D1.3: Data Management Plan

Due Date	14 December 2018
Delivery	
Submission of updated version	N/A
Lead Partner	
Dissemination Level	Public
Status	Final
Approved	By the Executive Board
Version	V1.0

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 800925.

DOCUMENT INFO

Date and version number	Author	Comments
16.11.2018 v1.1	Xuanye Gu	First Draft
16.11.2018 v1.1	Hugh Martin	Updated the draft with changes
19.11.2018 v1.1	Xuanye Gu	Updated the draft with changes suggested by David Wright and Derek Groen
06.12.2018	Xuanye Gu	Updated the draft with changes suggested by the reviewers Tomasz Piontek, Emily Lumley and David Coster
10.12.2018	Xuanye Gu	Updated by the editor David Wright
11.12.2018	Xuanye Gu	Addressed Peter Coveney comments annotated in the final document

CONTRIBUTORS

- Robin Richardson
- David Wright
- Derek Groen
- Xuanye Gu
- Hugh Martin
- Peter Coveney

Disclaimer

The content of this deliverable does not reflect the official opinion of the European Union. Responsibility for the information and views expressed herein lies entirely with the author(s).

Table of Contents

1 EXECUTIVE SUMMARY

This deliverable, D1.3: Data Management Plan, acts as a detailed and comprehensive document on the data management plan that are being followed to guide the use of various type of data by the project. This deliverable is linked to VECMA's Work Package 1: Management that includes this deliverable.

This management plan is a 'living document' that will be updated throughout the project, as required.

2 DATA MANAGEMENT

2.1 Introduction

This deliverable responds to the standard questions that must be answered to produce an initial VECMA project data management plan. The data management plan presented in this document was produced using the DMP Online tool available at: https://dmponline.dcc.ac.uk/ [1] and follows the H2020 DMP template [2].

2.2 Data Summary

Provide a summary of the data addressing the following issues:

- **State the purpose of the data collection/generation**
- **Explain the relation to the objectives of the project**
- **Specify the types and formats of data generated/collected**
- **Specify if existing data is being re-used (if any)**
- **Specify the origin of the data**
- **State the expected size of the data (if known)**
- **Outline the data utility: to whom will it be useful**

The purpose of the VECMA project is to enable a diverse set of multiscale, multiphysics applications -- from fusion and advanced materials through climate and migration, to drug discovery and the sharp end of clinical decision making in personalised medicine -- to run on current multi-petascale computers and emerging exascale environments with high fidelity such that their output is "actionable". That is, the calculations and simulations are certifiable as validated (V), verified (V) and equipped with uncertainty quantification (UQ) by tight error bars such that they may be relied upon for making important decisions in all the domains of concern. The central deliverable will be an open source toolkit for multiscale VVUQ based on generic multiscale VV and UQ primitives, to be released in stages over the lifetime of this project, fully tested and evaluated in emerging exascale environments, actively promoted over the lifetime of this project, and made widely available in European HPC centres. All data collected, used and generated by the project is done in support of this objective.

VECMA is a large consortium, comprising not just funded core partners, but also a network of associate partners who seek to participate in the project's activities. As such the list of types and formats of data generated within VECMA will be extensive and dynamic, and includes but is not limited to:

- Formatted/unformatted text
- Mov
- MP4

- Binary
- HDF5
- Xlsx
- Jpg
- VTK
- PDB
- PSF
- PRMTOP
- XTC
- PDF
- PNG
- EPS
- DICOM
- C3D
- VTK

VECMA is not actively involved in assembling initial datasets and has a policy of using data brought into the project by project partners.

The data originates from many different sources. Non-simulation data, used to build models generally, can originate from clinical data management systems or DICOM image stores.

Simulation results are generated from computational models, with the focus of the project being on running these models on high performance computing resources around Europe. We distinguish at least three kinds of data: 1) to produce input for simulations, 2) to verify results of simulations and 3) results of simulations.

This exact extent of the data the project will need to store is unknown but anticipated to be in excess of 300TB in total.

The project includes a fast track that will ensure applications are able to apply available multiscale VVUQ tools as soon as they are available, while guiding the deep track development of new capabilities and their integration into a wider set of production applications by the end of the project. The deep track includes the development of more disruptive and automated algorithms, and their exascale-aware implementation in a more intrusive way with respect to the underlying and pre-existing multiscale modelling and simulation schemes. The data managed and produced within the project is of immediate use to alpha users and researchers in these areas, and in the longer term to industrial researchers and a wider scientific community across various domains. The data generated by the project will typically be generated by software and workflows developed in the project, and therefore correspond to specific versions of that software. In addition to our data management infrastructure, the project has developed a software repository, which acts as a central store of our project's software tools. We will use the metadata associated with data objects to reference the specific version of the code or workflow used to generate the data, using its software repository URL.

2.3 FAIR DATA

2.3.1 MAKING DATA FINDABLE, INCLUDING PROVISIONS FOR METADATA

Making data findable, including provisions for metadata:

- **Outline the discoverability of data (metadata provision)**
- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?**
- **Outline naming conventions used**
- **Outline the approach towards search keyword**
- **Outline the approach for clear versioning**
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how**

Much of the initial data, at least that used to build models, is held by the project partners as the result of other projects and research endeavours. As such, VECMA does not have control over how this data is published and made available.

Where data is generated by research conducted within the project, we will mandate that the final results of a simulation can be made discoverable. UCL has been a participant in the EUDAT and EUDAT2020 projects and became the first higher education institutional partner to join the EUDAT CDI. We will therefore leverage the best practice and services which EUDAT provides to make data discoverable (including the issuing of unique identifiers through the Handle system or Digital Object Identifiers). This will allow us to exploit the EUDAT B2FIND catalogue to make data keyword searchable. In addition to EUDAT resources we will also exploit local institution or other standard repositories where possible or mandatory to use. Versions of code associated with publications will additionally be uploaded to the Zenodo repository which can then be referenced by DOI in metadata.

The EUDAT Consortium follows the OpenAIRE guidelines for Data Archives by mandating standard minimal metadata and publication of metadata using the OAI-PMH protocol. Simulation results will be deposited in the B2SHARE service, and as such VECMA researchers will be compelled to provide a basic metadata record that complies with the OpenAIRE application of the DataCite Metadata Schema. In addition, data will be documented with a content- or discipline-specific metadata record. The data generated by the project will arise from a number of different interrelated fields, therefore not a single metadata standard will apply to all the cases, but we will work with data generators to identify suitable standards from the Research Data Alliance Metadata Standards Directory [3].

Where appropriate we will use established community metadata schemas (such as the Common Information Model, developed by ENES, for climate models). However, there are no general standards for multiscale models, so we propose that we will develop and internal VECMA schema to mandate a minimal set of metadata that must accompany all communicated datasets (the VECMA project name and grant number, application area, link to code used and version number where appropriate). All project related deposits will use keywords that clearly take into account all multi-disciplinary and multiscale aspects of the generating application.

2.3.2 MAKING DATA OPENLY ACCESSIBLE

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so**
- **Specify how the data will be made available**
- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?**
- **Specify where the data and associated metadata, documentation and code are deposited**

- **Specify how access will be provided in case there are any restriction**s

Data that relates to published work will be made available after a suitable embargo period (as defined by the relevant journal). Where specific data is identified as having legal, ethical or IPR barriers, the VECMA project will work with the data owners to identify whether the data can be made open after a period of embargo. We will make use of the features of EUDAT that allow depositors to choose to keep data private and apply embargo periods.

Data will be made openly available via the B2SHARE repository. This is a user-friendly, reliable and trustworthy way for researchers to store and share research data from diverse contexts. It guarantees long-term persistence of data and allows data, results or ideas to be shared worldwide.

All data hosted within the EUDAT CDI will be advertised through the central B2FIND catalogue and assigned a persistent identifier. The B2FIND service is a web portal allowing researchers to easily find and access collections of scientific data and allowing them to access the data using a web browser. As well as the metadata mandated by EUDAT, we will provide links to software used to generate the data (generally VECMA modelling tools), which are listed in the software catalogue featured on the VECMA project website.

VECMA intends to make use of the B2DROP service provided by EUDAT for sharing live data internally in the project, which will ease the transition of making data openly available in future. B2DROP is a tool to store and exchange data with collaborators and to keep data synchronized and up-to-date. VECMA will take advantage of the free storage space provided for research data within the B2DROP framework.

2.3.3 MAKING DATA INTEROPERABLE

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.**
- **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?**

In general, data used and created by the VECMA project is stored in standard formats such as DICOM and PDB. Data will be annotated with the metadata standards mandated by EUDAT when it is deposited, along with appropriate standard from the Research Data Alliance Metadata Standards Directory.

Because of the vast array of data types arising from the VECMA project, it is impossible to define a single interoperability standard, while the project does not have sufficient human resources available to enforce ontological annotation. However, we will produce guidance for researchers to annotate their data using popular ontologies such as SNOMED [4].

2.3.4 INCREASE DATA RE-USE (THROUGH CLARIFYING LICENSES)

- **Specify how the data will be licenced to permit the widest reuse possible**
- **Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed**

- **Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why**
- **Describe data quality assurance processes**
- **Specify the length of time for which the data will remain re-usable**

We expect core project partners to deposit their data openly using a Creative Commons version 4.0 licence or equivalent. Unless there is a publication requirement, IPR or data protection issue, we would expect data to be made available at the conclusion of the relevant work package within VECMA. We will also encourage our associate partners to adopt similar policies and promote these policies at VECMA training events.

The EUDAT B2SHARE service allows data shared openly or kept private. Regardless of whether deposited data are made open or kept private, metadata records submitted as part of a data deposit are made freely available for harvest via OAI-PMH protocols. Accessible data is made available directly to users of EUDAT CDI services through graphical user interfaces and application programming interfaces.

We will make published data available for third-party use as long as the EUDAT platform is able to host it.

The use of open standard formats, metadata annotation and workflow documentation (on the VECMA software portal) will be used to help ensure data quality prior to deposit.

2.4 ALLOCATION OF RESOURCES

Explain the allocations of resources, addressing the following issues:

- **Estimate the costs for making your data FAIR. Describe how you intend to cover these costs**
- **Clearly identify responsibilities for data management in your project**
- **Describe costs and potential value of long-term preservation**

As outlined in section 6, we will largely build on the services provided by the EUDAT project to make our data FAIR compliant. The lead partner UCL already pays a membership subscription to participate in the EUDAT CDI, which will be beneficial to the whole consortium, so we don't anticipate incurring any further costs to use these services.

Project data management is primarily the responsibility of individuals leading tasks that generate data within the project but is being overseen by the Project Technical Manager (Dr Derek Groen) and the Project Applications Manager (Dr Olivier Hoenen).

We will leverage facilities offered by EUDAT for the long-term presentation of data.

UCL has previously developed a relationship with EUDAT data nodes RZG and EPCC to provide long term B2SHARE and B2SAFE provision, which we will aim to make use of in this project. PSNC, one of VECMA consortium partners, is also a member of EUDAT. PSNC will provide all the physical storage required for the project for the partners to store their data such as simulation results. To facilitate this, PSNC has asked each partner to provide what storage size is needed and they will allocate a total physical storage to be available for all partners.

2.5 DATA SECURITY

Address data recovery as well as secure storage and transfer of sensitive data

Internally within the project, file-based data will be shared using the B2DROP service, which uses the HTTPS protocol for secure transfer. Other types of data, such as DICOM image data, will be stored at a data centre at UCL, making use of the access control and secure transfer features provided by the service in question, and taking advantage of UCL's central data centre management policies. Other partners including HPC centres, PSNC and LRZ have considerable data storage, some of them free of charge. Therefore, the VECMA project can use storage resources provided from both EUDAT, PSNC storage and LRZ storage.

Data shared and published via the EUDAT CDI will be stored at one or more partner sites, according to applicable service level agreements and policies. Backup of data is performed at two levels using the B2SAFE service: multiple replicas of data are stored at different sites (i.e. geographically and administratively different); and data may additionally be backed up at an individual site. Responsibility for the storage and backup at any individual site lies with the designated site manager.

All EUDAT CDI core sites are large, national or regional data and computing centres and operate according to good IT governance and information security principles. Some sites are accredited through the ISO 27001 information security process and/or have certifications of trustworthiness such as the Data Seal of Approval, while others are working actively towards it.

2.6 ETHICAL ASPECTS

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

VECMA does not actively collect data from individuals, and simulation scenarios are largely based on publicly obtainable/consented data that has been provided to project partners. For further details on how the project is handling any ethical issues which arise refer to VECMA D4.5 Ethics Report [5].

Regarding data governance, VECMA is not intended as a facility for the routine processing of live, identifiable clinical data; it will operate in the research domain, and all data introduced by users will be required by the VECMA conditions of use to be pre-processed to render it non- personal, and so excluded from consideration under current and anticipated future research governance regulations. VECMA will however act as a Data Controller for the information relating to the registration and access control of its users, and such data will be handled in full accordance with appropriate pan-European legislation.

Regarding data ethics, the VECMA framework is designed to support independent users in their access to large-scale computational facilities and does not carry out patient-related research; as a, consequence the VECMA project does not itself acquire or handle patient-specific clinical data. Rather, it enables users to work with models, applications and data for which they are responsible, in the pursuit of their own research goals. Users sharing data must do so under the terms granted by the data's original ethical sanction, and again users will be required by the VECMA conditions of use to reach documented agreement that the terms of ethical sanction have been met. It is the case

however that ultimately VECMA cannot take responsibility for the provenance or ethical compliance of data share through its infrastructure, nor can it take account of the diverse legislation and the variable interpretation of European directives that may occur in the various Member States.

However, situations may arise where VECMA will have access to clinical data. In these situations, VECMA should be considered a *Data Manager*, which is delegated by the *Data Provider* (typically a hospital) to handle clinical data, for which the data provider has received from the *Data Owner* (the patient) the necessary permission to allow the treatment to be accessed by one or more *Data Consumers* (typically modelling experts) in order to fulfil a certain treatment scope. In order to be legally compliant, clinical data require two things: the permission to treat from the data owner (the patient), and an adequate protection of confidentiality. This in turn implies:

A: VECMA can handle only clinical data for which access has been granted. All users are fully responsible for ensuring that the necessary permission has been acquired. VECMA will assist not-for-profit users such as research hospitals or universities by providing them with informed consent templates (written by an expert) that provide the type of permission necessary for a given treatment using the project's tools and services.

B: Full anonymisation: when the processing of the data does not require the distinguishing of one individual patient from another, if necessary VECMA will provide a server, to be installed behind the hospital firewall, that will automate the replication of selected data to VECMA storage, while providing automated semantic annotation according to popular ontologies, and irreversible anonymisation according to agreed rules. This server will be managed by the hospital staff.

C: Pseudo-anonymisation via a trusted third party: if the identity of the patient cannot be entirely removed (for example, for personalised clinical treatment), the type of infrastructure is the same as (B) above but this time the data are annotated with a PatientID that remains, within the safety of the hospital secure network, associated with the patient's actual identity.

3 OTHER

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)
N/A

4 CONCLUSIONS

This data management plan will help VECMA project partners to identify the correct decisions that must be made regarding the use our data throughout our project. The plan is a living document and it will be updated at various project stages, as required.

5 ANNEXES

Acronyms and Definitions

Acronyms	Definitions

CDI	Collaborative Data Infrastructure
DICOM	Digital Imaging and Communications in Medicine (DICOM) is a standard for handling, storing, printing, and transmitting information in medical imaging
EUDAT	An EU funded collaborative data infrastructure
OpenAIRE	The Open access infrastructure for research in Europe
ENES	European Network for Earth System Modelling
FAIR (data)	To make data Findable, Accessible, Interoperable, and Reusable

6 REFERENCES

[1] Data management - H2020 Online Manual - European Commission. Link to the manual: http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm.

[2] H2020 templates: Data management plan v1.0 – 13.10.2016.

[3] Research Data Alliance Metadata Standards Directory: https://www.rd-alliance.org/groups/metadata-standards-directory-working-group.html .

[4] SNOMED: https://www.rd-alliance.org/groups/metadata-SNOMED-directory-working-group.html .

[5] VECMA D4.5 Ethics Report.