NEARDATA

Funded by the European Union

HORIZON EUROPE FRAMEWORK PROGRAMME

NEARDATA

(grant agreement No 101092644)

Extreme Near-Data Processing Platform

D1.3 Data Management Plan

Due date of deliverable: 30-06-2023
Actual submission date: 29-06-2023

Start date of project: 01-01-2023 Duration: 36 months

Summary of the document

Document Type	Data Management Plan
Dissemination level	Public
State	v1.0
Number of pages	11
WP/Task related to this document	WP1 / T1.1
WP/Task responsible	URV
Leader	Vanesa Ruana (URV)
Technical Manager	Pedro Garcia (URV)
Quality Manager	Marc Sanchez (URV)
Author(s)	Vanesa Ruana (URV)
Partner(s) Contributing	URV
Document ID	NEARDATA_D1.3_Public.pdf
Abstract	Document describing project data management strategies. The different experiments, workloads, benchmarks, and results will be delivered as Open Research Data for the community.
Keywords	Data Management Plan, Open Access, Open Research Data, FAIR data, ORDP

History of changes

Version	Date	Author	Summary of changes
0.1	11-04-2023	Vanesa Ruana	First draft.
0.2	25-05-2023	Several partners	Add the datasets.
1.0	29-06-2023	Vanesa Ruana	Final version.

Table of Contents

List of Abbreviations and Acronyms

API	Application Programming Interface
CC	Creative Commons
Cholec80	Endoscopic video dataset containing 80 videos of cholecystectomy surgeries from University of Strasbourg
CSV	Comma-separated values
DMP	Data Management Plan
DOI	Digital Object Identifier
DSAD	Dresden Surgical Anatomy Dataset
HeiChole	Surgical Dataset for surgical workflow and skill analysis from University Hospital Heidelberg und NCT

1 Executive summary

NEARDATA project is committed to good data management. In an effort to provide a management life-cycle of the data needed to validate results in scientific publications, this version of the Data Management Plan (DMP) has been provided as deliverable D1.3. This DMP describes how the research data will be made findable, accessible, interoperable and reusable. This DMP also presents a summary of the existing datasets that are currently known to be used over the course of the project.

2 Data Summary

NEARDATA project wants to enable open access and reuse of the research data generated by Horizon Europe projects. NEARDATA has the commitment to:

- Develop a Data Management Plan (DMP).
- Deposit the project's data in a research data repository.
- Ensure third parties can freely access, mine, exploit, reproduce and disseminate our data.
- Provide related information and identify (or provide) the tools needed to use the raw data to validate our research.

In particular, the project applies to:

- The data (and associated metadata) needed to validate the results presented in scientific publications.
- Other curated and/or raw data (and associated metadata) that is specified within this Data Management Plan.

The main goal of the NEARDATA project is to design an Extreme near-data processing platform to enable consumption, mining and processing of distributed and federated data without needing to master the logistics of data access across heterogeneous data locations and pools. We go beyond traditional passive or bulk data ingested from storage systems towards next generation near-data processing platforms both in the Cloud and in the Edge. In our platform, Extreme Data Types include both metadata and trustworthy data connectors enabling advanced data management operations like data discovery, mining, and filtering from heterogeneous data sources.

Table 1 presents a summary of the existing datasets that will be processed to validate the results of the NEARDATA project.

Table 1: Used datasets

UD1	
Name:	Resource for Genetic Epidemiology Research on Aging (GERA)
Origin:	The Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH)
Access:	The database of Genotypes and Phenotypes (dbGaP) - phs000788.v2.p3
Volume:	255G
Variety:	Genotype and Phenotype
Frequency of update:	NA
UD2	
Name:	Finland-United States Investigation of NIDDM Genetics (FUSION)
Origin:	Finland-United States Investigation of NIDDM Genetics (FUSION)
Access:	The database of Genotypes and Phenotypes (dbGaP) - phs000867.v1.p1
Volume:	9.2G
Variety:	Genotype and Phenotype

Frequency of update:	NA
UD3	
Name:	Wellcome Trust Case Control Consortium (WTCCC)
Origin:	The Wellcome Trust Case Control Consortium (WTCCC)
Access:	European Genome-Phenome Archive (EGA) - EGAS00000000067
Volume:	34G
Variety:	Genotype and Phenotype
Frequency of update:	NA
UD4	
Name:	Gene Environment Association Studies initiative (GENEVA)
Origin:	The trans-NIH Genes, Environment, and Health Initiative (GEI)
Access:	The database of Genotypes and Phenotypes (dbGaP) - phs000091.v2.p1
Volume:	43G
Variety:	Genotype and Phenotype
Frequency of update:	NA
UD5	
Name:	Northwestern University NUgene project (NUgene)
Origin:	Northwestern University
Access:	The database of Genotypes and Phenotypes (dbGaP) - phs000237.v1.p1
Volume:	6.1G
Variety:	Genotype and Phenotype
Frequency of update:	NA
UD6	
Name:	Dresden Surgical Anatomy Dataset (DSAD) [1]
Origin:	University Hospital Dresden and National Center for Tumor Diseases Dresden
Access:	Dresden Surgical Anatomty Dataset
Volume:	11 different organs, 33 patients (20.5GB)
Variety:	Organ Segmentation
Frequency of update:	NA
UD7	
Name:	HeiChole [2]
Origin:	University Hospital Heidelberg and National Center for Tumor Diseases Dresden
Access:	HeiChole Dataset
Volume:	33 patients, $\sim$200GB
Variety:	Surgical Workflow Analysis
Frequency of update:	NA
UD8	

Name:	Cholec80 [3]
Origin:	University of Strasbourg
Access:	Cholec80
Volume:	80 patients, ∼80GB
Variety:	Surgical Workflow Analysis
Frequency of update:	NA
UD9	
Name:	Synthetic video data [3]
Origin:	National Center for Tumor Diseases Dresden
Access:	Synthetic Video Data
Volume:	∼30GB
Variety:	Surgical Workflow Analysis and Segmentation
Frequency of update:	NA
UD10	
Name:	Datasets available through the METASPACE platform
Origin:	The METASPACE platform
Access:	Public datasets (over 8000, from over 500 METASPACE users) and selected private datasets belonging to our (EMBL) group
Volume:	10+ TB
Variety:	The size of an individual dataset varies from 100 MB to 200 GB. The sample types and metadata vary hugely.
Frequency of update:	Multiple times per day

Aside from these datasets and benchmarks, the NEARDATA project will generate other data to validate the results presented in scientific publications (test data, APIs, source code used to perform analysis, etc.). All this data will be made available as open data and its re-use will be encouraged. As the project progresses and data is identified and collected, further information on data details will be provided. Table 2 presents a summary of the already generated datasets in the process of validating the results of the NEARDATA project.

Table 2: Generated datasets

GD1	
Name:	Synthetic genotype and phenotype
Description:	Simulated genotype and phenotype for thousands of European individuals using EpiGEN
Access:	NA
Volume:	237M
Variety:	NA
DOI:	https://doi.org/10.1093/bioinformatics/btaa245
GD2	
Name:	Surgical Training Phantom Data (currently not public)
Description:	Continuously recorded phantom data from the Experimental Operating Room at the NCT
Access:	Data will be made accessible to project partners up request
Volume:	Increasing (currently $\sim$ 20GB)
Variety:	
DOI:	NA

NEARDATA data will not only be useful for the current and future generation of big data and cloud technologies researchers, but also big data practitioners and companies (from SMEs to multinationals) with a vested interest in new programming models for data analytics.

3 FAIR data

In general terms, research data should be **FAIR**, that is **findable**, **accessible**, **interoperable** and **reusable** [4].

3.1 Making data findable

Used data In order to ensure that the data used in the project is easily findable, we will make an effort to include standard identification mechanisms in all our publications, source code and tutorials. Although not all datasets used in the project provide these identification mechanisms, we will take special care to provide the necessary instructions, metadata and tools for locating and processing those datasets.

Produced data NEARDATA is expected to deposit generated data in an open online research data repository. We have selected Zenodo as our data repository of choice. Zenodo is an OpenAIRE and CERN collaboration that allows researchers to deposit both publications and data, providing tools to link related items through persistent identifiers and data citations. Zenodo automatically assigns a Digital Object Identifier (DOI) to each item to make them easily and uniquely citable. Moreover, Zenodo is set up to facilitate the finding, accessing, re-using and interoperating of data sets, which are the basic principles that ORD projects must comply with.

To this end, we have created a NEARDATA community in Zenodo[1] to gather all the open data contributions of the project. The repository allows to assign specific keywords to each dataset as well as a minimum of the DataCite's Metadata Schema [5] recommended terms.

Whenever possible (according to publisher copyright policies regarding open access), research publications will also be uploaded to this repository to ensure the maximum dissemination of the results of the project. Publications will be linked to its associated research data.

[1] `https://zenodo.org/communities/neardata-eu/`

Source code. To make the source code open to the general public, we have created a code repository in GitHub for NEARDATA[2]. GitHub is currently one of the most popular code management systems due to the advanced features and easy management that it provides to developers. This has various potential benefits to the management and dissemination of NEARDATA source code: for instance, GitHub is well-known across developer communities, which facilitates the access to the source code of NEARDATA. Moreover, GitHub offers a plenty of options to fork/branch/merge versions of a software project that enables third-parties to easily extend the source code developed in NEARDATA (even for internal use). Additionally, we'll also make source code citable and uniquely identifiable by automatically archiving software releases in Zenodo [6].

As of the last release of this document, the NEARDATA Github profile contains 1 individual repository hosting software results.

Finally, the NEARDATA web page[3] will list all project results and provide links to their respective repositories in Zenodo or GitHub.

3.2 Making data openly accessible

It is our intention that all data produced during the NEARDATA project is openly accessible as the default. Pre-existing datasets used in the experiments are mostly public and openly available (see Table 1).

Potential users will find out about the data through publications and the NEARDATA website. Data will be made available on publication of the associated paper and will be made accessible through the Zenodo repository.

3.3 Making data interoperable

Interoperability of data produced within the NEARDATA project is promoted through best practices. Data formats should adhere to widely used standards and should be compliant with available software applications. Where possible, standard codes will be followed (e.g.: ISO 639 for language codes, ISO 3166 for country codes, NUTS for region codes, ...).

As the project progresses and data is identified and collected, further information on making data interoperable will be outlined. Specifically, information on data and metadata vocabularies, standards or methodology to follow to facilitate interoperability and whether the project uses standard vocabulary for all data types present to allow interdisciplinary interoperability.

3.4 Increase data re-use (through clarifying licenses)

Data will be made accessible, and therefore available for re-use, within one month of the publication of the related peer-reviewed scientific article. Data will be shared under the Creative Commons Attribution 4.0 International Public License (CC BY 4.0) [7]. This license guarantees the widest possible re-use and redistribution while only requiring that appropriate credit is given.

The shared data will remain re-usable after the end of the project by anyone interested in it, with no access or time restrictions.

3.5 Management principles

The protocol below summarizes the management principles behind making generated research data FAIR:

[2] `https://github.com/neardata-eu/`

[3] `http://NEARDATA.eu`

PROTOCOL: Storing generated research data in NEARDATA project and making it FAIR

Beneficiaries will follow these procedures for each dataset collected or generated during the NEARDATA project:

- Store and make findable the dataset in the NEARDATA community of the Zenodo repository.
- Ensure that publications and research data behind them are cross-referencing each other through standard identification mechanisms.
- Ensure that each dataset provides metadata, particularly regarding access rights, licenses, and funding information.
- Each Work Package Leader is responsible for storing relevant research data to the repository.
- Data will be made accessible within one month of the publication of the related peer-reviewed scientific article.

Beneficiaries will follow these procedures for source code generated during the NEARDATA project:

- Store the source code under the NEARDATA organization in GitHub repository.
- Provide a comprehensive README file with instructions to run the code.
- Store each release of the source code to Zenodo repository and cross-reference related datasets and publications.
- Each Work Package Leader is responsible for storing relevant source code to the repository.

4 Allocation of resources

Regarding Open Access to research data, archiving at Zenodo is free of charge. Storing source code at the GitHub repository is also free of charge. Therefore, no costs are currently foreseen regarding the long term preservation of data.

The project coordinator has the ultimate responsibility for the data management in the project.

5 Data security

As NEARDATA delegates the archiving of data to Zenodo, their policies regarding data security apply:

- **Replicas**: All data files are stored in CERN Data Centres, primarily Geneva, with replicas in Budapest. Data files are kept in multiple replicas in a distributed file system, which is backed up to tape on a nightly basis.
- **Retention period**: Items will be retained for the lifetime of the repository. This is currently the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least.
- **File preservation**: Data files and metadata are backed up nightly and replicated into multiple copies in the online system.

- **Fixity and authenticity**: All data files are stored along with a MD5 checksum of the file content. Files are regularly checked against their checksums to assure that file content remains constant.
- **Succession plans**: In case of closure of the repository, best efforts will be made by CERN to integrate all content into suitable alternative institutional and/or subject based repositories.

6 Ethical aspects

There is no sensitive ethical issue of collecting, storing, processing and archiving data raised by the research of the NEARDATA project. Any potential ethical issue raised during the life of the project may be reported to the NEARDATA project board, which would, if necessary, raise immediate awareness of internal consortium members' executives, in order to take appropriate actions to resolve this issue.

Concerning potential ethical conflicts all issues will be resolved through the procedures depicted in relative legal documents (e.g., Consortium Agreement) and Commission guidelines.

7 Conclusions

This document is the unique version of the NEARDATA Data Management Plan. It presents the status of reflection within the NEARDATA consortium about the research data used, collected or generated alongside the project. This DMP describes how the research data has been, and will be made findable, accessible, interoperable and reusable.

References

[1] M. Carstens, F. M. Rinner, S. Bodenstedt, A. C. Jenke, J. Weitz, M. Distler, S. Speidel, and F. R. Kolbinger, "The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science," Scientific Data, vol. 10, no. 1, pp. 1–8, 2023.

[2] M. Wagner, B.-P. Müller-Stich, A. Kisilenko, D. Tran, P. Heger, L. Mündermann, D. M. Lubotsky, B. Müller, T. Davitashvili, M. Capek, et al., "Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark," Medical Image Analysis, vol. 86, p. 102770, 2023.

[3] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "Endonet: a deep architecture for recognition tasks on laparoscopic videos," IEEE transactions on medical imaging, vol. 36, no. 1, pp. 86–97, 2016.

[4] M. Wilkinson and et al, "The FAIR Guiding Principles for scientific data management and stewardship," Nature Scientific Data, no. 160018, 2016.

[5] DataCite Metadata Working Group, "DataCite Metadata Schema for the Publication and Citation of Research Data. Version 4.2. DataCite e.V.." https://doi.org/10.5438/bmjt-bx77, 2019.

[6] GitHub, "Making your code citable." https://guides.github.com/activities/citable-code/, 2016.

[7] Creative Commons, "Creative Commons Attribution 4.0 International Public License." https://creativecommons.org/licenses/by/4.0/legalcode, 2013.