Funded by
the European Union

ADDIT-CE

Alzheimer's Disease Diagnostics Innovation and Translation to Clinical Practice in Central Europe

HORIZON-WIDERA-2022-ACCESS-04

D7.1 Data Management Plan

Work Package:	WP7
Task:	T7.1
Responsible partner:	MU
Deliverable number:	D7.1 (D22)
Deliverable type:	DMP
Dissemination level:	PU
First Created:	26/06/2023
Last Updated:	02/09/2024
Version:	2.0

Content

Executive summary

Deliverable 7.1 Data Management Plan (DMP) is a detailed guideline of the treatment of all types of data associated with the project utilizing the Horizon Europe best practices on data management. DMP was created in cooperation with all partners. DMP is a vivid document describing the actual state of the data treatment and it will be updated on a regular basis.

Contributors

The following contributors are related to the project of this DMP:

- Jozef Hritz
  jozef.hritz@ceitec.muni.cz, ORCID: https://orcid.org/0000-0002-4512-9241
  Roles: Project Leader
  Affiliation: Central European Institute of Technology – Masaryk University
- Zuzana Trošanová
  trosanova.zuzana@mail.muni.cz, ORCID: https://orcid.org/0000-0002-0247-4168
  Roles: Editor, Researcher
  Affiliation: Central European Institute of Technology – Masaryk University
- Hana Nedozrálová
  Roles: Editor, Researcher
  Affiliation: Central European Institute of Technology – Masaryk University
- Markéta Procházková
  marketa.prochazkova@ceitec.muni.cz
  Roles: Researcher
  Affiliation: Central European Institute of Technology – Masaryk University
- Pavla Hlaváčková
  pavla.hlavackova@ceitec.muni.cz, ORCID: n/a
  Roles: Contact Person, Project Manager
  Affiliation: Central European Institute of Technology – Masaryk University
- Daniela Tršová
  daniela.trsova@ceitec.muni.cz, ORCID: https://orcid.org/0000-0002-7019-8310
  Roles: Editor
  Affiliation: Central European Institute of Technology – Masaryk University
- Vendula Kocandová
  vendula.kocandova@fnusa.cz
  Roles: Project Member
  Affiliation: St. Anne´s University Hospital
- Petra Kozlová
  petra.kozlova@fnusa.cz
  Roles: Project Manager
  Affiliation: St. Anne´s University Hospital

Projects

We will be working on the following projects and for those are the data and work described in this DMP.

Alzheimer's Disease Diagnostics Innovation and Translation to Clinical Practice in Central Europe

Acronym
ADDIT-CE

Start date
2023-01-01

End date
2026-12-31

Funding

- European Commission

  : 101087124 (project: EXCELLENCE HUBS, call: HORIZON-WIDERA-2022-ACCESS); https://cordis.europa.eu/project/id/101087124 (granted)

More than 55 million people worldwide suffer from dementia. Alzheimer disease (AD) is the main cause of this fatal disorder, without any effective disease modifying therapy. Early diagnosis and lifestyle modifications can significantly reduce the costs of care and treatment. There is no conceptual plan implementing modern diagnostic methods in the clinical practice in Czechia and Slovakia. The interaction between universities and private sector developing molecular diagnostic tools is fragmented and lacking. Limited number of talented students are invested in applied AD-focused research. The aim of ADDIT-CE is to interlink two ecosystems in Brno and Bratislava region, embracing the full quadruple helix of innovation driving actors: excellent scientific teams from Masaryk University and Slovak Academy of Sciences, collaborating with top biotech companies: Geneton, BioVendor, and MultiplexDX. Societal actors will be represented by organisations such as Slovak and Czech Alzheimer Societies, Memory Center and Czech Brain Aging Study. The regional government will be involved via Ministry of Health Slovak Republic, and South Moravian Innovation Centre. The joined ecosystems will unite R&I activities focusing on new diagnostic methods and their applications and further interlink academia and business spheres by creating a pilot industrial PhD programme. ADDIT-CE will generate a joint cross-border strategy covering basic and applied research activities aiming on accelerating the development of new tools for preclinical AD diagnostics and lifestyle/pharmacological intervention monitoring. New cutting-edge technologies will be transferred into clinical practise. Results of ADDIT-CE will be used to develop the Slovak National Plan to Combat Dementia, to enrich the Czech National Plan for AD, and will be widely disseminated to end users and society. ADDIT-CE will join forces of the involved ecosystems to revolutionise diagnostic approaches in both countries.

1. Data Summary

Instrument datasets

The following instrument datasets will be acquired in the project:

- **Sequencing Data**
  This dataset will be collected by experts in the project, with our own equipment.
  The equipment is very well described and known.
  Other researchers working in the same field of research could be interested in using this data.
- **Gel Visualisation Data**
  This dataset will be collected by experts in the project, with our own equipment.
  The equipment is very well described and known.
  Other researchers working in the same field of research could be interested in using this data.
- **FISH Data**
  This dataset will be collected by experts in the project, with our own equipment.
  The equipment is very well described and known.
  Other researchers working in the same field of research could be interested in using this data.
- **DNA/RNA sequencing**
  This dataset will be collected by experts in the project, with our own equipment.
  The equipment is very well described and known.
  Other researchers working in the same field of research could be interested in using this data.
- **Microscopy data (cryo EM, AFM)**
  This dataset will be collected by experts in the project, at a specialized infrastructure.
  The equipment is very well described and known.
  Other researchers working in the same field of research could be interested in using this data.
- **NMR data**
  This dataset will be collected by experts in the project, at a specialized infrastructure.
  The equipment is very well described and known.
  Other researchers working in the same field of research could be interested in using this data.
- **Xray data**
  This dataset will be collected by experts in the project, at a specialized infrastructure.
  The equipment is very well described and known.
  Other researchers working in the same field of research could be interested in using this data.
- **Mass Spectrommetry data**
  This dataset will be collected by experts in the project, with our own equipment.
  The equipment is very well described and known.

Other researchers working in the same field of research could be interested in using this data.

- **Biophysical and interaction data (DLS, DSF, ITC, AUC, FTIR, FRET...)**
  This dataset will be collected by experts in the project, with our own equipment.
  The equipment is very well described and known.
  Other researchers working in the same field of research could be interested in using this data.
- **Molecular simulation**
  This dataset will be collected by experts in the project, at a specialized infrastructure.
  The equipment is very well described and known.
  Other researchers working in the same field of research could be interested in using this data.
- **ELISA**
  This dataset will be collected by experts in the project, with our own equipment.
  The equipment is very well described and known.
  Other researchers working in the same field of research could be interested in using this data.
- **Immunoblotting**
  This dataset will be collected by experts in the project, with our own equipment.
  The equipment is very well described and known.
  Other researchers working in the same field of research could be interested in using this data.

Non-equipment datasets

We also collect data from questionnaires, case report forms, and electronic patient records. The non-equipment datasets are:

- **LIBRA** – "Lifestyle for a Healthy Brain: Intervention with Nutrition and Exercise LIBRA REmind" is designed for volunteers who subjectively experience memory decline but are still able to function independently. The aim of the study is to determine and compare the effects of two treatments developed to slow brain ageing.

Re-used datasets

We have found the following reference datasets that we have considered for re-use:

- reference genome GRCh38.p14
  It is available via:
  https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.40/. It is used in the project.
  Owner of this dataset: National Library of Medicine, https://www.ncbi.nlm.nih.gov/.
  The dataset can be used in the provided format without any conversion needed.
  We will keep a copy of the dataset and make it available with our results for the reproducibility.

We will use the dataset as follows: We will use it for sequencing (design of nucleotides).

We have found the following non-reference datasets that we have considered for re-use:

- Data from a diploma and bachelor thesis
  Owner of this dataset: Jozef Hritz. The owners of the dataset will collaborate on this project.
  The dataset can be used in the provided format without any conversion needed.
  We will download or get a copy.
  It is a fixed dataset, changes will not influence reproducibility of our results.
  We will make sure the selected subset will be available together with our results.
  We will use the dataset as follows: For comparison with newly acquired data.
- CBAS
  It is available via: http://www.cbas.cz/.

There is no need to harmonize different sources of existing data in our case.

Data formats and types

We will be using the following data formats and types:

- Tabular data (.xlsx; .csv; .dat)
  It is a standardized format. This is not a suitable format for long-term archiving; however, we plan to convert it to a suitable format before the end of the project. We expect to have 10000 GB of data in this format.

- Image data (.tiff; .qptiff)
  It is not a standardized format. The format ".qptiff" is a proprietary format for fish scans, generated by a software for the instrument "vectra polaris". This is not a suitable format for long-term archiving; however, we plan to convert it to a suitable format before the end of the project. We expect to have 8000 GB of data in this format.

- Statistical analysis (.pzfx)
  It is a standardized format. This is not a suitable format for long-term archiving; however, we plan to convert it to a suitable format before the end of the project. We will have only a small amount of data stored in this format.

- Scripts (.r)
  It is a standardized format. This is a suitable format for long-term archiving. We will have only a small amount of data stored in this format.

- Sequence Alignment/Map Format (SAM, BAM - compressed in BGZF format)
  It is a standardized format. This is a suitable format for long-term archiving. We expect to have 500 GB of data in this format.

- Mass spectrometry data (.mzxml)

It is a standardized format. This is a suitable format for long-term archiving. We expect to have 300 GB of data in this format.

- Sequence Data (.fasta; .seq)
  It is a standardized format. This is a suitable format for long-term archiving. We expect to have 200 GB of data in this format.
- Text (.txt; .docx)
  It is a standardized format. This is a suitable format for long-term archiving. We expect to have 200 GB of data in this format.
- cryo-electron microscopy and electron tomography data (.mrc)
  It is a standardized format. This is a suitable format for long-term archiving. We expect to have 1500 GB of data in this format.
- Macromolecular Crystallographic Information File (mmCIF, or PDBx/mmCIF)
  It is a standardized format. This is a suitable format for long-term archiving. We expect to have 1500 GB of data in this format.

2. FAIR Data

2.1. Making data findable, including provisions for metadata

There are no 'Minimal Metadata About ...' (MIA...) standards for our experiments. However, we have a good idea of what metadata is needed to make it possible for others to read and interpret our data in the future.

We will use other solution than (electronic) lab notebooks to make sure that there is good provenance of the data analysis: Some partners' labs are using lab notebooks, some labs are using electronic lab notebooks.

We will be keeping the relationships between data clear in the file names. All the metadata in the file names also will be available in the proper metadata.

2.2. Making data accessible

We will be working with the philosophy *as open as possible* for our data.

The data cannot become completely open because of:

- legal reasons
- non-patent business reasons: Some partners of the project are commercial companies. Some data may be part of their trade secret.
- we want to publish a paper first

Concerning the legal reasons, a data sharing agreement will be required.People can apply to an existing data access committee.

Data will be released only as soon as restrictions are falling away.

Metadata will be openly available without instructions how to get access to the data. Metadata will not be available in a form that can be harvested and indexed.

We have a consortium agreement that arranges Intellectual Property.

For the reference and non-reference data sets that we reuse, conditions are as follows:

- reference genome GRCh38.p14
  It is freely available for any use (public domain or CC0).
- Data from a diploma and bachelor thesis
  It is available under specific restrictions, which we will follow in our project: Data will be available under CC-BY after releasing the publication.

2.3. Making data interoperable

We will be using the following data formats and types:

- Tabular data (.xlsx; .csv; .dat)

It is a standardized format.

- Image data (.tiff; .qptiff)
  It is not a standardized format. The format ".qptiff" is a proprietary format for fish scans, generated by a software for the instrument "vectra polaris".
- Statistical analysis (.pzfx)
  It is a standardized format.
- Scripts (.r)
  It is a standardized format.
- Sequence Alignment/Map Format (SAM, BAM - compressed in BGZF format)
  It is a standardized format.
- Mass spectrometry data (.mzxml)
  It is a standardized format.
- Sequence Data (.fasta; .seq)
  It is a standardized format.
- Text (.txt; .docx)
  It is a standardized format.
- cryo-electron microscopy and electron tomography data (.mrc)
  It is a standardized format.
- Macromolecular Crystallographic Information File (mmCIF, or PDBx/mmCIF)
  It is a standardized format.

We will be using the following standards (encodings, terminologies, vocabularies, ontologies):

- Alzheimer's Disease Ontology
  (https://doi.org/10.25504/FAIRsharing.ckd4rf)

2.4. Increase data re-use

The following instrument datasets acquired in the project will use the following quality processes:

- **Sequencing Data**
  - Calibrating measurements
  - Repeat samples / measurements
  - Standardized data capture / recording
  - Use of controlled vocabularies
- **Gel Visualisation Data**

  - Calibrating measurements
  - Repeat samples / measurements
  - Standardized data capture / recording
  - Use of controlled vocabularies
- **FISH Data**
  - Calibrating measurements
  - Repeat samples / measurements
  - Standardized data capture / recording
  - Use of controlled vocabularies
- **Molecular simulation**
  - Calibrating measurements
  - Repeat samples / measurements
  - Standardized data capture / recording
  - Data Entry validation
  - Data peer review
  - Use of controlled vocabularies

As explained in Section 2.2, our data cannot become completely open.

There are no restrictions on where the data need to be stored related to the privacy reasons. We can use anonymization and data aggregation to make the data more openly available.

There are IP reasons why our data can not be open. It is clear who owns data and documents.

Someone will be given the decision power to move documents or data to a new place after the project has finished.

We will be archiving data (using so-called *cold storage*) for long term preservation already during the project. The data are expected to be still understandable and reusable after a long time.

To validate the integrity of the results, the following will be done:

- We will run a subset of our jobs several times across the different compute infrastructures.
- We will use independently developed duplicate tools or workflows for critical steps to reduce or eliminate human errors.
- We will run part of the data set repeatedly to catch unexpected changes in results.

3. Other research outputs

We use Data Stewardship Wizard for planning our data management and creating this DMP. We benefit from data stewardship guidance (e.g. FAIR principles, openness, or security) and it is reflected in our plans with respect to other research outputs.

4. Allocation of resources

FAIR is a central part of our data management; it is considered at every decision in our data management plan. We use the FAIR data process ourselves to make our use of the data as efficient as possible.

- **Alzheimer's Disease Diagnostics Innovation and Translation to Clinical Practice in Central Europe**

  Making our data FAIR is not a cost that can be separated from the rest of the project.

We will be archiving data (using so-called 'cold storage') for long term preservation after the project but also already during the project. The costs for archiving data will be paid out of departmental budgets from one or more of the project participants. The minimum lifetime of the archive is 5 years. The archival period can be extended – library or archive staff will decide. The decision whether or not to extend the renewal be based on available budget. Data formats of data in cold storage will not be upgraded over time. Archived data will be migrated regularly to more modern storage media (e.g. newer tapes).

We have a reserved budget for the time and effort it will take to prepare the data for publication. For making data or other research outputs FAIR, we budgeted: 16000 €.

Pavla Hlaváčková and Petra Kozlová are responsible for implementing the DMP, and ensuring it is reviewed and revised.

To execute the DMP, additional specialist expertise is required and we have such trained support staff available.

We do not require any hardware or software in addition to what is usually available in the institute.

5. Data security

Project members will not store data or software on computers in the lab or external hard drives connected to those computers. They will not carry data with them (e.g. on laptops, USB sticks, or other external media). All data centers where project data is stored carry sufficient certifications. All project web services are addressed via secure HTTP (https://...). Project members have been instructed about both generic and specific risks to the project.

The risk of information loss in the project or organization is acceptably low. The risk of information leak in the project or organization is acceptably low. The risk of information vandalism in the project or organization is acceptably low.

All personal information will be processed in pseudonymized form only. We pseudonymize inside the project, only limited people can access the keys.

We are running the project in a collaboration between different groups and institutes. However, there is no collaboration agreement in the project that describes who can have access to what data.

6. Ethics

Ethical approvals

The following project require ethical approval:

- **Alzheimer's Disease Diagnostics Innovation and Translation to Clinical Practice in Central Europe** (ADDIT-CE)
  - case number: *Masaryk University Research Ethics Committee, approval no.: EKV-2022-006; St. Anne´s University Hospital Brno, approval no.: 46V/2016-AM; Biomedical Research Centre of the Slovak Academy of Sciences, Bratislava Self-Governing Region, approval no.: 10287/2021/HF; Institute of Neuroimmunology of the Slovak Academy of Sciences, Bratislava Self-Governing Region, approval no.: 6265/2023/HF.*, status: *granted*

Data we collect

We will collect data connected to a person, i.e. "personal data". We explored General Data Protection Regulation (GDPR) considerations and relevant materials. We ask the data subjects for their consent. We collect consent for our use as well as for reuse of the data. The consent form will not be available for re-users. The procedure for obtaining consent from data subjects is set as follows: Patients sign an informed consent, which for Czech patients is established together with the patient card, in Slovakia the consents are established together with the documentation for the given study. The purpose of processing the personal data can be described as follows: Due to conducting a longitudinal study.

The data collection is subject to ethical legislation. It is covered by ethical review. It involves human subjects.

7. Other issues

We use the Data Stewardship Wizard with its *Life Sciences DSW Knowledge Model* (ID: dsw:lifesciences:2.6.7) knowledge model to make our DMP. More specifically, we use the https://dsw.muni.cz:443/wizard DSW instance where the project has direct URL: https://dsw.muni.cz:443/wizard/projects/6b93813a-5c0b-4f7d-a0e1-e9719f538433.

We will be using the following policies and procedures for data management:

- **Measure of the CEITEC MU Director no. 4/2020 "Research Data Policy"**
  https://is.muni.cz/do/ceitec/uredni_deska/opatreni_reditele/opatreni_reditele_2020_04_-_research_data_policy/107763757/Directors_Measure_2020-04_Research_Data_Policy_EN.docx
  This Research Data Policy aims to provide CEITEC MU researchers with basic definitions, rules, responsibilities, and conditions of data ownership and Data Management. This Measure is in accordance with relevant legislation and MU Directive No. 6/2013 "Research Data".
- **Open Science strategy at MU**
  https://openscience.muni.cz/en/open-science-at-mu/strategie-open-science-na-mu-2022-2028
  The link goes to university webpage, there can be downloaded two documents: MU Open Science Strategy 2022-2028, and Action plan for years 2022-2028. The ADDIT-CE project coordinator (Masaryk University) will apply this Policy when creating this DMP, as well as when working with project data.
- **FNUSA-ICRC Data Management Standard Operating Procedure**
  https://is.muni.cz/de/242921/ICRC_Data-management_v1r0_2019.pdf
  Purpose of this SOP is to describe the agenda of data management in both interventional and non-interventional clinical studies, and to define process flow from data manager appointment to database lock.