# Introduction

The development of the data management plan for the MP4D project was based on the information provided by the MP4D case study document. It was identified that the project was to be facilitated by a partnership between a large non-government organisation and an Australian university, AusUni.

The main objective is to carry out research project investigating the use of mobile technologies for women in agriculture in rural areas. Even though not fully specified in the case study, some innovative technologies are expected to be developed in order to improve the participants’ well-being and livelihoods. No indication of the ownership of resultant technologies was identified.

# Challenges and Rationale

The challenges faced in developing a data management was numerous, especially due to the limited scope of works that was given. Where there was an information gap in the case study, some works solutions we’re assumed using educated guesses but in an arbitrary fashion. There were a number of challenge elements which had to be addressed, namely, data infrastructure, intellectual property, ethical, legal, metadata and preservation.

## Project Structure

Prior to developing the data management plan, it was important to understand how the structure of the stakeholders are related as well as how the project timeline would flow into events.



Figure 1: Project Facilitator Hierarchy (double click to view larger)

In terms of project facilitators, there were five components. AusUni is identified as the research partner, who will ultimately be responsible for the research and output of the project. Non-government organisation (NGO), is largely the capital provider of the project. They will maintain a number of assets, including the Media Outreach Company, Data Services Contractor and the Mail Services Contractor. NGO also provides file storage services for any project related files and maintains the data generated by the project.

Data services and mail services contractors are facilitators of data transportation between the remote project location and the data repositories in Australia. Data services contractor is also responsible for providing backup and data centre services to the project.

The media outreach company is responsible for capturing project related mobile data. They have also been tasked with the de-identification of this data and filtering out of any data that is not project related.

Largely, project was identified is to be of waterfall type, with some tasks running in parallel.



Figure 2: Project Key Events Flow (double click to view larger)

It is identified that the close of the project is once the publication has been made and any research data is either persisted or destroyed as per requirements.

## Generated Data

It was identified that a number of data types would be generated taking various forms. The two main forms are physical forms and digital forms. Each event in the project timeline would generate at least one form of these data.

### Physical Data

Physical data took the shape of any physical documents that would be utilised in the projects. Mainly, these were feedback forms, survey questionnaires, training documents that participants used and other project related documentation. Due to the difficulties in persisting physical data, especially in storage and protection, as well as the difficulties in making physical data available to other researchers, post completion, it was decided to digitise all data that may take a physical form and store them in digital files.

### Data formats

It was decided that the preservation of data that is generated by the project would be of importance. It was identified that this research would be of wide public interest, as the research would result in the ability to help people in agriculture coming from disadvantaged areas aided by the unique technique that was to be utilised for the first time.

By inspecting <Figure 1>, the scope of the project was quite large with a large number of stakeholders. As a result, the reconstruction of such a project would be costly and difficult. It could also be identified that the research outputs would be of significant persistent value for researchers in the field of agriculture and data science. In addition, it perhaps may be the intention of the NGO to develop a usable product which would be aimed at the agricultural industry.

Therefore, it was necessary to record the data in formats which are resistant to obsolescence and enables backward compatibility (ANDS, 2016). Therefore, the following data format choices were made.

|  |  |
| --- | --- |
| File Type | Format |
| Documents (forms, training sheets, etc.) | DOCX |
| Images | TIFF |
| Audio | WAV |
| Tabular Data | CSV |

The ‘DOCX’ is a XML based file format for what was traditionally ‘DOC’, a Microsoft proprietary technology. This means the documents are universally accessible and has reduced risk of the files being damaged (Frank Rice, 2006). The UK Data Archive has found this format to be an acceptable file format for storing textual data for preservation purposes (University of Essex, 2002-2017).

The ‘TIFF’ format is a platform independent file format for data storage commonly for image data storage and is perhaps the most versatile and diverse bitmap format available (Murray & Van Ryper). Specifically, version 6 uncompressed is identified to as being the preferred format of storage, but other TIFF formats are also accepted for data preservation (University of Essex, 2002-2017).

The Waveform Audio File Format or ‘WAV’ format is a format designed to meet the requirements for multichannel sound in broadcasting and archiving (Library of Congress, 2013). This format is widely used and is also accepted as a format for data preservation (University of Essex, 2002-2017).

Comma Separated Values or ‘CSV’ format files are commonly used to transport large amounts of tabular data between entities (CSV Reader, 2017). With simple encoding and wide usage, this is a natural choice for storing tabular data.

During the research phase, the mobile phone usage data was stored in a data warehouse. This enabled the easy access of researchers in Australia to real time data as it was collected. This would later be transported in to tabular data or ‘CSV’ files and archived in a data repository.

### Transfer, Storage and Backup

The storage of the data was deemed to be handled by the NGO. As a private entity, they are best position to provide enterprise level data stores and security to the research data via engaging the necessary specialists.

Two main aspects of data transfer identified, namely, physical data transfer and digital data transfer. Both forms had security concerns given that at least, at some point, data would have private information. Example of the data flow has been shown below.



This meant a secure a transfer for both physical and digital data was required, and therefore, the NGO was tasked with engaging the relevant entities to ensure this requirement. All physical data was to be transported to a collection point at the mail services contractor and transferred in bulk every month to NGO headquarters.

Given that the NGO was to take the responsibility of the data storage, they we’re tasked with carrying out necessary backups of digital data and the security of the physical data. It was deemed that all digital data would be backed up daily in at least 2 geo-replications. All physical data is to be securely stored within the NGO headquarters location in Australia.

## Ownership, Copyright and Intellectual Property

Given the partnership of an NGO organisation and the university, the ownership/intellectual property rights of data and outputs can be quite complicated. Whilst many researchers would want their work to be referenced, used and extended, it may be the intention of the NGO partner that their intellectual property be used to gain some market advantage and therefore be kept concealed.

To ensure that future researchers and parties of interest are not handicapped without exposing the NGO to undue market risk it was decided that an Attribution-Non-commercial 4.0 International licence would cover the research data.

The usage of this licence, it would be possible for individuals and organisations to share and adapt the data as they see fit, given the conditions that appropriate attribution is given and is not used for non-commercial purposes (CreativeCommons, 2017).

It is to note that this does not apply for any software or applications developed during the project.