European Research Council
Established by the European Commission

ERC OPEN RESEARCH DATA MANAGEMENT PLAN (DMP)

Project Acronym	Project Number
H-MIP	853271

VERSION 1.0, 21 January 2022

This DMP is based on the ERC Data Management Plan Template. It is a living document, which will be updated at least once per year until the completion of the project.

SUMMARY

This project is generating a number of datasets, some of which will be made open access, and some of which will be kept confidential to protect the privacy of the human research subjects involved. The datasets include:

Reference: OHDNAB
Name: Open Human DNA Bloodmeal Data
Origin and brief description: Processed data on the human DNA found in mosquito bloodmeals. This data includes: (1) Time, date, and location of mosquito capture, (2) Species, morphological characteristics, and estimated age of captured mosquito, (3) Number of different men and women represented in the DNA, (4) Records of any other mosquitoes in which these same people's DNA was detected.
Expected size: 10-100 MB
Data types and formats: The dataset includes datetime, numeric, and character data, structured and stored as a json file and accompanying metadata.

Reference: OHM
Name: Open Human Mobility Data
Origin and brief description: Processed data on the human mobility patterns in Catalonia obtained from volunteer's using the Space Mapper mobile phone research app, from publicly available geolocated Twitter data, and from publicly available mobile phone metadata (released by the Spanish National Statistics Office). This data includes estimated fluxes between areal units at different times of data and other aggregate mobility statistics (e.g. radius of gyration). It excludes any information that could be used to identify or reidentify any person.
Expected size: 10-100 GB
Data types and formats: The dataset includes datetime, numeric, and character data, structured and stored as a set of compressed csv files with accompanying metadata.

Reference: OMB (https://doi.org/10.5281/zenodo.597466)
Name: Open Mosquito Bite Data
Origin and brief description: Reports of mosquito bites sent through the Mosquito Alert citizen science platform. This data includes: date, time, and location of report, approximate time of bites, number of bites and approximate location on body, whether the bites occurred indoor, outdoors, or in a vehicle, and information on mosquito species if available. The data is part of a larger dataset that is updated and released on Zenodo daily.
Expected size: 10-100 MB
Data types and formats: The dataset includes datetime, numeric, and character data, structured and stored as a set of compressed json files with accompanying metadata.

Reference: OHMBN
Name: Open Human-Mosquito Biting Network Data
Origin and brief description: Synthetic data on potential degree distributions of human-mosquito biting networks in Catalonia, generated using an agent-based model that incorporates information on human and mosquito population distributions and human mobility patterns.
Expected size: 10-100 GB
Data types and formats: The dataset includes numeric, and character data, structured and stored as a set of compressed csv files with accompanying metadata.

Reference: OHCN
Name: Open Human Contact Network Data
Origin and brief description: Estimates of age-specific contact patterns in Spain during the Covid-19 pandemic. This data was generated from the CSIC Distancia-Covid survey (https://distancia-covid.csic.es/) using methods and insights from H-MIP. It includes mean and median number of contacts between age groups as well as full population estimates of the distribution of by age group, disaggregated by Autonomous Community.
Expected size: 10-100 GB
Data types and formats: The dataset includes datetime, numeric, and character data, structured and stored as a set of compressed csv files with accompanying metadata.

Reference: CHDNAB
Name: Confidential Human DNA Bloodmeal Data
Origin and brief description: Raw data recorded from human DNA sampled in mosquito bloodmeals. This data will be stored securely, analyzed, and securely deleted within two years. The data includes: (1) Time, date, and location of mosquito capture, (2) Species, morphological characteristics, and estimated age of captured mosquito, (3) Number of different men and women represented in the DNA, (4) Records of any other mosquitoes in which these same people's DNA was detected, (5) DNA microsatellite lengths.
Expected size: 10-100 MB
Data types and formats: The dataset includes datetime, numeric, and character data, structured and stored as a json file and accompanying metadata.

Reference: CHDNAS
Name: Confidential Human DNA Saliva Data
Origin and brief description: Raw data recorded from human DNA sampled directly from volunteers' saliva and/or buccal swabs. This data will be stored securely, analyzed, and securely deleted within two years. The data includes: (1) Time, date, and location of sampling, (2) IDs of CHDNAB records in which this person's DNA was detected, (3) DNA microsatellite lengths.
Expected size: 10-100 MB
Data types and formats: The dataset includes datetime, numeric, and character data, structured and stored as a json file and accompanying metadata.

Reference: CHM
Name: Confidential Human Mobility Data
Origin and brief description: Raw data from volunteer's using the Space Mapper mobile phone research app. This data includes date, time, and location of mobile phone geolocations shared with informed consent.
Expected size: 10-100 GB
Data types and formats: The dataset includes datetime, numeric, and character data, structured and stored as a set of encrypted and compressed csv files with accompanying metadata.

1. MAKING DATA FINDABLE *(dataset description: metadata, persistent and unique identifiers e.g., DOI)*

All of the open datasets listed in the summary either have been or will be placed in open access data repositories and assigned DOIs. They are all accompanied by metadata using Google's Schema.org metadata standard following the guidelines from SOSO published by the Earth Science Information Partners and the draft document "DCAT-AP to Schema.org Mapping" published by the *European Commission's Joint Research Centre*.
The metadata information relative to each dataset is stored in .json-ld files (JSON for Linking Data) that are validated with Google' Structured Data Testing Tool.

2. MAKING DATA OPENLY ACCESSIBLE *(which data will be made openly available and if some datasets remain closed, the reasons for not giving access; where the data and associated metadata, documentation and code are deposited (repository?); how the data can be accessed (are relevant software tools/methods provided?)*

All of the open datasets listed in the summary either have been or will be made accessible by placing them, along with their metadata and other supporting information, on *Zenodo* as well as in UPF's digital e-Repository (http://repositori.upf.edu). This latter repository is designed to collect, disseminate, and provide persistent and reliable, long-term open access to the research and scholarship of UPF's faculty, staff and students. The e-Repositori is managed by UPF's University Library in partnership with its IT Services.

Each dataset is accompanied by its metadata in json format (described above). In addition, a public portal will be created on GitHub pages that includes all of the metadata and code examples for downloading and processing the data.

Some of the datasets remain closed (as noted in the Summary), in accordance with the project's ethics review process and Data Protection Impact Assessment, to protect the privacy of the human subjects involved.

3. MAKING DATA INTEROPERABLE *(which standard or field-specific data and metadata vocabularies and methods will be used)*

Efforts are being made to use field-specific vocabularies to ensure interoperability. This is being done in coordination with other researchers involved in citizen-science-based mosquito studies through the Global Mosquito Alert initiative that the PI helped to found (https://globalmosquitoalert.com/).

4. INCREASE DATA RE-USE *(what data will remain re-usable and for how long, is embargo foreseen; how the data is licensed; data quality assurance procedures)*

The open data will remain re-usable permanently with no embargo. It is licensed as CC0 and/or CC-BY (depending on the dataset). Quality is checked when data is initially generated, and the datasets can be corrected if users identify problems (which they can communicate through Github or via email). Corrected datasets are documented and published as new versions, with earlier versions remaining permanently available so that researchers can always turn back to them, link published results to them etc.

5. ALLOCATION OF RESOURCES and DATA SECURITY *(estimated costs for making the project data open access and potential value of long-term data preservation; procedures for data backup and recovery; transfer of sensitive data and secure storage in repositories for long term preservation and curation)*

As Zenodo and the UPF Institutional Repository will be used for dissemination and long-term preservation, there are no extra costs for making the data compliant with FAIR standards. H-MIP's budget includes 21,450 euros for gold open access publications. The potential value of long-term data preservation is high, as it will enable future researchers to replicate project analysis, to conduct novel analysis, and to make comparisons with future data sources.

Sensitive data is collected and stored separately using strong encryption. Access to sensitive data is granted only to project members who have completed research ethics training and signed non-disclosure agreements. Data transfer is done using secure protocols (SCP, SFTP and/or HTTPS).

DISCLAIMER. Please note that the ERC Data Management Plan is not a part of the Ethics Review. It is the responsibility of the Principal Investigator to inform the ERCEA Ethics Team of any ethics issues/concerns regarding the collection, processing, sharing and storage of data in relation to the project. The Principal investigator can also be asked to submit an Ethics Data Management Plan (Ethics DMP).