Project[1] Number: [873185]
Project Acronym: [FALAH]
Project title: [Family farming, lifestyle and health in the Pacific]

DATA MANAGEMENT PLAN

[1] The term 'project' used in this template equates to an 'action' in certain other Horizon 2020 documentation

1. Data summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

Data will be processed in order to answer to research issues (9 research questions in the FALAH project). These issues will be exposed in papers and reports. According to specifications, new knowledge will be openly shared as far as possible.

What types and formats of data will the project generate/collect?

In this project, we dissociate two kinds of data: private data and public data.
Private data are data that are accessible exclusively for researchers involved in the current project. These data include personal data, pseudonymised data and more generally raw data.
Public data are data that are accessible for persons not involved in the project. These data include anonymised data and more generally consolidated data.
Data should be easily usable, that is we plan to widely use metadata in each of these five times in the life cycle of data.
We expect using an SDMX model and other suitable standards for each specific kind data. For instance, for geographic data, data will be disseminated through OGC standards (WMS, WFS, WCS). If needed, we will provide sensor data through the OGC SWE (O&M, SOS). Metadata should enable to easily find and understand the data produced and shared in the frame of the FALAH project.

Will you re-use any existing data and how?

We will re-use some relevant data. Indeed, we did get seed funding to build the network before we get the FALAH project and this research was used as proof of concept for FALAH proposal.
Some data we collected during previous research projects can add information or can be used to compare the current situation and the situation some years ago. Moreover, the research team could use existing open data if this is found relevant in the frame of FALAH project.

What is the origin of the data?

Researchers will get information about family agriculture through observations obtained from household interviews.
Data will be collected through digitalized questionnaires on Recap application. Questionnaire include: diet, well-being, digital use, sleep, ethnicity, socio economic status variables related to WP2 and WP3.
Anthropometric data will be obtained with measurements obtained directly from participants.
Moreover, we will use the iRecall 24 Pacific app to gather micro and macro nutrients intake in participant.
Regarding physical activity, participants (around 200) will wear an accelerometer continuously during 7 days.

Geographic and spatial data will be obtained by using satellite images to map cultivated areas in urban and peri-urban landscapes and neighbourhoods. This will focus in particular on the extension of informal settlements and the urban growth of informal settlements on the outskirts of major cities and the surrounding rural environment.

What is the expected size of the data?

We expect getting data from at least 800 participants (400 adolescents and 400 adults).
All the participants should be involved in data collection through questionnaires. Data volume should not exceed 5 Mo.
We expect getting accelerometric data from at least 200 adolescents, i.e. a volume of about 200 Go.
20 participants will be interviewed with 1 h audio recordings, i.e. a volume of about 1200 Mo (60 Mo per hour).
Some resolution satellite images will be used, probably a volume of about 20 Go.
We expect getting about 200 / 300 [depending on the satellite images] Go data.

To whom might it be useful ('data utility')?

These data will be useful for researchers of the consortium and be reused for future projects in the Pacific region where data regarding family farming lifestyle and health are very sparse.
The satellite data used for the mapping will be reused in urban planning and urban growth monitoring projects in the major Melanesian cities, which have experienced particularly rapid growth in recent years.

2. FAIR Data

2. 1. Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

As part of the research, it is recommended that FALAH researchers publish the data they use to write scientific papers. To do this, researchers should use dedicated platforms specialized in data dissemination. Of course, such platforms have their own systems of data and metadata management. We acknowledge that it is better the data is published on dedicated platforms providing persistent identifier such as DOI, but it is the responsibility of the researcher to wisely choose according to constraints and their own disciplines.

What naming conventions do you follow?

While we are aware of each discipline can have its own naming conventions for describing data contents, we will follow the criteria of the Declaration of Helsinki. The names of variables, fields or any other data dimensions will be sufficiently and widely described in metadata to unable future users to properly understand the content of the data.

Will search keywords be provided that optimize possibilities for re-use?

To our knowledge, all the platforms (or all those we considered) used for dissemination of data or knowledge provide keywords associated to resources. Team members will publish data including keywords.

Do you provide clear version numbers?

We plan to provide access mainly to consolidated data, i.e. data that are in their final version for the project. However, we recognize that even such data can be updated. That is why a version number of data will be assigned to each update, starting at 1.0.0 and with increments according to the extent of the changes.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

When publishing data, it is recommended that FALAH researchers choose wisely the platform so that it provides rich metadata and it properly describes data. As previously described, regarding geography data and sensor data, we plan to use OGC standards. Statistical metadata could be described following the SDMX standards, which is a very general standard for most of data.
For other social sciences disciplines such as geography, anthropology, sociology, linguistics general standards will be followed.
In case in no specific standards exist in disciplines, metadata based on Dublin-Core entries will describe as well as possible data: authorship, title, description or abstract, keywords, process of data collection, date of collection, description of data processing from raw data (data firstly collected to data currently available), description of fields (for instance rows and columns for tabular data), description of versions, potentially license or rights associated to data, how to access data, how to reference data, etc.

2.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

Private data (mainly non-anonymised data) will be accessible only for researchers involved in FALAH project. These data will be stored in a secure server but they will not be publicly accessible due to legal and ethical restrictions.
During the publication process, data (anonymised when necessary) used for writing the papers will be deposited in a trusted repository (probably Zenodo: EU Open Research Repository). It includes the feature consisting in providing a DOI.
After the ten-year-embargo-period, the anonymized full dataset will be openly accessible in a trusted repository and in an archive centre.

How will the data be made accessible (e.g. by deposition in a repository)?

Private data will be "locally findable" for researchers involved in the FALAH project according to their needs. Public data will be stored in an approved archive centre (CINES for instance). During the publication process, we will apply for services providing persistent identifier, preferably Digital Object Identifier (DOI).
During the publication process of data, metadata will be created to fully describe and understand data. With an understandable title, a complete abstract and relevant keyword, metadata should also describe:

- Authors (with a personal identifier such as ORCID-ID when possible)
- Methods to collect data
- Processes applied on data
- Rights associated (preferably CC-BY)

We will do our best to use international standards, such as SDMX, to describe and disseminate data. That would help users who would like to re-use data.
The specialised approved repositories we plan to use to disseminate data (Zenodo: EU Open Research Repository) support harvesting.

In this project, we will tend to manage data according to the following principle: "Data should be as open as possible and as closed as necessary". In the life cycle of data, we will distinguish five steps:

- Collection: In some cases, personal information will be needed. This step will be in accordance with GDPR. Participants will be informed about the project and its aims, how to contact DPO if necessary, how long their personal data will be stored, etc.

- Pseudonymisation/Anonymisation: When data involve human individuals, data will be pseudonymized in order to analyse and apply process on non-personal information as far as possible. Such data will not be published but will be used by the teams involved in the project. Only anonymized data (i.e. impossible or very extremely difficult to identify) will be published.

- Storage: Data will be on a secure remote server of Exodata, an organisation external of UNC. Data will be available to the researchers involved in the project for their analyses. It is expected that personal data will be stored 5 years after the end of the FALAH project with an extension till the end of the period of re-use of data.

- Archive: Data and anonymised personal data will be archived in an approved archive centre, such as CINES (https://www.cines.fr).

- Dissemination: During the processes of publishing in research publications, it will be expected to disseminate data used for the specific studies through specialised approved repositories such as Zenodo (https://zenodo.org/).

What methods or software tools are needed to access the data?

It depends on the kind of data. As described before, most of the data we expect to collect will be tabular or texts. It is plan to only provide data in "well-known"/usual formats (csv, doc, txt, etc.) for this data, according to the disseminating platform features and researcher practices. Regarding data that is more specific, such as geography data, a particular software may be indispensable to open and read shapefile or raster data for instance. However, even in this case we plan to use very common and widely used formats (shp, gml, geojson, ASCII grid, geotiff, etc.), generally readable with open and free software.

Is documentation about the software needed to access the data included?

No, it is plan to only provide data in "well-known"/usual formats (csv, etc.). However, should special software be required to open a particular file type or format, this will be described in the file's metadata.

Is it possible to include the relevant software (e.g. in open source code)?

It is basically spreadsheet and word-processing software. Geography data can be viewed and processed with openGIS and accelerometer data can be read with a free software (https://activinsights.com/technology/geneactiv/), for instance. Although it is highly unlikely regarding the aims of the current project, if the FALAH team develops a new software or a new data format, it is recommended that an open source code would be provided.

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

On dedicated repositories (see 2.4 Increase data re-use). We expect mainly using "Zenodo: EU Open Research Repository" that supports open access and allows data restrictions when needed.

Have you explored appropriate arrangements with the identified repository?

Open access GDPR compatible are recommended for researchers. As far as possible, we plan to use "Zenodo: EU Open Research Repository" to store FALAH data. This repository seems to suit with the EU recommendations. That is why, for now, we have not explored specific arrangements with any repository.

If there are restrictions on use, how will access be provided?

If restricted data are stored on a private repository, restrictions will be managed through the repository user management system. If restricted data are only available on a private server, access will be given on demand after an examination of the request.

Is there a need for a data access committee?

A data access committee would be the gold standard. However, humans' resources do not allow it at this tage. The GDPR Research contact from the University of New Caledonia can be solicited as an advisor (as well as the UNC DPO).

Are there well described conditions for access (i.e. a machine-readable license)?

Published data will be accessible through standardized protocols as much as possible (see 2.1). It will also depend on the features provided by the chosen data repositories. Non-published, especially non-open data will be accessible on demand and discussed on a case-by-case basis.

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

We will verify the identity of the person on a private repository through user management system and elsewhere, on demand.

2.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

As far as possible the data produced in the project will be interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. As previously described, we intend to use OGC, SDMX or other well-known standards when dealing with data dissemination when it is appropriate.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

Since the FALAH project involve many research teams from several institutions, we plan to use standards in order to make easy the exchanges of data. As previously highlighted, the standards we expect to use already include specific vocabularies.
As far as possible, we would use existing concepts and vocabularies to build data structure definition. Generally speaking, we would be able to use SDMX cross-domain concepts and code lists for major part of data collected in the project. For specific geographic data we would use INSPIRE concepts and ontologies. When necessary, we will create new domain concepts and code lists whether we will not use the existing ones. These new concepts will be fully documented and open in that case.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

In this project, we do not use ontologies or specific vocabularies. The themes covered by the FALAH project are common in the daily activities and practices of local people and in their relations with the environment. If this should be the case, we would provide mappings and openly publish the generated ontologies or vocabularies with the agreement of their authors. Ontologies and vocabularies could be described through Web Ontology Language (WOL) and Resource Description Format (RDF), making easier the dissemination of this potential new knowledge.

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?

The ontologies or vocabularies produced as well those already existing highly depends on the research field. As far as possible, we will do our best to produce mappings and satisfy re-usability of the produced knowledge. But it is impossible for now, with our current knowledge, to give an unequivocally positive answer.

2.4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

As previously described, private data will be stored in non-open specific repositories for archiving. However, public data will be deposited on open repositories (providing DOI) in order to permit the widest re-use possible.

It is expected to document data through metadata all along the life cycle of data. Metadata would fully describe:

- Collection methodology (including sampling, references to questionnaires, measuring tool, targeted population, etc.)

- Analysis methodology (including variable definition, data cleaning, processes used to create new variables, units of measurement, etc.)

- Related publications when necessary (publications, research papers, etc.) with an identifier and a link (preferably DOI)

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

It is plan and it is recommended that researchers provide the data they used in this project. The recommended license is CC-BY.
After the end of the project an embargo of 10 years will apply on the full dataset. After the embargo and the research publication process, each dataset gathered during the project and publicly available (anonymised for personal data) will be deposited in the chosen repository. However, it is asked to researchers to publish the datasets used for any accepted manuscript publication as soon as possible in a trusted and appropriate repository.
According to GDPR recommendations, personal data will be removed. However, we will store pseudonymised data in an approved archive repository (CINES for instance). Metadata of pseudonymised data will be openly accessible but pseudonymised data will not be open.

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

It will make sense if data collected during FALAH project would be re-used. It is planned to reuse FALAH data, especially in projects related to the South Pacific region. Of course, published data will be comprehensible and re-useable by third parties.
Restrictions can occur regarding pseudonymized but non-anonymized data because of ethical considerations, such as GDPR or participants who did not accept all the conditions describing the data life cycle. That is why, this kind of data will be available uniquely on demand at the researcher's discretion.

How long is it intended that the data remains re-usable?

It is intended that data remains reusable during at least 10 years after the end of the project (period generally suggested by the repository platforms) but we expect it will be reusable for a larger period with an archiving process (history or retrospective data for instance).

Are data quality assurance processes described?

In the FALAH project we follow a process in three main steps including:

- Professional quality management to implement your project.

To ensure high quality of our project, it presupposes good planning and requires professional quality management throughout. Thanks to Mr. Guillaume WATTELEZ, we have been developing and improving project management tools and procedures tailored to research projects for many years. We are ready to support the role as a coordinator and the consortium in setting up a functional governing structure and communication, monitoring risks, preparing and submitting reports and deliverables in a timely manner, and reaching our milestones.

- Adherence to scientific quality requirements specific to the field of research.

Our field of research comes with distinct quality aspects. Indeed, in health-related projects the adherence to protocols is curtail and we have standard procedure, to ensure that studies in communities is conducted identically at every site and make data comparable. The collection methods have already been published and widely used in the past. The new methods will be peer-reviewed and published on dedicated platforms and during the publication of results they will also be described in papers (i.e. peer-reviewed as well). This should guarantee good quality data.

- Operational quality assurance procedures at participant level.

Our standards and policies in place that we adhere to are very important. Good quality management at the participant level facilitate our daily work and internal procedures including a reliable realization of our tasks within the project.

Further to the FAIR principles, DMPs should also address:

Data security: Data from questionnaires (possibly non-anonymized data) is stored on a secured remote server (Exodata).

Ethical aspects: For data involving human participants, consent is systematically requested. When data is non-anonym, GDPR principles should be followed. As much as possible, only pseudonymized data is processed and identifying tables should be removed at the end of the project or sometime later.

3. Allocation of resources

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

Private data will be stored in a dedicated secure remote server managed by a private society (Exodata). Researchers will get access to data when necessary. This represents a total cost of 10 000 €/year.
Archiving is an appropriate way for long term preservation of data. It cloud be done with CINES (https://www.cines.fr), an approved archive centre. This is a cost of 0 €/year.
Public data will be made available through Zenodo. This represents a cost of 0 €/year.
Open access publication is evaluated to 1500 € per publication. This will be covered by European Union via Open Research Europe Journal or via cOAlition S.

How will these be covered?

These costs related to open access to research data will be covered through the University of New Caledonia or the cOAlition S conditions.

Who will be responsible for data management in your project?

Mr. Guillaume WATTELEZ is the data manager in the current project.

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)?

The preservation could be ensured with storages on a server, in a repository, that ensures a guaranteed durability of 10 years after the end of the project. For long term preservation, we could contact archive centers such as CINES.

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.).

In this project, no other digital nor physical outputs would be produced. In the case of a new software or protocol, it will be asked to deliver it on a dedicated platform such as **Github** or **protocols.io** (https://www.protocols.io/) respectively. In other case, researchers should wisely choose the most appropriate platform.

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.

The main outputs of this research should be datasets, interviews, scientific papers and methodologies as well as scientific and probably local knowledge. As described in the project overview, it is not expected to produce new digital, physical or other kind of outputs. However, in case of other kind outputs, the team project will do its best to work in line with the FAIR principles so that these outputs will be properly managed, shared and available for re-use according to scientific ethic. Moreover, the DMP will be modified according to changes in the new situation observed.

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Data from questionnaires is stored on a secured server (Exodata). Data from accelerometer is currently stored on external hard drives (and duplicated once). It is plan to store it on a remote server accessible through sft and scp protocols. Accelerometric data will also be stored as quickly as possible after their collection on dedicated repository with a restricted access like on Zenodo (Open repository for EU-funded research).

Is the data safely stored in certified repositories for long term preservation and curation?

The secured server is rented for 5 years with an extension till the end of the period of re-use of data.
Contacting CINES, which is a specialized centre in archiving, for a long-term preservation would give guarantees in terms of safety and preservation.

5. Ethical aspects

Are there any ethical 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). When dealing with personal data, an informed consent will describe the life cycle of data from collection to dissemination and including anonymisation. The University of New Caledonia (UNC) is the coordinator of FALAH project. As such, the DPO of UNC is the contact specified in questionnaires when collecting personal data.
It is not expected to share personal data because they will be mainly used to manage the process of data collection as well as to merge the data from multi-sources. We will preferably share anonymized data (i.e. totally non-identifiable data) but we may share pseudonymized data with other research teams when necessary. In this case, each staff having to access to the pseudonymized data will be informed about what it is expected in terms of data access, security and restrictions. In particular, they will be warned that it is important to guarantee a high security level regarding this kind of data and it is not allowed to share this pseudonymized data with non-authorized persons / staffs. During the sharing process, a first step will be to send a document mentioning authorized and prohibited actions regarding the data to be shared. Then, it will be asked the research staff agrees these conditions by signing the access agreement.

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

Yes, informed consent includes long term preservation of personal data.

6. Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

No.