D1.2: Updated DMP RP1

Dissemination level:	PU
Document type:	R — Document, report
Version:	1.0
Date:	28-04-2023

Funded by the European Union
European Innovation Council

The HYBRAIN project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101046878.

Document Details

Project Number	101046878
Project title	HYBRAIN
Title of deliverable	Update DMP RP1
Due date of deliverable	30 April 2023
Work package	WP1
Author(s)	Gabi Maris (UT)
Reviewer(s)	Iris Schutte (UT)
Approved by	Coordinator
Dissemination level	PU: Public
Document type	R — Document, report
Total number of pages	14

Partners

- University of Twente (UT)
- University of Heidelberg (UHEI)
- Trust-IT SRL (Trust-IT)
- IBM research GMBH (IBM)
- The Chancellor, Masters and Scholars of the University of Oxford (UOXF)

Abstract

This deliverable represents the Data Management Plan (DMP) describing how HYBRAIN's research data is handled and the measures that will be taken to ensure the data availability and re-use. We make use of the DMP template for Horizon Europe provided by the European Commission and provide information about the research data collected and generated within the project.

Table of Contents

List of tables

Symbols, abbreviations and acronyms

D	Deliverable
DMP	Data Management Plan
EC	European Commission
EISMEA	European Innovation Council and SMEs Executive Agency
IBM	IBM research GMBH
RP	Reporting Period
Trust-IT	Trust-IT SRL
UHEI	University of Heidelberg
UOXF	The Chancellor, Masters and Scholars of the University of Oxford
UT	University of Twente
WP	Work Package

1 Introduction

This deliverable describes HYBRAIN's first version of DMP developed during the reporting period 1 (RP1) outlining the measures that will be taken during the project in order to maximize access and re-use of the data for further purposes and applications. We make use of the Horizon Europe DMP template provided by the European Commission and include information on the collected data in WP2, WP3, WP4 and WP5 by the HYBRAIN's partners. This DMP will be updated in the deliverables D1.3 *Updated DMP RP2* and *D1.3 Updated DMP RP3*.

2 Objectives

The aims of this deliverable presenting the first version of the HYBRAIN's DMP are:

- to describe the way data will be handled during the project.
- to ensure the availability and utility of HYBRAIN's research data.

3 HYBRAIN's data management plan

3.1 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.

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

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

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

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

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

We will use data that has been previously collected in the partner's labs and the information available in publications. No constraints regarding re-use of this data exist except confidential data from specific projects or data for which intellectually property rights and commercial interests are relevant. Furthermore, for testing the applicability of our hardware we will make use of publicly available benchmark tasks and data bases, such as the MNIST handwritten digit data set.

The generated and re-used data can be divided in the categories presented below. The corresponding experiment type/tools the formats and the sizes of the files, are summarized in Table 1. The size of the data files depends on the number of collected points, the type of experiment or calculation method and the data format.

- **Fabrication data**: generated for the fabrication of electronic and nanophotonic devices.
- **Measurements data**: generated from the different electrical and optical characterization measurements.
- **Model data**: training of neural network models used in the project.
- **Programming data:** python scripts.
- **Analysis data** generated from the measured raw data. It is data on which operations have been done.
- **Miscellaneous data** generated when a new setup is built, or parts of a setup are changed.

- **Publication data**: all the raw, processed data and illustrations belonging to publications and transferred to the external data repository for open access.
- **Management data:** consortium meetings documents, periodic and continuous reports for the European Commission.
- **Communication data** generated for the communication and activities of the project.

Table 1 Type of data generated in HYBRAIN and the corresponding experiment type/tools, parameters, file formats and sizes.

Type of data	WP#	Type experiment/ Activity/Tools	Parameters	File types/ formats	File sizes
Fabrication	2,3,4,5	Microfabrication technologies - Optical microscopy - Electron microscopy - Scanning Probe Microscopy	- Distances - Heights	tiff, png, jpg	1-20 MB per microscopy image
Measurement	2,3,4,5	Electrical characterization:	- voltage - current	txt (ascii)	Individual measurement: kBs – Gbs (single line traces - long measurements)
Model	2,3,4,5	Training of neural network models	Various training hyperparameters	Parameters of the trained networks.	1-500MB per network.
Programming	2,3,4,5	Python scripts	Not applicable	py	kBs
Analysis	2,3,4,5	- Data smoothing - Filtering	- voltage - current	txt (ascii), pdf	kBs-GBs
		-MVMs (Matrix vector multiplication) performed using in-memory computing hardware	Various techniques for programming the crossbar array - Various weights used for MVMs	txt (ascii), pdf	kBs
Miscellaneous	2,3,4,5	- CAD drawings - design files - printed circuit boards - setup schematics and images - manuals for instruments	- dimensions - material types - electronic elements	dxf, dc2, jpg, pdf	several MB/file
Publication	2,3,4,5	Not applicable	Not applicable	pdf, jpg, png, tiff	below 1 GB

Management	1	- meetings documents: presentations, agenda, minutes. - Templates deliverables, presentations, posters. - Financial, technical reports and deliverables for the EC.	Not applicable	pdf, jpg, png, doc, xls, ppt	A few GB
Communication	6	News items, newsletters, annual bulletins, videos, promotional materials, press releases, etc.	Not applicable	jpg, mp4, pdf, png, tiff	A few GB

The data is collected by all project partners conducting research within the work packages WP2, WP3, WP4 and WP5 with the aim to fulfil the project's tasks and to reach the objectives summarized in table 2.

Table 2 HYBRAIN objectives per work package.

WP#	Objective
2	Objective 1: To realize a photonic convolution processor (PCP) for ultralow-latency feature extraction. Photonic analog computing enables carrying out matrix-vector multiplications in a single time step and will allow for parallelization via wavelength division multiplexing (WDM) at multi-GHz modulation frequencies.
3	Objective 2: To interface the PCP with AIMC where the initial convolution layers are implemented via PCP and the later convolution layers and the final fully connected linear layers are implemented using AIMC.
4	Objective 3: To interface the PCP and AIMC with DNPUs for nonlinear classification. By nonlinear transformation of the electrical input signals into a high-dimensional space, linearly inseparable data become linearly separable, reducing the required complexity of the PCP and AIMC.
5	Objective 4: To demonstrate ultra-low latency AI edge inference applications based on a hybrid electronic photonic brain-inspired architecture.
1	Objective WP1: To ensure strategic, financial and contractual management of the consortium and to ensure the day-to-day operational project management.
6	Objectives WP6: To enable external stakeholders to become aware of, learn from and implement project results; To prepare for exploitation of IP protected project results.

The estimation of the total expected size of the collected data is between 1 and 10 TB.

All the generated and re-used research data originate from the partners laboratories and research groups. The raw measurement data is collected through the digital reading of commercial or home-built instruments and electronic measurement apparatuses. In addition, we re-use data from publicly available benchmark tasks and data bases, such as the MNIST handwritten digit data set.

The data belonging to publications will be stored in a trusted repository (see section 3.2.2) allowing anyone to download and use the data in their research. The raw measurement data is expected to be useful to other researchers and scientists within the in-memory computing research community. Analysis and publication data will be useful to a broader audience ranging from scientists to industry and end

users. It is expected that the generated data will be used by other researchers within the research groups of our partners.

The management data is useful to all the consortium partners. For official documents like deliverables, financial, periodic and continuous reports, the utility is extended to the European Commission. The communication and dissemination datasets are useful to all HYBRAIN stakeholders, allowing them to become aware of and to learn about how to benefit from the project's results.

3.2 FAIR data

3.2.1 Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?

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.

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

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

The trusted repositories provide a DOI number to uploaded data sets which make the datasets identifiable.

Metadata files are generated along the chain from raw data collection to the final analysis by means of README files (.html and .pdf) describing the parameters, settings, analysis steps. These are stored together with the corresponding files within network drive of each partner organization. We follow the Dublin Core metadata schema for data description and aim to document all "relevant" data in such a way that they are understandable to researchers in the field.

- Raw data collected from experimental instruments have rich digital metadata attached to the generated data files.
- Analysed data will be documented including information about the source data, analysis type, protocols, etc.
- For all data belonging to publications, we will follow the metadata scheme used by the zenodo.org and 4TU Centre for Research Data, both based on Dublin Core metadata schema.

The metadata schemes used by the trusted repository include appropriate keywords making the data sets belonging to publications findable.

By making use of the metadata scheme provided by the by the zenodo.org and 4TU Centre for Research Data, the metadata are indexed and searchable. The repositories support harvesting of metadata via a standard protocol.

3.2.2 Making data accessible

Repository:

Will the data be deposited in a trusted repository?

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

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

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.

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.

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

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

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

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

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?

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

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)?

Repository:

All raw and analysed data belonging to publications will be made available through Open Access (OA) publication. We have identified two candidates for data repositories that are most suitable for the project: zenodo.org and 4TU Centre for Research Data. These repositories generate digital object identifiers (DOI) and provide long-term storage for the deposited datasets. zenodo.org is an OA trusted repository based at CERN, Switzerland and is free of charge. The 4TU repository has a CoreTrustSeal, is based in the Netherlands and for the UT researchers is free of charge for storages up to 1 TB per year. It generates DOIs and provides long-term storage for the deposited datasets.

Data:

All raw and analysed data underlying a journal publication will be made available for verification/replication and reuse, i.e., will be made freely available together with the relevant metadata (including README.txt files describing the meaning of all numerical values in provided data files) on the zenodo.org and 4TU public repositories fulfilling the FAIR requirements.

Exception from Open Access are exploitable results that require careful intellectual property and justify an embargo period. Access will be restricted for IPR protection until the patent publication date or after the embargo period elapses. The 4TU.ResearchData and zenodo.org allow setting an embargo.

The chosen repository offers a free and standardized protocol to access the data. Data will be available upon open access principles without the need for a user account.

Metadata:

Data and metadata will remain available for at least 15 years at 4TU and 20 years at zenodo.org, Metadata is always openly available even for data under an embargo.

3.2.3 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?

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?

Will your data include qualified references1 to other data (e.g. other data from your project, or datasets from previous research)?

The research data together with the required metadata are stored in open standard file formats (see Table 1). We will use the meta data standard of the zenodo.org and 4TU repository which are similar to Dublin Core. This facilitates the exchange and re-use of data inside and outside the consortium. Furthermore, the interoperability of the data and metadata is insured by making use of common vocabularies used in the research field literature. New or uncommon vocabularies necessitating explanations, i.e. for interdisciplinary interoperability, will be clarified. Full text will be used before using any abbreviations.

3.2.4 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.)?

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?

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

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

Describe all relevant data quality assurance processes.

All the data is accompanied by rich metadata and documentation describing all steps in the chain from collection to final analysis. This is done by means of README files describing the used methodology, parameters, settings, analysis steps. Detailed documentation enables that all steps in the chain can be verified by others, ensuring the quality of the work.

The datasets that are deposited into the trusted repositories will be associated with CC-BY 4.0 Creative Commons licenses to protect the authorship of the data. The Attribution license CC-BY allows re-distribution and re-use of a licensed work on the condition that the original creator is appropriately credited.

The data will remain openly accessible in the chosen open-access repository for at least 20 years for zenodo.org 15 years for 4TU Centre for Research Data.

For a through data description, we follow the Dublin Core metadata schema and README files (see point a).

[1] *A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/)*

3.3 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.

Code/software will be stored as soon as possible on GitHub with open access and will be version controlled. GitHub is integrated with the zenodo.org and 4TU repository which will promote the findability and re-use of the code. The codes developed on GitHub are easily shared with the interested externals via the repositories. Language of the code/software is in open format (such as Python) which fosters interoperability, and re-usability.

The communication data related to the research data and findings will be made public and will be published online mainly via the website and social media channels. The DOI number of the data sets will be added to the website promoting their findability. These data will be allowed for reuse, provided that the reference is made to the source of the information, e.g., the URL of the website: https://hybrain.eu.

3.4 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.) ?

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)

Who will be responsible for data management in your project?

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)?

The amount of data storage needed during the research and at long term are available at the partners organizations offering a secure archive at no costs. Sufficient storage free of charge is available in the lab for the physical samples. zenodo.org is free of charge. The 4TU repository for data is free of charge for storages up to 1 TB per year for UT researchers and covers sufficiently the needed storage capacity.

The costs for the partners' network storages are covered by the research group of each partner.

The main PI of each partner is responsible for the data management. The partners PIs and research staff are supported by own institution staff specialized in data management, FAIR data, and open science. These services are free of charge.

All the research data generated during the project will be preserved at the partner's network drive or archive system for at least 10 years. The data generated will be preserved in the archive system of the partners' institutions, preserving data at long term. The data on the archive is immutable (no removal within 10 years and no change) and is searchable for the research group in which has been generated. The data belonging to publications will remain openly accessible in the chosen open-access repository for at least 15 years at no costs.

3.5 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)?

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

Data is stored on local computers of each partner research group and transferred as soon as possible to the partners institutions network storage. These network storages are protected by login and password credentials tied to the institution's accounts. The institution server has automatic daily backups to hard disks on a physically different location. The data can be recovered by contacting each institution's IT services. The UT network storage is ISO27001 and NEN7510 certified and is daily backed up. Backups are kept for 28 days. The IT department can restore data within 1 day in case of an incident.

The data will be stored in zenodo.org, 4TU repository. The data archive of each institution is also available for long-term archiving.

3.6 Ethics

Are there, or could there be, any ethics 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).

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

The objective of HYBRAIN is to develop an energy-efficient physical-substrate based platform for brain-inspired computing using electrons and photons. One of the main application areas of HYBRAIN's technology we expect to be artificial intelligence (AI), in particular machine learning. For testing the applicability of our hardware in the field of AI we will make use of publicly available benchmark tasks and data bases, such as the MNIST handwritten digit data set. The potential impact of our activities is to make AI more energy efficient and therefore more environment friendly. There is absolutely no reason to expect any stigmatization of particular social groups, political or financial adverse consequences, misuse etc. In testing our technology, we will make use of publicly available and privacy-non-sensitive data bases. Since we do not collect sensitive data (no studies with humans and animals), there are no immediate ethical issues to take care of.

All communication actions carried out will comply to the General Data Protection Regulation (EU) 2016/679 (GDPR).

3.7 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 University of Twente has its own Research Data Management policy:

https://www.utwente.nl/en/lisa/library/miscellaneous/docs-ru/research-data-policy-ut.pdf

The UT partner is using the guidelines on research data management specifically developed for the Nano Electronic Materials (NEM) cluster based on the Science & Technology faculty guidelines. The NEM cluster regroups research groups specialising in creating and charactering thin films and in designing, modelling and constructing nanomaterials for electronic and optical applications.

Acknowledgements

The HYBRAIN project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101046878.