D8.1 Data Management Plan

Document information table

Contract number:	800898
Project acronym:	ExaQUte
Project Coordinator:	CIMNE
Document Responsible Partner:	CIMNE
Deliverable Type:	Report
Dissemination Level:	PUblic
Related WP & Task:	WP 8 Task 8.1
Status:	Final version

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No **800898**

Authoring

Prepared by:				
Authors	Partner	Modifications	Version	Comments
Cecilia Soriano	CIMNE	create document		
Contributors				
Cecilia Soriano	CIMNE			
Riccardo Rossi	CIMNE			
Quentin Ayoul-Guilmard	EPFL			

Change Log

Versions	Modifications	Comments
1		First version
2		Review by EPFL
3- Final		Final review by WP leader

Approval

Approved by:				
	Name	Partner	Date	OK
Task leader	Cecilia Soriano	CIMNE	29/11/2018	OK
WP leader	Cecilia Soriano	CIMNE	29/11/2018	OK
Coordinator	Riccardo Rossi	CIMNE	29/11/2018	OK

Table of contents

Introduction 4

1. Data Summary 5

2. FAIR Data 11

3. Allocation of resources 13

4. Data Security 13

5. Ethical aspects 14

6. Other issues 14

Introduction

The ExaQUte project participates in the Pilot on Open Research Data launched by the European Commission (EC) along with the H2020 program. This pilot is part of the Open Access to Scientific Publications and Research Data program in H2020. The goal of the program is to foster access to research data generated in H2020 projects. The use of a Data Management Plan (DMP) is required for all projects participating in the Open Research Data Pilot, in which they will specify what data will be kept for the longer term. The underpinning idea is that Horizon 2020 beneficiaries have to make their research data findable, accessible, interoperable and re-usable (FAIR), to ensure it is soundly managed.

This initiative aims to improve and maximize access to and re-use of research data generated by Horizon 2020 projects and takes into account the need to balance openness and protection of scientific information, commercialization and Intellectual Property Rights (IPR), privacy concerns, security as well as data management and preservation questions.

Although open access to research data thereby becomes applicable by default in Horizon 2020, during the ORDP it applies primarily to the data needed to validate the results presented in scientific publications, although other data can also be provided by the beneficiaries on a voluntary basis

Data Management Plans (DMPs) are a key element of good data management, providing an analysis of the main elements of the data management policy that will be used by the consortium with regard to the project research data. A DMP describes the data management life cycle for the data to be collected, processed and/or generated by a Horizon 2020 project. As part of making research data findable, accessible, interoperable and re-usable (FAIR), a DMP should include information on:

- the handling of research data during and after the end of the project;
- what data will be collected, processed and/or generated;
- which methodology and standards will be applied;
- whether data will be shared/made open-access, and
- how data will be curated and preserved (including after the end of the project).

This document is the first version of ExaQUte project's DMP and has been elaborated within the first 6 months of the project. If significant changes arise during the course of the project (such as new data, changes in consortium policies, etc.), the DMP will have to be updated.

This DMP has been produced following the *Horizon 2020 FAIR Data Management Plan (DMP) template*, and includes thee following sections as suggested by the aforementioned guide:

1. Data Summary
2. FAIR Data
3. Allocation of resources
4. Data Security
5. Ethical aspects
6. Other issues

The ExaQUte Management Plan will be updated as the project progresses.

1. Data Summary

The ExaQUte project aims at constructing a framework to enable Uncertainty Quantification (UQ) and Optimization Under Uncertainties (OUU) in complex engineering problems, using computational simulations on Exascale systems. The methods and simulation tools developed in ExaQUte will be applicable to many fields of science and technology.

In particular, the chosen application focuses on **wind engineering**, a field of notable industrial interest. The problem to be solved has to do with the quantification of uncertainties in the simulation of the **response of civil engineering structures to the wind action**, and the shape optimization taking into account uncertainties related to wind loading, structural shape and material behavior.

The project entails the numerical simulations of heavy real engineering problems though the use of different codes and solvers that, given some input data, produce a file including the values of the relevant parameters that describe the results of the simulation of the original problem. Thus, the use and/or generation of large data sets is inherent to the nature of the project, making it very exigent regarding the amount of data involved.

Having said that, we have identified five main types of data sets that will be used and/or generated during the span of the project:

- data related to the management of the project (such as GA and CA documentation, review reports, rinutes of meetings, deliverables, papers in journals and communications in conferences, documentation of audits, etc.);
- data related to the geometry of the structure to be simulated;
- data produced as outcome of the numerical simulation;
- data for validation of the simulations;
- software.

Specific datasets may be associated to scientific publications, public project reports and other raw data or curated data not directly attributable to a publication. Datasets can be both collected, unprocessed data as well as analyzed, generated data.

Research data linked to exploitable results will not be put into the open domain if they compromise its commercialization prospects or have inadequate protection, which is a H2020 obligation. The rest of research data will be deposited in an open-access repository.

ExaQUte has created an intranet, organized under GitLab repository at `https://gitlab.com/principe/exaqute`, a snapshot of which is shown in Fig. (1). At the same time all the developments of ExaQUte will be integrated at the GitHub page of Kratos: `https://github.com/KratosMultiphysics/Kratos` (Fig. 4), which includes a wiki with the documentation of the project.

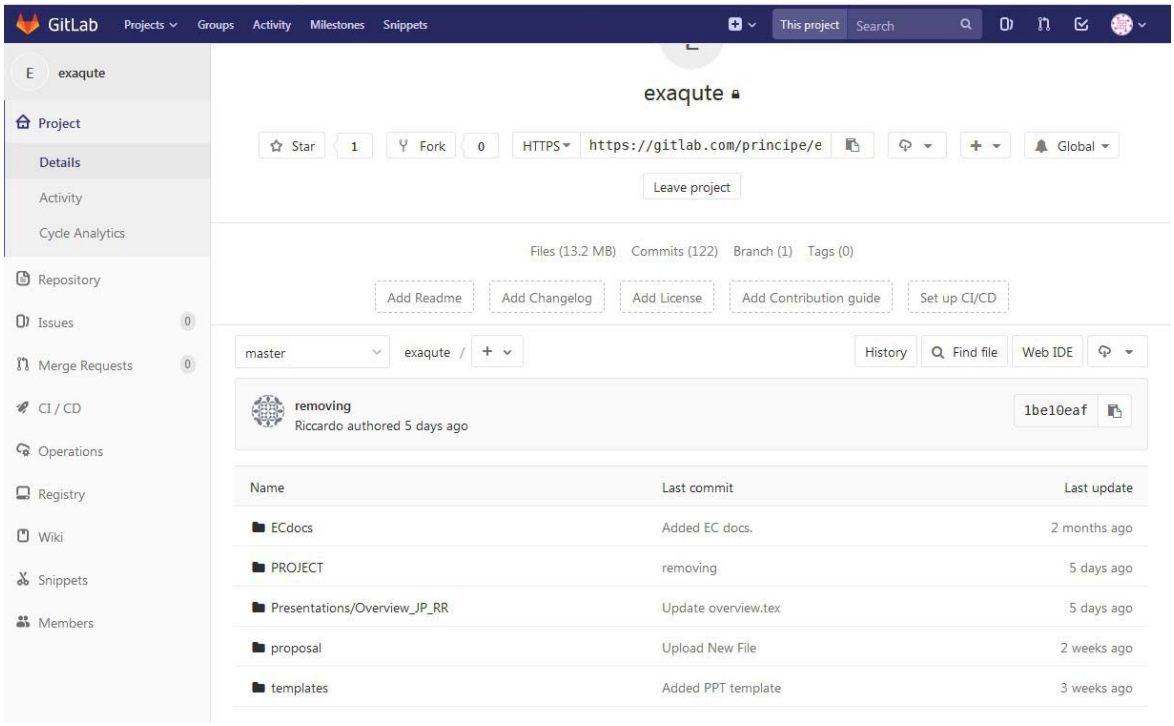

Figure 1: Git repository to share documents between partners

In parallel, PU documents related to this project will be uploaded to the ExaQUte customized repository created under the Open Science Platform Scipedia, available at `https://www.scipedia.com/institution/exaqute.eu`, a snapshot of which is shown in Fig. (2).

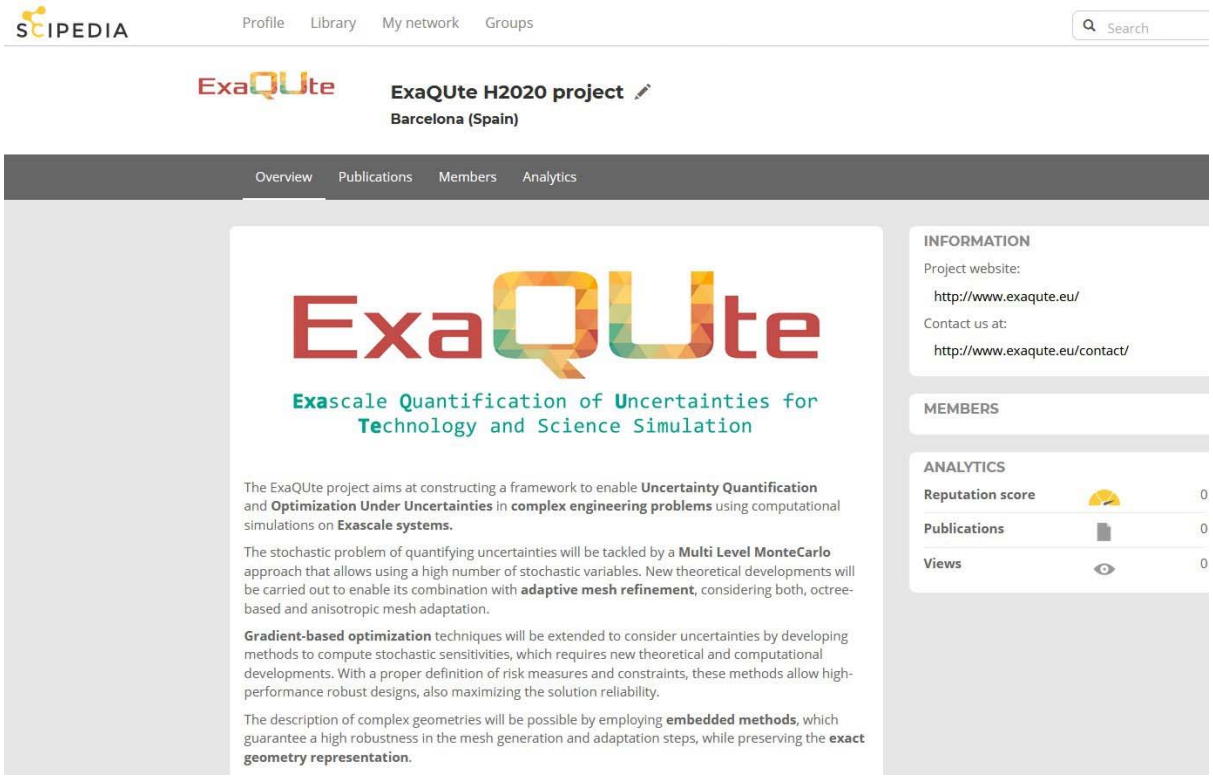

Figure 2: Scipedia repository to share open documents

The project has also created a dedicated webpage for ExaQUte (www.exaqute.eu) where all the public reports and deliverables will be uploaded as they are produced (Fig. 3)

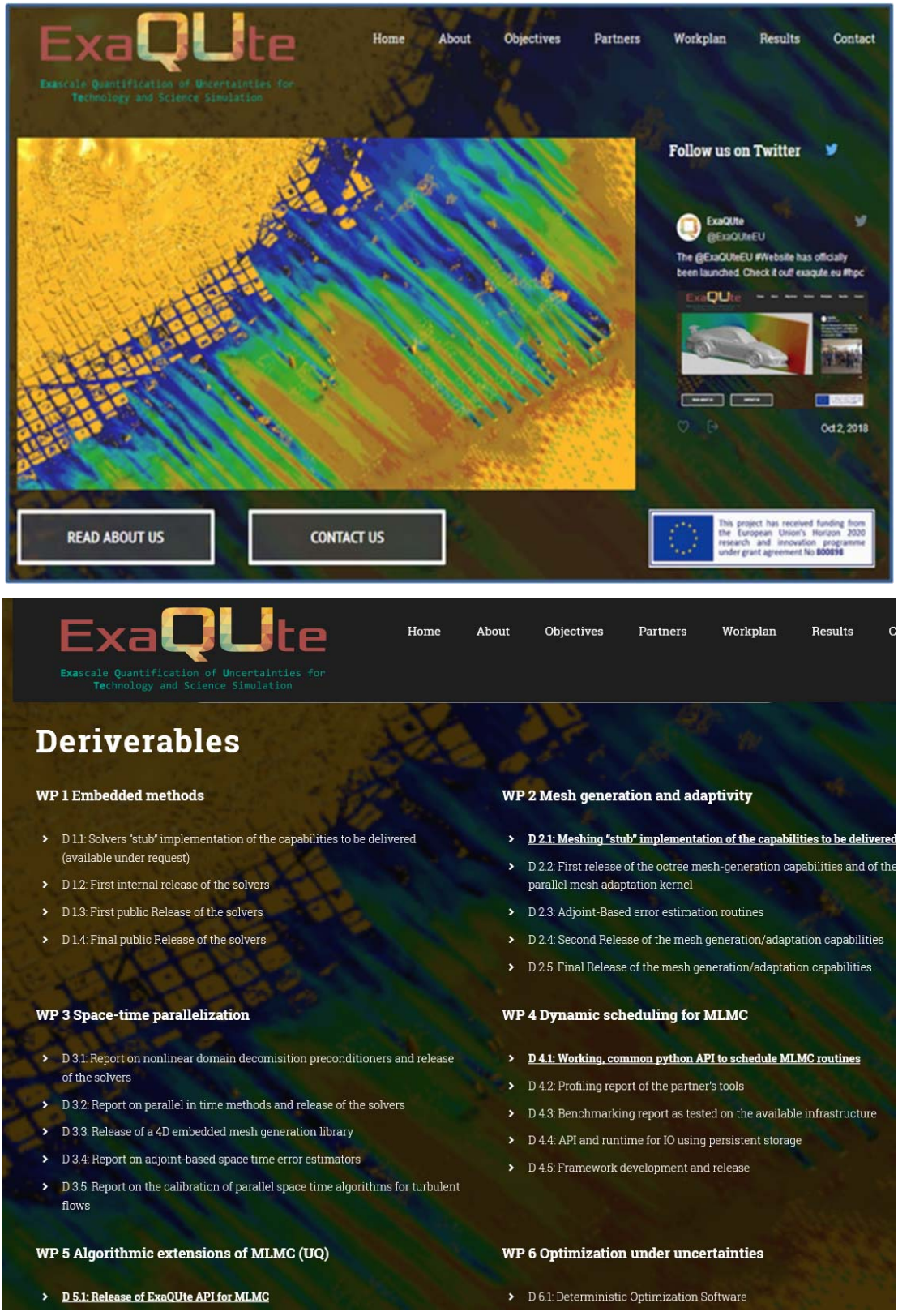

Figure 3: ExaQUte webpage with the list of deliverables to be uploaded as they are produced during the span of the project

All the code from Kratos is publicly available at the GitHub page: https://github.com/KratosMultiphysics/Kratos (Fig. 4). The same platform also includes a wiki with the documentation of the project. On this platform, all the developments of ExaQUte will be integrated.

Kratos adopts open standards for input and output formats, thus simplifying the exchange of data. In particular a JSON (Java Script Object Notation) format is employed in the definition of the parameters defining the simulation. Simulation results can be stored either in proprietary ".post.bin" format (which can be opened by the GiD software) or in HDF5 format.

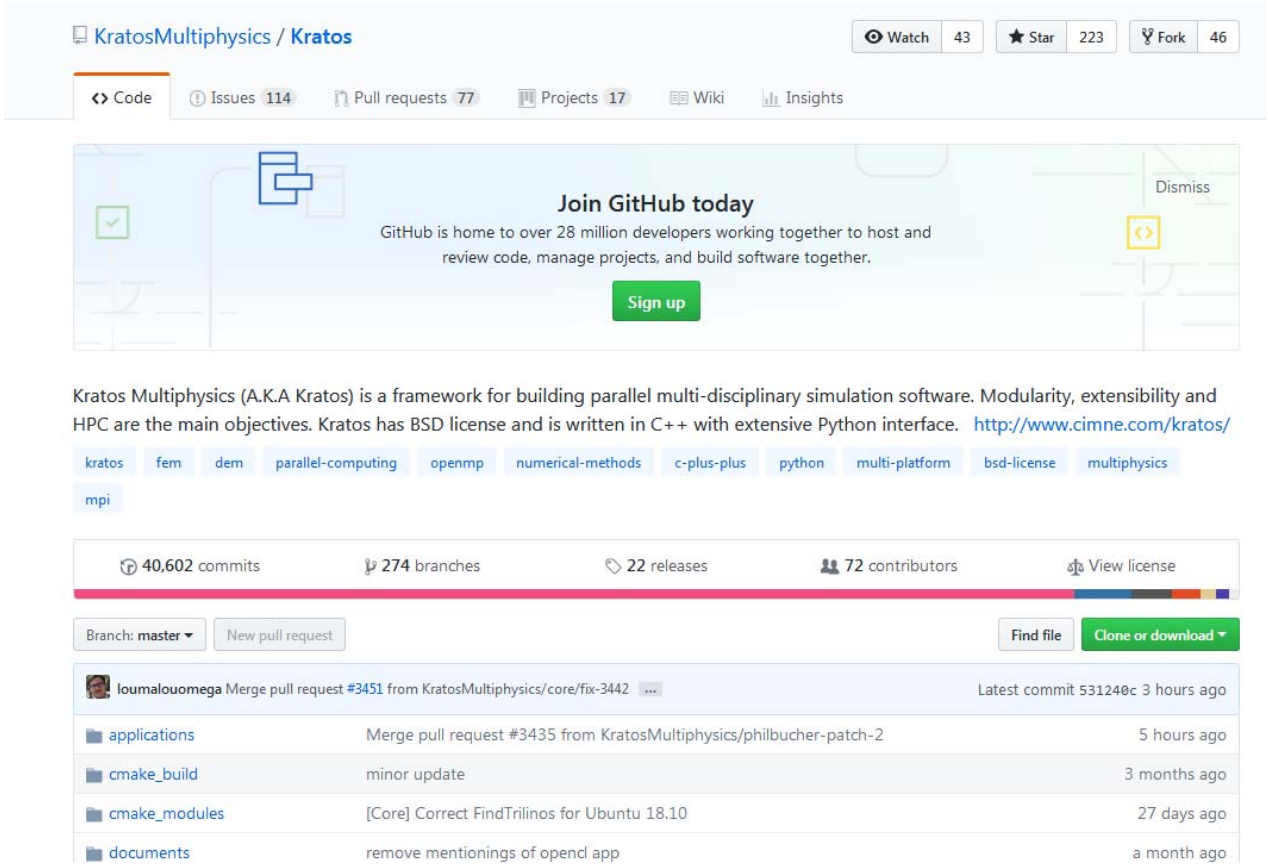

Figure 4: ExaQUte Code repository at GitHub

1.1. Documents and Dissemination material

Documents will consist of all the reports generated during the project, including all deliverables, publications and internal documents. Microsoft Word (DOCX) and PDF(preferred) and will be used for final versions, while intermediate versions can consider the usage of TeX (or LaTeX) files.

ExaQUte will produce dissemination material in a diversity of forms: flyers, newsletter, public presentations (DOCX, PPTX, PDF or OpenDocument formats), and videos demonstrating the performance of solvers, algorithms and plugins (widely used video file formats for distribution, such as MOV or AVI will be used)

We expect this data to be in the order of dozens of gigabytes, given the size of the videos (the lion's share of this type of data) to be included in the dissemination material

This data will be useful for those who want to learn about the outcomes of the project. From the point of view of Project Management, the documentation will be useful to EC officers and the consortium to assess the progress of the project.

Specific Provisions for Research publications:

Project Partners are responsible for the publication of relevant results to scientific community by Scientific Publications. The data (including associated bibliographic metadata) needed to validate the results presented in scientific publications will be deposited in a research data repository. This data is needed to validate the results presented in the deposited scientific publication and is therefore seen as a crucial part of the publication and an important ingredient enabling scientific best practice.

Metadata will maximize the discoverability of publications and ensure the acknowledgment of EU funding. Bibliographic data mining is more efficient than mining of full-text versions. The inclusion of metadata is necessary for adequate monitoring, production of statistics, and assessment of the impact of H2020.

In addition to basic bibliographic information about deposited publications, the following metadata information is expected.

- EU funding acknowledgement:
    - Contributor: "European Union (EU)" & "Horizon 2020".
- Peer Reviewed type (e.g. accepted manuscript; published version).
- Embargo Period (if applicable):
    - End date.
    - Access mode.
- Project Information:
    - Grant number: “800898”
    - Name of the action: “Research and Innovation action”
    - Project Acronym: “ExaQute”
    - Project Name: “EXAscale Quantification of Uncertainties for Technology and Science Simulation”
- Publication Date.
- Persistent Identifier:
    - Authors and Contributors. Wherever possible identifiers should be unique, non-proprietary, open and interoperable (e.g. through leveraging existing sustainable initiatives such as ORCID for contributor identifiers and DataCite for data identifiers).
    - Research Outcome
- License. The Commission encourages authors to retain their copyright and grant adequate licences to publishers. Creative Commons offers useful licensing solutions.

The ExaQUte project will support the open-access approach to Scientific Publications (as defined in article 29.2 of the Grant Agreement). Scientific publications covered by an editorial copyright

will be made available internally to the partners and shared publicly through references to the copyright owners' websites.

Whenever possible, a scientific publication, as soon as possible and at the latest six months after the publication time, will be deposited in a machine-readable electronic copy of the published version (or final peer-reviewed manuscript accepted for publication) in a repository for scientific publications. Moreover, the beneficiary should aim at depositing at the same time the research data needed to validate the results presented in the deposited scientific publications.

CIMNE (through its Spin-off Scipedia S.L.) has developed the Scipedia Publications repository, which is an open-access repository. The repository is indexed by Google and fulfills international interoperability standards and protocols to gain long-term sustainability.

1.2. Data related to the geometry of the structure to be simulated

Simulation geometries will be prepared using GiD or other CAD/Preprocessing software.

Exact geometries will be stored in the open format described in:

https://link.springer.com/article/10.1186/s40323-018-0109-4

The format employs a JSON notation and is hence readily readable and interchangeable.

Whenever possible (not proprietary geometries) geometries written in this format will be made available through the website.

1.3. Data produced as outcome of the numerical simulation

The ExaQUte project targets the solution of UQ and optimization problems by the use of variations of Monte Carlo techniques. This essentially implies running very many simulations and extracting statistical data from the outcome of each simulation sample.

For this very reason, the intermediate results are never stored, since it is preferred to generate and analyze new data on the fly rather than to store and later analyze the results.

The computation outcome, to be stored and made available to the end user, is thus a "normal" postprocessing output enriched with a statistical characterization of the results.

Kratos supports a multiplicity of formats for postprocessing output. Within ExaQUte, output to native GiD format and to the open HDF5 format will be used.

We note in any case that final results will be made available to the general public only for selected benchmarking cases. Outcome of other simulations would not be of interest to the general public.

1.4. Data for validation of the simulations

Validation data is typically available in the form of tables of data recorded by sensors or possibly as video footage.

Whenever possible, sensor input will be stored in HDF5 format so as to maximize its encapsulation and to make it portable. Videos will be stored using commonly available codecs.

JPG and PNG will be used to store static images.

Only our industrial partners (Str.ucture) could have some testing data that could be described as restrictive regarding their industrial interests, and thus that data could not be available to the

general public. However, they would make it available to the consortium when necessary, under their preferred conditions.

1.6 Software

ExaQUte produces open-source software, which can be readily downloaded and compiled from the source repository. Point releases corresponding to the deliverable (containing both a snapshot of the source and the compiled object for Linux64) will be made available through the project's GitLab account. Released software will be packaged in ZIP format.

The possibility of packaging the software so that it can be automatically installed as a Linux package or as a pip package will be explored. However no guarantee of success can be made in this sense.

2. FAIR Data

2. 1. Making data findable, including provisions for metadata

To facilitate discoverability (the degree to which something, especially a piece of content or information, can be found in a search of a file or database) of the data produced in the project, ExaQUte will establish a taxonomy for the data generated during the duration of the project.

The ExaQUte project will generate data resulting from the simulation results during the development of the different simulation tools and the final validation experiments. The data and associated software produced and/or used in the project should be discoverable (and readily located) and identifiable by means of a standard identification mechanism (e.g. Digital Object Identifier). This provision clearly refers to data designed for publication.

Produced data files, plugins and research data will be accompanied by a README file including who created or contributed to the data, its title, date of creation and under what conditions it can be accessed. Documentation will also include details on the methodology used, analytical and procedural information, any assumptions made, and the format and file type of the data. In the case of software, it may also include installation instructions and usage examples. All this information will be inside the manuscripts as well, unless structure of the document inhibits it (e.g. a journal/conference paper).

Releases are identified by the Git hash tag associated to the snapshot from which they were generated. The name also takes into account the type of compilation (Release, Debug, etc.). Such data can also be queried when launching the program, for example:

```
>>> from KratosMultiphysics import *
 |  /           |
 ' /   __| _` | __|  _ \   __|
 . \  |   (   | |   (   |\__ \
_|\_\_|  \__,_|\__|\___/ ____/
           Multi-Physics 6.0.0-17e3c693fe-FullDebug
```

In the case of manuscripts, the owner/responsible of the document will be the one controlling the version of the document, while files created by partners adding contributions to the original will be named by appending "_initials" to the filename.

2.2. Making data openly accessible

All documents and data that compromise neither IPR nor licensing rights will be available to the public on the different platforms and repositories described in Section 1.

Information about the modalities, scope and licenses (e.g. licencing framework for research and education, embargo periods, commercial exploitation, etc.) in which the data and associated software produced and/or used in the project is accessible should be provided.

The data and associated software produced and/or used in the project should be assessable by and intelligible to third parties in contexts such as scientific scrutiny and peer review (e.g. the minimal datasets are handled together with scientific papers for the purpose of peer review, data are provided in a way that judgments can be made about their reliability and the competence of those who created them).

2.3. Making data interoperable

Interoperability is the ability to access and process data from multiple sources without losing meaning, and then integrate that data for mapping, visualization, and other forms of representation and analysis.

The data and associated software produced and/or used in the project should be interoperable allowing data exchange between researchers, institutions, organisations, countries, etc. (e.g. adhering to standards for data annotation, data exchange, compliant with available software applications, and allowing re-combinations with different datasets from different origins).

2.4. Increase data re-use (through clarifying licenses)

Data re-use will be facilitated thought the repositories of the project.

The consortium has set up quality procedures for internal documents, deliverables and software. Publications are not considered in the procedure as they already go through an external refereed process.

Images and videos to be used, and those acquired in the project, will go through a natural quality control by the RTD partners as they will monitor that minimum quality requisites are obtained in the shootings to be able to run their algorithms. Quality of images and videos produced during the project will be assessed by end-user partners who will control that the obtained material is compliant with standards in the industry.

In the case of the software produced, the quality is guaranteed by several means: continuous integration performed by partners, and the integration of tests to confirm the right performances.

3. Allocation of resources

Each ExaQUte partner has to respect the policies set out in this DMP. Datasets have to be created, managed and stored appropriately and in line with applicable legislation. The Project Coordinator has a particular responsibility to ensure that data shared through the ExaQUte website are easily available, but also that backups are performed and that proprietary data are secured.

CIMNE, as Project Coordinator of ExaQute, will ensure dataset integrity and compatibility for its use during the project's lifetime by different partners.

Validation and registration of datasets and metadata is the responsibility of the partner that generates the data. Metadata constitutes an underlying definition or description of the datasets, and facilitates finding and working with particular instances of data.

Backing up data for sharing through open access repositories is the responsibility of the partner possessing the data. Quality control of this data is the responsibility of the relevant WP leader where the data was generated, supported by the Project Coordinator.

If datasets are updated, the partner that possesses the data has the responsibility to manage the different versions and to make sure that the latest version is available in the case of publicly available data. WP1 will provide naming and version conventions.

Last but not least, all partners must consult the concerned partner(s) before publishing data in the open domain that can be associated to an exploitable result.

All dissemination material produced during the project will be preserved and made public as soon as possible to let the research community know about ExaQUte solutions and results at the earliest date.

For the public reports and dissemination material, no much extra effort is considered for its preservation beyond the act of publishing them in public repositories (see Section 2.2).

It is agreed that this data has to be preserved a minimum of 3 years after the project's end.

4. Data Security

Storage and maintenance of ExaQUte data will be handled according to the data category, privacy level, need to be shared among the consortium, and size. This section covers the storage selections for data, independently of whether the data is to be shared externally. For that purpose, specific storage systems allowing public access will be selected.

Software data and source code will be stored on a **GitHub** server: a project management web application offering multiple-project support, version control (SVN and Git), issue tracking, file management, activity feeds, wiki and forums. Allowing installation on a partner's server is an important feature as it is a project requisite for internal sharing of software.

The use of Git guarantees that a distributed copy of all the data is available on all the computers who cloned the repository, thus removing the need for backup procedures.

Maintenance of datasets stored in partners' servers will be carried out according to the partners' backup policy.

We do not envision any sensitive data to be produced/transferred during ExaQUte. Only our industrial partners (Str.ucture) could have some restrictive data regarding their industrial interests,

that they would make available to the consortium when necessary and under their preferred conditions.

5. Ethical aspects

ExaQUte will neither make use of nor produce any type of data that could be described as either “sensitive” or raising any ethical issue.

6. Other issues

N/A