www.im-twin.eu

IM-TWIN: from Intrinsic Motivations to Transitional Wearable INtelligent companions for autism spectrum disorder

a European funded project

ORDP: Open Research Data Pilot report 2

Deliverable 6.18

This project has received funding from the European Union's Horizon 2020 Research and Innovation Program Under Grant Agreement No 952095.

Project duration 24 months (November 2020, October 2022), Consortium: Consiglio Nazionale delle Ricerche (ITA), Universiteit Utrecht (NLD), Centre de Recherches Interdisciplinaires (FRA), Università degli Studi di Roma La Sapienza (ITA), Plux-Wireless Biosignals S.A. (PRT).

Deliverable data

Work Package:	6 Management and dissemination
Work Package leader:	CNR
Deliverable beneficiary:	CNR
Dissemination level:	Public
Due date:	31th July 2022 (Month 21)
Type:	ORDP: Open Research Data Pilot
Authors:	F. Montedori, V. Sperati, G. Baldassarre

Acronyms of partners

CNR-ISTC	Consiglio Nazionale delle Ricerche, Istituto di Scienze e Tecnologie della Cognizione (Italy)
UU	Universiteit Utrecht (The Netherlands)
CRI	Centre de Recherches Interdisciplinaires (France)
LA SAPIENZA	Università degli Studi di Roma La Sapienza (Italy)
PLUX	Plux - Wireless Biosignals S.A. (Portugal)

Table of contents

1. Overview of the deliverable

This document updates the *Data Management Plan* (DMP), provided in the previous deliverable D6.17 *Open Research Data Pilot, report 1*[1].

Compared to the first DMP version, the following changes have been made:

- by “data” is meant only experimental data collected during the experiments with human participants, namely the data needed to validate the results presented in scientific publications (or the “underlying data”), including the associated metadata;
- the DMP has been compiled using the online H2020 template[2] available at https://dmponline.dcc.ac.uk/. The pdf produced by the online tool is attached at the end of this document.

2. Conclusions

The DMP is a dynamic document: it will be updated alongside the project progress, taking into consideration all the significant changes affecting the data management. Updates on the DMP will be given in the last deliverable:

- D6.19: ORDP, Open Research Data Pilot, report 3, due at M35 (30 September 2023).

[1] https://im-twin.eu/wp-content/uploads/2021/03/DELIVERABLE_08_D6.17_Data_Pilot_report_1.pdf

[2] See link Digital Curation Centre's DMP Online tool in the page https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

Plan Overview

A Data Management Plan created using DMPonline

Title: IM-TWIN: from Intrinsic Motivations to Transitional Wearable INtelligent companions for autism spectrum disorder

Creator: Valerio Sperati

Affiliation: Other

Funder: European Commission

Template: Horizon 2020 DMP

Project abstract:

Research into autism spectrum disorder (ASD) is important since the condition affects about 1 in 10 newborn children in developed countries. Previous EU-funded research resulted in the development of a prototype wearable companion robot called PlusMe for ASD treatment and daily support. The EU-funded IM-TWIN project now aims to furnish PlusMe with intelligent behaviour, give it extra embedded biosensors and cameras for detecting a child’s affective state and integrate all components into an Internet of Things system itself called IM-TWIN. It will also validate the device and its components with target stakeholders and perform activities to advance the system components to a higher technology readiness level. The project’s work will help to meet the needs of ASD therapy centres and families with children with ASD.

ID: 106066

Start date: 01-11-2020

End date: 31-10-2023

Last modified: 06-09-2022

Grant number / URL: https://cordis.europa.eu/project/id/952095

IM-TWIN: from Intrinsic Motivations to Transitional Wearable INtelligent companions for autism spectrum disorder - Detailed DMP

1. Data summary

State the purpose of the data collection/generation

The experimental data collected on human participants (children with a diagnosis of Autism Spectrum Disorders, ASD, and with Typical Development, TD) are used to run a first validation of IM-TWIN technological tools, to detect the children's affective states.

Explain the relation to the objectives of the project

Video data, collected by environmental cameras, and physiological data, collected by sensorised t-shirt, will be used as a dataset to train an Artificial Intelligence, through Machine Learning (ML) algorithms, to categorise the child's affective states during the therapeutic activity.

Specify the types and formats of data generated/collected

Three types of experimental data will be collected during the project:

- **Personal data**: personal information about children and their parents (e.g., age, address, contact details), collected through written informed consent form.
  - format: paper support and corresponding digital support (scanned PDF);
  - purpose: mandatory documents in clinical research;
- **Clinical data**: questionnaires and psychological tests for behavioural evaluations.
  - format: paper support and corresponding digital support (scanned PDF);
  - purpose: creation of clinical dataset, relevant for the purpose of the research
- **Biometric data**:
  - **video recordings** of the experimental activities with children
    - format: .mp4 (standard video file);
    - purpose: creation of dataset for *Facial Expression Recognition* (FER), to refine existing dataset used by the ML tools, used to categorise the participant's facial expressions.
  - **physiological data** such as Heart Rate (HR), Electrodermal Activity (EDA), Skin Temperature (ST) collected through the sensorised t-shirt;
    - format: .h5 and .txt files (generated by the sensorised t-shirt software "OpenSignals");
    - purpose: creation of dataset to train a ML tool to categorise the affective/emotional states of the participant.

Specify if existing data is being re-used (if any)

Database used for computer vision algorithms:

- for the *Facial Expressions Recognition* (FER) algorithm, the consortium used the *Real-world Affective Faces Database* (RAF-DB), freely available at www.whdeng.cn/raf/model1.html to train the ML tool.
- to refine the ML performance of the FER algorithm, the consortium requested to the *Georgia Institute of Technology* the access to the *Multimodal Dyadic Behaviour Dataset* (MMDB), https://cbs.ic.gatech.edu/mmdb/dataset.php .

Specify the origin of the data

- **Biometric data** (video recordings and physiological data) are collected from children with Autism Spectrum Disorders and with Typical Development, aged between 30 and 60 months, during the experimental activities;
- **Personal data** collected through the signed informed consent;
- **Clinical data** collected during the experimental activities by the researchers.

State the expected size of the data (if known)

- **Personal data**: around 1MB per participant (scanned pdf)
- **Clinical data**: around 1 MB per participant (scanned pdf)
- **Biometric data**:
  - video data: around 800 MB mp4 file per 10 minutes activity
  - physiological data: around 30 MB h5 file per 10 minutes activity

Outline the data utility: to whom will it be useful

Given the sensitivity of biometric data (videos cannot be anonymised, and the physiological signals are relatively unhelpful without the related videos), the access to raw data will be granted only to the authorised consortium members.

The consortium will evaluate the usefulness to share aggregated physiological data on public repositories, such as GitHub or Zenodo:

- IM-TWIN GitHub page: https://github.com/IM-TWIN
- IM-TWIN Zenodo page: https://explore.openaire.eu/search/project?projectId=corda__h2020::02f7552bef74ef4e919727a60530c9e0

Clinical and personal data are not to be shared with the consortium members.

2.1 Making data findable, including provisions for metadata [FAIR data]

Outline the discoverability of data (metadata provision)

The following metadata will be associated to biometric data:

Video data:

- filename
- date
- camera name (es: ....)
- frame rate (es: 25 fps)
- frame resolution (es: 1980 x 1020)
- length (es: 12 min, 4 sec, 12 frames)
- description (es: subject with High Functioning ASD, aged 36 months)

Physiological data:

- filename
- date
- sampling rate (es: 200 Hz)
- description (es: subject with High Functioning ASD, aged 36 months)
- associated video (es: name of the video file associated to the data)

The metadata will be provided in the file iself if possible (about videos the VLC software allows to save custom metadata in the field "Media Information"), or alternatively in an associated README.txt file.

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?

N/A

Outline naming conventions used

N/A

Outline the approach towards search keyword

N/A

Outline the approach for clear versioning

N/A

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

N/A

2.2 Making data openly accessible [FAIR data]

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

Raw biometric data (videos and physiological signals) will not be open, as sensitive and strictly confidential.The data will be shared only within the authorised personnel in the consortium, after a pseudonymisation procedure.
During the course of the project, the consortium will evaluate the usefulness to openly share aggregated data.

Specify how the data will be made available

N/A as raw data will remain confidential. In case of disclosure of aggregated data, they will be available in the project dedicated pages in GitHub or Zenodo:

- IM-TWIN GitHub page: https://github.com/IM-TWIN.
- IM-TWIN Zenodo page: https://explore.openaire.eu/search/project?projectId=corda__h2020::02f7552bef74ef4e919727a60530c9e0

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 (e.g. in open source code)?

N/A.

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

Biometric data (including associated metadata), personal data and clinical data are stored in a secure server at ISTC-CNR, with restricted access.

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

The security measures adopted by CNR-ISTC to manage the access to the secure server are described in the confidential deliverables "7.3 *Protection of Personal Data POPD, requirement No. 3"*, and "*7.5 Protection of Personal Data POPD, requirement No. 5*", Work Package 7 *Ethics*.
Currently (September 2022), only authorised reasearchers within the project can access the data.

2.3 Making data interoperable [FAIR data]

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

N/A

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?

N/A

2.4 Increase data re-use (through clarifying licenses) [FAIR data]

Specify how the data will be licenced to permit the widest reuse possible

N/A

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

N/A

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

N/A

Describe data quality assurance processes

N/A

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

Biometric data will be stored on a secure server at ISTC-CNR for 6 years from the beginning of the project; after this persiod, data will be erased. This information is depicted in the *Data Protection Impact Assessment (*DPIA), a document included in the confidential deliverable 7.5 POPD, requirement no. 5.

3. Allocation of resources

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

N/A

Clearly identify responsibilities for data management in your project

The following ISTC-CNR personnel are responsible for data management:

- Dr. Valerio Sperati (valerio.sperati@istc.cnr.it)
- Dr. Flora Giocondo (flora.giocondo@istc.cnr.it)

Describe costs and potential value of long term preservation

N/A

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

Sensitive data are stored in a secure server at ISTC-CNR. Access to server is restricted to authorised personnel only, within the project. Access regulated by personal password and authorised profile.

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

Biometric data will undergo a pseudo-anonymisation procedure (the identity of the participant will be substituted with an alphanumerical code).

Data will be collected only upon release of a signed informed consent (document already approved by the reference ethical committee).

All aspects relating to data protection have been assessed in the project ethical deliverables and in the ethic documentation supporting the informed consent for data treatment.

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

Data Protection Impact Assessment (DPIA) approved by the Data Protection Officer of CNR (rpd@cnr.it).

DPIA reviewed by the *CNR, Research Ethics and Bioetichs Commitee*, and then by the *Italian National Ethic Committee* (the ethci committe which released the ethical clearance for the project experimental activities).