DELIVERABLE D7.3

# Data Management Plan

**Lead Beneficiary**: 1-UNIPD

**Author**(s): Eloisa Di Sipio[1], A. Manzella[2], E. Trumpy[2], R. Pasquali[3]

**Author(s) affiliations**: [1]UNIPD, [2]CNR, [3]GEOSERV

**Date**: 31/08/2022

***Dissemination Level***

| | | |
|---|---|---|
| **PU** | Public, fully open | X |
| **SEN** | Sensitive - limited under the conditions of the Grant Agreement | |
| **CI** | EU classified - RESTREINT-UE/EU-RESTRICTED, CONFIDENTIEL-UE/EU-CONFIDENTIAL, SECRET-UE/EU-SECRET under Decision 2015/444 | |

**This research is funded by the European Union (G.A. 101046937). The views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EISMEA. Neither the European Union nor the granting authority can be held responsible for them.**

### *Document History*

| Version | Date | Authors | Description |
|---|---|---|---|
| 1 | 19/08/2022 | E. Di Sipio (UNIPD) | Creation of the document |
| 2 | 25/08/2022 | E. Di Sipio (UNIPD) | First draft ready |
| 3 | 30/08/2022 | R. Pasquali (GEOSERV) | Revision draft |
| 4 | 30/08/2022 | A. Manzella, E. Trumpy (CNR) | Revision draft |
| 5 | 31/08/2022 | E. Di Sipio (UNIPD) | Final document |
| | | | |

### *Disclaimer*



This publication was completed with the support of the European Innovation Council and SMEs Executive Agency (EISMEA) under the HORIZON-EIC-2021-PATHFINDEROPEN-01 programme. This research is funded by the European Union (G.A. 101046937). However, the views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or EISMEA. Neither the European Union nor the granting authority can be held responsible for them.

# TABLE OF CONTENTS

**EXECUTIVE SUMMARY** ..... **5**

**1. INTRODUCTION** ..... **6**

**2. DATA SUMMARY** ..... **7**

**3. FAIR DATA** ..... **9**

3.1 MAKING DATA FINDABLE, INCLUDING PROVISIONS FOR METADATA ..... 9
3.2 MAKING DATA ACCESSIBLE ..... 11
*3.2.1 Repository* ..... *11*
*3.2.2 Data* ..... *11*
*3.2.3 Metadata* ..... *13*
3.3 MAKING DATA INTEROPERABLE ..... 13
3.4 INCREASE DATA RE-USE ..... 15

**4. OTHER RESEARCH OUTPUTS** ..... **17**

**5. ALLOCATION OF RESOURCES** ..... **17**

**6. DATA SECURITY** ..... **18**

**7. ETHICS** ..... **19**

**8. OTHER ISSUES** ..... **19**

**REFERENCES** ..... **19**

**APPENDIX A** ..... **20**

QUESTIONNAIRE ON RESEARCH DATA ..... 20

# TABLE OF FIGURES

FIGURE 1: RD MANAGEMENT PROCESS SCHEME ..... 7

# LIST OF TABLES

TABLE 1: DATA TYPE AND DATASETS THAT ARE EXPECTED TO BE PRODUCED BY THE PROJECT ..... 8

# ABBREVIATIONS AND GLOSSARY OF ACRONYMS

| | |
|---|---|
| CA | Consortium Agreement |
| DeepU | Deep U-tube heat exchanger breakthrough: combining laser and cryogenic gas for geothermal energy exploitation |
| DMP | Data Management Plan |
| DoA | Description of Action |
| DOI | Digital Object Identifier |
| EC | European Commission |
| FAIR | findable, accessible, interoperable and reusable |
| GA | Grant Agreement |
| IPR | Intellectual Property Rights |
| OA | Open Access |
| ORD | Open Research Data |
| PC | Project Coordinator |
| PC-Co | Project Co-coordinator |
| RD | Research Data |
| SC | Steering Committee |
| T | Task |
| WP | Work Package |

# EXECUTIVE SUMMARY

The D7.3 “Data management plan” is a confidential document delivered in the context of WP7, Task 7.1: General project direction, management and communication with regard to data management, collection, generation, storage and preservation.

This document describes the Data Management Plan (DMP) for “Deep U-tube heat exchanger breakthrough: combining laser and cryogenic gas for geothermal energy exploitation (DeepU)” Project.

The use of a Data Management Plan is required for all participating projects supported by the European Innovation Council and SMEs Executive Agency (EISMEA) under the HORIZON-EIC-2021-PATHFINDEROPEN-01 programme.

This document defines the policy adopted for the management of data produced or collected during the project activity, including their storage in a suitable online repository to assure optimal preservation, security and sharing performances.

This DMP also describes the data management life cycle for the collected, processed and/or generated data during the DeepU project. On this regard, it will include information on: the handling of research data during and after the end of the project, what data will be collected, processed and/or generated, which methodology & standards will be applied, whether data will be shared/made open access and, how data will be curated & preserved (including after the end of the project).

This deliverable represents the DMP at the present stage of the project: during the project life cycle, new information will update the DMP as described in following sections. Indeed, the datasets are not fixed; quite the opposite, they will evolve during the lifespan of the project, particularly whenever significant changes arise such as changes in Consortium policies. Future modifications will be tracked during each periodic report. In detail, two DPM updates are foreseen at month 12 (D7.4) and month 36 (D7.8).

# 1. INTRODUCTION

DeepU consortium adopt Open Science practices to increase the chances of delivering the results to a wide community of researchers. Whilst respecting the IPR policy the data/research outputs will be made findable, accessible, interoperable and reusable (FAIR).

The Horizon Europe Model Grant Agreement requires that a data management plan ('DMP') is established and regularly updated. The requirements for research data management of Horizon Europe as described in article 16-17 and Annex 5 of the Grant Agreement, are addressed.

A Data Management Plan (DMP) has been developed using FAIR data principles. The DMP outlines what datasets the project will generate and compile, and how these datasets will be made accessible and stored. The DMP also describes measures that have been taken to safeguard and protect sensitive data and emphasizes that the produced results must be easily located and accessible.

DeepU has chosen to use the template provided for the Data Management Plan. At present, very little data has been collected by the project. The DeepU DMP is intended to be a 'living' document that will outline how the DeepU research data will be handled during and after the project, and so it will be reviewed and updated over the course of the project.

UNIPD will be responsible for disseminating this DMP to all project partners. Each project partner will be responsible for managing their data, metadata, and insuring their data meets the quality standard set out in the DeepU Quality Assurance Plan (D7.1).

Open access (OA) can be defined as the practice of providing on-line access to scientific information that is free of charge to the reader. In the context of R&D, open access typically focuses on access to 'scientific information' or 'research results', which refers to two main categories:

- Peer-reviewed scientific research articles (primarily published in academic journals)
- Research data

There is a clear distinction between Open Access to scientific peer-reviewed publications and Open Access to research data in terms of policy and commitments. In fact, while Open Access is an obligation in Horizon 2020 to scientific peer-reviewed publications, regarding Research Data the commitments are less definite, and their management is more flexible. Although Open Access is always preferable, it is possible to opt for a different level of shareability of the research data if actually needed.

Research Data comprehend the information, generally consisting in datasets of parameters, generated in the action by the Consortium during the life span of the project, collected to be examined and considered, and to serve as a basis for reasoning, discussion or calculation.

Open access to research data consists of the right to access and reuse digital research data under the terms and conditions set out in the Grant Agreement.

Even though the Consortium greatly endorse the precepts of open science, nevertheless, data sharing in the open domain can be restricted as a legitimate reason to protect results that can reasonably be expected (e.g. commercial or industrial exploitation). Strategies to limit such restrictions could include anonymizing or aggregating data, agreeing on a limited embargo period or publishing selected datasets not completely open.

Most of the WPs and the relative deliverables indirectly influence this document, due to the data they contain. For this reason, this deliverable contains a section indicating for each dataset the corresponding WPs, deliverables or documents produced by the project that make use of such data.

# 2. DATA SUMMARY

**Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.**

The presented DMP concerns the status of the data production among the Consortium as planned on the Description of the Action (DoA). Nevertheless, during the project life cycle, new data will be created and expected datasets will be expanded with new information. Consequently, in order to track the future production of data, datasets will be updated in the next months to reflect the project developments.

The whole RD management process is recapped in the following diagram (Figure 1):

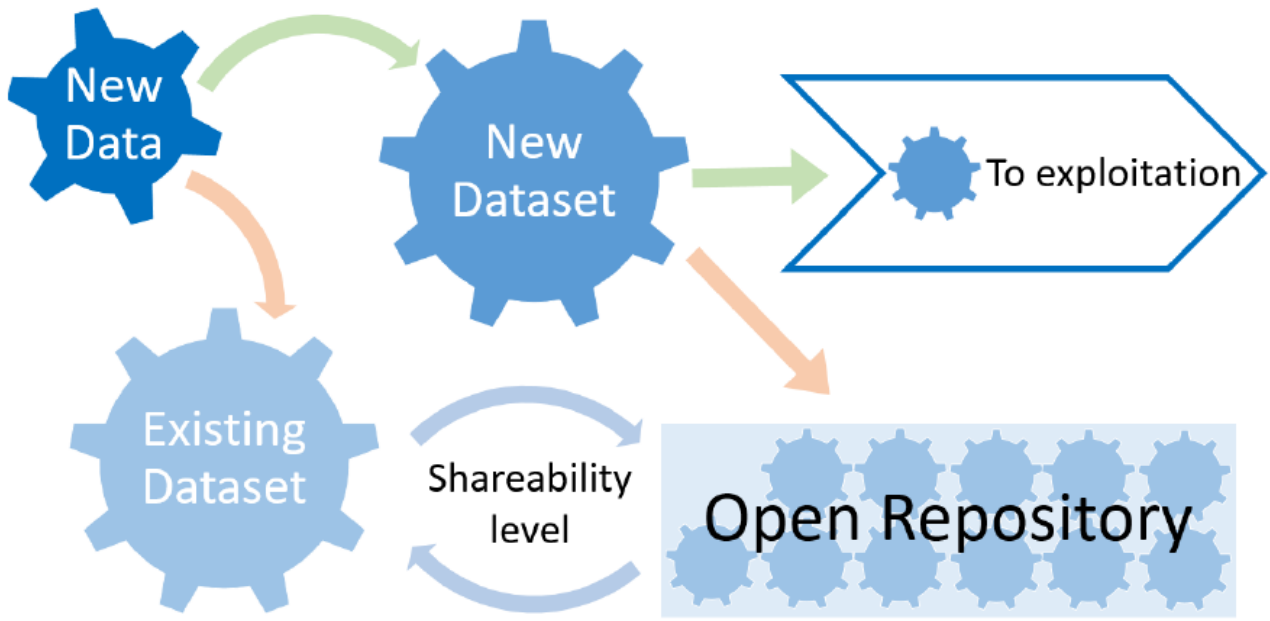


***Figure 1: RD management process scheme***

In each WP the re-use of any existing data will be carefully described taking advantage of the questionnaire on Research Data (Appendix A) already prepared that will be shared with partners at the first GA (6M).

**What types and formats of data will the project generate or re-use?**

Table 1 provides a preliminary summary of the Data type and the corresponding Datasets that are expected to be produced by the project. This brief outline of the datasets gives an instant overview on the expected basic information of the RD.

More precise information on RD will be obtained through a questionnaire that will be completed in the next months by all the partners involved in the project (see APPENDIX A).

***Table 1: Data type and Datasets that are expected to be produced by the project***

| No | Product | File formats |
|---|---|---|
| **1** | Text | Plain text, XML, HTML, PDF |
| **2** | Spreadsheet | CSV (Comma-separated values), Tab delimited values, PDF |
| **3** | Images | JPEG, TIFF; PNG, SVG |
| **4** | Geospatial | GML (Geographical Mark-up Language), KML (Keyhole Mark-up Language), ESRI Shapefile, Geo-referenced TIFF |
| **5** | Numerical | NetCDF, CSV, JSON |
| **6** | Video | MP4, MOV, WMV, AVI, AVCHD, FLV, F4V e SWF, MKV, WEBM o HTML5 |
| **7** | Audio | Waveform Audio File Format (WAVE) |
| **8** | Database | Delimited Flat File w/DDL |
| **9** | Archives | ZIP, TAR, GZIP, 7Z |

**What is the purpose of the data generation or re-use and its relation to the objectives of the project?**

The DeepU Data Management Plan (DMP) aims to provide a strategy for managing data generated and collected during the project and to optimize access to and re-use of research data. Data generated during the project can be divided into the following groups:

- Data collected from the outside of the project: used or modified data acquired / collected from existing databases, repositories, open publications because functional for the project activities planned in the different WPs.
- Original data generated by project partners during the laboratory and technological activities. These datasets consist of compiling raw data provided by partners working in the different WPs in the format of agreed metrics.

The overall aim of DeepU is to create a deep (>4 km), closed-loop, vitrified, waterproof, non-cracked U-tube heat exchanger, by combining laser and cryogenic gas into a single technological drilling solution at laboratory scale in different rock types.

The purpose of this data collection is to help achieve the technical, environmental and standardization objectives of DeepU project, that are:

(i) To select a cryogenic gas able to cool in a controlled manner the rock melted by a laser;
(ii) To develop an innovative lightweight drill string able to host the gas and the laser at the same time;
(iii) To develop specific temperature control analysis and innovative laser lenses able to convey the heat and to sustain multilateral drilling;

(iv) To determine the physical-thermal phenomena affecting different kind of rocks in order to assess the borehole wall vitrification and integrity;
(v) To evaluate the legislative aspects and potential environmental risk related to the innovation proposed;
(vi) To model the DeepU geothermal exploitation potential considering some virtual case studies;
(vii) To assess the possible interest of stakeholders in bringing DeepU from laboratory to field validation and explore possible market applications

**What is the expected size of the data that you intend to generate or re-use?**

The data generated through measurements and laboratory activities are expected to be saved mainly as comma-separated value (CSV format, readable over long time periods) format. The whole dataset for each DeepU WP is expected to comprise less than 10 MB. Open access publication on peer-review journal are expected to respect this size limits.

More specific information about each research dataset will be provided by project partners through the questionnaire shown in Appendix A.

Anyway, the maximum file size limit is expected to be lower than 50 GB, the maximum file size allowed by the repository selected by DeepU project and described in section 2.2 of this document.

**What is the origin/provenance of the data, either generated or re-used?**

Data collected during DeepU activities concern mainly laboratory and technological activities. A detailed list of research datasets to be expected for each WP will be provided at the first DMP update (D7.4, M12). Data from existing standards (CEN, ISO) and best practice guidelines on the project innovations will gathered and assessed on the basis of the replicability of the project technologies to the drilling and geothermal industries.

**To whom might your data be useful ('data utility'), outside your project?**

The data generated in the project will be very beneficial to a variety of stakeholders including: researchers, policy makers, public funders, device developers, utilities, private investors. The data collected will be relevant to key areas of geothermal and drilling research development including: technology development, consenting and project finance. Information gathered will help identify challenges in these areas and serve as an input for researchers, private stakeholders, drilling players and policy design at national and European level, especially in geothermal energy management. In detail, engineers, researchers and experts in geothermal energy are the main potential re- users of the DeepU dataset. Finally, the data will inform the wider public about the developments and potential of the ocean energy sector.

# 3. FAIR DATA

The DeepU DMP (D7.3) applies the Findable, Accessible, Interoperable, Reusable (FAIR) approach for the project's results.

## 3.1 MAKING DATA FINDABLE, INCLUDING PROVISIONS FOR METADATA

**Will data be identified by a persistent identifier?**

With the exception of RD that are not classifiable as ORD due to commercial exploitability reasons, the RD will be always included in the open access repository and they will be identified with a unique reference code (DOI), despite their exact shareability level. Shareability level and conditions may

vary during the project: any changes will be operated in the data repository system as a consequence.

**Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.**

DeepU will publish 3 annual reports which will be the primary method of disseminating the data collected. The annual reports will present data as meta or grouped data; individual sources of data will not be identifiable. The annual report will include references to the original data source

All DeepU documents will be identifiable based on a common naming convention. To ensure document and data control, each document and data set shall be uniquely identifiable. Each deliverable and data set must be associated unique document name to ensure version control. The deliverable and data identifier must be used in the deliverable filename.

The data identifier for the deliverable must be: <Deliverable identifier><Up-to-five-words-from the data name>_<project name acronym><followed by the version number (v01, v02, def)>

Example: D7.1_QualityAssurancePlan DeepU_def

The overall aim of DeepU is to create a deep (>4 km), closed-loop, vitrified, waterproof, non-cracked U-tube heat exchanger, by combining laser and cryogenic gas into a single technological drilling solution at laboratory scale in different rock types.

Relevant data will be collected annually to inform project partners and the European Commission (EC) on DeepU progress. The metadata, once ready, will be disseminated through the Annual Report. These reports will be published in full as appropriate in the following locations:

- The DeepU project website https://www.deepu.eu/
- A research data repository, compatible and compliant with OpenAIRE guidance. The repository will host scientific publications and all datasets associated with such publications.

The Confidential project data sets and reports will be hosted on:

- The DeepU Consortium private file sharing folder (VRE section of the website), this allows secure data share across partners. This area provides a space for information exchange and an archive for all the documentation produced along the Project lifespan.
- Individual partner's institutional online repositories will host and preserve data until the end of the project.

In accordance with Annex 5 of the grant agreement (Appendix A), the project partners are obliged to protect the results where these can be expected to be commercially or industrially exploited.

**Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?**

Keywords will be provided in the metadata in order to clearly identify the type of data contained in the report and increase the possibility for discovery and potential re-use (see section 2.3).

**Will metadata be offered in such a way that it can be harvested and indexed?**

Metadata Harvesting is favoured by the use of Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), a low-barrier mechanism for repository interoperability. In addition, as described in detail in section 2.3, some of the metadata is used to index the records to make retrieval easier.

## 3.2 MAKING DATA ACCESSIBLE

### 3.2.1 Repository

According to the procedure defined and agreed in the Consortium Agreement, the Consortium will comply with the Grant Agreement clause for the research data generated from the project, striving for maximum openness of data collected.

Research data should be 'FAIR', that is findable, accessible, interoperable and re-usable.

Having in mind that the "Guidelines for FAIR Data Management in Horizon 2020" [1] recognize the need to balance openness and protection of scientific information, commercialization and Intellectual Property Rights (IPR), privacy concerns, security as well as data management and preservation questions, DeepU Consortium has selected the Zenodo repository (https://zenodo.org) [2] for storing and preserving the project data. A reference to the DeepU project will be attributed to each uploaded dataset, and the same will be done for the new data that will be produced and organized as coherent datasets.

**Will the data be deposited in a trusted repository?**

Zenodo is hosted by CERN and is funded, among others, by the European Commission via the OpenAIRE projects

- FP7: OpenAIRE (246686), OpenAIREplus (283595)
- Horizon 2020: OpenAIRE2020 (643410), OpenAIRE-Connect (731011), OpenAIRE-Advance (777541), and OpenAIRE-Nexus (101017452).

Zenodo is developed and supported as a marginal activity, and hosted on top of existing infrastructure and services at CERN, in order to reduce operational costs and rely on existing efforts for High Energy Physics. CERN has some of the world's top experts in running large scale research data infrastructures and digital repositories in order to deliver a trusted digital repository.

**Have you explored appropriate arrangements with the identified repository where your data will be deposited?**

The Zenodo service ("Zenodo") is offered by CERN as part of its mission to make available the results of its work.

Access to Zenodo's content is open to all, for non-military purposes only. Content may be uploaded free of charge by those without ready access to an organized data center.

The term of use are available at the following link https://about.zenodo.org/terms/.

**Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?**

Zenodo assigns all publicly available uploads a Digital Object Identifier (DOI) to make the upload easily and uniquely citeable. Zenodo further supports harvesting of all content via the OAI-PMH protocol. For further information, please check the "DOI versioning" at the following link https://help.zenodo.org/.

### 3.2.2 Data

**Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and**

**contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.**

A mixed situation exists in terms of dataset openness. Although all datasets are shared within the Consortium, only a part of that amount is labelled as ORD at this moment. Furthermore, data sharing related to the technological innovation in WP1 and WP2 will be restricted as a legitimate reason to protect results that can reasonably be expected to be commercially or industrially exploited. Strategies to reduce such restriction will include agreeing on a limited embargo period and will be the focus also of WP5.

As soon as WP1 and WP2 data outcomes will be produced, each dataset will be examined in order to evaluate its exploitability potential. After that, it will be possible to relabel each dataset with the appropriate shareability policy, always aiming for the maximum openness and reusability of data.

**If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.**

Materials generated under the project will be disseminated in accordance with partners policies.

Publication of data shall occur during the project, if appropriate, or at the end of the project, consistent with normal scientific practices. Research data which documents, supports and validates research findings will be made available after the main findings from the final research dataset have been accepted for publication and/or 6 months after the laboratory measurements, the field data collection and numerical simulation are completed, according to the project workplan.

When an embargo status is defined, the partners will provide and end date for the embargo. The repository will restrict access to the data until the end of the embargo period; at which time, the content will become publically available automatically.

**Will the data be accessible through a free and standardized access protocol?**

Access to databases and associated software tools generated under the project will be available for educational, research and non-profit purposes related to geothermal energy developments. Such access will be provided using web-based applications, as appropriate. ZENODO digital repository tool, a simple and innovative service that enables researchers, scientists, EU projects and institutions to share, preserve and showcase multidisciplinary research results (data and publications) will be used as storage system for DeepU Project.

Access to data objects: Files may be deposited under closed, open, or embargoed access. Files deposited under closed access are protected against unauthorized access at all levels. Access to metadata and data files is provided over standard protocols such as HTTP and OAI-PMH. Use and re-use is subject to the license under which the data objects were deposited.

**If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?**

Restricted Access is also foreseen for data containing potential for patentability (see CA and GA agreement). These files will not be made publicly available and sharing will be made possible only by the approval of the creator of the original file.

The metadata about the restricted data will be shared in a repository in order to increase findability/discoverability and explain to others what they have to do to request access.

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

The access to restricted/sensitive data will be regulated by the SC and it is subjected by the approval of the creator of the original file. A direct request must be sent to the Consortium, so the identity of the person accessing these data can be ascertained. The identity of the project partner accessing the data will be ascertained via website login through the private section.

The access to files deposited under open access condition is free and will follow the Zenodo protocol.

**Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?**

The Steering Committee meeting is in charge to determine the data access policy of DeepU project. Every 6 months, during the General Assembly, the data access policy could be reviewed.

### 3.2.3 Metadata

**Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?**

Metadata is licensed under CC0, except for email addresses. All metadata is exported via OAI-PMH and can be harvested. The metadata format is described in section 2.3.

**How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?**

The research is stored safely for the future in CERN's Data Centre for as long as CERN exists. Therefore, data and metadata are guaranteed to remain available after the end of the project.

**Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?**

**When data will be prepared in a format different from the standards described in Table 1, the creator of the data will be responsible to provide reference to the software needed to access or read data and, in case, to include the relevant open source code, if available. More detail will be provided by partners once the questionnaire in Appendix A will be filled in.**

When data will be prepared in a format different from the standards described in Table 1, the creator of the data will be responsible to provide reference to the software needed to access or read data and, in case, to include the relevant open source code, if available. More detail will be provided by partners once the questionnaire in Appendix A will be filled in.

## 3.3 MAKING DATA INTEROPERABLE

**What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?**

Data will be collected and shared in a standardised way using a standard format for that data type (see Table 1). As required, reference will be made to any software required to run it. Given the scope of this project it is anticipated that publicly available software will be used to store data. Barriers to access through interoperability issues are not anticipated.

The metadata format will follow the convention of the hosting research data repository. A draft metadata format is set out below and this is subject to review in the next DMP update.

General Information

- Title of the dataset/output
- Dataset Identifier (using the naming convention outlined in Section 2.1)
- Responsible Partner
- Work Package
- Author Information
- Date of data collection/production
- Geographic location of data collection/ production
- The title of project and Funding sources that supported the collection of the data i.e. European Union's Horizon 2020 research and innovation programme under grant agreement No 840651.

Sharing/Access Information

- Licenses/access restrictions placed on the data
- Link to data repository
- Links to other publicly accessible locations of the data
- Links to publications that cite or use the data
- Was data derived from another source?

Dataset/Output Overview

- What is the status of the documented data? – "complete", "in progress", or "planned"
- Date of production
- Date of submission/publication
- Are there plans to update the data?
- Keywords that describe the content
- Version number
- Format - Post Script (PDF), Excel (XLSX, CSV), Word (DOC), Power Point (PPT), image (JPEG, PNG, GIF, TIFF).
- Size - MBs

Methodological Information

- Used materials
- Description of methods used for experimental design and data collection
- Methods for processing the data
- Instruments and software used in data collection and processing-specific information needed to interpret the data
- Standards and calibration information, if appropriate
- Environmental/experimental conditions
- Describe any quality-assurance procedures performed on the data
- Dataset benefits/utility

**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? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?**

The use of uncommon or specific ontologies or vocabularies is not foreseen. More information on this regard are expected once the questionnaire (see APPENDIX A) will be completed and will be updated, in case, in the next DMP.

**Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?**

Any reference to other data and datasets from previous research will be clearly stated.

## 3.4 INCREASE DATA RE-USE

**How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?**

Data used for validate scientific open access publication will be available for data generation and validation/interpretation /re-use.

There are different ways to validate data analysis and facilitate data re-use. In detail, DeepU project is expected to make use of:

- 'readme' file: any information that cannot be recorded in a structured way (i.e. as the values of fields in a data or metadata file) are recorded as free text within a readme file.
- file formats within the data file: it is possible to record information in addition to the main data content for example in XML or CSV standard files, that provide a way of recording sampling strategies and procedures as well as measurement values, variables and units of measurements.
- separate metadata file: some disciplines have developed special file formats or data structures for recording supporting information.
- published journal article: some of the information needed to understand data would normally be provided in a journal article reporting the research. In order to prevent duplication of effort, it is possible to refer to an article to provide more information about a dataset, but before doing so you should be sure that (a) the article provides sufficient detail and (b) that the article will be available as open access.

**Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?**

The digital research data generated in the action ('data'), in line with the FAIR principles, as soon as possible will be deposited in a trusted repository. Open access to the deposited data will be ensured via the repository, under the latest available version of the Creative Commons Attribution International Public License (CC BY) or Creative Commons Public Domain Dedication (CC 0) or a licence with equivalent rights, following the principle 'as open as possible as closed as necessary', respecting the beneficiary's legitimate interests, including regarding commercial exploitation, or any other constraints, justifying it in the DMP.

Metadata of deposited publications will be open under a Creative Common Public Domain Dedication (CC 0) or equivalent, in line with the FAIR principles (in particular machine actionable) and provide information at least about the following: publication (author(s), title, date of publication, publication venue); Horizon Europe funding; grant project name, acronym and number; licensing terms; persistent identifiers for the publication, the authors involved in the action and, if possible, for

their organizations and the grant. Where applicable, the metadata must include persistent identifiers for any research output or any other tools and instruments needed to validate the conclusions of the publication.

Open access to peer-reviewed scientific publications states that at the latest at the time of publication, a machine-readable electronic copy of the published version or the final peer-reviewed manuscript accepted for publication, is deposited in a trusted repository for scientific publications. In addition, immediate open access is provided to the deposited publication via the repository, under the latest available version of the Creative Commons Attribution International Public Licence (CC BY) or a licence with equivalent rights. The licence may exclude commercial uses and derivative works (e.g. CC BY-NC, CC BY-ND).

**Will the data produced in the project be useable by third parties, in particular after the end of the project?**

Even though the Consortium greatly endorse the precepts of open science, nevertheless, data sharing in the open domain can be restricted as a legitimate reason to protect results that can reasonably be expected (e.g. commercial or industrial exploitation). Strategies to limit such restrictions could include anonymizing or aggregating data, agreeing on a limited embargo period or publishing selected datasets not completely open.

When no restrictions are specified and justified, the data could be used by third parties, in particular after the end of the project. The data will remain re-usable at undetermined time

**Will the provenance of the data be thoroughly documented using the appropriate standards?**

Data provenance is the documentation of where a piece of data comes from and the processes and methodology by which it was produced. Put simply, provenance answers the questions of why and how the data was produced, where, when and by whom. This kind of information will be recorded as metadata (see section 2.3) to confirm the authenticity of data and to enable it to be reused.

In its simplest form, provenance can be recorded in a single README text file that describes the data collection and processing methods used. Provenance can also be recorded in a more structured way using specific elements in very generic metadata standards such as Dublin Core, to discipline-specific metadata standards such as ISO 19115-2.

**Are all relevant data quality assurance processes well described?**

Data quality assurance is the process of identifying and eliminating anomalies by means of data profiling and cleansing. Data quality control will be performed by the partner responsible for the RD creation. Duplicates, outliers, errors, and missing information will be detected in order to obtain accurate, complete, and consistent data, essential to track the progress of current projects and proposed initiatives. The following practices will be applied to each RD creation:

- Relevance: the data should be interpretable. This means that the partner has appropriate data processing methods, that the data format is interpretable by the partner software and that the legal conditions allow the partner to use such data.
- Accuracy: Ensuring the accuracy of the data by techniques like data filtering and outlier detection.
- Consistency of data: By checking internal and external validity of the data consistency can be ensured.
- Timeliness: The more up to date data suggests more precise calculations / organization.
- Compliance: it is important to check whether the data used complies with DeepU project progress or not.

# 4. OTHER RESEARCH OUTPUTS

**In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).**

**Beneficiaries should consider which of the questions pertaining to FAIR data above can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.**

In case of digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) project outputs, the partner responsible for it (i.e. Prevent for D1.3 Drill strings prototypes) will provide a plan for its management answering to the Questionnaire provided to all Beneficiaries during the 6M GA. The beneficiary's legitimate interests, including regarding commercial exploitation or any other constraints, will be respected.

# 5. ALLOCATION OF RESOURCES

**What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.)?**

The activities related to making the data/outputs open access are anticipated to be covered within the allocated budget for each work package. Further investigation of potential cost related to a repository need to be done. The repository will ensure that data is stored safely and securely and in full compliance with European Union data protection laws and in accordance with Annex 5 (Article 16-17) of GA.

**How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions).**

The costs related to open access to research data are eligible as part of the DeepU Horizon Europe grant. The costs of making scientific publications, hosting a project website and the partners and open access data repositories are contained within the DeepU budget as eligible costs.

**Who will be responsible for data management in your project?**

UNIPD has Project Coordinator, in collaboration with consortium partners and especially CNR (WP6 leader), is responsible to manage the data generated during the project. The PC will identify an appropriate data repository to store and safeguard the datasets but ensure that data is readily accessible. All WPs leaders will be responsible of the quality of data generated during the project and will assists the PC in organizing the different datasets.

CNR, in charge of WP6: Dissemination and Communication, in collaboration with consortium partners, will oversee the identification of which datasets will be disseminated and the most appropriate means of disseminating this data.

**How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long).**

Resources for long term preservation, associated costs and potential value, as well as how data will be kept beyond the project and how long, will be discussed by the Consortium's General Assembly

(GA) at the DeepU Steering Committee (SC) meeting. The use of Zenodo repository is useful in dealing with the long-term preservation of data related to DeepU project.

The PC is responsible for all data management during and after data collection, while each WP leader is responsible for preparing the datasets related to the corresponding WP, indicating which data can be made public and which are bound by IPR agreements. Copies of all datasets will be saved by WP leaders to local servers and external drives, while the final version will be updated in Zenodo.

# 6. DATA SECURITY

**What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?**

According to the best practices in the matter of data preservation, all files uploaded to Zenodo will be stored in CERN's 18 petabytes disk cluster. In that infrastructure, each file copy has two replicas located on different disk servers. For each file, two independent MD5 checksums are stored. One checksum is stored by Invenio (a Free Open Source Software providing repository/document management platform, an integrated library system and a code library to build large-scale information systems) and used to detect changes to files made from outside of Invenio. The other checksum is stored by EOS (the primary low latency storage infrastructure for physics data from the Large Hadron Collider), and used for automatic detection and recovery of file corruption on disks.

From the security standpoint, physical disks in which uploaded data will be stored are set at CERN Data Centre located on CERN premises, all physical accesses will be restricted to a limited number of staff with appropriate training, and who have been granted access in line with their professional duties. Remote access to the servers is restricted to Zenodo staff and the operating system and installed applications are kept updated with latest security patches via the automatic configuration management system Puppet. CERN Security Team runs both host, network based intrusion detection systems, and monitors the traffic flow, pattern and contents into and out of CERN networks in order to detect attacks. All access to zenodo.org happens over HTTPS, except for static documentation pages which are hosted on GitHub Pages.

Zenodo stores user passwords using strong cryptographic password hashing algorithms. Users' access tokens to GitHub [3] and ORCID [4] are stored encrypted too.

Zenodo allows users to upload files under closed access. Closed access means that zenodo.org users will not be able to access the files you uploaded. The files are however stored unencrypted and may be viewed by Zenodo operational staff under specific conditions. This means that "closed access" on Zenodo is not suitable for secret or confidential data.

In case of sensitive "closed access", "secret" or "strictly confidential" datasets, data owner and the involved DeepU partners will evaluate together the most suitable ad hoc solution to protect and secure such information, including hosting in DeepU website [5].

**Will the data be safely stored in trusted repositories for long term preservation and curation?**

The use of Zenodo repository comply with the request of long-term preservation and curation of project data, ensuring also the data security over time.

Using standard or open data formats ensures longer-term usability of data. Therefore, project files will be converted into standard open formats for long term preservation of the data (see Table 1).

# 7. ETHICS

A dedicated section (section 4) has been included in the DoA of DeepU project to ensure that ethical principles are used in the framework of the project, including among others, specific aspects about data collecting in real cases and scientific moral belief [7].

There are no known ethical or legal issues that can have an impact on data sharing, because research activities within the project do not directly involve the collection of any personal data or other sensitive information. Furthermore, metadata too are not relevant from this standpoint, as the data collection process will include anonymizing or aggregating data strategies, whenever such actions will be necessary.

In the whole processes of data collection, data analysis, dataset creation, dataset storing and dataset access, the Consortium will make certain that the General Data Protection Regulation (GDPR), which has entered into force in May 2018, is ensured, especially in regards to protection of private personal data.

# 8. OTHER ISSUES

**Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?**

The lead partner and all beneficiaries are subject to the CA and GA. All data must be collected, stored and disseminated in accordance with these Agreements.

# REFERENCES

[1] H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020 (2016). EUROPEAN COMMISSION Directorate-General for Research & Innovation.

[2] http://zenodo.org/ - ZENODO data repository website.

[3] https://github.com/ - GitHub Inc. web-based hosting service for computer code.

[4] https://orcid.org/ - ORCID website.

[5] DeepU EU Project: https://www.deepu.eu/

# APPENDIX A

## QUESTIONNAIRE ON RESEARCH DATA

The Questionnaire on Research Data includes an overview of the datasets that are expected to be generated by the project. At this purpose, a survey amongst the Consortium will be conducted and all partners are asked to complete the following questionnaire about DMP.

In addition, a Data format template structure is provided for future use in the project.

| TOPIC | Q. No | |
|---|---|---|
| General information | A1 | Please, enter your DeepU partner name |
| | A2 | At this time, did your actions during the project generated any research data? |
| | A3 | At this time, did you used or modified any data acquired from the outside of the project? |
| | A4 | At this time, did you published any open access scientific peer-reviewed publication that includes research data? |
| | A5 | At this time, how many dataset have you generated,used or modified? |
| | A6 | Do you plan to generate even more research data during the lifespan of the project? |
| | A7 | Do you plan to use or modify data acquired from the outside of the project during the lifespan of the project? |
| | A8 | Do you plan to publish open access to sceintific peer-reviewed publications that make use of research data during the lifespan of the project? |
| | A9 | How many dataset do you plan to generate, use or modify during the whole lifespan of the project? |

| TOPIC | Q. No | |
|---|---|---|
| Data summary | B1 | What is the short name that you would use to describe this specific dataset? |
| | B2 | What is the purpose of the data collection/generation and its relation to the tives of the project? |
| | B3 | What type and formats of data will the WP generate/collect? |
| | B4 | Will you re-use any existing data and how? |
| | B5 | What is the origin of the data? |
| | B6 | What is the expected size of the data? |
| | B7 | To whom might it be useful ('data utility')? |

| TOPIC | Q. No | |
|---|---|---|
| Data summary | C1 | 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. |
| | C2 | How will the data be made accessible? |
| | C3 | What methods or software tools are needed to access the data? |
| | C4 | Is documentation about the software needed to access the data included? |
| | C5 | Is it possible to include the relevant software (e.g. in open surce code)? |
| | C6 | If there are restrictions on use, how will access be provided? |
| | C7 | How will the identity of the person accessing the data be ascertained? |

| TOPIC | Q. No | |
|---|---|---|
| Making data interoperable | D1 | 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)? |
| | D2 | What data and metadata vocabularies, standards and methodologies will you follow to make your data interoperable? |
| | D3 | 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? |

| TOPIC | Q. No | |
|---|---|---|
| Increase data re-use (through clarifying licences) | E1 | How will you the data be licensed to permit the widest re-use possible? |
| | E2 | When will the data 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 avaialbel as soon as possible. |
| | E3 | 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. |
| | E4 | How long is intended that the data remains re-usable? |

Data format template structure

| | |
|---|---|
| DATASET SHORT NAME | Unique short yet descriptive name of this dataset |
| DATASET REFERENCE IN REPOSITORY (i.e. DOI) | Digital Object Identifier assigned by Zenodo |
| DATASET DESCRIPTION | Long description of this dataset, including detail on variables and parameters, material and methods, time, location and condition of the collecting process (if applicable), reasons that lead to data acquisition actions. |
| STANDARDS AND METADATA | If applicable, standard of measurement, and data exchange protocols are descripted. |
| DATA SHARING | Information regarding shareability and access right of this specific dataset according to Zenodo (i.e. Open Access, Embargoed Access, Restricted Access, Closed Access). Indication of License of use. |
| AFFECTED DOCUMENTS (WPs, Deliverables, and OA Publications) | WPs and Open Access Publication that make use of this specific dataset. |