Data Management Plan

Modelling and Orchestrating heterogeneous Resources and Polymorphic applications for Holistic Execution and adaptation of Models In the Cloud

H2020-ICT-2018-2020
Leadership in Enabling and Industrial Technologies: Information and Communication Technologies

Grant Agreement Number
871643

Duration
1 January 2020 –
31 December 2022

www.morphemic.cloud

Deliverable reference
D9.1

Date
1 May 2020 – 30 June 2020

Responsible partner
University of Piraeus Research Center

Editor(s)
Alexandros Raikos, Dimosthenis Kyriazis, Jean-Didier Totow

Reviewers
Chris Kachris, Jørgen Kirksæther

Distribution
Public

Availability
morphemic.cloud

Executive summary

This report reflects the deliverable of Task 9.1 “Data Management” and describes the Data Management Plan (DMP) of the MORPHEMIC project. The purpose of the DMP is to present a plan for the management, collection, generation, storage, protection, sharing and preservation of data related to all MORPHEMIC activities and to clarify the project’s data management policy in agreement with the MORPHEMIC consortium.

The MORPHEMIC DMP complies with the guidelines specified in the European Commission's Data Management Template. It reflects the status of the data that is collected, processed, or generated, and the respective methodology and standards, whether and how these data will be shared and made open, and how it will be curated and preserved.

The DMP defines the general policy and approach to data management in the MORPHEMIC project, covering the data management related issues both on the administrative and on the technical level. This includes topics like application reconfiguration logs and monitoring metrics collection, publication and deposition of open data, the data repository infrastructure and compliance with the Open Access Infrastructure for Research in Europe (OpenAIRE). The current document also defines the foreseen resources needed for the data openness, and the ethical aspects that are taken into consideration in the context of MORPHEMIC.

Author(s)
Alexandros Raikos, Dimosthenis Kyriazis, Jean-Didier Totow, Christina Souvatzi, Chris Kachris, Aleksandra Skwara, Ferath Kherif, Robert Gdowski, Sebastian Geller

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 871643

Table of Contents

List of Tables

Document History

Edition	Description	Date
1.0	Initial edition of the document	06/2020
1.1	Post-review language & semantics corrections	07/2020
1.2	Post-review *Data Summary* content restructuring	09/2020
1.3	Post-review language & semantics corrections	11/2020
1.4	Use case provider dataset information updates	01/2021
1.5	Ethics chapter information updates and minor corrections	02/2021
1.6	Final review and minor corrections	03/2021

1 Introduction

1.1 Deliverable purpose

This document summarizes the MORPHEMIC Data Management Plan (DMP), which has been conceived to support the data management life cycle for all data that will be collected, generated, and processed by the project. The MORPHEMIC DMP identifies best practices for gathering information about the variety of data to be used in the project that will optimise the development, specific processes and repositories for the generated data and assess their suitability for the sharing and reuse in accordance with official guidelines.

1.2 Document structure

The structure of this document complies with the guidelines and recommendations specified in the European Commission's Data Management Template[1]. Chapter 2 delivers a summary on the datasets that will be collected and generated in the project. Chapter 3 provides the data management policy to be followed about the search, retrieval and access to data. Chapters 4, 5, 6 and 7 address the allocation of the available resources, the security of handled datasets, ethical aspects as well as other issues.

2 Data Summary

During the project, several types of data will be generated and collected. Such data include technical datasets, such as the aggregated runtime performance monitoring metrics from an application deployed and managed by the MORPHEMIC platform, configuration data generated from open source software utilities and performance-related data included in the project's deliverables and scientific journal publications and conference submissions derived from the partners research activities.

In the development phase of the MORPHEMIC platform, all use case application related input and output datasets will be made publicly available.

- In use case 1, IS-Wireless will generate application-specific monitoring data related to the deployment of services in a virtualized 5G cloud radio access network (RAN) context. IS-Wireless will be publishing all the use-case 1 experiments' relevant data which is not subject to the company's IPR.
- For the purposes of use case 2, CHUV will collect and use data for neuroimaging applications and generate application-specific monitoring data, pre-processed brain images, brain models, as well as statistical analysis results through SPM on the web, and the federated learning application. The results of the analysis will be made publicly available only in the development phase, where anonymized datasets are being processed.
- In the context of use case 3, ICON will use publicly available geometric object data and parameters regarding simulations in the development phase and generate application-specific monitoring data. Client related simulation data sets in the validation phase are proprietary and the results will not be published.

The validation phase of the MORPHEMIC platform, reflected in Work Package 6, involves the deployment of the use case applications and their evaluation for a reliable approximation of real-world performance and overall stability. This work will be carried out by members of the MORPHEMIC consortium during the entire course of the project, and it will result in valuable performance monitoring datasets, which provide crucial information about the quality of the software solution. Due to intellectual property concerns, and sensitive patient and client information in this phase, the resulting application-processed datasets from use cases 2 (e-BrainScience) and 3 (Computational Fluid Dynamics Simulation) respectively, will not be published. IS-Wireless will be publishing all the use-case 1 (Virtualized base station for 5G cloud-RAN) experiments' relevant data which is not subject to the company's IPR. The deliverable D6.2, the Validation Framework Design, will further define the validation phase in greater detail, while also adhering to the guidelines put in place by this document.

[1] The European Commission's Data Management Template can be found here.

The following sections detail the purpose and the available types of data, the data origin and the data size of the related datasets that will be handled by the MORPHEMIC project.

2.1 Data purpose

What is the purpose of the data collection/generation and its relation to the objectives of the project?

The collection and generation of all data within the confines of the project is crucial for receiving feedback on the development of a technically-sound overall solution, the MORPHEMIC solution, as well as the deduction of comprehensive scientific proof, in any technically representable form that favours automated multi-cloud and cross-cloud deployment in DevOps processes and beyond.

The purposes of the use case collected and generated data, as well as their relation to the project's objectives, are described in the following tables.

Table 1: Use Case 1 - Virtualized base station for 5G cloud-RAN (IS-Wireless): Data purpose

UC1 dataset	**Purpose of collected and generated data**
Infrastructure-specific and application-specific monitoring data	The generated monitoring data are required as parameters for the proactive adaptation mechanism to evaluate the application deployment's utility function. **Relation to the project's objectives:** *TO*[2]*2*: To predict confidently the application behaviour; KPI[3]3, KPI4. *IO*[4]*2*: To demonstrate the usefulness of MORPHEMIC for dynamic applications; IR[5]4, IR5; KPI13
RAN performance validation data	This data set is used to validate the feasibility of the deployment in a RAN context, both in a functional manner utilizing the 3GPP specification guidelines for ensuring communication QoS, and in a non-functional manner regarding total cost and scalability. **Relation to the project's objectives**: *IO2*: To demonstrate the usefulness of MORPHEMIC for dynamic applications; IR4, IR5; KPI13

Table 2: Use Case 2 - e-BrainScience (CHUV): Data purpose

UC2 dataset	**Purpose of collected and generated data**
Application-specific neuroimaging data	The collected neuroimaging data sets are processed as inputs of the image pre-processing pipeline, SPM on the web and federated learning use case applications. **Relation to the project's objectives**: *IO2*: To demonstrate the usefulness of MORPHEMIC for dynamic applications; IR4, IR5; KPI13, KPI14
Infrastructure-specific and application-specific monitoring data	This data is used as parameters for the proactive adaptation mechanism and for evaluating the application deployment's utility function. **Relation to the project's objectives**: *IO2*: To demonstrate the usefulness of MORPHEMIC for dynamic applications; IR4, IR5; KPI13, KPI14

[2] TO: Technical Objective
[3] KPI: Key Performance Indicator
[4] IO: Impact Objective
[5] IR: Impact Result

Table 3: Use Case 3 - Computational Fluid Dynamics (ICON): Data purpose

UC3 dataset	Purpose of collected and generated data
Application-specific CFD simulation data	Collected geometric data and simulation parameters will be used by ICON's use case applications as inputs for extracting post processed simulation data. **Relation to the project's objectives**: *IO2*: To demonstrate the usefulness of MORPHEMIC for dynamic applications; IR4, IR5; KPI13
Infrastructure-specific and application-specific monitoring data	This data is used as parameters for the proactive adaptation mechanism and for evaluating the application deployment's utility function. **Relation to the project's objectives**: *IO2*: To demonstrate the usefulness of MORPHEMIC for dynamic applications; IR4, IR5; KPI13, KPI14

2.2 Data collection

What types and formats of data will the project generate/collect?

The following kind of data will have to be handled:

- **Infrastructure-specific monitoring data sets**, which are time series database (TSDB) records aggregating primitive data types related to infrastructure-specific metrics concerning the use of resources. (e.g., processing load, memory usage, disk access etc.).
- The **data sets used and generated by the use cases** (either directly, such as in the case of IS-Wireless' Virtualized base station for 5G cloud-RAN, or indirectly through patient data, such as in the case of CHUV's e-BrainScience), to be utilised for validating the project outcomes. Each use case dataset is characterized uniquely.
- The **software artefacts of research value** obtained from the MORPHEMIC platform. Given the infrastructure-related focus of the MORPHEMIC project, several datasets of research value are expected to be generated (e.g., artefacts comprise generated configuration files, deployment logs, playbooks, real-time information about the QoS of running applications, etc.). As of the first year of the project, it is premature to state the extent in which such artefacts will be made available.

Each use case dataset's type and format is described in the following tables, respectively.

Table 4: Use Case 1 - Virtualized base station for 5G cloud-RAN (IS-Wireless): Data types and formats

UC1 dataset	Types and formats of collected and generated data
Application-specific monitoring data	Time series database (TSDB) records aggregating primitive data types related to application-specific activity (e.g., number of users currently connected).
RAN performance validation data	Records aggregating primitive data types related to RAN-specific performance metrics between components (e.g., latency).

Table 5: Use Case 2 - e-BrainScience (CHUV): Data types and formats

UC2 dataset	Types and formats of collected and generated data

Image pre-processing pipeline application data	**Input**: Anonymized Medical images of the brain **Output**: Pre-processed medical images **File formats**: dicom, nii, mif, nii.gz, csv, txt, mat, bvecs, bvals.
Image pre-processing pipeline monitoring data	Time series database (TSDB) records aggregating primitive data types related to the application execution's: a) number of users b) computation time for each pipeline task c) memory usage d) storage usage
SPM on the web application data	**Input**: Brain models **Output**: Statistical analysis results **File formats**: nii, mat.
SPM on the web monitoring data	Time series database (TSDB) records aggregating primitive data types related to the application execution's: a) number of users b) system response time c) storage usage
Federated learning application data	**Output**: ML analysis results **File format**: json
Federated learning monitoring data	Time series database (TSDB) records aggregating primitive data types related to the application execution's: a) number of users b) system response time c) storage usage

Table 6: Use Case 3 - Computational Fluid Dynamics (ICON): Data types and formats

UC3 dataset	**Types and formats of collected and generated data**
CFD simulation application data	**Input**: 3D surface representation object, simulation parameters **Output**: Images, simulation data and logs. **File formats**: obj.gz, json, jpg, tsv, txt
Application-specific monitoring data	Time series database (TSDB) records aggregating primitive data types related to the application execution's: a) *simulationTime* measured runtime; b) number of workers per *simulationType*; c) number of cases in the queue per *simulationType;* d) number of cases running per *simulationType* **e)** maximum duration of the simulations per type

2.3 Existing data re-use

Will you re-use any existing data and how?

The MORPHEMIC project builds upon and extends the H2020 EU MELODIC project. From a technical perspective, there are many MELODIC OSS components in the architecture which will be re-used, extended and/or improved upon.

This means that a number of selected infrastructure monitoring datasets generated during the course of the MELODIC project may be re-used in the development phase of MORPHEMIC by members of the consortium.

Regarding the use case collected data sets:

- Use case 2 (e-BrainScience by CHUV), will be using as input images from the freely publicly available anonymized MRI data sets: OASIS[6], UK biobank[7], IXI[8]. The extent of their use concerns only the development phase of the MORPHEMIC platform.
- Use case 3 (Computational Fluid Dynamics Simulation by ICON), will be using publicly available geometric data for the CFD simulation's execution.

2.4 Data origin

What is the origin of the data?

All application and infrastructure related monitoring data sets, any occurring configuration files and other software artefacts of research interest will be extracted from the MORPHEMIC platform's persistent storage components. Further details regarding the microservice architecture of the platform and the included components from the MELODIC architecture will become available in the deliverable D4.1, Architecture of pre-processor and proactive reconfiguration.

2.5 Expected data size

What is the expected size of the data?

The size of the application and infrastructure related monitoring data set is highly unpredictable and relies heavily on the technical requirements of each use case application and the execution context. Examples of parameters affecting the resulting monitoring data set size are the complexity of each use case application's microservice architecture, the distinct number of computational, networking and storage nodes and services required, the total runtime of each use case application, the scraping intervals set by the MORPHEMIC platform at execution time for all the actively monitored components, etc. Due to the scale of the platform's execution, it would be appropriate to assume that there will be at least thousands of lines in the semi-structured format of the extracted monitoring data sets.

As for any configuration files, playbooks and other software artefacts of research value derived from the MORPHEMIC platform's execution, the data size also depends on the same parameters. Since these artefacts are expected to be generated in a much smaller scale in time than the time-series monitoring data mentioned previously, they will prove more compact in size.

2.6 Data utility

To whom it might be useful?

The collected data content may be of interest to both the commercial sectors from which they were collected, as well as to a wider community of data scientists, or students of data science, to carry out machine learning research. DevOps operators may be interested in the application and infrastructure monitoring data as a means of comparison against their own practices.

Regarding data generated and published from the use case application validation phase, use case 1 (Virtualized base station for 5G cloud-RAN by IS-Wireless) will publish the RAN performance validation data set, which could be useful for simulating network performance over time for a Cloud-RAN deployment for further research purposes.

The usefulness of other derivative execution data which will be published, as well as platform performance data included in scientific journal publications, conference paper submissions and other scientific dissemination activities will resonate more with researchers and organizations looking to expand their understanding of some technical challenges in hybrid cloud computing.

[6] The Open Access Series of Imaging Studies (OASIS) official website.
[7] UK Biobank's Brain Imaging Documentation.
[8] The IXI Dataset is available here.

3 FAIR

The MORPHEMIC project's general data management policy that is presented in the subsequent sections has been developed in accordance to Horizon 2020 FAIR principles[9], open data requirements[10] and implementation guidelines[11].

3.1 Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers, such as Digital Object Identifiers)?

The data produced (as of the types described previously in this deliverable) will be associated with relevant metadata. These defer per case:

- The **data sets used and generated by the use cases** will be accompanied with metadata regarding the actual data (both technical, such as: types of datasets, size, collection date, etc.; and non-technical, such as the context, the usage within the scenario, etc.).
- The **artefacts of research value**, such as the respective underlying infrastructure resources used (e.g., edge resources), the mechanisms generating the datasets (e.g., monitoring), the exploitation value (e.g., for proactive adaptations), and generated configuration files will also be linked with metadata (as keywords).

What naming conventions do you follow?

The naming convention to be followed is the following format:

<type>.<name>.<version>,

where: (i) *type* can be UC (for the use cases datasets), SP (for scientific publications), OS (for open source artefacts), RD (for research datasets), (ii) *name* will briefly describe the dataset (e.g., the use case name, the publication name, etc.), and (iii) *version* will reflect the version of the given dataset, with all internal versions being v0.x and the first public version being v1.0, with updates increasing the decimal part of the version number.

Will search keywords be provided that optimize possibilities for re-use?

As described previously, a rich set of keywords will be linked with all potential datasets to ensure that they can be searched and retrieved based on the content.

Do you provide clear version numbers?

Yes, as specified in the naming convention. Moreover, each version will be accompanied with a short description reflecting the changes / updates for that version comparing to the previous one.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Besides the metadata described above for publication related datasets, open source artefacts, etc. the key aspect refers to the metadata for the data sets used and produced by the use cases. These will be identified and documented following the specification of the use case scenarios that will also highlight the generated datasets and the data to be processed and made available.

[9] Access the Report and Action Plan from the European Commission Expert Group on FAIR Data here.
[10] The open data requirements can be found in this EC Document.
[11] Implementation guidelines are examined here.

3.2 Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

Application process-specific datasets which stem from the final validation process of the Computational Fluid Dynamics Simulation use case by ICON and the e-BrainScience use case by CHUV on the MORPHEMIC platform are confidential for legal and contractual reasons, given that they are client and patient data, respectively.

All other types of data generated or collected from the MORPHEMIC project, including data sets used and generated by the use-case provider, open source software artefacts and artefacts of research value – that are not subject to use case provider's company IPR - will be openly accessible as the default.

In addition, according to the H2020 framework programme, publications (e.g., conference and journal peer review papers) are required to be openly accessible. All the partners have been made aware of this and are expected, given also that relevant provisions have been made to the project DoA, to pay the fees to publishers needed so that their publications are openly accessible.

How will the data be made accessible (e.g. by deposition in a repository)?

All of the publicly available data derived from the MORPHEMIC project will be made accessible through the MORPHEMIC OpenAIRE page[12], as well as through the project's website. Additionally, data will also be made available through Zenodo, an open research repository created and supported by OpenAIRE and CERN.

What methods or software tools are needed to access the data?

Closed-source tools will not be required to access the data. Open source solutions, such as Docker, git and web technologies are utilised (e.g. several software components will be released as docker images that can be pulled by interested parties, Git repositories will be utilized, as well as the website for the respective datasets).

Is documentation about the software needed to access the data included?

Not required given the methods that have been selected for making the data available.

Is it possible to include the relevant software (e.g. in open source code)?

The open source contributions are one of the key goals of the MORPHEMIC project. All open source software components developed and enhanced by the consortium members for the MORPHEMIC project will become publicly available as Docker images.

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

Certified and widely used repositories will be utilized. These refer for example to OpenAIRE, Zenodo, GitHub, Bitbucket and Docker Hub.

Have you explored appropriate arrangements with the identified repository?

More information about the specific arrangements with the identified code repositories to be utilised will be published in the scheduled Project Periodic Reports.

Is there a need for a data access committee?

[12] The link to the MORPHEMIC OpenAIRE page will be available in a scheduled Project Periodic Report, once the datasets have been processed by the OpenAIRE Validator.

A data access committee is not required, since all published data sets will be openly available and under no specific access conditions.

Are there well described conditions for access (i.e., a machine-readable license)?

Regarding the software artefacts and the application monitoring datasets, these will be accompanied by the respective licenses for re-use.

How will the identity of the person accessing the data be ascertained?

It is not within the interests of MORPHEMIC project to identify the person accessing the data, since these will be publicly available and the respective data analytics tools (e.g., Google Analytics) will be exploited to obtain information and statistics on an aggregated level.

3.3 Making data interoperable

All data interoperability matters are considered as *non-applicable* since MORPHEMIC will mainly focus on the integrity and the publication of infrastructure-related datasets that will be collected during the project.

In use case 2, e-BrainScience applications provide tools for pre-processing neuroimaging data and comparing neuroimaging models. CHUV is building the tools on an established standard (e.g., SPM, neuromorphometric atlas, etc.), so that data produced for the neuroscience community is easily exchangeable within that community.

3.4 Increase data re-use (through clarifying licenses)

How will the data be licensed to permit the widest re-use possible?

All datasets regarding application monitoring data will be published under the CC0 license, a straightforward and effective way to make it possible for others to mine, exploit and reproduce the data.

Following the definition of the scenarios and compilation of the partners exploitation plans (as well as the overall exploitation plan of the project) the licensing schemes for any other data assets to be published will be further elaborated in the deliverable D8.1, Exploitation Strategy.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

All datasets regarding application monitoring data will be available as soon as the OpenAIRE Validation process is completed for each provided dataset in the MORPHEMIC OpenAIRE repository. The scheduled MORPHEMIC Project Periodic Reports will include formal updates to their online availability in the OpenAIRE and Zenodo repositories.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

All datasets regarding application monitoring data and other software artefacts will be usable by third parties as soon as they are publicly available.

How long is it intended that the data remains re-usable?

There is no specified timeframe after which any published data will stop being considered re-usable. The intent of the MORPHEMIC project is to keep the aforementioned datasets online, openly accessible and re-usable as the default for as long as the respective hosting platforms permit it.

Are data quality assurance processes described?

For all produced artefacts of the project, relevant data quality assurance processes have been identified (e.g. data profiling, completeness and precision checks in large data sets, thorough reviewing of documents, publications which

include derived performance data and extended source code testing) and put in place by the project management committee. These will be applied to all generated outcomes.

Quality assurance processes regarding application monitoring data will be defined on a per-case basis for each published dataset in the scheduled Project Periodic Reports.

4 Resource Allocation

As of June 2020, the sole envisioned costs for making data FAIR in MORPHEMIC are possible fees to be paid to publishers to make data related to publications open. In this case, these fees, which are eligible expenses, will be reported as regular project expenses. Furthermore, there are no further expenses in keeping any data related to the publications available.

The allocation and management of time-series and experimentation data is defined in the corresponding Confluence page[13], provided by 7bulls.

Long-term preservation of data will be solved thanks to the MORPHEMIC OpenAIRE page and is not expected to generate any costs. Since OpenAIRE is an EU site, the length of the preservation will automatically be in line with EU directives. The link to the MORPHEMIC OpenAIRE page will be available in a scheduled Project Periodic Report, once the datasets have been processed by the OpenAIRE Validator.

5 Data Security

Technical information regarding data storage and software component communication practices for ensuring the security and the protection of data from third party access, disasters and other risks will be defined in the deliverable D4.2, Security Design and Implementation.

All use case collected and generated datasets with sensitive client and patient data, stemming from the project's validation phase of use cases 2 (e-BrainScience) and use case 3 (Computational Fluid Dynamics Simulation) will not be published. IS-Wireless will be publishing all the use-case 1 experiments' relevant data which is not subject to the company's IPR.

There will be no other EU-classified information either as background or results stemming from any MORPHEMIC activities.

6 Ethical Aspects

In order to carry out the work envisaged in MORPHEMIC, as described in this project, partners will be involved in research and development of technologies that have the capacity to collect, store, analyse and process vast amounts of data. On top of that, the activities planned in work package 6 will include experimentation with data, some of which can be considered sensitive, as described further in this chapter. As such, the project touches on research ethics related to privacy and data protection issues.

6.1 Human Participants

The project involves the indirect involvement of human participants, as is apparent with use case 2 (e-BrainScience), for the data processing of neuroimaging data sets of the brain. During the entire course of the MORPHEMIC project, CHUV will only utilise publicly available anonymized neuroimaging datasets of the brain from the OASIS[14] project. The nature of the participants' disorder is such that special procedures for obtaining informed consent is needed. As required by the Ethics Summary Report, the description of the research participants' recruitment procedure, informed consent procedure and template will be submitted as deliverables in Work Package 10, fulfilling all ethics requirements upon the start of use case execution. Thus, no adults who are unable to give informed consent will be involved in any part of research carried out in MORPHEMIC.

[13] The Confluence page provided by 7bulls can be accessed here by creating a free 7bulls Confluence account.
[14] The official Open Access Series of Imaging Studies (OASIS) website: https://www.oasis-brains.org/.

Considering the execution of use case 1 (Virtualized base station for 5G cloud-RAN), IS-Wireless will approach deployment of the network based on the set of requirements, which will be done purely in made up (artificial) environment only and not in actual scenario; therefore, human participation is excluded. As human participation is explicitly excluded in the execution of use case 1, this use case does not entail any ethical considerations related to the participation of human subjects.

Use case 3 (Computational Fluid Dynamics Simulation by ICON) will be executed by using the MORPHEMIC architecture and tools for running CFD simulations on semi-industrial cases relevant to the automotive industry and processing data coming from these simulations. No human participants will be involved in the requirements elicitation and evaluation of the use case and no personal data will be collected or reused. The use case will also not require recording any personal data or monitoring the activities of the researchers involved in the simulations. As a result, this use case does not entail any ethical considerations related to the participation of human subjects.

6.2 Personal Data

Due to the extent of the platform's operations (i.e., in cloud application profiling, monitoring, optimization and higher-level runtime management across various cloud infrastructures), there are varying degrees of complexity in any sensitive data processed across the entire stack depending on each given use case application. As a result, the MORPHEMIC project has defined appropriate sensitive data management practices for the following phases:

- The **Development phase**, in which some of the project's consortium members actively collaborate to define all requirements, and design and develop the necessary software components of the MORPHEMIC platform.
- The **Validation phase**, in which some of the project's consortium members actively collaborate to successfully validate the operation of the platform in a production environment using real world use cases.

The following sub-sections provide further information as well as a GDPR assessment regarding all activities under each phase of the project.

6.2.1 Data processing overview

6.2.1.1 Development phase

In the development phase of the project, all use case related data will stem from publicly available anonymized datasets. This means that the data collected and processed has been rendered anonymous in such a way that the data subjects can no longer be identified (and therefore is no longer personal data and thus outside the scope of data protection law).

6.2.1.2 Validation phase

In the scope of work package 6, the Validation Framework, partners will engage in the validation of the MORPHEMIC tools in three realistic scenarios. As a general rule, each part of each use case demonstration will be done by partners independently, meaning that no data will be transferred between the participants. As a result, each consortium member providing a use case application will be responsible for handling any sensitive data with appropriate measures and will follow their own privacy policy.

Use cases 1 (Virtualized base station for 5G cloud-RAN) and 3 (Computational Fluid Dynamics) will not be processing personally identifiable data, as detailed in section 6.1. Use case 2 (e-BrainScience) will be processing publicly available anonymized neuroimaging data sets of the brain from the OASIS project.

6.2.2 GDPR compliance

In the collection of data, the project addresses the overall legal framework that is stated in the General Data Protection Regulation (Regulation 2016/679) on the protection of natural persons regarding the processing of personal data and on the free movement of such data. Also, it considers the procedures for electronic data protection and privacy conforming to the Directive 2002/58/EC (and its amending acts, such as 97/66/EC and 2009/136/EC). The following assessment[15] tables provide a higher-level overview into how the project complies with the GDPR in practice.

[15] Assessment provided by https://gdpr.eu/, a website co-funded by Project REP-791727-1 of the Horizon 2020 Framework Programme of the European Union.

Table 7 – GDPR: Lawful basis and transparency assessment

Condition	**Legal reference**	**Relevant actions**
Conduct an information audit to determine what information is processed and who has access to it.	Regulation (EU) 2016/679 – *Article 30*	(a) Task 9.3 (*Data Management*) (b) Work Package 10 (*Ethics requirements*)
Have a legal justification for your data processing activities.	Regulation (EU) 2016/679 – *Article 6*	Applies, as stated in p.1-b: *Processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract.*
Provide clear information about your data processing and legal justification in your privacy policy.	Regulation (EU) 2016/679 – *Article 12*	(a) Task 9.3 (*Data Management*) (b) Work Package 10 (*Ethics requirements*)

Table 8 – GDPR: Data security assessment

Condition	**Legal reference**	**Relevant actions**
Take data protection into account at all times, from the moment you begin developing a product to each time you process data.	Regulation (EU) 2016/679 – *Article 5*	(a) Described in D9.1 (*Data Management Plan*) (b) Described in D4.2 (*Security Design and Implementation)*
Encrypt, pseudonymize, or anonymize personal data wherever possible.	Regulation (EU) 2016/679 – *Article 32*	(a) The development phase of the project utilizes publicly available anonymized data sets. (b) Encryption and data security information is described in D4.2 (*Security Design and Implementation*).
Create an internal security policy for your team members and build awareness about data protection.	Regulation (EU) 2016/679 – *Recital 78*	(a) Internally introduced *Feature 6 (Security Concepts)* for functional approach on developing with security in mind. (b) Established regular meetings on development status including platform security planning.
Know when to conduct a data protection impact assessment and have a process in place to carry it out.	Regulation (EU) 2016/679 – *Article 35*	See subsection 6.2.3 for further details.
Have a process in place to notify the authorities and your data subjects in the event of a data breach.	Regulation (EU) 2016/679 – *Article 33 & 34*	It is not applicable in this case, since there is no personal data involved and all data sets previously including personal information have been anonymized.

Table 9 – GDPR: Accountability and governance assessment

Condition	**Legal reference**	**Relevant actions**
Designate someone responsible for ensuring GDPR compliance.	Regulation (EU) 2016/679 – *Article 25*	UPRC, as the lead partner of the Task 9.3 (*Data Management*), will ensure GDPR compliance across the project's activities.
Sign a data processing agreement between your organization and any third parties that process personal data on your behalf.	Regulation (EU) 2016/679 – *Article 28*	A data processing agreement is deemed as not applicable in the case of the MORPHEMIC project, as there are no third parties involved outside of the project's consortium partners.
If your organization is outside the EU, appoint a representative within one of the EU member states.	Regulation (EU) 2016/679 – *Article 27*	UPRC, as the lead partner of Task 9.3 (*Data Management*), acts as a representative for all matters related to the management of data during the entire course of the project.
Appoint a Data Protection Officer (if necessary).	Regulation (EU) 2016/679 – *Article 38*	Considering that all data processed for the platform's development and validation are not personally identifiable in any way, there is no need for a Data Protection Officer role in the project.

The MORPHEMIC project consortium is not directly responsible for matters related to privacy rights of the data subjects as described in Regulation (EU) 2016/679, Chapter 3. Since each validation procedure will be performed solely and independently by the use case partners, each sensitive data procedure is subject to each use case partner's privacy policy.

Moreover, in the case of Centre hospitalier universitaire Vaudois (CHUV) where patient data are involved and it's a non-EU member state organization, the deliverables of Work Package 10, *Ethics Requirements*, provide more detail regarding sensitive data collection and informed consent.

6.2.3 Data protection impact assessment

The enactment of a Data Protection Impact Assessment Process (DPIA) would be required for the data processing pipeline of use case 2 (*e*-BrainScience) during the validation phase of the project. The use case includes large-scale processing of patient data (neuroimaging data sets of the brain from patients with neurocognitive disorders) which, from the point of view of the GDPR (Regulation (EU) 2016/679 – Article 9(1)), is considered a special category of personal data which is likely to "*result in a high risk to the rights and freedoms of natural persons*".

In reality, the neuroimaging data sets of the brain used by CHUV for their validation of the MORPHEMIC platform are publicly available anonymized data sets by the OASIS project. The aforementioned sets have undergone specific anonymization procedures to ensure the complete removal of any personally identifiable information. The corresponding publications[16] for each data set include further details on the exact procedures implemented to remove sensitive patient information from the neuroimaging data sets.

Any other procedures in the development and the validation phase of the MORPHEMIC project also do not include the processing of any special category of personal data.

6.3 Involved Third Countries

One of the demonstration scenarios, use case 2, is owned by the Lausanne University Hospital (CHUV) and it will include processing publicly available anonymized data from populations of patients with neurocognitive disorders and making the data available to clinicians, neuroscientists, epidemiologists for diagnosis and research. While Norway and

[16] Refer to the publications for each available data set: https://www.oasis-brains.org/#data.

Switzerland participate in Horizon2020, from the point of view of the G.D.P.R they are not EU member-states, therefore considered non-EU countries.

As a general rule, each part of use case demonstration will be done by partners independently, meaning that no data will be transferred between the participants. As a result, each partner will be responsible for handling any sensitive data with appropriate measures and will follow his own privacy policy. Therefore, CHUV will be responsible for acquiring relevant ethical approvals according to the local regulations in Switzerland.

As required by the Ethics Summary Report, a declaration on compliance for collecting and processing personal data; detailed information on the procedures for data collection, storage, protection, retention, and destruction in compliance with national Swiss law ("*la Loi fédérale relative à la recherche sur l'être humain*" (LRH, RS 810.30)) and the EU legislation; as well as detailed information on the informed consent procedures in regard to the collection, storage, and protection of personal data will be submitted as deliverables in the context of the Work Package 10, upon the start of use case execution. CHUV will not further process previously collected personal data for the use case execution.

7 Other Issues

MORPHEMIC has no usage of other national, funder, sectorial or departmental procedures for data management.