Data Management Plan

Admin Details

Project Name: Prof. Green Data Management Plan

Principal Investigator / Researcher: Prof. Green

Institution: Dalhousie University

Part I: Data Collection

What types of data will you collect, create, link to, acquire and/or record?

The data collected for the research consists of audio and text files from interviews to industry members. The data is mostly from surveys conducted to industry members and other open data sets from healthcare sources. The data sets include information regarding healthcare, healthcare expenditures, demographic data, healthcare outcomes.

The formats previously used are MP3 for audio and pdf, excel, word and txt. As part of the data management plan, the data will be transformed into MPEG for audio files, csv for numeric information and txt and xml for anything regarding notes.

What file formats will your data be collected in? Will these formats allow for data re-use, sharing and long-term access to the data?

The formats chosen for this research are:

Numeric datasets: CSV

notes and transcripts: Txt, word

Audio recording: MPEG

These formats were chosen since they're industry standard and allow ease of use, distribution and can be machine readable. These formats can be shared if necessary. However, Prof. Green needs to keep the information for himself, given the sensitive nature of it. According to Prof. Green's needs the information will be archived for the long term making it possible to access it whenever necessary. As stated, he will share the quantitative information regarding the study, but not the interview nor transcripts.

What conventions and procedures will you use to structure, name and version-control your files to help you and others better understand how your data are organized?

To name and organize the data files already existing and the ones that will be generated, a naming convention has been designed. The main guidelines are the following:

- File names will remain under 32 characters.
- There will be a clear classification of the types of files in record (transcript, photo, etc.).
- No spaces nor special characters.
- Underscores will be used instead of periods or spaces.
- File names must be descriptive
- Dates of entry must be included (international standard for date notation YYYY_MM_DD or YYYYMMDD)

To assure that no information would be lost and make sure of the reliability of the data for this project, the data management plan has established the following version control program:

- Every week there will be a data migration to the data storage. Therefore, there must be a version control file for each week's final file.
- These file's versions will be maintained for 6 months after a milestone has been accomplished.
- Each version will be organized through a software such as GitHub to simplify the data management process. The GitHub account will belong to Prof. Green and it will be a private account. The team members can only upload information to it. However, will need to request to Prof. Green the latest file to start working in it at the beginning of each week.
- There is a naming convention set up as part of the data management plan. The naming convention will work around dates. Using the latest date as the ultimate file for a milestone.
- Every time a new file is created or updated there will be a record / comment on the latest update as to identify the recent changes.
- Synchronize files in all locations at the end of each week, so there is an update version of the latest file.
- All the final milestone and master versions will be stored on the designed server, at Dalhousie or any other commercial institution chosen by Prof. Green. The back up sever, will contain a copy of everything store in the main server and the archives will only contain the final information of the research that Prof. Green wants to share.

PART II: Documentation and Metadata

What documentation will be needed for the data to be read and interpreted correctly in the future?

As part of the research, it was advised to follow the DDI standard. As part of it, a XML file will be included containing all the necessary information of the research. Thus, making it easier for researchers to know the topic and main details of the document and for machines to read at the same time. Below, it is possible to find a structure of the metadata to be included:

- Principal investigator(s) [Dublin Core -- Creator]
- Title [Dublin Core -- Title]
- Funding sources
- Data collector/producer
- Project description [Dublin Core -- Description]
- Sample and sampling procedures.
- Substantive, temporal, and geographic coverage of the data collection [Dublin Core -- Coverage].
- Data source(s) [Dublin Core -- Source]
- Unit(s) of analysis/observation
- Variables
    - Summary statistics
    - imputation and editing information
    - Universe information
- Related publications.
- Technical information on files
- Data collection instruments
- Flowchart of the data collection instrument.
- Index or table of contents
- List of abbreviations and other conventions
- Interviewer guide.

How will you make sure that documentation is created or captured consistently throughout your project?

To make sure that the requirement in terms of measurements and data creation are consistent, each participant will have the obligation to fill out a weekly progress report. Such progress report will have a list and workflow description of all the procedures that the researcher must take to maintain the consistency within the data capture and creation process.

If you are using a metadata standard and/or tools to document and describe your data, please list here.

As mentioned before, to maintain the relevancy and reputation of the research, DDI standard has been selected as metadata standard. The DDI is in XML format, which will enable the effective exchange of information.

PART III: Storage and Backup

What are the anticipated storage requirements for your project, in terms of storage space (in megabytes, gigabytes, terabytes, etc.) and the length of time you will be storing it?

This project is calculated to last an average of 10 years. So far it has used about 25 gbs. However, the final amount of these data should be about 100-200 gbs by the end of the project. The way the data will be stored will be the following:

- Main Data server 200 GB
- Back up Hard Drive 200 GB
- Final archives 200 GB
- External hard drive 200 GB

How and where will your data be stored and backed up during your research project?

The data will be stored once at the end of each week. Once the final version of the data is ready to be stored and it has been saved according to the version control standards the data will be migrated from the local computer to Dalhousie's data storage server or any chosen commercial server.

At the same time, there will be a copy stored on a back up server that will be located at a different institution. A commercial data back up service can be hired for this task.

Another copy will be maintained in an external hard drive, wich will be on possession of the researcher at all times. This is hard drive will also contain the latest file submitted by the version control standards. So, the researcher has one file storage for his own safety in case of an extreme situation.

Finally, once the project is finished, the final results and metadata will be stored in Dalhousie's data verse or any other commercial storage institution that the researcher fits suitable, according to the proposed budget.

How will the research team and other collaborators access, modify, and contribute data throughout the project?

The rest of the team will be able to access the information that Prof. Green has created through a web-based application that will request the information to the data base through an I.D. that only Prof. Green will be able to create. Once they request the information, the data base will allow the team-members to download it to their computer and work with it.

As to return the information in a safe way, the team-members will post their latest update of information to Prof. Green's git-hub account where they'll follow the version control standards that were chosen. Each team-members will have his/her own branch to update the research. The research members will not have the access to download any information from GitHub, only upload it and it will be Prof. Green the one in charge of assembling the information together and reach each created milestone. Once that is done, the information will be passed onto the respective server and back-up systems.

PART IV: Preservation. -

Where will you deposit your data for long-term preservation and access at the end of your research project?

At the end of the research project, the data can be preserve in a few data-storage places. Depending on the needs or preferences of the researcher. The following is a list of the proposed places for data-storage:

- Dalhousie's Data verse
- Microsoft's data storage services
- Amazon cloud storage
- Oracle data storage services

Indicate how you will ensure your data is preservation ready. Consider preservation-friendly file formats, ensuring file integrity, anonymization and de-identification, inclusion of supporting documentation.

As to maintain the quality and ease of preservation for this research, a suggestion of the right formats to store the information has been shown to Prof. Green. The recommendations for the formats were given on the format type are of the data management plan.

Part V: Sharing and Reuse. -

What data will you be sharing and in what form? (e.g. raw, processed, analyzed, final).

Given the highly classified nature of the information, the data will only be shared with the people related to the research.

The users of the research will be primarily Dr. green's research team and they'll only use it for research purposes. Once the research is finished, Prof. Green agreed to share the findings of his quatitative data in the formats mentioned before.

The types of data that Prof. Green will share with his team members are:

- Processed Data
- Analyzed Data
- Final Data

It is important to highlight that these data types were chosen to be shared given that most of the names and sensitive information are already taken out and/or protected.

Have you considered what type of end-user license to include with your data?

The licensing for this research will be creative commons. The exact type of the license is:

- CC BY-NC-ND

This type of licensing will allow other people to take the results of Prof. Green and use them only if they acknowledge Prof. Green as the creator of it, and it will allow other researchers to create on top of his findings. However, it won't allow anyone to take the findings and use them for commercial purposes. The findings will only be allowed to be used for research purposes.

This type of licensing also allows other researchers to apply new type of licensing from the finding that were a result of this research and their owns.

What steps will be taken to help the research community know that your data exists?

The findings of this research will proof valuable to anyone on the primary care medical community. To make this research easy to find, the following strategies have been placed through the data management plan:

- Metadata that is easy to find and machine readable
- Repository that assigns a DOI (Dalhousie's Data verse)
- Standard citation

Part VI: Responsibilities and Resources. -

Identify who will be responsible for managing this project's data during and after the project and the major data management tasks for which they will be responsible.

The responsible for managing the data management project will be the team in charge of the creation of it. As part of the responsibility, the data management team has not only developed the data management plan but has also created the main milestones for its implementation. Once the data management implementation is completed, the data management will fall under Prof. Green and whoever he names to be his data manager consultant. The main schedule for the task to be developed in the data management plan are the following:

- PART I (Data classification)/ Julie Timm
    - Quality of data
    - Data format
- PART II (Data storage)/ Yichao Ling
    - Version control (Weekly)
    - Weekly check-up of metadata standards (Progress report)
    - Weekly Storage
    - Back up
    - Hard Drive
- PART III Data archives/(Kiranteja Kolli)
    - Final version data & metadata

Part VII: Final commercial storage. -

How will responsibilities for managing data activities be handled if substantive changes happen in the personnel overseeing the project's data, including a change of Principal Investigator?

Given the nature of the project, Prof. Green is the final responsible person for the data and the research. However, if he was not able to finish it for any reason, he would have to designate within the faculty members at Dalhousie University to continue with the research. Such individual will be chosen exclusively by Prof. Green and must understand the nature of the research and the data management procedures in case he needs to continue it. Both Prof. Green and his successor must sign a non-disclosure agreement of the information they handle within the research.

If one of Prof. Green's team-members is not able to continue in the research he has to inform with a 4 weeks span, so Prof. Green can find someone suited for the job. The person leaving the team must deliver all of the information he had on his possession and will be checked by Prof. Green to ensure that he is no taking with him any vital information. Anyone leaving Prof. Green's team must sign a non-disclosure agreement and have a brief legal explanation on the repercussion of any disclosed information on their part.

What resources will you require to implement your data management plan? What do you estimate the overall cost for data management to be?

As part of this project we've proposed to use Dalhousie's servers. However, if the researcher needs to have more available information on other options. The following tables we'll help as a guide.

AMAZON . -

Amazon Main Storage	
First 50 gb	$ 1.25
Next amount	$ 3.60
Total	$ 4.85
Amazon Final Archives	
Data storage	$ 2.20
Retrieval (x 1000 retrieval)	$ 55.00
Data Transfer Pricing	$ 10.00
Total	$ 67.20
Amazon Back-up	
Data storage	$ 2.76
Retrieval (x 1000 retrieval)	$ 5.94
Data Retrievals	$ 10.00
PUT, COPY, or POST Requests	$ 10.00
GET and all other Requests	$ 10.00
Lifecycle Transition Requests into Standard – Infrequent Access	$ 10.00
Total	$ 48.70
DATA ASSISTANTS	

2ETF	$ 4,800.00
Total	$ 4,920.75

MICROSOFT

Microsoft Main Storage (Cool storage)	
First 50 gb	$ 0.67
Next amount	$ 2.01
Write Operations* (per 10,000)	$ 1,216.00
List and Create Container Operations (per 10,000)	$ 669.00
Read Operations (per 10,000)**	$ 122.00
All other Operations (per 10,000), except Delete, which is free	$ 54.00
Data Retrieval (per GB)	$ 2.44
Data Write (per GB)	$ -
Total	$ 2,066.12
Microsoft Back up Azure	
Storage	$ 12.16
LRS	$ 6.42
Total	$ 18.58
Microsoft Archives (Final storage)	
First 50 gb	$ 1.25
Next amount	$ 3.60
Total	$ 4.85
DATA ASSISTANTS	
2ETF	$ 4,800.00
TOTAL	**$ 6,889.55**

Note: The type of database is "cool". Since it is the type with the best price that covers the needs for the project.

Note: LRS was chosen for this since it's the cheapest and most reliable option for the research.

Oracle. -

Oracle Main Storage (Cool storage)	
Block Volumes	$ 11.46
Object Storage - Storage	$ 6.88
Object Storage - Requests	$ 46.00
Data Transfer Service (HDD)*	$ 11.46
File Storage	
Total	$ 75.80
Oracle Back up Azure	
Archive Storage	$ 8.00
Total	$ 8.00
Microsoft Archives (Final storage)	
Archive Storage	$ 0.70
Total	$ 0.70
DATA ASSISTANTS	
2ETF	$ 4,800.00
TOTAL	**$ 4,884.50**

Part VIII: Ethics and Legal Compliance. -

If your research project includes sensitive data, how will you ensure that it is securely managed and accessible only to approved members of the project?

As part of the securities created for this research, only a selected group of people have access to the information. The web-based portal access gives Prof. Green's team-members the right to request and download the information needed, provided that they possess the necessary I.D. created by Prof. Green.

As part of the sharing of information, the team-members can only update the information through git-hub. However, they cannot download it. This is used to have an easy to understand version-control. Nevertheless, each participant will have one branch on a private account managed by the data manager of the project. Nothing can be download, only uploaded. If there is any download, it will be considered a security breach. Thus, allowing the University to take the necessary legal measures.

A possible issue with the usage of GitHub for version control, would be the fact that anyone could use any other online resource to resent any sensitive information to another server or online sharing device. As to avoid this circumstance, there will only be one computer that can have internet connection for this research. This computer will be Prof. Green's which will be attached to Dalhousie's secured VPN , which will only permit to use Prof. Green's online access to Git-hub and all other resources will be banned. Thus, allowing to only use one computer to upload the latest information.

Every time his team-members need to access the information they will have to download it through Prof. Green's computer and then send it to any other personal device used to gathered information. Once they finish with it, they must send the information back to Prof. Green's computer and upload it to Github. No record of data must be kept in any other device.

If applicable, what strategies will you undertake to address secondary uses of sensitive data?

To address secondary uses of sensitive data, it will be explained to all participants the importance of the information gathered and will explained in a simple, yet concise way how their identities and all other personal details will not be shown nor deductible through the research. This would make easier to obtained all of the right permissions on the consent forms. The research cannot proceed without making this point clear and getting the right consent from the participants. As stated before, the information shared will only be the quantitative one, personal details are not to be shared and it is mandatory to explain to the participants how unlikely it is to recognize anyone from statistical data.

How will you manage legal, ethical, and intellectual property issues?

As stated in previous part of the data management plan. The research contains sensitive information that has been carefully handled, making sure that the personal information of the participants is never disclosed and managed in a way that the result doesn't lead to any assumption about the personal identity of the participants.

As part of the ownership of the research, the property rights for this research are:

CC BY-NC-ND. This type of licensing allows to take the findings of the research if they're attributed to Prof. Green directly and only in a research environment. This copyright will not allow anyone use it for commercial purposes. Thus, allowing further research to be built on top of this by reusing the info, with

the condition that Dr. Green is attributed for it and its only for non-commercial purposes. However, Once the new research is done, the researcher can set up a new type of licensing for their new findings. As part of the research, Prof. Green made every participant to give sign a consent agreement for this research and a confidentiality form protecting their personal information.