Network of excellence in make Digital Technologies and Artificial Intelligence (AI) solutions for electromechanical and power systems applications (DiTArtIS)

D1.2 Data management plan

Due date of deliverable	M3
Actual submission date	18.12.2022
Organisation name of lead beneficiary for this deliverable	UTCN
Dissemination Level	PU
Start date of project	1.10.2022
Duration	36 months

Responsible organization:	Technical University of Cluj-Napoca
Authors:	Sorin Iulian Cosman
Version:	Final

Contents

1. HISTORY OF CHANGES

No.	Chapter	Change	Page

2. DATA SUMMARY

2.1. Purpose of the data collection/generation

The goal of the DiTArtIS project is to boost the recipients' reputation for excellence in research and innovation while also raising the staff members' digital literacy levels. This will be accomplished by creating elite teams to generate fresh concepts and tackle problems, ensuring that UTC and its twinning partners sustainably increase their capacity for scientific innovation and excellence in the fields of electromechanical applications and power systems, and energy, by integrating digital technologies and artificial intelligence (AI) techniques and solutions.

The DMP describes the types of data that will be generated or gathered during the project, the standards that will be used, the ways how the data will be exploited and shared for verification or reuse, and how the data will be preserved.

This document has been produced following these guidelines and aims to provide a consolidated plan for DiTArtiS partners in the data management plan policy that the project will follow. This document is the first version delivered in M4 of the project. The DMP will be updated during the lifecycle of the project.

2.2. Types and formats of data that will be generated/collected

Being in line with the EU's guidelines regarding the DMP, this document should address for each data set collected, processed and/or generated in the project the following characteristics: Dataset description, reference and name, Standards and metadata, Data sharing, archiving and preservation. Now, an estimation of the size of the data cannot be given. To this end, the consortium develops a number of strategies that will be followed in order to address the above elements.

2.2.1. Data description, reference and name

In order to be able to distinguish and easily identify data sets, each data set will be assigned with a unique name. This name can also be used as the identifier of the data sets.

All data files produced, including emails, include the term "Ditartis", followed by file name which briefly describes its content, followed by a version number (or the term "FINAL"), followed by the short name of the organisation which prepared the document (if relevant).

Each data set that will be collected, processed or generated within the project will be accompanied by a brief description.

2.2.2. Standards and metadata

This version of the DiTArtiS DMP does not include a compilation of all the metadata about the data being produced in DiTArtiS project, but there are already several domains considered in the project which follows different rules and recommendations. This is a very early stage identification of standards:

- Microsoft Word 2010 for text based documents (or any other compatible version). doc, .docx, .xls,

- .xlsx, .ppt, .pptx. Also, especially where larger datasets need to be dealt with, .csv and .txt file formats will be used. All finished and approved documents will also be made available as .pdf documents.
- Illustrations and graphic design will make use of Microsoft Visio (Format: .vsd), Photoshop (Format: different types possible, mostly .png), and will be made available as .jpg, .psd, .tiff and .ai files.
- MP3 or WAV for audio files.
- Quicktime Movie or Windows Media Video for video files.

These file formats have been chosen because they are accepted standards and in widespread use. Files will be converted to open file formats where possible for long-term storage.

Metadata will be comprised of two formats – contextual information about the data in a text based document and ISO 19115 standard metadata in an xml file. These two formats for metadadata are chosen to provide a full explanation of the data (text format) and to ensure compatibility with international standards (xml format).

2.2.3. Re-use of the existing data

The digital data created by the project will be diversely curated depending on the sharing policies attached to it. For both open and non-open data, the aim is to preserve the data and make it readily available to the interested parties for the whole duration of the project and beyond.

With the agreement of all project partners, the data generated by DiTArtiS will be made publicly accessible. As for the detailed simulation files, the information will only be shared within the project partners up until they are published at a conference.

2.2.4. Origin of the data

Specific datasets may be associated to scientific publications (i.e. underlying data), public project reports, presentations and other raw data or curated data not directly attributable to a publication. Other datasets will be related to any public report or be useful for the research community. They will consist of parts of the general datasets generated or full datasets (i.e. up to 2 years of key operating data), and be published as soon as possible.

2.2.5. Data utility

The Consortium has identified the following target groups and end-users who will especially benefit from EnerMaps' data offerings:

- Researchers (including staff in universities and non-university research organizations working in the field of electromechanical and power systems applications) through access to data resulting from the collaboration of the DiTArtiS research working groups;
- Industry (especially in the field of electromechanical and power systems applications), will have free and easy access to databases and related insights that they have selected as crucial for the development of their activities;
- Research and administrative staff in research organizations will have access to presentations and material developed for the training events;

- General public can take advantage of the valuable insights provided by the website.

3. FAIR DATA

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)? Open access data (generic models and documentation, properties of studied electromechanical systems, scientific papers...) will be made available on the DiTArtiS website, hosted by TUC, under the link "Dissemination and public involvement." and in ZENODO, the general-purpose open repository developed under the European OpenAIRE program and operated by CERN. DOI numbers, ISBN numbers, and other publication information will be communicated. Partners will have access to restricted data on a secure platform. A link to this platform (to which they will have access with their username and password) will be available and sent to partners for ease of access. Metadata will be available either on the download platform or in a downloadable "readme" file alongside the data. Additional subfolders or links can be created as data is collected and produced in the project to better structure and identify the data.

What naming conventions do you follow? The following naming convention is proposed for final versions of scientific papers:

- "Authors Paper title Conference/journal name/acronym (date of publication)"

Other documents are named as follows:

- "Authors Title (Vx) (date of publication)"

In those conventions (Vx), II represents the version number (if needed). Because the data will be collected and produced as part of the project, the partners may change the naming convention as needed. The date will be entered in this form LL, ZZ, YYYY, meaning month, day, and year.

Will search keywords be provided that optimize possibilities for re-use? The metadata included with the data, as well as the DITartis website, will provide the information required to discover and reuse the data that will be available online. When possible, relevant keywords (for example, from the IEEE keyword list) should be listed at the beginning of each document for scientific papers.

Do you provide clear version numbers? With the exception of final versions of scientific papers (which have only one version), the version number of each document circulating between partners is clearly stated in the file name if necessary, with v.1, v.2, v.3, … representing version 1, version 2, version 3, and so on. To avoid confusion between documents (especially if they are written, read, and corrected iteratively), the date is also mentioned next to the version of the document, allowing you to find the most recent 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. For easy categorization and searching, the main metadata (title, author, date, version...) will be provided in the document title. Other metadata (e.g., keywords, abstract, related URL, contact person, access constraints, related references, data structure, software/test model and configuration, and working and testing conditions environment...) may be associated with the data as needed. The metadata's nature and structure will be determined as

the data is generated. The Director Interchange Format is a metadata standard relevant to this discipline (DIF)

3.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? As the project is not a research focused project, the scientific articles, presentations, teaching material will be made public.

How will the data be made accessible (e.g., by deposition in a repository)? All the open access data will be made available on the DiTArtiS website directly under the "Public Dissemination and Engagement" link and ZENODO. A dedicated download platform will be also set up for larger files. Access-restricted data will be shared on a platform by project partners. This secure platform is accessible via a specific link on the DiTartis website but requires a username and password to access.

If there are restrictions on use, how will access be provided? A Creative Commons license will be considered for peoople involved in this project once the platform is fully operational and privacy issues have been resolved.

Is there a need for a data access committee? A data access committee is unnecessary. Because UTC will host the secure platform, and UTC is responsible for platform access just as it is for open access data.

Are there well described conditions for access (i.e. a machine readable license)? Restricted access data is protected by a login and password, whereas open access data is available for free. There are no other access requirements.

How will the identity of the person accessing the data be ascertained? Each individual, work group, or beneficiary (at the institutional level) will be assigned a unique authentication code. The identities of those who have access to sensitive data will be determined. The person who accesses open access data will not be identified.

3.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)? What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable? To ensure interoperability, data will be readable using open-source software and/or widely used formats such as.csv,.mat,.doc,.txt,.pdf, etc. Furthermore, the data will be structured in accordance with the metadata so that the various values can be easily identified. IEEE keywords can also be used in metadata.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability? A glossary can be created if necessary to describe project-specific terms or abbreviations.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Yes, if necessary, there will be maps to those ontologies used more frequently

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

How will the data be licensed to permit the widest re-use possible? If the data must be protected for reuse, the method of licensing the data (using creative commons license) will be decided by unanimous agreement among all partners at the same time as the storage platform is chosen.

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. Data undermining publications will be available for re-use right after publications are released. When writing the agreement to make the data public, an embargo will be specified if necessary. In case of embargo period, data will be available for re-use right after this period has ended.

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. Third parties will have direct access to and use of open access data, which will be available even after the project is completed. Possible restrictions on third-party access to intellectual property data (during and after the project's completion) must be the subject of a future non-disclosure agreement or license agreement with the third party in question.

How long is it intended that the data remains re-usable? The data will be kept on the repository for at least ten years after the project ends, unless otherwise specified or requested by partners or third parties.

Are data quality assurance processes described? The method by which the data is generated will be specified in the metadata to ensure data reproducibility. For model data, references and assumptions, as well as the software version, must be specified.

4. ALLOCATION OF RESOURCES

What are the costs for making data FAIR in your project? Because the metadata and data structure will be generated concurrently with the results, the costs of making data fair will be partially integrated into the research costs. In terms of the data repository and its management, internal servers are used for sensitive data (at UTC) as well as open-access data on the DiTartis website, resulting in no costs a priori. Costs will be reported by the relevant partners if necessary.

Who will be responsible for data management in your project? TUC owns the DiTartis website but is also in charge of data management.

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)? Unless differently specified in the specific data sharing agreements the data will be kept available online for 1 year after the end of the project. The associated costs are therefore the ones related to the maintenance of the DiTartis website and of secure repository during that period. If need be, the preservation of specific data for a longer time after the end of the project may be discussed among the partners.

5. DATA SECURITY

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)? All data will be stored on TUC servers, but each partner will also keep local backups on hard disks to allow for easier recovery in the event of an unexpected server shutdown. To gain access to a secure platform for confidential data, a login and password are required. While open-access data can be transferred using common data-transfer websites (e.g., We transfer, etc...), sensitive data must be transferred only through a secure platform.

Is the data safely stored in certified repositories for long term preservation and curation? The status of the TUC servers determines whether or not data is stored in certified repositories.

6. ETHICAL ASPECTS

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA). For the time being, there are no ethical or legal concerns regarding data sharing.

Is informed consent for data sharing and long-term preservation included in questionnaires dealing with personal data? If personal data questionnaires must be written, informed consent for data sharing and long-term preservation will be include