HISTORY / SUSTAINABLE DEVELOPMENT – DATA MANAGEMENT PLAN

DESCRIPTION

This is a fictional exemplar data management plan developed in May 2020 by Alex Guindon (Geospatial and Data Services Librarian, Concordia University) for educational and guidance purposes. It is mostly based on information taken from a data curation profile that describes an oral history project used "to establish a longitudinal data set that will look at the long term impact of sustainable development in the Nnindye community located in the Mpigi District in Uganda." (Sapp Nelson & Beavis, 2013).

Sapp Nelson, M., & Beavis, K. (2013). History / Sustainable Development - Purdue University. *Data Curation Profiles Directory*, *5*(1). https://doi.org/10.7771/2326-6651.1032

DATA COLLECTION

WHAT TYPES OF DATA WILL YOU COLLECT, CREATE, LINK TO, ACQUIRE AND/OR RECORD?
The original data will include audio and video recordings of the interviews. Photographs of the participants will also be included. In a second stage, transcripts (textual data) will be produced in the original language (Luganda) and English translations (textual) will also be created. Finally, the files will be coded using a qualitative data coding software such as ATLAS.ti or NVivo and the annotations will be added to the textual files.

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 video recordings will probably be in .wmv format (Microsoft proprietary format). These will later be converted to an open format such as mp4 for archival and dissemination. The audio recordings will use the wav format and can be converted to flac if compression is needed for archival. The photographs files will be in jpg format. The transcripts and annotation files will be created in .doc format but a .txt or .rtf will be created for dissemination and archival.

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?
All members of the research team will use the same directory structure and file naming conventions. The directory structure will include separate folders for: audio files, video files, photo files, transcripts, translations and annotations. Each participant will be assigned a number. File names will include the participant number, the date of the interiew (YYYYMMDD) and, if appropriate, the version number. Example: ID_04_Interview_20200422.wav

DOCUMENTATION AND METADATA

WHAT DOCUMENTATION WILL BE NEEDED FOR THE DATA TO BE READ AND INTERPRETED CORRECTLY IN THE FUTURE?
A document describing the precise methodology of the research project will be created. The instrument used for the interviews (interview protocol) will be made available.
A spreadsheet will be created to record all the relevant metadata linked to the interviews, photographs and videos. The columns will include such information as: participant ID number, age, gender, date, time, name of the interviewer. The codebook used by the researchers to annotate the interviews will be shared. It will make explicit all codes or acronyms used in the annotation process and the methodology employed to create themes or categories. Any taxonomy or metadata standard used for coding used will be identied.

HOW WILL YOU MAKE SURE THAT DOCUMENTATION IS CREATED OR CAPTURED CONSISTENTLY THROUGHOUT YOUR PROJECT?
All forms and collecting tools will be shared and discussed with members of the research team before the beginning of the interviews. All necessary documentation will be available to individual researchers onsite. Given the uncertainty regarding electricity and internet access during the field work, each reseracher will need to record the meatadata individually on their own computer. However, each day, or as often as technically possible, the metadata from each researcher will be consolidated on a master spreadsheet on the PI's computer. A graduate student will be responsible for a daily quality check of the collected metadata.

IF YOU ARE USING A METADATA STANDARD AND/OR TOOLS TO DOCUMENT AND DESCRIBE YOUR DATA, PLEASE LIST HERE.
A yet to be determined qualitative data software will be used to code the transcripts. Although no specific metadata standards will be used for the coding of the interviews, we will consider using the Text Encoding Initiative (TEI) (a

standard for the representation of texts in machine-readable digital form) to mark up the final, annotated transcripts. But this will depend on an appropriate budget to pay for research assistants at the end of the project.

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?
We estimate that we will have approximately 13 video files (wmv) (half of the participants will be video-recorded) of approximately 1 hour each as well as 25 audio files (wav), also one hour long. We estimate the number of photographs to approximately 200. The tables below show a rough estimate of the storage needs during the project (working files) and the storage need for preservation. Working files should take around 80 GB. To err on the side of caution, we should make sure to have at least 100 GB of storage space, plus an equal amount for backups.
The total data size for long-term preservation should be lower, in the area of 30 GB. Since this data has historical value it should be preserved indefinitely.

Working files				
	Video (wmv)	Audio (wav)	Photographs (jpg)	Textual (docx)
Quantity	13	25	200	50
Duration (minutes)	60	60	N/A	N/A
File size (MB)	5000	600	3	2
Total size	65,000	15,000	600	100

Preservation files				
	Video (mp4, H.265)	Audio (flac)	Photographs (jpg)	Textual (txt)
Quantity	13	25	200	50
Duration (minutes)	60	60	N/A	N/A
File size (MB)	1500	350	3	0.5
Total size (MB)	19,500	8,750	600	25

HOW AND WHERE WILL YOUR DATA BE STORED AND BACKED UP DURING YOUR RESEARCH PROJECT?
Daily, each PI will transfer the data form the recording instruments to their laptop. In addition, they will perform a daily backup on an external harddrive as well as weekly backup on DVDs.
Ideally, we would like to be able to copy the most recent working copy of the data to PURR (Purdue University Research Repository), but this may prove to be difficult given the limited bandwith and scarce internet connections in Uganda. Also, in order for this remote copy to be done we would need to devise a secure transfer protocol with the IT department in Purdue.

HOW WILL THE RESEARCH TEAM AND OTHER COLLABORATORS ACCESS, MODIFY, AND CONTRIBUTE DATA THROUGHOUT THE PROJECT?
Given the IT circumstances in Uganda and the fact that each PI will collect and store their own working data, there will not be collaboration based on the data during the field work period. Once the PIs are back at Purdue, data will be placed on PURR. During the pre-publication period, only the PIs and authorized research assistants will be given passwords required to access the data. Data shall not be transferred via email and will not be placed on an external platform.

PRESERVATION

WHERE WILL YOU DEPOSIT YOUR DATA FOR LONG-TERM PRESERVATION AND ACCESS AT THE END OF YOUR RESEARCH PROJECT?
The data will be deposited in PURR (Purdue University Research Repository) for preservation and access. Ideally, the files should be preserved indefinitely as oral history data remains useful in the long-term.

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.
Video files will be converted to mp4 format for preservation. Audio files will be converted to flac and photographs will be availble as jpg. Regular fixity checks (cheksum) will be done by PURR.
Interview transcripts will be carefully de-identified by removing all direct identifiers (name, age, address, etc.) as well as indirect identifiers such as mentions of specific people and places.
The documentation will include a user guide describing the project and the methodology, including the complete information required to interpret the coding of interviews.

SHARING AND REUSE

WHAT DATA WILL YOU BE SHARING AND IN WHAT FORM? (E.G. RAW, PROCESSED, ANALYZED, FINAL).
The interview transcripts pre and post coding (in English and Luganda) will be made publicly available after an embargo period of 4 years to allow the researchers to complete all necessary publications.
The interview audio files as well as videos and photographs will be made available to the interviewed participants and their familly as agreed in the consent forms. We will try to obtain the rights to share these recordings with other researchers/students through the consent forms. Recordings for which we have not obtained explicit sharing consent will be restricted.

HAVE YOU CONSIDERED WHAT TYPE OF END-USER LICENSE TO INCLUDE WITH YOUR DATA?
All transcripts will be made available under a CC0 license.
Audio and video recordings and photograpgs for which sharing consent has been obtained, will be shared upon request only. Researchers will need to agree to the terms and sign a license that will specify that the data can only be used for research purposes and must be destroyed after completion of the research projecy. No redistribution of the data will be allowed.

WHAT STEPS WILL BE TAKEN TO HELP THE RESEARCH COMMUNITY KNOW THAT YOUR DATA EXISTS?
The data will be visible in the PURR repository which is fully indexed by search engines. The dataset will also be identified with DOI. All the articles published by team members based on this project will link to the dataset in PURR. The transcripts will be directly downloadable from PURR.

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.
During the field-research part of the project, each PI will be responsible for his/her own data. They will have to abide by the data management protocol devised before data collection that govern file naming conventions and necessary backups.
After the completion of field research, a research assistant, upon supervision from one of the PIs, will be responsible for data cleanup and for veryfing that all data files follow the data management protocol. Once the data cleanup and verification is completed, the research assistant will communicate with PURR to start the data ingest procedures.

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?
The three PIs are collectively responsible for the data, so it is unlikely that they will all leave the project. The research assistant who will manage the data before it is transferred to PURR, will never take the laptops, the harddrives or the DVD backups outside of the lab. He will document any modification of the data files on a daily basis and that information would be transferred to another research assistant in the event he leaves the team before the data has been transferred to PURR.

WHAT RESOURCES WILL YOU REQUIRE TO IMPLEMENT YOUR DATA MANAGEMENT PLAN? WHAT DO YOU ESTIMATE THE OVERALL COST FOR DATA MANAGEMENT TO BE?
We plan to employ a research assistant for a few hours (less than 50) to clean up and manage the data before it is transferred to PURR . At $25/hr this should amount to approximately $1,250. There is no cost for hosting up to 1TB of data (for funded projects) in PURR.
We will purchase three 1TB hard drives (100 $ each) for data storage while doing field research.
All costs will be included in the funding request.

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?
The data files (transcripts, videos, audio recordings and photographs) will be safely hosted on PURR. During the embargo period (4 years) they will only be accessible by members to the research team who will be identified by their university credentials (username and password).
After the embargo, all de-identified transcripts will be made publicly available. Audio and video recordings and photographs for which explicit sharing consent was obtained will made available upon reuqest and subject to a research-only license. The other files will remain restricted and be kept in PURR.

IF APPLICABLE, WHAT STRATEGIES WILL YOU UNDERTAKE TO ADDRESS SECONDARY USES OF SENSITIVE DATA?

Secondary users (exclusively researchers and students) will be required to sign a specific license. The terms will require that the data is used stricly for research or education purposes. The data will not be redistributed. Researchers will need to maintain the anonymity of participants in all their research outputs (publications, conference presentations, etc.)

HOW WILL YOU MANAGE LEGAL, ETHICAL, AND INTELLECTUAL PROPERTY ISSUES?

All data will be managed according to the ethics rules and policies of Purdue University and the research project will get hrough an IRB approval.

The intellectual property rights of all recordings will remain with the three PIs. Transcripts will be available under a CC0 license.