Aruna (Aaron) Tillerkeratne
S27345483

REFLECTION ON ISSUES

MP4D Data Management Plan

Contents

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.

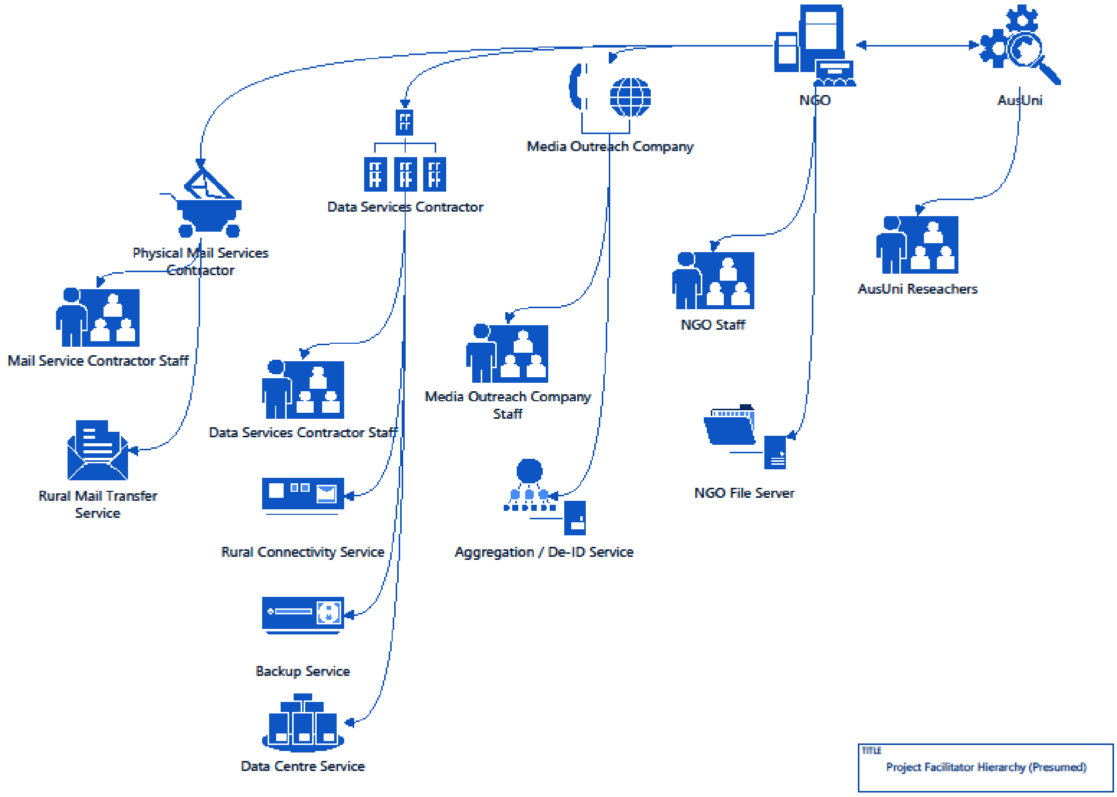

NGO
AusUni
Media Outreach Company
Data Services Contractor
Physical Mail Services Contractor
Mail Service Contractor Staff
Rural Mail Transfer Service
Data Services Contractor Staff
Rural Connectivity Service
Backup Service
Data Centre Service
Media Outreach Company Staff
Aggregation / De-ID Service
NGO Staff
NGO File Server
AusUni Reseachers
TITLE
Project Facilitator Hierarchy (Presumed)

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.

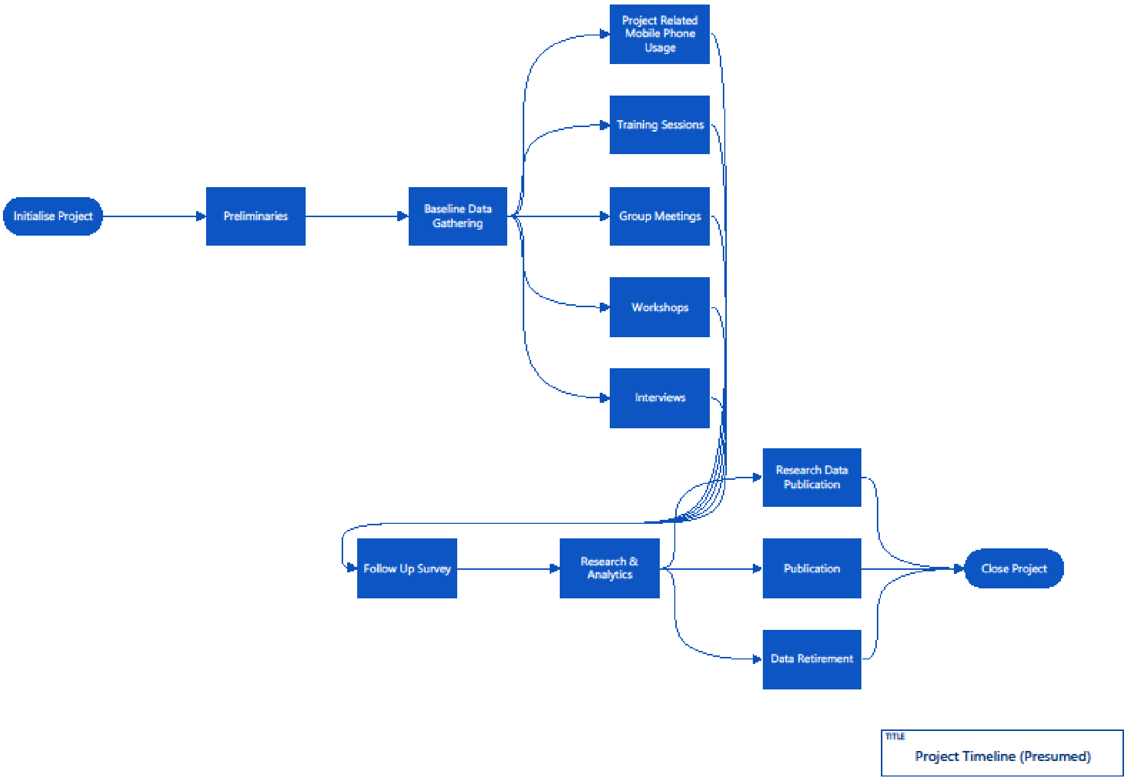


Figure 2: Project Key Events Flow (double click to view larger)

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).

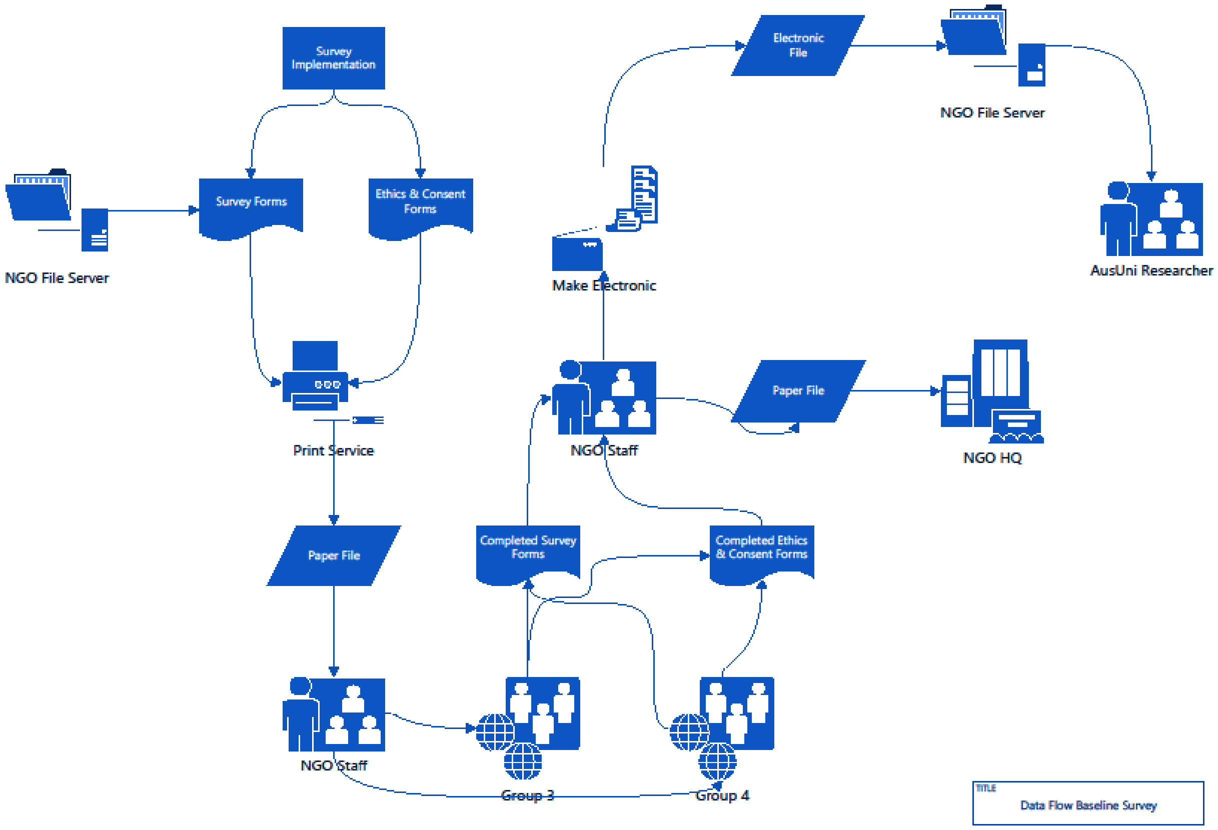
Comma Separated Values or 'CSV' format files are commonly used to transport large amounts of tabular data between entities (CSV Reader, 2017). With simple encoding and wide usage, this is a natural choice for storing tabular data.

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. This is mainly due to the fact that Creative Commons licences do not contain specific terms about the distribution of source code, which is essential to the distribution of a software (CreativeCommons, 2016).

Due to the experimental nature of the project, it was identified that any software generated should be openly distributed among the larger community for further development. However, it is also important that the NGO's financial commitment to the project is substantiated and the distribution of developed software does not become modified by other companies hindering the market standing of the NGO.

It was identified, most of the software that would be generated would be server based, since the development of hub would require some server side element to the system. It was necessary to protect the rights of a server side application (developed in the project) from being used by other companies for market advantage. To ensure that developers are not handicapped and to achieve the above goal, it was decided that a GNU Affero General Public License would be used.

The GNU Affero GPL specifically targets server side applications from being modified by a party but not distributed, potentially providing a better service and disadvantaging the NGO's market share. With this licence if a third party acquires the server software developed in the project, and modifies it and implements in a server that everyone can use, then the source code must be available for download (Free Software Foundation Inc., 2015).

For any standalone software that is generated as a result of the project will carry a GNU General Public License. This ensure that future developers and researchers are free to improve upon the results but cannot use it to gain commercial advantage over the NGO by improving and distributing a non-free copy of the software or hiding the source code from others.

The ownership of intellectual property is to be jointly shared by the NGO and AusUni. This is a certifiable balance between project funding (NGO) and analysis (AusUni), both which are important to achieving the mission statement of improving the livelihoods of women in agriculture in rural areas. The ownership of all generated data and intellectual property will be jointly owned by the NGO and AusUni, however software development, which is understood to be undertaken only by the NGO, NGO shall have ownership of all intellectual property in the development of the software. However, any software developed and source code shall be made available freely under the GNU GLP 3.0 license.

Ethical Requirements and Issues

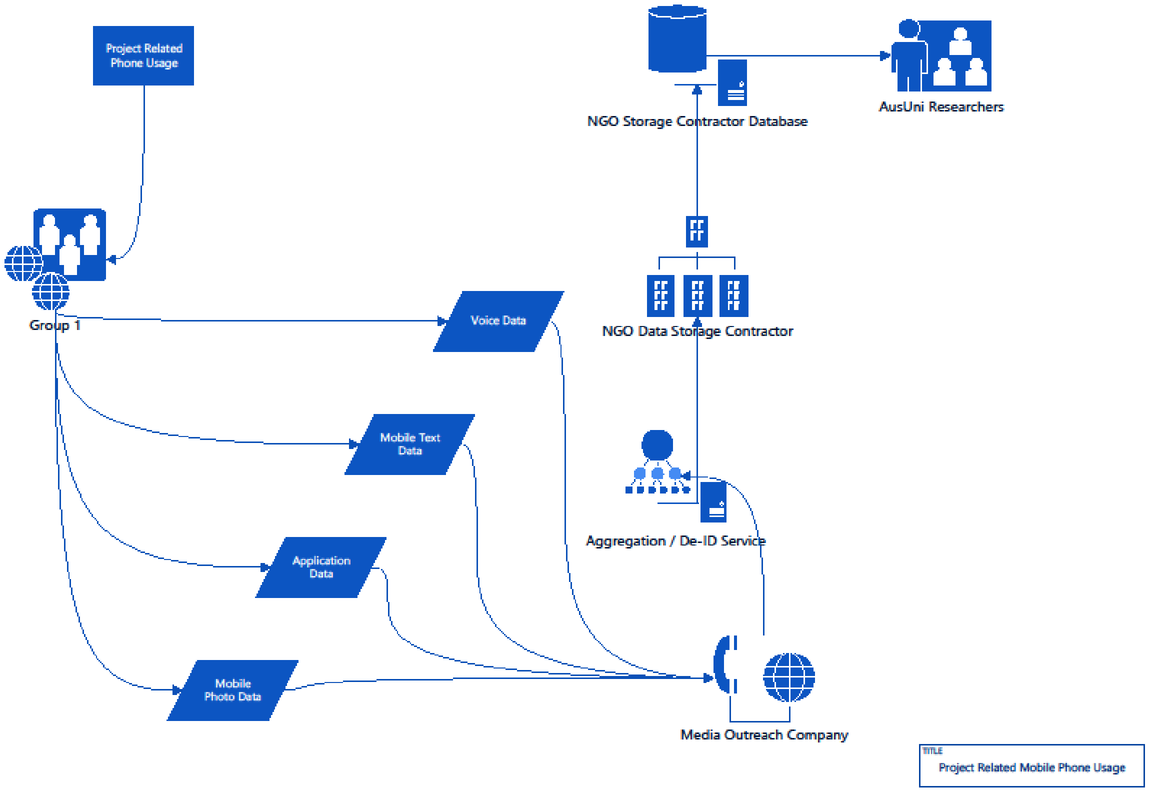
The project raises a number of ethical issues relating to privacy and confidentiality, and to a lesser extent cultural sensitivity. As the research involves data gathering on human subjects, it is of upmost importance to ensure that participants are aware of the extent and the use of the gathered data.

Therefore, prior to the start of the project, in the baseline survey phase, all participants will be briefed on the project, the extent of the data gathering and the data governance principles utilised.

All participants are also required to complete a consent form and questionnaire to demonstrate their understanding and willingness to participate in the project.

Privacy

Given the project will aim to record personal phone calls and text messages, it is imperative that individuals are not identifiable. To further increase security, the de-identification must be carried out at as early as possible in the data gathering process.


Project Related Phone Usage
NGO Storage Contractor Database
AusUni Researchers
Group 1
Voice Data
NGO Data Storage Contractor
Mobile Text Data
Aggregation / De-ID Service
Application Data
Mobile Photo Data
Media Outreach Company
TITLE
Project Related Mobile Phone Usage

The media outreach company, the contractor for gathering mobile phone related data is tasked with immediate di-identification of the data as it passes through to the project team. This also reinforces security of the data in the case of any data leaks whilst in transfer from the project local location to NGO headquarters.

The aggregation and de-identification service is responsible for generating keys which can provide source uniqueness, however the keys are assigned anonymously to participants and no human readable knowledge will be persisted after the completion of the project, ensuring that participant anonymity is maintained. This is then in accordance with the Privacy Act 1988, and the de-identified sensitive data can be shared legally (ANDS, 2017).

Confidentiality

There are also some notable confidentiality issues arising from the project. These are namely the leakage of project related documentation to non-project stakeholders, as well as information from one project event inadvertently becoming exposed to another. There are number of key exposure points where there may be a confidentiality risk.

On site at the research location, confidentiality risks are present where survey documentation and other physical data may be exposed prior to being de-identified. Some or many of these documents may contain information that may be used to identify participants and other location information. Other documents may relate to the performance/interaction of NGO staff with participants and each other and is imperative they are not tampered or openly discussed. To minimise the on-site confidentiality risks, the physical files would be kept in a securely locked cabinets and periodically use the mail services to move them to NGO headquarters. NGO staff assets for recording digital information, such as laptops and mobile phones are to be kept password protected and with personnel at all times.

Once the files arrived at NGO headquarters, there was also another exposure risk of team members outside of the project gaining access to the project materials. Some sensitive documentation may include project budgeting and strategy which would present some confidentiality risk to the project team. To address this issue, all physical items are to be kept in a separated area with access being controlled by access card to the project team. All physical files are to be digitised into appropriate formats and access is to be restricted to the project team.

Post completion of the project, all physical data is to be retired via an appropriate destruction method. Maintaining physical files present a number of preservation challenges and since all files are digitized, a duplication of information.

Sharing Data and Access Control

There are a number of data transportation, sharing and access challenges, some of which have already been discussed. As with data sharing, access was discussed to be limited to the project team during project proceedings.

All data digital data from the project location is to be securely transported via the data transmission services contractor. This ensures the secure and low loss data transmission between the researchers based in Australia and the NGO personnel on location.

All data sharing is to be done within the project team only. NGO will provide necessary servers (and access control) to the AusUni research team where they are free to access the data. Any data updates and project updates are communicated via email, but no project generated data is to be communicated.

All digital data is password protect to ensure security and any voice and text data are also encrypted to ensure in the event of a breach, voice data cannot be examined to determine identity of the participants.

Documentation and Metadata

To ensure the long term usability of the data, a rich metadata system is to be implemented to all persistent files (digital files). This not only enables the project stakeholders to easily understand the scope and context of a particular set of data, it is imperative to integration with open research data repositories. Types of metadata issued are highlighted below.

Data File	**Metadata Type**
Physical files (e.g. survey forms and workshop forms, etc.)	Descriptive – NGO employee ID numbers must be attached although removed from public viewing.
Images	Descriptive, Technical – A clear description of the event, subject as well as image dimensions must be available.

Phone Conversations	Technical – Standard metadata to be attached (see standard below).
Text Messages	Technical – Standard metadata to be attached (see standard below).
Mobile phone images	Technical – No data to identify the users, but standard image metadata only to be kept (see standard below).
Aggregated and Analytic outputs	Descriptive – Full description of algorithm(s) or manipulations to be provided.

It is worthy to note that all physical files will be digitised and appropriately described with meta data tags. AGLS meta data standard is adopted here for metadata collation due to its ability to capture wide range of projects. The standard is based on the Dublin Core Metadata initiative and is localised for use in Australia, which was of particular appeal (National Archives of Australia, 2010).

Retention and Archival

Only de-identified and aggregated data will be retained after the completion of the project. Any physical data will be destroyed securely to ensure anonymity. All digitised data will be deposited to the AusUni repository and Registry Interchange Format – Collections and Services schema will be adopted for the storage. Based on the international standard ISO 2146:2010, RIF-CS provides a generic information standard that is not unique to a particular research domain (ANDS, 2016).

References

ANDS. (2016, December). *File Formats.* Retrieved from ANDS Guide: http://www.ands.org.au/__data/assets/pdf_file/0003/731775/File-Formats.pdf

ANDS. (2016, November 18). *Research Data Australia Content Providers Guide : Learn about RIF-CS Schema.* Retrieved from ANDS: http://guides.ands.org.au/rda-cpg/rifcsschema/#whatisrifcs

ANDS. (2017, January). *Data sharing considerations for Human Research Ethics Committees.* Retrieved from Australian National Data Service: http://www.ands.org.au/__data/assets/pdf_file/0009/748737/HRECS.pdf

CreativeCommons. (2016, August 04). *Frequently Asked Questions - Creative Commons.* Retrieved from Creative Commons: https://creativecommons.org/faq/#can-i-apply-a-creative-commons-license-to-software

CreativeCommons. (2017, 02). *Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).* Retrieved from Create Commons: https://creativecommons.org/licenses/by-nc/4.0/

CSV Reader. (2017). *What is a CSV file ?* Retrieved from CSV Reader: https://www.csvreader.com/csv_format.php

Frank Rice, M. C. (2006, May). *Introducing the Office (2007) Open XML File Formats.* Retrieved from Microsoft Developer Network: https://msdn.microsoft.com/en-us/library/aa338205(v=office.12).aspx

Free Software Foundation. (2016, 11 18). *The GNU General Public License .* Retrieved from GNU Operating System and the Free Software Movement: https://www.gnu.org/licenses/gpl.html

Free Software Foundation Inc. (2015, 05 10). *Why the Affero GPL.* Retrieved from GNU Operating System and the Free Software Movement: https://www.gnu.org/licenses/why-affero-gpl.html

Library of Congress. (2013, 10 17). *WAVE Audio File Format.* Retrieved from Sustainability of Digital Formats, Planning for Library of Congress Collections: http://www.digitalpreservation.gov/formats/fdd/fdd000001.shtml

Murray, J. D., & Van Ryper, W. (n.d.). *Encyclopedia of Graphics File Formats, 2nd Edition.* OReilly.

National Archives of Australia. (2010, July). *AGLS Metadata Standard Part 1 : Reference Description.* Retrieved from Australian Government Locator Service & AGLS Metadata Element Set: http://www.agls.gov.au/pdf/AGLS%20Metadata%20Standard%20Part%201%20Reference%20Description.PDF

University of Essex. (2002-2017). *Formatting Your Data.* Retrieved from UK Data Archive: http://www.data-archive.ac.uk/create-manage/format/formats-table