# Digital Manufacturing and Design (DiManD)

**European Training Network.**

**Grant agreement No 814078– H2020-MSCA-ITN**

# Deliverable 6.12

# Data Management Plan

**March 2021**

**Lead parties for Deliverable:** STIIMA
**Deliverable due date:** M30
**Actual submission date:** M23
**Dissemination level:** Public





*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant No. 814078*

# Table of Contents





*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant No. 814078*

## Summary

The aim of this document is to introduce the initial version of the DiManD Data Management Plan. It answers standard questions concerning the management of data collected and generated in the framework of the DiManD project.

**Team involved in deliverable writing:** All beneficiaries



*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant No. 814078*

# 1 Introduction

This deliverable presents the data management plan (DMP) to manage the data associated to the scientific publications and the data inherent to the research of DiManD project: manufacturing lifecycle data including control data, monitoring and diagnosis data, context data, user interaction data, etc. The data will be made available via the Zenodo repository as open access and licensed under CC-BY. In case of potentially patentable inventions, sharing will be delayed to investigate patent protection first. The Dissemination and Exploitation Coordinator (DEC) will be the responsible for preparing and maintaining the Data Management Plan.

# 2 Data Summary

The following issues will be presented in this section:

- The purpose of the data collection/generation
- The relation of the data collection/generation to the objectives of the project
- The types and formats of data that will be generated/collected
- The origin of the data
- The expected size of the data (if known)
- The data utility

The main objective of DiManD project is to provide high-quality multidisciplinary, multi-professional and cross-sectorial research and training to high-achieving early stage researchers in the area of Industrie 4.0 and intelligent informatics driven manufacturing

The purpose of data/collection/generation in the framework of the DiManD project is to store and share the project data that might be useful for the project partners and that would be reusable by researchers in the field. And that will support the main objective of the project - to educate the next generation of Industrie 4.0 practitioners.

In the project, data management is essential due to the importance of data in Industrie 4.0. One of the key research challenges is big data analytics. And two of the objectives of the project are strongly linked to data:

- To design and develop a control concept and underpinning **data models** for autonomous behaviour adaptation of distributed manufacturing systems based on context-aware autonomous systems (WP4).
- To analyse and apply new ICT trends, such as **Big Data,** Cyber Physical Systems and **Data Mining**, in manufacturing systems to enable more efficient processing of data for control and configuration and advanced diagnostics and monitoring purposes that at the same time provide security and privacy by design without compromising the need to share data between different organisations in the manufacturing chain (WP5).

The project addresses understanding how new product design and manufacturing will be influenced by **lifecycle data**, etc. And marrying the concepts of industrial informatics and **data analytics** with globalised manufacturing and distribution system science.



*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant No. 814078*

The acquisition of skills and training in big data, data analytics, data mining, data visualization human-machine interactions and data are also relevant.

Data will be used in the individual research projects:

- ESR1: A concept for open evolvable assembly systems. The project will develop a model for evolvable assembly systems that will define and capture the dynamics of the product-process-system evolution including what semantic data models are needed to capture this evolution and enable data-driven system intelligence and response.
- ESR2: Self-learning for Optimum Manufacturing Equipment (Individual & Collective Response). The learning process will utilise the **status data**, past experiences and key parameters to analyse possible scenarios and propose actions for achieving the individual and collective goals.
- ESR3: Cyber-Physical Systems and User Interaction Experience into Industrie 4.0. **Semantic of data** and new **User Interaction Experiences with data technology** will be addressed.
- ESR4: Human Centred Design for Industrie 4.0. Advance service innovation. The goal will be to incorporate Human Centred Design skills to link deep customer knowledge, with resources and digital **data** flows in one single system.
- ESR5: Simulation-based Runtime Testing and Adaptation of Cyber Physical Systems using digital twins. With usage of historical **operation data** to create the living digital simulation models.
- ESR6: Cyber-Physical Systems and End of life management in home automation. Using AI-based on **big data**
- ESR7: Precision manipulation and assembly of electro-optical components. The project will exploit **cyber-physical data** to develop a **functional model** that will be validated in a real case scenario.
- ESR8: Design and development of cost-effective solutions for High throughput, mixed model electronic assembly and packaging.
- ESR9: Investigation of Transition Technologies to support Assembly Station Reconfiguration in the automotive industry. The project will develop a methodology to enrich the information contained in **product models**, specifically **CAD files** and its application to achieve reconfiguration of assembly systems.
- ESR10: Self-learning Cyber-Physical Production Systems. ESR10 will focus on the development of agent based cyber physical systems with a strong emphasis on self-learning, requiring learning, context and **big data** to implement it.
- ESR11: Developing Energy Saving Techniques and Tools in Production Systems. ESR11 will focus on **big-data** techniques to implement energy saving solutions for industrial systems based on agent based cyber physical components.
- ESR12: Flexible Robotics. Adaptive control models will provide the robot's trajectory for each inspection case by using the **digital information** for the piece and by linking it with the input **data** from the deep learning module.
- ESR13: Artificial Intelligence applied to Oil & Gas. The **data** of the production processes will be used.
- ESR14: Development of **data** models and adaptation strategies for intelligent products. ESR14 will develop intelligent (test) instruments for application within assembly systems using, for example, plug-and-produce concepts, product-driven production, **data collection and analytics** through cloud-based services.



*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant No. 814078*

In regard to the types of the generated/collected data - due to the scientific objective of DiManD, this project will mainly generate and collect the following type of data:

- Monitoring data of production and manufacturing processes, obtained by sensors
- Control data of Cyber-physical systems and devices
- Operational data, including manufacturing resource status information.
- Context data
- Human-machine interaction data
- Product data, including product designs and specific product instance data and identifiers.
- Simulation models
- Data models
- Semantic models and ontologies.
- Images and videos
- Software
- Proposed trajectories for simulated pieces
- …

The formats of the data depend on the type of the data collected/generated. One of the challenges in preserving data for long term is the choice of file format. In the DiManD project whenever possible data will be saved in open formats or widely popular formats which do not depend on proprietary software. For text files this can be .txt, for tabular data - .csv, for image- .tif, for audio- .mp3 and for video- .avi. And for semantic representation XML, RDF or OWL will be used.

Some of the ESRs have defined the format of the data they are going to use. As we are at the beginning of the project, others were not able to define their data format. In the next list, we can find some of the examples:

- ESR1
  - Product specifications, as XML or JSON format text files.
  - Operational data, as XML or JSON format text files.
  - Data for Big Data analysis in JSON / XML format
- ESR2
  - Sensor data in JSON format text files or CSV tabular data as appropriate.
  - Operational data, as XML or JSON format text files.
- ESR7
- Data for analysis in JSON textESR9
  - CAD files, or CSV files.
  - Images in PNG, BMP, JPG or similar format files.
  - ROS scripts in PYTHON or C++ formats and launch files in (STEP, XML format.)
  - Robot trajectories in *.mod files.
  - Product/process models (EXPRESS/XML/UML)
- ESR 11



*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant No. 814078*

- Data for Big Data analysis in JSON / XML format
- *.png for images

- ESR 12:
  - *.mod for robot trajectories and *.png for images.
- ESR 13:
  - *.xls and *.csv for experimental files.

While no external data will be reused for the project, some partners may reuse some of their own pre-existing data. Additional context data provided by Open Data platforms could be required for economic, social or commercial contrast or data enrichment.

We expect that the total volume of data generated by the consortium over the course of the project could reach around 100 terabytes.

As far as the data utility is concerned, the DiManD data will be useful first of all to the consortium and to the project partners. However, the objective is also to share the data that might be useful to the international scientific community.

# 3 FAIR (Findable, Accessible, Interoperable and Reusable) data

To give data greater value and enhance their propensity for reuse, by humans and at scale by machines, data should be Findable, Accessible, Interoperable and Reusable (FAIR)[1] to the greatest extent possible.

## *3.1 Making data findable, including provisions for metadata*

- **Discoverability of data (metadata provision)**

Unless there is an established metadata schema for the type of data published in the project the Dublin Core simple level schema will be used. This is a well-established general scheme which can effectively cover a large variety of data. The elements of the schema are as follow: title, creator (authors), subject (e.g. Engineering), description, publisher, contributor, date, type (e.g. sound), format, identifier (DOI), source, language, relation, coverage (e.g. location), rights, funding.

Semantic repositories might be constructed to semantically represent data. Ontologies will be used to represent those data. Ontology identification/production is part of the research objectives. Nevertheless, some of the ESRs have defined their needs at this initial stage:

- ESR1: Data will abide by semantic models and ontologies, defined in the OWL 2 and RDF standards. Data will be presented in JSON format and/or XML schemas as appropriate.

[1] **FAIR Data - Findable, Accessible, Interoperable, and Reusable**
Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi: 10.1038/sdata.2016.18 (2016)
https://www.nature.com/articles/sdata201618



*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant No. 814078*

- ESR2: Gathered data will be presented in CSV files or text files abiding by the JSON format and/or XML schemas as appropriate.

- **Identifiability of data and standard identification mechanism**

All archived publications will be associated with a DOI.

- **Naming conventions**

The following folder structure will be used:
*<institution>/<researcher name or initials>/<data identifier>/<date>*

If necessary the *date* folder can be broken down into to further subfolders. When a short phrase cannot be a clear data identifier a reference number will be used instead. However, in this case a separate *readme* text document with a short description of each data set and its reference number will be created in the *researcher's name* folder.

The YYYY-MM-DD date convention will be used for dates in folder and file names. This is the ISO 8601 standard and when in the beginning of the folder/file name, items are automatically sorted chronologically. For example the 1st of October 2019 will be written as 2019-10-01.

The following naming system will be used for files:
*YYYY-MM-DD_project name_institution_researcher intitials_measurement /simulation _number*
For example, if there was a document containing tabular data representing operational data taken by the author of this DMP on the 1st of October 2019, the file name of the first set for the day would be:
*2019-10-01_MGEP_LE_operationData_01.csv*

For every data set a separate text document with explanatory notes will be created and saved in the *date* folder together with all other files. The notes will contain full details of how and under what conditions the information was collected or generated and explanation of the file naming system. In the case of data generated from simulation models the exact input parameters and references to the simulation code will be included in the explanatory document.

- **Approach towards search keyword**

A "keywords" category will be systematically included in the metadata file.

- **Approach for clear versioning**

For archived objects that may exist in different versions (typically, software and models), all versions will be archived, including the metadata of the newer versions indicating the location of the previous one and containing a summary of the changes. Versions will be numbered according to standard conventions (e.g. with a suffix "v" followed by a number and possibly a letter for minor changes).

- **Standards for metadata creation**



*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant No. 814078*

It is not foreseen the creation of a standardized metadata format.

## 3.2 Making data openly accessible

- **Data openly available**

Whether the data will be made fully open immediately will be decided case-by-case. In any case, all data (excluding confidential data) will be made fully accessible after publication of the corresponding article. In addition, the data will be openly shared among partners of the consortium, except in some specific cases (confidential data, projects involving third parties with specific agreements). The DOI corresponding to the data will be provided in publications. For data made open prior to publications, the corresponding links will be found on the partners' institutional websites.

- **Methods or software tools needed to access the data**

As open formats or widely popular formats have been selected, public domain software will be used to access the data. If not, the specific tools required will be indicated in the metadata file.

- **Emplacement of data, associated metadata, documentation and code**

A specific DiManD archive will be created within the open Zenodo archive repository. Data will be deposited there associated with their metadata. In addition to the Zenodo archive, each partner will archive their own data on their institutional servers and/or external storage media, and may allow direct access via their laboratory homepage.

Furthermore, ROS software libraries will be upload as open-source repositories to the GitHub platform, as the main web-based/cloud hosting site suggested by the ROS community[2]. It not only allows full open access and total contribution to the ROS/ROS-Industrial ecosystem, but also code maintenance and consistency.

- **Access in case of restrictions**

For restricted datasets (e.g. confidential files), access will be managed by the owner partner responsible for the data generation. In general, efforts will be made to curate the data file in order to provide the relevant source data for other researchers who wish to use the data without compromising confidential information.

## 3.3 Making data interoperable

- **Interoperability of data (the use of data and metadata vocabularies, ontologies, standards or methodologies in order to facilitate interoperability)**

[2] ROS.org: Recommended Repository Usage - Link



*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant No. 814078*

Because of the diverse types of data and of tools, it will not be possible (nor necessarily desirable) to attain a full interdisciplinary interoperability within the project.

However, as specified above, efforts will be made, to use a common format (readable with public domain software) and have a common metadata format.

Ontologies will be also used for semantic representation.

## *3.4 Data re-use (through clarifying licenses):*

- **Licensing of data in order to permit the widest reuse possible**

DiManD data will be stored under one of the Creative Commons licenses. For figures, media, posters, papers and file sets CC-BY license will be used. CC-BY ensures the research will be openly available but it requires that a credit in the form of citation is given when used or referred to. In the case of complete databases or structured datasets with highly factual data CC0 license will be used.

CC0 similar to CC-BY is an open license but does not legally require users of the data to cite the source. However, the moral obligation of attribution remains same as in the case of any research journal paper citation. In the special case of programming or computer simulation code sharing the MIT license will be used. MIT provides full open access to the code but also removes any liability of the authors in the event of any legal claim or damage caused by the use of the code.

All data with potential commercial value will be embargoed for 12 months after the termination of the project to allow for patent filing.

- **Availability of data for re-use**

Data will be fully available for reuse after publication of the corresponding article. Decisions on making the data immediately available will be made by partners on a case-by-case basis.

- **Data quality assurance processes**

Data quality will be the responsibility of each partner. However, a collegial evaluation of the archived data will be made during consortium meetings to ensure a shared minimal quality standard.

- **The length of time for which the data will remain re-usable**



*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant No. 814078*

We will not impose limits on the length of reuse of the data. The main issue will be the continuity of data archival on the Zenodo platform, which is so far impossible to predict.

During the progress of the project, it will be discussed if there will a time when the data has outlived its usefulness and who will have own/manage the archive after the project.

# 4 Allocation of resources

In this section, the allocation of resources will be exposed by examining the following issues:

- **Estimation of costs for making the data FAIR**
- **Responsibilities for data management in your project**
- **Costs and potential value of long term preservation**

The estimated costs for making the DiManD data FAIR are mainly those of the time (estimated to be about 1 week/year for each partner) that the researchers will dedicate to the activities connected with this issue. These costs are thus covered by each partner using the project funds.

Every collaborator involved in the project will be responsible for the management of the data that he/she will collect/generate. The Dissemination and Exploitation Coordinator (DEC) and the Project Manager will make sure that the common agreed principles of management of the project data are respected. Both of them will be contact persons in a case any of the project partners has a request concerning the management of the data.

Regarding the costs of preservation of the digital data, DiManD will use a free data repository, Zenodo. The costs of long term preservation of data using a chargeable repository will be discussed later in the course of the project.

# 5 Data security

**Data recovery, secure storage and transfer of sensitive data**

DiManD will make sure that the all the data collected/generated is safely stored for long term preservation. As stated above, the general principle will be a central archival on the Zenodo platform for the whole consortium plus a storage of each partner's data on their institutional servers. This issue will be discussed later in the project and the proper measurements and actions will be taken toward this objective.

# 6 Ethical aspects

Personal data and the human interaction data will be stored and processed in the project. In order to protect the personal data and the human interaction with technical systems, participants will be given an informed consent form and detailed information sheets that:



*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant No. 814078*

- are written in a language and in terms they can fully understand,
- describe the aims, methods and implications of the research, the nature of the participation and any benefits, risks or discomfort that might ensue ,
- explicitly state that participation is voluntary and that anyone has the right to refuse to participate and to withdraw their participation, samples or data at any time — without any consequences,

The consortium must ensure that potential participants have fully understood the information and do not feel pressured or coerced into giving consent. Participants must normally give their consent in writing (e.g. by signing the informed consent form and information sheets).

Moreover, the data that will be shared with other members of the consortium or with the research community will include only curated data to remove any possibility to identify the participants.

# 7 Use of another national/funder/sectoral/departmental procedures for data management

A few partner institutions require that articles be deposited into their own open archives (HAL format). This will be one way to ensure fully open access in the cases where publications do not appear in Open Access journals. In general, Open Access journals will be preferred for publications; however this will have to be balanced with the need for high impact.

# 8 Conclusions

This deliverable presents the initial DMP of the DiMand project.

The Data types and formats as well as procedures for data management have been defined and agreed by participants at this stage of the project. However, they will be reviewed and updated if needed during the progress of the project.

The DMP is a continuously evolving document which will be updated during the project. Updates will be reflected in the D6.13 Data Management Plan deliverable document due in M30 (October 2021). All activities regarding data publishing will be coordinated and supported by the Dissemination and Exploitation Coordinator (DEC) which is also responsible to ensure that the current DMP is followed by all partners.



*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant No. 814078*

# 9 Versions

| D6.12 Data Management Plan | |
|---|---|
| **Version - Date** | **Comments & Recommendations** |
| V0.1 – 27/09/2019 | Table of content |
| V0.2 – 21/10/2019 | Inputs of partners |
| V0.3 – 28/10/2019 | New inputs of partners |
| V1.0 – 15/11/2019 | Final version |
| V2.0 – 05/10/2020 | Intermediate version |
| V3.0 – 14/12/2020 | Third version |
| V4.0- 16/03/2021 | Fourth version |



*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant No. 814078*