Document Identifier	COMETE_WP5_D5.14
Status	Final

COMETE

Next-Generation Computational Methods for Enhanced Multiphase Flow Processes

H2020-MSCA-ITN-2018
Grant Agreement N. 813948

Deliverable Rel. D5.14,
Deliverable Number D37
Data Management Plan

WP	5	Coordination

Dissemination level[1]	CO
Nature[2]	ORDP

Due delivery date	30/04/2019
Actual delivery date	06/02/2020

Deliverable lead beneficiary	UNIUD	Deliverable reference person	Cristian Marchioli
E-mail	marchioli@uniud.it		

Other contributors to the deliverable	All partners

Document version	Date	Author(s)	Comments
1	10/01/2020	Cristian Marchioli	First draft circulated within the consortium
2	06/02/2020	Cristian Marchioli	Final version incorporating all consortium's input

Disclaimer and acknowledgement
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No 813948.

---

[1] Dissemination level: PU = Public, CO = Confidential, only for members of the consortium
[2] Nature of the deliverable: R = Report, E = Ethics or, O = Other

Deliverable summary

The deliverable describes how the research data that will be generated in the frame of the COMETE project will be made findable, accessible, interoperable and re-usable. The DMP is available on the project's website at this link:
http://158.110.32.35/COMETE/DOCS/DMP.pdf

Table of Contents

1. Data Summary

The purpose of the data collection/generation is to provide a comprehensive repository of both numerical results and experimental measurements that can be used to examine the three-phase turbulent flows with two interface-separated continuous phases and one dispersed phase that were targeted in the research proposal.
The data collection/generation is crucial to achieve the objectives of the project because it will allow the development of a unitary framework of multiscale methods for accurate prediction of industrial multiphase flows involving particle-gas-liquid interactions in turbulent conditions.
The typical types and formats of data that the project will generate/collect depend on their origin: data coming from simulations can be stored in ASCII files if their size is limited to a few Mbytes, or in binary files if their size is significantly larger. For large, complex, heterogeneous data, the Hierarchical Data Format version 5 (HDF5), which is an open source file format specifically developed to support this kind of data, will be used. The HDF5 format also allows for embedding of metadata making it self-describing. All uploaded data files will be check-summed for the integrity check.
We are going to produce new datasets and repositories, so we do not plan to re-use any existing data previously generated by the project partners outside of the project itself.
As mentioned, the origin of the data will be either numerical simulations (from WP1, WP2 and WP3) or experimental measurements (from WP4), and the expected size of the data repository will typically ranges from few Mbytes, e.g. for most of the experimental measurements or for post-processed statistics extracted from the numerical simulations, up to hundreds of Gbytes for the raw data produced as output of the numerical simulations.
The numerical and experimental repository will be extremely useful within the COMETE network to the development and inter-operability assessment of the simulation tools that will be used in WPs 1 to 3 and might be useful beyond the COMETE network, e.g. to interested users in the multiphase flow community for benchmark purposes or to scientific writers and the press to produce high-quality infographics, demonstrating the impact potentials of the developed multiscale approach.

2. FAIR data

2. 1. Data findability, including provisions for metadata

2.1.1 Discoverability

The data produced and/or used in the project will be openly accessible and discoverable through the COMETE public website http://158.110.32.35/COMETE. The data will be indexed using Zenodo, general-purpose open repository developed under the European OpenAIRE program and operated by CERN (https://zenodo.org/). Note that, because of their huge size, it is not possible to store raw data files elsewhere and share them directly through public open platforms.
Since the open data support the quality and credibility of the open publications, all data will be discoverable through the scientific publications. Each scientific publication will include Digital Object Identifiers (DOI) that point to the associated open data sets, making them identifiable and locatable.
To maximise data accessibility and discoverability, the ercoftac.org platform will be used to store all the data files with suitable size (order of a few Mbytes).

2.1.2 Identification

Each data set will carry a DOI as unique and persistent identifier. Data sets will be referenced in scientific publications and if the open data platform permits, scientific papers based on the data will be linked on the open data platform.

2.1.3 Metadata

The data sets follow the EU Open Data Portal Metadata definitions (https://ec.europa.eu/digital-singlemarket/en/news/metadata-specificatons-eu-open-data-portal). From the complete set of fields, a minimum set will be provided. The Dublin Core Metadata Initiative will be followed (http://dublincore.org). For all characterizations at least the following metadata will be provided via the COMETE website upload form and the individual metadata files in the data folders:
□ Dataset identifier (points to subfolder with the same name)
□ Title (meaningful name of the dataset)
□ Alternative title (additional information in concise format)
□ Description (description of the dataset)

 Keywords
 Identifier (DOI reserved for this data set)
 Link to where data are stored (usually the path of the folder in the common storage system)
 Dataset type (raw data, post-processed data, formatted/unformatted)
 Documentation (description of the dataset content)
 Format (file type, usually a compressed archive in ZIP format)
 Publisher (the COMETE consortium)
 Contact name (full name of the scientist in charge for the production of the dataset)
 Contact full address
 Contact e-mail

The top folder of the repository contains a plain text file called REPOSITORY.TXT that provides the most important metadata, notably the project identifier of each specific data set that points to the folder in which the open data are stored:

 Contact name (full name of the scientist that should be contacted for this data set)
 Contact e-mail - E-mail address of the contact person
 Project identifier - a local identifier of the dataset in format LLL-YYMMDD
 Title (meaningful name of the dataset)
 Dataset type (raw data, post-processed data, formatted/unformatted)
 Format (file type, usually a compressed archive in ZIP format)

NOTE: It is understood that these fields are duplicates of those fields, which are also stored at lower dataset folder level. The repetition at higher level serves creating a simple to use catalogue in the project.

The metadata for each dataset will be stored in a file called METADATA.TXT, which will be placed in the folder of the corresponding dataset. This file contains the necessary fields (which depend on the type of data stored in the set) in different columns.

2.1.5 Naming conventions

The filename of a dataset must contain a clear, concise and very short name that identifies the contents. Words must be in lowercase and separated by underscores ("_"). Other naming conventions are discouraged.

Filename extensions are encouraged to ease the understanding of the folder contents.

Individual files are placed in subfolders according to this structure:

LLL-WP#-YYMMDD is the name of the folder for a given dataset where LLL is the three-letter abbreviation of the Consortium member institute at which the data set is created (e.g. TUW for Technische Universität Wien), WP# indicates the work package within which the data set is produced, YY stands for the last two digits of the current year (e.g. 20 for 2020), MM stands for the two digits of the current month (e.g. 02 for February), DD stands for the two digits of the current day of month (e.g. 6 for the sixth day). A complete example for a project identifier is TUW-WP2-200206.

The folder contains at least the following files:

 README.TXT - brief description of the folder contents, authors, other useful information
 LICENCE.TXT - brief description of the terms of use of the data and acknowledgments
 METADATA.ODS - the metadata of the dataset including the change track record

The folder contains subfolders that correspond to different raw data and/or post-processed statistics stored.

Search keywords are part of the metadata, following the appropriate (community) standards and will be provided that optimize possibilities for re-use. Each dataset will at least be tagged with the following keywords: COMETE, H2020, EID, 3PHASE.

2.1.6 Metadata standards and templates

To our best knowledge, no domain-specific metadata standard for the datasets to be produced within the COMETE project exist. Therefore, a column-oriented data format with an explanation of the columns will be created in the scope of this project. This is one tangible outcome of the COMETE Data Management Plan initiative.

Template metadata files can be found in the dedicated COMETE website section
http://158.110.32.35/COMETE/Research/Data/TEMPLATES

The entire template folder can be copied into the ./Data folder and can be renamed according to the naming convention LLL-WP#-YYMMDD for a new dataset. The folder contains exemplary

 README.TXT
 LICENCE.TXT
 METADATA.ODS
 a list of exemplary folders corresponding to different types of data (raw data or post-processed data).

2.1.7 Storage administration and access permissions

Data sets are managed via COMETE's website. Consortium members who need to use the storage platform need to comply with standard security rules and guidelines as well as appropriate training in the use of the data.

The website section for the COMETE datasets can be directly reached through standard internet browsers or on Windows/Linux/Mac operating filesystems via sftp connection to COMETE webserver. 3 TBytes of storage have originally been allocated. The system is backed up and implements automatic versioning based on file changes. Administration permissions are managed directly with the owner of the storage platform. Access permission requests, requests for dataset release and/or folder administration requests must be addressed to marchioli@uniud.it.

Three main groups are used to manage read-only and read-write permissions:

- cometeproject-developers - anyone in this group can create and edit data sets and their metadata files in the project space
- cometeproject-readers - anyone in this group can read in the project space
- cometeproject-writers - anyone in this group can read, write and delete in the project

Membership is decided on a case-by-case basis and is exclusively granted by consensus agreement of the data owners and by requesting access permission to marchioli@uniud.it.

2.2. Data accessibility

Experiment and simulation data might be openly accessible only after the end of the projects as soon as the results have been published. All publications will be open access. Accessibility and open availability will be granted only after approval of all persons who were involved in the production of a given dataset. This is particularly true for the experimental data that will be collected at EHP, which will be made available only upon permission of EHP in consideration of the intellectual property right (IPR) regulations related to the results achieved within the project, which are detailed in the Grant Agreement and in the Consortium Agreement linked to GA N. 813948, which establish that (1) the copyright and IPR for the data generated, collected or used during the project is owned by the Party that generates them and (2) the PhD Candidate does not acquire, due to his/her assignment, any right of industrial or commercial property on the research results, the equipment, the knowledge or the expertise of the laboratory in which he/she is hosted, including unpublished results.

In general, we plan to make all post-processed data produced and/or used in the project openly available to interested users once these data have been exploited to reach the scientific and research-related objectives of COMETE. If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

As mentioned in Sec. 2.1, all openly accessible data will be made available upon deposition in a repository. All data stored in formatted (ASCII) files can just be downloaded and edited with any text editor application. Data stored in unformatted (binary) files will be accompanied by proper read/write instructions (including the open source fortran/C/Matlab code). In general, no specific software tools will be needed to access the data.

Data will be indexed using the Zenodo Portal (https://zenodo.org/) and will be complemented by the associated metadata, documentation, and code, which will also be deposited in the COMETE website. Whenever data file size allows it, data will also be shared through certified open access repositories, e.g. the ERCOFTAC's Classic Collection Database (https://www.ercoftac.org/products_and_services/classic_collection_database/).

Once included in the open access repository, there will be no specific restrictions on use. Also, since each WP in which the data are produced is under the responsibility of one beneficiary, then accessibility of data will be decided directly by the beneficiary together with the Supervisory Board, which automatically acts as data access committee.

Since the data that originate from the simulations and experiments targeted in the COMETE project are relative to physical variables of importance for three-phase turbulent flows (e.g. fluid velocity, particle velocity, interface position and related post-processed statistics) there is no need to create complex procedures to ascertain the identity of the person accessing the data once these have been made openly accessible. It is enough to address any request of access permission, dataset release and/or folder administration requests to the project's coordinator by e-mail (see also item 2.1.7). The requestor needs to justify the request and specify the purpose (e.g. academic use, no further dissemination of details). If access is granted, the specific data file can be downloaded directly from the repository or, if not available there, communicated in electronic form to the requestor.

2.3. Making data interoperable

All COMETE data and metadata are designed for interoperability and in all cases follow open community standards and file formats (in particular ASCII, binary and hdf5), which are fully compliant with available software applications typically used within the fluid mechanics community. Controlled scientific terminology, established within the academic fluid mechanics and multiphase flow community, is also used to ease interoperability, data exchange and re-use between researchers, institutions and organisations. We do not foresee the need to use uncommon or generate project specific ontologies or vocabularies. However, in view of the specificity of the shared data (as mentioned, they refer to specific physical variables of importance for three-phase turbulent flows and are therefore expected to be used mainly for academic purposes being of potential interest for researchers in the fields of engineering and applied physics), we foresee little need for re-combinations with different datasets from different origins.

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

Experimental data test data will be made openly accessible after the end of the project or publication of the results, whatever comes first. Wherever possible, the data will be shared right after production.

2.4.1 License

Openly accessible data will be licensed under Creative Commons 4.0 International License with Attribution (see https://creativecommons.org/licenses/by/4.0/), which guarantees maximum re-use (and redistribution) while maintaining the traceability of the use and credit to the data providers and their sponsors.

Users are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material for any purpose, also commercially.

The licensor cannot revoke these freedoms as long as one follows the license terms.

Under the following terms:

Attribution — One must give appropriate credit, provide a link to the license, and indicate if changes were made. One may do so in any reasonable manner, but not in any way that suggests a licensor's endorsement.

No additional restrictions — One may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

2.4.2 Timing

As mentioned before, we foresee the need for an embargo to give time to publish (according to the deliverables' time schedule), but we plan to make research data available as soon as possible after publication, and definitely within the end of the COMETE project.

2.4.3 Re-use

Once made openly accessible, the data produced and/or used in the project will be useable by third parties, in particular after the end of the project, in the fields of engineering and applied physics, upon explicit request and approval of the dataset manager and the data producer.

2.4.4 Quality assurance

Data sets, metadata, measurement/numerical setup and procedure description will be reviewed by at least one peer prior to engaging the release procedure.

Data sets, metadata, measurement/numerical setup and procedure description will be marked RELEASED only after approval of the person in charge of the measurement campaign/numerical simulation and the project supervisor (in case the project supervisor is also the person in charge, an additional reviewer will be appointed).

Review and release includes a checksum validation of the data files, the measurement/numerical setup, conditions and procedure as well as sanity checks against similar studies and control of systematic errors.

In case of data quality uncertainties after release, a new version IN WORK is created and the released data set version is marked INVALID and removed from the public repository.

The description of measurement setup (materials and method) and of the numerical setup (governing equations and numerical implementation) are annotated with product references.

The measurement conditions are described, along with measurement location, date and time (periods) are noted. Any potential and known adverse effects (environmental influences, influences of the measurement equipment) are described in the metadata.

2.4.5. Validity

The data will remain usable at least 5 years after the end of the project, and until the repository withdraws the data or goes out of business.

3. Allocation of resources

3.1 Cost estimate

The project Coordinator will keep the data sets and perform their publication in the open data repository. The estimated effort is 10 hours per data set, 4 data sets (one for each WP) per year, i.e. 40 hours or 1 week per year over the entire project period. This resource is covered by the project management funds.
Note: the project Coordinator will track the actual efforts and regularly update this estimation.
Each researcher in the project is responsible to create the data sets using the adopted open data format, providing the metadata files, describing the measurement setup, anonymising the data, reviewing the data sets and performing the release process using the appropriate storage infrastructure/platform. The estimated effort is 20 hours per data set, 4 data sets per year, i.e. two to three working weeks per year over the entire project period. The institutes that carry out the measurements/simulations cover this resource.
Note: The participating institutes are strongly encouraged to track the time they are spending to prepare the data sets and to publish them and to report their actual estimates to the Coordinator.

3.2. Data management responsibilities

This data management plan is maintained by the project Coordinator, Cristian Marchioli (Univ. Udine). All work package leaders, deputies and ESRs commit to cooperate on the establishment of this DMP and to deliver the required information such that the associated deliverables and milestones can be produced in due time with the requirement quality levels:
Data storage and backup responsibilities are covered by the data repository providers. The COMETE project repository is managed by the project Coordinator in person, with the aid of his Department (DPIA). Cristian Marchioli is the site manager and, together with DPIA's IT services, provides support for the upload to the data storage system and performs a formal (file integrity, naming, metadata completeness) check. Long-term data preservation will be ensured by Univ. Udine at no additional cost, and will allow maintenance of a currently-unavailable public database for industrial applications.

4. Data storage and security

All data delivered to the COMETE project repository are backed up by DPIA's IT services. In addition, a copy of released data will be kept on the ERCOFTAC platform (provided that the size of the data file complies with the platform's rules and policies). Both services are intended for long-term storage of scientific research data. Upon unintentional loss of data (misuse of the collaborative workspace, accidental removal), the project Coordinator, Cristian Marchioli, needs to be contacted via email to marchioli@uniud.it. He will interact with DPIA's IT services to restore the latest known copy. No additional costs occur for storage, backup and restore activities.
Non-public data sets can be provided by the project members using HTTPS transfer protocol after authentication by asking to marchioli@uniud.it: Considering the type of data shared, this is a reasonably secure policy to counteract data manipulation. Since e-mail is not considered a secure communication channel for data and metadata files (data can be modified and it is unclear what fields have been modified with respect to the original data source), a link to the authentic data source shall be considered reliable information and will therefore be provided.
COMETE produces non-sensitive data (namely, all data made openly accessible will be anonymizes). No personal information is processes nor stored. There is no privacy policy issue related with the production of the data. For secure storage COMETE relies on the local infrastructure available at the institution of the project's coordinator, where the project's website is hosted, of EUDAT.
Every consortium member must inform the project Coordinator without delay if a person affiliated (associated or employed) with the institute and who has access to the project data, leaves the institute. In this case, the Coordinator will revoke as soon as technically possible and resources permitting (working hours) the access of the person to the data.

5. Ethical aspects

Ethical aspects are not expected to be relevant for the data to be produced and shared during

the COMETE project. Indeed, we do not foresee the need to access non-anonymized data and/or sensitive information. Should that happen, access will be managed by the project Coordinator in close cooperation with the organization that provides the data set: Non-anonymized data will only be communicated in encrypted fashion and digitally signed.
COMETE partners are of course required to comply with the ethical principles as set out in Article 34 of the Grant Agreement, which states that all activities must be carried out in compliance with:
a. ethical principles (including the highest standards of research integrity (as set out, for instance, in the European Code of Conduct for Research Integrity) and including, in particular, avoiding fabrication, falsification, plagiarism or other research misconduct) and
b. applicable international, EU and national law.
The COMETE project does not involve the use of human participants or personal data in the research and therefore there is no specific requirement for ethical review.

6. Other issues
Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

7. Further support in developing COMETE's DMP
As well as European Commission policies on open data management, all project partners must also adhere to their own institutional policies and procedures for data management:

IMP-PAN Gdansk:
https://www.imp.gda.pl/wasteman/EN/privacyPolicy.php?s=

TU Wien:
https://www.tuwien.at/en/research/rti-support/research-data/research-data-management/policy/

Univ. Udine:
https://www.uniud.it/it/ricerca/open-access/valorizzazione-e-promozione-open-access/open-access-informazioni-generali

EHP:
https://www.euroheat.org/legal/
Note: EHP has its own set of internal policies and procedures on data management, which are subject to periodic review

ESTECO:
https://www.esteco.com/technology/simulation-data-management
Note: ESTECO has its own set of internal policies and procedures on data management, which are subject to periodic review

Contact marchioli@uniud.it by e-mail for any questions concerning the data sets and their management in the scope of the COMETE project.

8. Summary of responsibilities

8.1 Data production Responsibility: Project Partner (PP) in charge for the WP within which data are produced
8.2 Metadata creation Responsibility: PP in charge for the WP within which data are produced
8.3 Quality assurance of data Responsibility: PP in charge for the WP within which data are produced and project Coordinator
8.4 Data security Responsibility: Project Coordinator
8.5 Data archiving & data sharing Responsibility: Project Coordinator
8.6 Policy compliance Responsibility: PP in charge for the WP within which data are produced and project Coordinator

9. Annex A – Example Metadata File Template

This metadata file was generated on <insert date> by <insert name>

GENERAL INFORMATION
---

1. Title of Dataset:

2. Dataset Identifier in Repository:

3. Responsible Partner:

4. Author Information:
Investigator Contact Information
Name:
Email:
Supervisor Contact Information
Name:
Email:
Co-Supervisor Contact Information
Name:
Email:

5. Date of data collection:

6. The title of project and Funding sources that supported the collection of the data:

SHARING/ACCESS INFORMATION
---

1. Licenses/access restrictions placed on the data:
2. Link to data Repository:
3. Links to other publicly accessible locations of the data:
4. Links to publications that cite or use the data:

DATASET & FILE OVERVIEW
---

1. This dataset contains X sub-dataset as listed below:
A. Datasheet name:
B. Datasheet name:
C. Datasheet name:

2. What is the status of the documented data? – “complete”, “in progress”, or “planned”
Are there plans to update the data?

METHODOLOGICAL INFORMATION
---

1. Description of methods used for experiment/simulation design and data collection:
<Include links or references to publications or other documentation containing experimental design or protocols used in data collection>

2. Methods for processing the data: <describe how the submitted data were generated from the raw or collected data>

3. Instruments and software used in data collection and processing-specific information needed to interpret the data:

4. Standards and calibration information, if appropriate:

5. Environmental/experimental conditions, if appropriate:

6. Describe any quality-assurance procedures performed on the data: