Grant agreement no. 823712

CompBioMed2

Research and Innovation Action
H2020-INFRAEDI-2018-1
Topic: Centres of Excellence on HPC

D 3.1 Data Management Plan

Work Package: 3

Due date of deliverable: M4

Actual submission date: 31 January 2020

Start date of project: 01 October 2019 Duration: 48 months

Lead beneficiary for this deliverable: *LRZ*
Contributors: *LRZ, UCL*

Disclaimer

Project co-funded by the European Commission within the H2020 Programme (2014-2020)		
Dissemination Level		
PU	Public	YES
CO	Confidential, only for members of the consortium (including the Commission Services)	
CI	Classified, as referred to in Commission Decision 2001/844/EC	

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **823712**"*

Table of Contents

1 Version Log 3
2 Contributors 3
3 Definition and Acronyms 4
4 Introduction 5
5 Data Summary 5
6 FAIR Data 6
6.1 Findable Data 6
6.2 Accessibility 7
6.3 Interoperability 8
6.4 Reuse 9
7 Allocation of Resources 9
8 Data Security 10
9 Ethical Aspects 10
10 Other 12

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **823712**"*

1 Version Log

Version	Date	Released by	Nature of Change
V0.1	13.01.2020	Gerald Mathias	First Draft
V0.2	24.01.2020	Gerald Mathias	Revised Draft
V1.0	31.01.2020	Emily Lumley	Final Draft, submitted to the EC

2 Contributors

Name	Institution	Role
Gerald Mathias	LRZ	Principal Author
David Wifling	LRZ	Co-Author
Marco Verdicchio	SURFsara	Reviewer
Mat Bieniek	UCL	Reviewer
Peter Coveney	UCL	Reviewer
Emily Lumley	UCL	Reviewer

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **823712**"*

3 Definition and Acronyms

Acronyms	Definitions
CDI	Collaborative Data Infrastructure
CoE	Centre of Excellence
DICOM	Digital Imaging and Communications in Medicine (DICOM) is a standard for handling, storing, printing, and transmitting information in medical imaging
DMP	Data Management Plan
DoA	Description of the Action
EUDAT CDI	An EU funded Collaborative Data Infrastructure
FAIR data	Findable, Accessible, Interoperable & Reusable Data
HDPA	High performance data analytics
HPC	High Performance Computing
IPR	Intellectual Property Rights
MIBBI	Minimum Information for Biological and Biomedical Investigations
OAI-PMH	The Open Archives Initiative Protocol for Metadata Harvesting
OpenAIRE	The Open access infrastructure for research in Europe
PDB	Protein Data Bank
VECMA	Verified Exascale Computing for Multiscale Applications (EU Project)
VVUQ	Verification, Validation, and Uncertainty Quantification

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **823712**"*

4 Introduction

This deliverable responds to the standard questions that must be answered to produce an initial H2020 data management plan (DMP). The DMP presented in the remainder of this document was produce using the DMPOnline tool available at: https://dmponline.dcc.ac.uk/.
It is based on the DMP of the predecessor Centre of Excellence (CoE), CompBioMed.

5 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**

CompBioMed2 is the second phase of the Computational Biomedicine Centre of Excellence (CoE), CompBioMed, an outward facing CoE comprising members from academia, industry and the healthcare sector. Its core purpose is to facilitate the uptake and exploitation of HPC-based Computational Biomedicine simulation approaches to a greater number of therapeutic areas.

Understanding the complex outputs of such simulations requires the convergence of High Performance Computing (HPC) and high-performance data analytics (HPDA). CompBioMed2 seeks to combine these approaches with the large, heterogeneous datasets from medical records and from the experimental laboratory to underpin clinical decision support systems. All data collected, used and generated by the project is done in support of this objective

The project's three core research strands focus on the areas of cardiovascular, musculoskeletal and molecular modelling. Each of these scientific communities and associated software packages employ different data formats. Major types of data comprise:

- **Imaging data** which stem mostly from clinical trials, experiments, or visualize results of the simulations. They may serve as initial data for data analytics and machine learning tasks. Typical file formats used for these data are JPG, PNG, DICOM, MP4, and MOV.
- **Other clinical and experimental data** which serve as reference for simulations. Here a broad variety of file formats are used from formatted/unformatted plain text, PDF and DOCX files, tabulated data formats like CSV and XLSX, as well as (raw) binary data.
- **Musculoskeletal data** which record the motion of joints, bones, muscles, etc. They are the key output of musculoskeletal simulations and use file formats such as C3D and XMDF.
- **Cardiovascular data** resulting from heart and blood flow simulations in the project are mostly recorded in HDF5 and VTK file formats.
- **Molecular modelling data**: Structures of complex biomolecules, assemblies thereof, smaller molecules and molecular dynamics trajectories serve as initial conditions or are output of simulations conducted in the project and may be targets of HDPA approaches. Most frequently used file formats are PDB, PSF, XTC, TRR. A variety of tools is available that are able to read and convert these different formats.

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **823712**"*

ComBioMed2 is a large Centre of Excellence, comprising not just funded core partners, but also a growing network of associate partners who seek to participate in the Centre's activities and bring in their own data and data formats. Thus, the above list is long but by no means complete.

CompBioMed2 builds on the data stock of the predecessor CoE, CompBioMed, but it is not actively involved in assembling initial datasets. These are brought to the project by its partners, e.g. from the cardiovascular, molecularly-based and neuro-musculoskeletal exemplary research, tasks 2.1-2.3, and the "Analytics for in silico augmented clinical trials", task 3.6. The data originate from many different sources: Non-simulation data, used to build models generally, can originate from clinical data management systems or DICOM image stores or from biochemical structure databases such as Protein Data Bank (PDB; http://www.rcsb.org/pdb/). Simulation results are generated from computational models that are run on HPC resources around Europe.

The size of the data the project will need to store largely depends on the time frame and the level of postprocessing and/or compression applied. A sufficient volume of immediate storage for input and output data is typically provided by the HPC centre, where the simulations are conducted. Long-term storage or archive systems are usually based on tape media, which are cheaper comparing to disk storages and therefore have less limitations on the size of data they store. The tape storage however, imposes a latency in retrieving the data and therefore limits its accessibility. We estimate the total size of data that will be generated by the project during its lifetime to exceed 4 PB, but much less may be required to be stored by CompBioMed2.

The data managed and produced within the project is of immediate use to medics, clinicians and researchers the key areas of the CompBioMed2 CoE, and in the intermediate term to industrial researchers and drug/medical device manufacturers and ultimately to patients. The data produced by the project will typically be generated by software and workflows developed in the project, and therefore correspond to specific versions of that software. Task 3.4 of the CompBioMed2 project is dedicated to data curation. It will promote the reuse of data, through the application of annotation and metadata, which will make the data more amenable to analytics technologies and make data more easily found and reused by other researchers.

6 FAIR Data

Core partners that produce new data within simulations and the analysis of existing data are urged to publish their data according to FAIR (Findable, Accessible, Interoperable, and Re-usable) principles. With respect to the associate partners, CompBioMed does not have control over how this data is published and made available. Nevertheless, CompBioMed will encourage partners to follow the FAIR principles. These efforts are promoted by task 3.4 "Data Curation" of the project, which will offer support to the core and associate partners of the CompBioMed2 CoE.

6.1 Findable Data

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**

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **823712**"*

- **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**

For data generated by research conducted within the project, we will mandate that the final results of any simulation can be made discoverable. UCL has been and is a participant in the EUDAT and EUDAT2020 projects, and has joined the EUDAT Collaborative Data Infrastructure (CDI) among other CompBioMed2 partners, such as BSC, EPCC, and SURFsara. EUDAT CDI operates the important platforms B2DROP and B2SHARE to upload, share and publish data, as well as B2FIND, which provides a search engine to find stored data. We will therefore leverage the best practice and services which EUDAT provides to make data discoverable.

The EUDAT Consortium follows the OpenAIRE guidelines for Data Archives by mandating standard minimal metadata and publication of metadata using the OAI-PMH protocol. We have already added CompBioMed as a community to B2SHARE in which data can be published and searched later based on the community. For now, we are using the generic metadata schema which is based on the Dublin Core Schema. Ideally, we would want to have a community-specific metadata schema defined and implemented in B2SHARE so that the CompBioMed2 researchers can publish their data.

Data that is being published in B2SHARE will be harvested by B2FIND, which is a metadata portal for data discovery. 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, which will include PDBx/mmCIF and MIBBI.

All the data that we publish via the EUDAT B2SHARE data repository service will be assigned a persistent identifier through the Handle system. This unique dataset reference can be used in data citations and to enhance discoverability. EUDAT B2SHARE provides an automatic versioning with ascending numbers and time indices that allows a clear identification of the dataset version.

This will allow us to exploit the EUDAT B2FIND catalogue to make data keyword searchable, either via the B2FIND web interface the application programming interfaces OAI-PMH, JSON-API, CSW2.0 as outlined in http://b2find.eudat.eu/guidelines/harvesting.html.

6.2 Accessibility

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 restrictions**

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **823712**"*

In general, all data that relates to published work and has been generated within CompBioMed2 will be made available after a suitable embargo period (as defined by the relevant journal). However, legal, ethical or IPR barriers may prohibit specific data to be published with unrestricted access. Here, the CompBioMed2 project will work with the data owners to identify whether the data can be made open after a period of embargo or if other measures can be applied to alleviate such restrictions. We will make use of the features of EUDAT that allow depositors to choose to keep data private, password protected, or to 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. CompBioMed2 has implemented the dedicated task 3.2 “Data Storage System Deployment and Maintenance” for this purpose.

All CompBioMed2 data published in the EUDAT B2SHARE data repository service, 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 metadata on scientific data, and redirect the users to the origin of the data. Additionally, the metadata will contain references to the software and version that was used to generate the data, which are generally CompBioMed2 modelling tools. A comprehensive list of these tools is featured on the CompBioMed2 project website at https://www.compbiomed.eu/services/software-hub/.

CompBioMed2 will also make use of the B2DROP service provided by EUDAT for sharing live data internally in the project. B2DROP is an online cloud storage tool to store and exchange data with collaborators and to keep data synchronized and up-to-date. This will ease the transition to making data openly available in future, as B2DROP is linked to the B2SHARE data repository. CompBioMed2 will take advantage of the free storage space provided for research data within the B2DROP framework.

6.3 Interoperability

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 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 within the CompBioMed2 CoE, it is impossible to define a single interoperability standard, and the project does not have sufficient resources available to enforce ontological annotation. However, task 3.4 “Data Curation” will supply guidance for researchers to annotate their data using popular ontologies.

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Grant Agreement No **823712**“*

6.4 Reuse

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 expect data to be made available at the conclusion of the relevant tasks of CompBioMed2. In the course of the project we will regularly inquire the individual tasks which data has been published and offer help with metadata annotation. We will also encourage our associate partners to adopt similar policies, and promote these policies at CompBioMed2 training events.

Data reuse is also promoted within the project, when data generated by simulations is analysed with deep learning and machine learning techniques by different tasks across work packages. Similarly, these data may be of interest and could be analysed by related CoEs focusing on computational biomolecular research such as BioExcel2 (https://bioexcel.eu/).

The EUDAT B2SHARE service allows data to be 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 B2DROP platform is able to host it. If relevant data exceeds this limit, we will explore other possibilities within task 3.2 “Data Storage System Deployment and Maintenance”. The use of open standard formats, metadata annotation and workflow documentation (on the CompBioMed2 software portal) will be used to help ensure data quality with respect to documentation and completeness. For the reliability of the data generated by simulations, CompBioMed2 will closely collaborate with the VECMA project to provide VVUQ for the methods used. Here, VVUQ stand for **V**alidating that the model underlying the simulation is sufficiently accurate, **V**erifying that the model is implemented correctly and **Q**uantifying the statistical **U**ncertainty to e.g. the finite time span of the simulation or the limited resolution of the model. All these points are fundamental for the quality and reliability of the data.

7 Allocation of Resources

Explain the allocation of resources, addressing the following issues:

- **Clearly identify responsibilities for data management in your project**
- **Describe costs and potential value of long term preservation**

As outlined in section 0, we will largely build on the services provided by the EUDAT project to make our data FAIR compliant. EUDAT CDI services generally come for free. Furthermore, the

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Grant Agreement No **823712**“*

CompBioMed2 partners UCL, BSC, EPCC, and SURFsara are members of EUDAT, 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. With task 3.4, CompBioMed2 has allocated dedicated PMs to support core and project partners enforcing FAIR principles.

We will leverage EUDAT B2SHARE for the long-term preservation of data. In addition, SURFsara is a partner in the EUDAT consortium, leading the work package that develops and maintains the B2SAFE EUDAT service. Furthermore, the HPC centres, which provide compute time allocations for CompBioMed2 projects, offer tape archives for long term storage of large data sets; such providers are SURFsara, EPCC, BSC, and LRZ. Currently, no additional costs arise for CompBioMed2.

8 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 EUDAT B2DROP and B2STAGE services for secure transfer. Other types of data, such as DICOM image data, will be stored at data centres of the core partners such as UCL and UOXF, 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. Support for the project partners is implemented in task 3.3 "Data Staging Systems" of CompBioMed2.

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.

9 Ethical Aspects

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

CompBioMed2 as a CoE 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 (core and associate, as well as via other collaborations).

Regarding data governance, CompBioMed2 is not intended as a facility for the routine processing of live, identifiable clinical data; it operates in the research domain, and all data

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **823712**"*

introduced by users will be required by the CompBioMed2 conditions of use to be pre-processed to render it non-personal, and so to be excluded from consideration under current and anticipated future research governance regulations. CompBioMed2 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 CompBioMed2 framework is designed to support independent users in their access to large-scale computational facilities. 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 CompBioMed2 conditions of use to reach documented agreement that the terms of ethical sanction have been met. It is the case, however, that ultimately CompBioMed2 cannot take responsibility for the provenance or ethical compliance of data shared 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. Some tasks within CompBioMed2 will access and analyse clinical data. Here, the corresponding CompBioMed2 task 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:

1. CompBioMed2 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. CompBioMed2 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 CoE's tools and services.

2. Full anonymisation: when the processing of the data does not require the distinguishing of one individual patient from another, if necessary CompBioMed2 will provide a server, to be installed behind the hospital firewall, that will automate the replication of selected data to CompBioMed2 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.

3. 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 (2) 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.

Any further ethical and corresponding legal issues regarding handling of sensitive data that may arise during the project will be addressed by the "Ethics Panel" implemented within the task 3.7 "Ethics" of work package 3 "Data Management and Analytics."

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **823712**"*

10 Other

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

N/A

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **823712**"*