validation) and has extensive experience in applying the guide in other NLeSC projects. The guide is also so called "live document" that is continuously updated and refined, making it easy to encompass lessons learned from the PROCESS work in a way that benefits automatically a larger group of projects.

The scope of the guideline documentation extends beyond the software quality aspects, extending to areas covered by other PROCESS documents (e.g. publishing of the results) and covers some details that need to be adjusted (e.g. exact repository used for the software). In the process context other PROCESS documents will naturally have a precedence - and in case there is a danger of misunderstanding the exact approach is documented in the project Wiki (described earlier in this section).

Some of the key principles stemming from the NLeSC guide are:

1. In case of doing proof-of-concept/prototyping work that doesn't comply with the software development process, state this explicitly in all the communications
2. Version control - apply consistent practices from the beginning of the project
3. Arrange formal code reviews as part of the development process
4. Automate testing as much as possible
5. Apply standards and language-specific implementation guides is available
6. Do an in-depth assessment of the IPR-related issues at least in two stages:
   a) Finalising the design of the software
   b) Before making software publicly available

The implementation of these approaches will be reviewed in the executive board meetings, based on information collected by the WP8 leader.

1.6 Dissemination and exploitation quality issues

1.6.1 DoA dissemination aspects

The DOA includes a list of potential dissemination channels and KPIs that project partners identified at the time of writing the proposal. These will be reviewed and complemented during the project lifetime, with the first update documented in the deliverables D9.1 ("Initial DEP and market research Report"). The dissemination-related Key Performance Indicators (KPIs) are defined as follows:

Table 1 Dissemination-related Key Performance Indicators

Target area	Indicator	Expected progress (cumulative numbers unless otherwise stated)		
		After M12	After M24	After M36
Scientific	Number of publications, talks, presentations in conferences and workshops	4	12	20
Scientific	Number of lectures, courses or training events (including extreme scaling workshops)	2	8	16
Other projects	Number of meetings with other project presence (either hosted or participated)	5	12	20
Website	Unique monthly visitors (best three-month average)	200	400	600
Website	Returning monthly visitors (best three-month average)	50	100	150
Press	Number of mentions in paper press, online media, TV/Radio	4	12	30
Social media	Number of followers/friends on social media networks (across all platforms)	80	150	450

Developers	Monthly downloads of technical documentation: White papers, architecture descriptions or software releases (best three month average)	50	150	200

From the quality perspective, the scientific goals require balancing the type of output and its impact: for example a talk or a presentation in a small workshop that targets an ideal niche audience is of much higher value than a publication that is perhaps more prestigious on the surface but doesn't provide similar targeted audience. This concrete approaches to solve this challenge will be discussed in the D9.1 and in case they require changes to the overall quality processes of the project this deliverable will be updated to reflect the next practices.

1.6.2 Internal guidelines

The project has an internal guideline document complementing the plans presented in the DOA and providing rapid guidance to supplement the procedures and methods presented in the previous chapters. The document contains reflections, best practices and templates for identifying, refining and promoting project success stories. This covers both direct activities of the project team as well as activities that support dissemination and exploitation indirectly, such as forming alliances with entities that have synergies with PROCESS goals, building and managing communities and so one.

1.7 IPR-related quality assurance issues

PROCESS needs to take IPR issues into account in several parts of its activities:

- Publishing the software solutions through the GitLab installation
- Integrating software components with other open source solutions
- Publishing derived datasets (to ensure that the constraints of the original license dataset is published in respected)
- Submitting publications to scientific journals (to ensure at least green open access)
- Preparing presentation material (e.g. ensuring that photographs used as an illustration do not infringe licensing terms)
- Dealing with potential infringement of IPR generated by the project

Hence, the IPR issues form a part of Software Quality Assurance, Dissemination and exploitation activities as well as forming an integral part of the Data Management Plan of the project. Thus several groups and individuals are dealing with the issues and need to act with relatively high degree of autonomy. The overall coordination of the IPR issues is the responsibility of the Innovation Manager, who will ensure that the relevant IPR-related information is made available to everyone involved in the day-to-day IPR management.

1.8 Privacy issues

While the datasets used by the use cases and pilots are not planned to include data that would have privacy issues, any changes to this practice need to be reviewed by the Innovation Manager. The work package leader of WP9 will ensure that any contact information collected will be used only for the purposes the consent was obtained for (and in a way that is compliant with GDPR). WP9 leader will report on the privacy issues in the PEC meetings at the minimum twice a year.

2 Data Management Plan

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" (N.B. need to check the text hasn't changed in PROCESS GA before submitting the deliverable!)*

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 assets 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.

2.3 Potential Data assets

The potential reusable datasets will emerge from the following primary sources:

- 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 use case-related data assets will almost certainly represent the majority of the assets that are used. The technical platform development may develop tools and technologies that are e.g. used for testing, benchmarking or validation of the PROCESS solution. The primary data management approach will be based on the software quality process, leveraging as much as possible metadata and repository structure used for the software releases and link the services storing the physical datasets to the PROCESS software repository.

2.4 Common approaches to data management

As in all of the five use cases the foundation of the development is the use of existing datasets, the decision to store and publish a derived set is based on assessment of the potential value this derived dataset might represent for other users. The assessment is currently based on the following abstract "checklist" that will be refined and formalised during the project lifetime based on the experiences gained in its application:

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?

Regarding the last point, LMU has secured storage space for 20TB of raw data at LRZ, with a minimum commitment of providing managed, high-availability access at least for three years after the end of the project. It is assumed that if the datasets will be used, this time period can be extended, or the dataset migrated to a community-specific data repository.

PROCESS will aim at complying with the FAIR principles[8] with all of its data publication activities. Any deviations from these principles will be documented, together with the reasons for them (e.g. constraints imposed by the practices of the specific community).

2.5 Use-case specific data management aspects

The following sections will present *a priori* assessment of 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, so this deliverable will present only very brief summary of the potential reusable data assets and the possible approaches their use could be supported. This section is expected to be refined considerably in the future editions of the PROCESS DMP.

2.5.1 UC#1

Background datasets

The UC#1 uses the following published datasets as background material, each of them already published in repositories that are well known by the medical informatics community.

Table 3 Background datasets of use case 1

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	

Data generated

The UC#1 will generate two types of data assets of potential interest:

- Derived datasets based on one of the published ones (Camelyon17, Camelyon16 ...) that support more comprehensive training of machine learning algorithms. The methods include rotation of images, adding noise or simulated processing artefacts (e.g. foreign bodies like hairs in the scanned tissue slide).
- Actual neural networks trained by the datasets.

Publishing approach

In case of the derived datasets, it is likely that in most cases it would be most appropriate to publish the method for generating the derived dataset (software and artefact images). A small sample of trained networks could also be of interest.

[8] https://www.rd-alliance.org/fair-guiding-principles-scientific-data-management-and-stewardship.html

In the former case publishing the "recipe" for a derived dataset would ideally be done in the context of the original repository, whereas in the latter the software repository might be the most natural location.

2.5.2 UC#2

Background datasets

The work in the UC#2 will rely on accessing the data from the LOFAR Long Term Archive (LTA - https://lta.lofar.eu/). The publishable data assets would likely be a minimal set for validation testing of the software.

2.5.3 UC#3

Background datasets

The work in the UC#3 is based on using the UNISDR community as a pilot community for advanced PROCESS tools. The project will support the data management of the community, however for now the work does not result in generation of publishable datasets by the project.

2.5.4 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 compared to publishing of the generation algorithm.

2.5.5 UC#5

Background datasets

The datasets will be based on the Copernicus data service.

Data generated

Tools, documentation and parameter files for accessing Copernicus data, possibly specialised derived sets of data (e.g. time series of specific location).

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
3. Providers of generalised Copernicus access services

Evaluating (based on the experiences of the first pilot versions) which of these channels will be most promising channels for the project data assets will be one of the key focus areas of the update of this DMP.