Data Management Plan
-Deliverable-

Deliverable number: D1.3

Due date: 31.03.2021

Nature: ORDP (Open Research Data Pilot)

Dissemination Level: Public

Work Package: WP1

Lead Beneficiary: West University of Timisoara (UVT or WUT)

Other contributors: UGENT and UNIMIB

Version 1

Abbreviations and Acronyms

WUT / UVT = West University of Timisoara

UGENT = Ghent University

UNIMIB = University of Milan-Bicocca

DPO = Data Protection Officer

1. Data summary

a. Purpose of data collection

The main purpose of data collection is to answer our research questions regarding the way in which neuroticism and its aspects function as risk factors for developing emotional problems. As a novel approach to this topic, we will mainly examine the possibility that this relationship is due to the influence of neuroticism on learning through distinctive evaluative processes (liking or disliking learned stimuli) or though attributive processes (assigning descriptive labels to the learned stimuli). Other conditioning paradigms like classical and fear conditioning could also be considered. Our main source of data will be human participants from which we will gather information about their reactions to certain stimuli, efficiency of their learning habits, and self-report measures of different aspects of personality. Less often we will resort to gathering eye-tracking data (information about gaze orientation) in a laboratory environment, and possibly to intensive gathering of data through Experience Sampling (short questionnaires completed multiple times per day, during several days or weeks).

b. Relation of data collection to the objectives of the project

Objective 1 Increasing the scientific profile of the early-stage researchers through joint research activities, advanced training, and short-term mobility.

Working with data sets that comprise scientific psychological data is a core skill required in psychological research. Early-stage researchers will develop their skills in multiple ways which involve working with data, and requires collection of data from human participants:

- Designing experimental and self-report procedures through which data will be collected to answer their research question
- Organizing data sets
- Statistical analysis of data

Objective 2 Strengthening the WUT's research excellence at the crossroad between personality psychology and experimental clinical psychology by increasing the international visibility and reputation of the early-career and mid-career researchers from this field.

By research excellence we mean better skills and knowledge in the domains of relevance, originality and complexity of explored scientific questions; procedures addressing these questions (study designs); and ultimately, statistical analysis of data resulted after these procedures. We aim to achieve this through multiple scientific studies implemented through the course of our project, which will also envision collaboration between all project's partners, and the transfer of expertise from more experienced partners to WUT. Each research project will have a distinct research question related to the main objective of the project which will absolutely require data collection from real human participants.

Objective 3 Enhancing the scientific capacity of WUT by transforming the positive lessons learned through this partnership into institutional and departmental policy measures regarding research management.

The expertise transferred to WUT project members through collaboration on scientific psychological studies which require data collection, will be perpetuated outside the project's context. WUT members will include the acquired expertise in their individual future research project and implement policies at the departmental level based on activities which turned out to be most efficient and profitable during the project.

c. Types and formats of data

Types of data:
- Data resulted from self-report psychological questionnaires
- Data resulted from cognitive measures (liking of disliking of previously conditioned stimuli, measurement of memory, measurements of attention, measurements of attributed labels to conditioned stimuli)
- Eye-tracking data, which represents information of eye-movements and gaze fixation
- Ecological momentary assessments based on intensive self-report measures (multiple daily assessments)

Formats of data:
- Comma separated values (.csv)
- Microsoft Excel (.xls)
- Text (.txt)
- SPSS (.sav)
- R (.R)
- R syntaxes and SPSS syntaxes in order to make transparent our data management and data analysis procedures

d. Re-use of existing data

We will use only original data collected solely to address the scientific objectives of our project. However, we will make available the data we collected on Open Science Framework web application (osf.io), PsychArchives and on project's website https://h2020learnvul.uvt.ro/research.html.

e. Origins of the data

Data will be collected from human participants mostly from Romania. On possibility, data collection will also will also be carried on in Italy and Belgium. These are residence countries for the three partner universities working on this project.

f. Expected size of the data

The size of the data depends on study design that the data is required for, and will be estimated as for each study individually as soon as the design is agreed on by the working team. A rough estimation of data size is 150-500 participants for each study.

g. To whom data will it be useful?

- LEARNVUL consortium (for their own research, to test alternative similar hypothesis and estimate feasibility of future studies)
- Scientific research communities (mainly other research groups working on related subjects)
- Peer-reviewers (to assess the quality of data and data analysis while reviewing submitted manuscripts)
- For PhD students (as training data for their research ideas)

2. FAIR data

To ensure the best research practices, the data and metadata used in LEARNVUL will be managed according to the principles of FAIR data management (*Findable*, *Accessible*, *Interoperable* and *Re-usable* data).

As a key conduit leading to knowledge discovery and innovation, we will make data "FAIR" by addressing each one of the specific points outlined in the Guidelines on FAIR Data Management in Horizon 2020.

2.1. Making data findable, including provisions for metadata

a. Outline the discoverability of data (metadata provision)

LEARNVUL will use PsychArchives as the main data repository with the goal of making the research data findable in accordance with the H2020 Open Access Mandate. It is important to note that the facilities and management of the repository structure are in compliance with FAIR data principles.

PsychArchives allows the publication of digital research objects (DROs) quickly and free of charge. It also provides easily access and metadata quality. This feature is assured by formal control procedures run by PsychArchives staff, which enhances the online findability of the research. Also, to increase the international visibility of PsychArchives DROs, they are indexed in public search engines such as GoogleScholar and PubPsych. Another advantage of this data repository is that all the research related output is available in one place and linked within PsychArchives, so that users can easily find what they need.

The LEARNVUL Consortium members will establish a community on PsychArchives website. And because by adding comprehensive and detailed metadata, other researchers could easily find our data and benefit from its use, we will provide the creation and permanent actualization of metadata quality.

As described in the project Grant Agreement (Article 29.2), the bibliographic metadata will include all of the

following:

- o the terms "European Union (EU)" and "Horizon 2020";
- o the name of the action, acronym and grant number;
- o the publication date, and length of embargo period if applicable;
- o a persistent identifier.

To maximize findability and to comply with the H2020 Open Access Mandate, the Consortium will also integrate the public datasets, deliverables and results in the European Commision Funded Research (OpenAire) community.

Open Science Framework will be used as an open-source software and a project-management system. The OSF is free and has no storage limit. And once a project is registered, users can click a button to generate a DOI for that project and then share or cite those data or materials.

Because the OSF is intended to be collaborative, the LEARNVUL Consortium will also use it as a collaboration tool. Data will be made accessible for the other members of the Consortium and users will be able to easily add contributors to project.

In order to promote transparency in research, the OSF will also be used as a preregistration tool.

b. Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

The metadata and data will be findable through two standard identification mechanisms: we will use keywords and Digital Object Identifiers.

PsychArchives allows the user to perform simple and advanced search queries using specific keywords. Every upload in PsychArchives is assigned a Digital Object Identifier (DOI) to make them easily and uniquely citable. In the repository, DROs are safely stored and made available for the long term.

c. Outline naming conventions used

A consistent naming convention will be used for all the data and metadata found in the data repository. The name convention will be versioned and structured as follows:

DataType_UNI_Description_Controller_H2020_LEARNVUL_ Version_UniqueNr

Explanation of the naming convention (in accordance to the DRO examples and their corresponding PsychArchives collections):

DataType = Research Data (RD) / Questionnaires, tests (Test) / Software script (Code) / Other files (Other)

UNI = study conducted by: UVT (West University of Timisoara) / UNIMIB (University of Milan-Bicocca) / UGENT (Ghent University)

Description = what kind of data?

1. Data resulted from self-report psychological questionnaires
2. Data resulted from cognitive measures (liking of disliking of previously conditioned stimuli, measurement of memory, measurements of attention, measurements of attributed labels to conditioned stimuli)
3. Eye-tracking data, which represents information of eye-movements and gaze fixation
4. Ecological momentary assessments based on intensive self-report measures (multiple daily assessments)

Controller = the name (initials) of the person who collected and handled the data

Version = version number (01, 02, 03...)

UniqeNr = a unique number will be given to each research data and metadata

Example of dataset name: *RD_UVT_2_HF_H2020_LEARNVUL_02_0001*

d. Outline the approach towards search keyword

Further information on a set of general keywords that should apply to all public datasets, public deliverables and scientific publications, will be outlined in the subsequent versions of the Data Management Plan. These keywords will be determined based on LEARNVUL's main interests.

e. Outline the approach for clear versioning

As presented in the naming convention, each file will contain a version number. Also, by using the PsychArchives features, each revision could be integrated in the DOI versioning of the data.

f. Specify standards for metadata creation (if any). If there are no standards in your discipline describe what type of metadata will be created and how

Structural (how data is formatted and assembled) and descriptive (how data is identified, keywords, DOIs) metadata will be created by using the PsychArchives data repository.

2.2. Making data openly accessible

a. Specify which data will be made openly available? If some data is kept closed provide rationale for doing so

The data and metadata used in LEARNVUL will be shared within and beyond the Consortium, according to the principle stated in H2020 Open Access Mandate ("as open as possible, as closed as necessary").

Types of data which will be made openly available:

- o Research Data
- o Questionnaires, tests, assessment instruments (only if they are not copyrighted materials)
- o Software script
- o Deliverables

- o Results
- o Publications
- o Other files (images, videos - only if they are not copyrighted materials)

Due to privacy concerns and security reasons, we will protect personal and sensitive data that could lead to the identification of a participant or that could cause any psychological harm.

b. Specify how the data will be made available

As specific ways employed to make data available, we will use the PsychArchives data repository, the Open Science Framework software and OpenAire programme.

With regard to manuscripts publications, the Gold Open Access route will be adopted. So that, the published papers will be available in open access mode immediately upon publication. Obviously, open access to all data related to each publication will be made available through repositories immediately.

Supplementary, the data will be made available for the public by disseminating information through the official website (https://h2020learnvul.uvt.ro/) and through the Twitter account (https://twitter.com/learnvul).

c. Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software?

Data and metadata stored in the data repository will be accessible for the public, free of charge. All the employed software is user friendly and provides usage information. Datasets will be made accessible in .csv format, and data analysis syntaxes will be distributed in .R format. These formats can be viewed with any basic text editing program and can easily be imported into R for data analysis. Potentially, we could use .xls and .sav formats to store data which can also be accessed with R. For storage of images, we will use .jpg, .tiff, and .png formats which do not require any specific software to be viewed. We will also provide Inquisit syntaxes used to program computerized experiments in .exp format. Even though this format requires a paid Inquisit license to run .exp format is accessible through text editing programs and can be viewed in case our experimental procedures are of interest to be transferred into other software.

d. Specify where the data and associated metadata, documentation and code are deposited

All data and associated metadata will be deposited on PsychArchives, Open Science Framework and OpenAIRE. We will also use the official website (https://h2020learnvul.uvt.ro/), mainly the two special sections (Research results and Open Access Data) to disseminate data to the public.

e. Specify how access will be provided in case there are any restrictions

All data deposited on PsychArchives is accessible without any restrictions for public. For more important data, such as results, potential users will contact the data owner in order to gain access. In particular cases we will also use Creative Commons license to protect the copyrights of the LEARNVUL project.

2.3. Making data interoperable

a. Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.

All data deposited in the aforementioned depositories will be described in a standard way, using accepted metadata standards and vocabularies which permit to combine them with other sets of similar type of data. All data will be provided in a "machine-readable" format that can be used and understand by computers.

The data/ metadata stored in the depositories will be represented by: (accepted) manuscripts/ publications, measurements/ observations, experimental data, results of experiments/ statistic analyzing scripts, experimental scripts, materials involved in the experimental scripts. All these data/ metadata will be provided in a standard way that permit to the research community to access them by reading and combining with other data sets.

- The manuscripts/ publications will be provided using the more common PDF format which is a free readable format.

- The measurements/ observations/ experimental data will not be provided in the "raw format", but in an analyzable format. That means experimental data will be documented, described and provided using informative data that will be required by the chosen depositories. For instance, all experimental data is supposed to be provided using the .xls and/or .csv file format which facilitates the access to the whole data for visualising, reading, combining with other sets of similar data, and uploading easily to any statistic analysing software.

- The code for the statistic analysing will be provided via a free statistic analysing software which permit to the scientific community to access, verify the analyses, and use the code in further similar statistic data analyses. A free statistic analysing programme option is R Programming (https://www.r-project.org/).

- The experimental scripts designed to build up the experiments/ collect the data will be provided initially via the Inquisit Lab Program (https://www.millisecond.com/), a software used in designing psychological experiments. The main disadvantage in using this software consists in limiting the interoperability of the experimental codes. The scientific community would have to use the same software in order to verify and further utilize the scripts for collecting similar data. The software is free for a limited time (30 days). Thereafter, the "Monkey mode" will be implicit used and the data collection will be stopped. Otherwise, Inquisit Lab Program is one of the most used software in psychological experiments. A fair decision regarding the interoperability of this type of data will be held after a discussion with the Consortium.

The information provided in this section will be updated with respect to the interoperability requirements foreseen by the chosen depositories where all the data will be stored. Moreover, for each data format uploaded on the depositories will be mentioned the software programs/ computer applications which facilitate a suitable access and, further, the interoperability.

LEARN
VUL

Funded by
the Horizon 2020
Framework Programme
of the European Union

b. Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

For this moment, a standard vocabulary for all data types provided by LEARNVUL is limited. Standard vocabularies will be used (see the information mentioned above) to facilitate the interoperability of data with regard to the statistical analyses and the format in which experimental data will be provided. Regarding other types of data mentioned in the previous section, the interoperability description and implementation will be updated. Other forms of providing standardized data/ metadata will be mentioned according to the interoperability requirements of the chosen repositories. A further analysis of the existing mature interoperability resources provided by ELIXIR website (https://elixir-europe.org/platforms/interoperability) will be performed in order to choose the suitable ones.

2.4. Making data openly accessible/ Increase data re-use (through clarifying licences)

a. Specify how the data will be licensed to permit the widest reuse possible

LEARNVUL will provide information via the chosen repositories on how each material/ dataset and so on can be reuse, including term-of-use or the license under which all data can be accessed and re-use. Given that each project material will receive a persistent identifier code via the repository storing, the reuse will be possible as long as the identifier will be associated with material and mentioned it in each reusable format.

For instance, experimental scripts could be reused in the same format just for a study replication, whereas the code for data analysis could be reused in the same format with few modifications or specifications with regard to the variables used in the analysis. Given that these types of data are made within the research groups from the Consortium, they can by widely reused only through code identifier specification. Other types of data, for instance personality measurements scales or different kinds of stimuli represent copyrighted materials/ data, so the reuse of them will be strictly restricted by third parties, but not for the Consortium's research groups.

More information on this topic will be added/updated during the project evolving, specifying exactly how each type of data will be licensed for reuse.

b. Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed.

The data will be made available for re-use immediately after the publication or even during the pre-registration period, depending on the type of data. For instance, the experimental scripts (but without the copyrighted materials) could be provided during the pre-registration period. Data analysis could be provided for re-use after the publication. All this information will be updated with regard to LEARNVUL project evolving.

c. Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

The third parties can access and use some types of data provided and stored on the repositories. For instance, the rationale of the experimental scripts and the coding for data analyses could be used by third parties. Whereas the code for data analysis can be used in the whole format, the scripts for experiments will not provide specific materials which involve copyrighted elements (images/ stimuli, items from personality inventories).

Again, more information on this topic will be added during the project evolving. In particular cases, LEARNVUL project can request access to the legal tools regarding the solutions for licensing (e.g., Creative Commons; https://creativecommons.org/licenses/#_blank).

d. Describe data quality assurance processes

The quality assurance processes of the data will be expended to all types of data involved in the LEARNVUL project.

The primary type of data is represented by the data collected through experimental procedures which involve writing experimental scripts. So that, the first step in assurance high quality of data collected consists in designed and implemented well-documented experiments. This procedure involves documentation of research existing with respect to the research topics of interest and appropriate methodologies used, and checking the correctness of the experimental scripts according to each research objective. Moreover, the data will be collected using platforms for data collection. To ensure a high quality of data collection, inclusion and exclusion criteria for participants will be provided to the platforms in order to facilitate a high-quality data received.

Once the data will be collected and cleaned, they will be stored on the aforementioned repositories.

e. Specify the length of time for which the data will remain re-usable

The data will remain available for re-use for unlimited time according to the repositories policies, but at least five years according to the LEARNVUL Grant Agreement.

3. Allocation of resources

a. What are the costs for making data FAIR in your project?

The costs for making data FAIR depend on the type of data.

The accessibility, interoperability and re-usability will be assured by the free repositories chosen. So, the access of the project data by others/ third parties will be support any costs. Specific access requirements will be provided for each material, if some particularly cases will arise.

The publications will be available via the Gold Open Access route. The payment for this type of publication is funded under the Horizon 2020 Grant.

b. How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

As it was mentioned before, only publication process will support costs. To cover these costs, the payment requirements were specified in project proposal, so the publication costs are funded under the Horizon 2020 Grant.

c. Who will be responsible for data management in your project?

The data management process will be assured by dr. Andrei Rusu, one of the senior researchers from West University of Timisoara.

d. Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

All project data will be uploaded on the aforementioned repositories. The data will remain available there for unlimited time, according to the repositories preservation policies, but at least five years according to the LEARNVUL Grant Agreement. All the repositories require not costs. The time for which data will be kept available is set-up according to the Grant Agreement (at least five years/ any updates will be made on this document if the policies for data preservation will change). All data will be available, less those which suppose copyrights or sensitive characteristics. Again, any updates will be made on this document if particular cases will arise.

4. Data security

LEARNVUL will follow a set of principles to ensure data security:
- Use Dropbox Professional to transfer data between project members and short-term storage. Dropbox has a set of security policies that assures protection against unauthorized access.
- Use Open Science Framework for long-term storage of data
- Storage of data in multiple locations

5. Ethical aspects

- No potential ethical issues have been identified
- We will not store or collect any personal data which would allow to identify participants.
- We do not work with any vulnerable population
- WUT as the host institution has appointed a DPO whose contact details will be transmitted to participants during each study
- LEARNVUL consortium has complied to all ethical principles regarding the project

6. Other

No other data management aspects have been identified so far.