Grant agreement no. 675451

CompBioMed

Research and Innovation Action
H2020-EINFRA-2015-1
Topic: Centres of Excellence for Computing Applications

D1.3 Data Management Plan

Work Package: 1

Due date of deliverable: Month 06

Actual submission date: 31 / March / 2017

Start date of project: October, 01 2016 Duration: 36 months

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

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 **675451**"*

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

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

1 Version Log

Version	Date	Released by	Nature of Change
V1.0	8/3/2017	Stefan Zasada	First draft
V1.1	20/3/2017	Marco Verdicchio	Review comments
V1.2	24/3/2017	Stefan Zasada	Draft with Matt Harvey review comments
V1.3	24/3/2017	Peter Coveney	Final review commands

2 Contributors

Name	Institution	Role
Hugh Martin	CBK	Author
Stefan Zasada	UCL	Editor
Marco Verdicchio	SARA	Reviewer
Matt Harvey	ACE	Reviewer
Peter Coveney	UCL	Reviewer

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

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
EUDAT CDI	An EU funded Collaborative Data Infrastructure
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

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

4 Introduction

This deliverable answers the standard questions that must be answered to produce an initial H2020 data management plan. The data management plan presented in the remainder of this document was produce using the DMPOnline tool available at: https://dmponline.dcc.ac.uk/.

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

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**

The CompBioMed project seeks to exploit the third pillar of science in order to render predictive models of health and disease more relevant to clinical practice by providing a personalized aspect to treatment. CompBioMed exists to facilitate the uptake and exploitation of high performance computing within the biomedical modelling community, in order to bring benefits to many areas, from education and training of the next generation of computational biomedicine researchers, through new access mechanisms best suited to the community, to the impact of the methods on healthcare delivery within industry and the clinical context, especially in advancing personalised and precision medicine.
All data collected, used and generated by the project is done in support of this objective.

ComBioMed is a large consortium, comprising not just funded core partners, but also a growing network of associate partners who seek to participate in the Centre's activities. As such the list of types and formats of data generated is long, 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

CompBioMed is not actively involved in assembling initial datasets, and has a policy of using data brought to 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.

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

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.

The exact size of the data the project will need to store is unknown, but anticipated to be in excess of 2PB overall.

The project's three core research strands focus on the areas of cardiovascular, musculoskeletal and molecular modelling. The data managed and produced within the project is of immediate use to clinicians and researchers in these areas, 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. In addition to our data management infrastructure, the project has developed a software repository[1], 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.

In addition, the project's associate partners are bringing further simulation scenarios to the centre, and the data generated through these simulation scenarios will be of use to similar groups of stakeholders.

[1] http://www.compbiomed.eu/software-hub/

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

6 FAIR Data

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**
- **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, the results of other projects and research endeavours. As such, CompBioMed 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 and is a participant in the EUDAT and EUDAT2020 projects, and is the first institutional partner to join the EUDAT CDI. 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. Simulation results will be deposited in the B2SHARE service, and as such CompBioMed 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, which will include PDBx/mmCIF and MIBBI.

All the data that we host within the EUDAT Collaborative Data Infrastructure (CDI) will be assigned a persistent identifier through the Handle system. This unique dataset reference can be used in data citations and to enhance discoverability, and it includes clear versioning capabilities which we will leverage.

This will allow us to exploit the EUDAT B2FIND catalogue to make data keyword searchable.

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**

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

- **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**

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 CompBioMed 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 CompBioMed modelling tools), which are listed in the software catalogue featured on the CompBioMed project website.

CompBioMed will also make use of the B2DROP service provided by EUDAT for sharing live data internally in the project, which will ease the transition to 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. CompBioMed 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 the CompBioMed project, it is impossible to define a single interoperability standard, and the project does not have sufficient resources available to enforce ontological annotation. However, we will produce guidance for researchers to annotate their data using popular ontologies such as SNOMED.

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

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 would expect data to be made available at the conclusion of the relevant work package within CompBioMed. We will also encourage our associate partners to adopt similar policies, and promote these policies at CompBioMed training events.

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 platform is able to host it.

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

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

7 Allocation of Resources

Explain the allocation 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. UCL is paying a membership subscription to participate in the EUDAT CDI, 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. Stefan Zasada).

We will leverage facilities offered by EUDAT for the long-term presentation of data.
UCL has previously developed a relationship with the 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. In addition, SURFSara is a partner in the EUDAT consortium, leading the work package that develops and maintains the B2SAFE EUDAT service.

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

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 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.

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.

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

9 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

CompBioMed 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.

Regarding data governance, CompBioMed 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 CompBioMed 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. CompBioMed 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 CompBioMed 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 CompBioMed 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 CompBioMed conditions of use to reach documented agreement that the terms of ethical sanction have been met. It is the case, however, that ultimately CompBioMed 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 CompBioMed will have access to clinical data. In these situations, CompBioMed 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. CompBioMed 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. CompBioMed 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 CompBioMed will provide a server, to be installed behind the hospital firewall, that will automate the replication of selected data to CompBioMed 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.

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

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.

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

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 **675451**"*