INFO 6540: Data Management

Group Project

RCASE 2: Professor Green

Admin Details

Project Name: High-stress environments
Principal Investigator / Researcher: Professor Green
Institution: Dalhousie University

Data Collection

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

A variety of data types will be collected, including textual, tabular, and audio data.

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?

Currently there are a mix of data file formats. We recommend utilizing Plain Text Data (.txt) for textual data, Comma-Separated Values (.csv) for tabular data, and Free Lossless Audio Codec (.flac) for audio files. These open, standard, interchangeable, long lasting formats are important to use because they allow others to reuse the data in the future. They also allow for long-term access to the data since they are sustainable digital file formats.

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?

We recommend creating a standard file naming system that everyone involved in the project will use (ie: YYYY-MM-DD-time-NameOfResearcher). Using software such as Git is useful in version control, making it possible to go back to an older version of the file in the event of mistakes. With such a wide range of data formats, it is important to practice maintaining organized data.

Documentation and Metadata

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

Using a metadata standard is good practice and very important. Metadata standards help sum up datasets by providing data about them (ie: title, date, size, etc.) We recommend using the Data Documentation Initiative (DDI) standard, as it is a metadata standard commonly used for the social sciences.

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

We recommend that everyone involved in the project be aware of the Data Documentation Initiative guidelines and all follow them. Maintaining consistency in documentation allows for easy searchability later. We recommend having a meeting before beginning the project to discuss these documentation standards and the importance of following them.

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?

Current storage needs are about 24 GB. In anticipation of future data collection, we estimate a total need for at least 72 GB of storage. We anticipate the necessity of this storage of data for ten years, since we expect it to be published after is use for current research.

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

We recommend storing at least 3 copies of the data, in at least 2 locations, with 1 location being offsite. For the offsite storage, we recommend using a cloud storage system. Your institution provides cloud storage through OneDrive. As you mentioned, you are currently storing data in Zotero, DropBox, and Google Docs; however, it is best practice to store your data in one place to maintain organization and simplicity. For onsite storage we recommend using encrypted external hard drives. We also recommend digitizing the physical copies of spreadsheets from previous data projects. It will assist in organization and accessibility for you and your colleagues.

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

The research team and collaborators will be able to access, modify, and contribute data through their institution’s access to SharePoint.

Preservation

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

We recommend storing your data on your institution's Dataverse for long term-preservation and access at the end of your research. Dataverse is an open source web application to share, preserve, cite, explore, and analyze research data.

Indicate how you will ensure your data is preservation ready.

We recommend anonymizing all sensitive data (i.e. healthcare datasets, interviews) by removing direct identifiers (i.e.names, addresses, postal codes, telephone numbers or pictures). Indirect identifiers (i.e. workplace, occupation, or exceptional values of characteristics like salary or age) also need to be removed because when they are placed together with "publicly available information sources" (Corti, Van den Eynden, Bishop, & Woolard, 2014), people can be identified. The audio files and transcripts should be especially encrypted and password protected as they contain more obvious identifiers for the participants. Additionally, using plain text or CSV machine readable format is superior for long term preservation since other softwares update regularly and older versions can no longer be accessed.

Sharing and Reuse

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

We recommend sharing all analyzed data that has been anonymized. You need to share your data so that people can reuse it. Reuse of data is important because it can help provide historical context in the future, be used for comparative research, restudy or follow up and secondary analysis. Additionally, it can be used for replication or validation, research design and methodological advancement, and for teaching and learning purposes (Corti, Van den Eynden, Bishop, & Woolard, 2014).

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

MIT license. This is a proprietary free software license.

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?

We recommend only publishing the anonymized data, and using encryption during the analytical phases so only people with the passwords can access the data.

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

Sensitive information will not be published. If a researcher feels they require access to the sensitive data, they will be required to submit a request. Requests will be judged on a case by case basis.

RCASE 3: Professor Pinkerton

Admin Details

Project Name: Excel Rescue
Principal Investigator / Researcher: Professor Pinkerton
Institution: Dalhousie University

Data Collection

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?

We do not recommend working solely in Excel. 88% of all Excel spreadsheets have errors in them, and due to the nature of Excel, they are very hard to detect (Olshan, 2013). Simple text based machine-readable formats, such as Plain Text (.txt) or Comma-Separated Values (.csv), make it easier to identify errors. Excel also has very poor version control making it hard to compensate for mistakes. Finally, Excel is often forced to do things more complex than it was intended for. Keeping the data in excel files on your own computer is fine for manipulating them since you are more comfortable with that, but in order to share your data with others, it is beneficial to use open source formats. Open, standard, interchangeable, long lasting formats are important because they allow others to reuse the data in the future. They also allow for long-term access to the data because they are sustainable digital file formats.

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

We recommend creating a standard file naming system that everyone involved in the project will use (ie: YYYY-MM-DD-time-NameOfResearcher). Using software such as Git is useful in version control, making it possible to go back to an older version of the file in the event of mistakes. Additionally, with the massive number of spreadsheets, it is important to practice maintaining organized data. We recommend you maintain variables in the columns and observations in the rows; this allows for easy manipulation of the data and is best practice to reduce errors.

Documentation and Metadata

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

Using a metadata standard is good practice and very important. Metadata standards help sum up datasets by providing data about them (ie: title, date, size, etc.) We recommend using Dublin Core as it is an interoperable online metadata standard.

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

We recommend that everyone involved in projects be aware of the Dublin Core guidelines and all follow them. Maintaining consistency in documentation allows for easy searchability later. We recommend having a meeting before beginning any project to discuss these documentation standards and the importance of following them.

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?

The anticipated storage space requirements will be 60.8 GB for your 17,384 spreadsheets that average 3.5 MB as you stated. In the anticipation of future data collection, we estimate storage needs growing exponentially. As you mentioned you are storing some data for future projects, we anticipate needing the data to be stored indefinitely.

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

With the data you plan to analyze later, we recommend storing at least 3 copies of the data, in at least 2 locations, with 1 location being offsite. For the offsite storage, we recommend using a cloud storage system. Your institution provides cloud storage through OneDrive. For onsite storage we recommend using encrypted external hard drives. We recommend purchasing a 1TB external hard drive, which can be found for $60-$100 at any electronics store, and purchasing more as needed. We recommend being consistent when using the shared cloud server instead of emailing spreadsheets back and forth; this is poor practice and leads to disorganized data and a lack of version control and security.

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

The research team and collaborators will be able to access, modify, and contribute data through their institution's access to SharePoint.

Preservation

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

We recommend storing your data on your institution's Dataverse for long term-preservation and access at the end of your research.

Indicate how you will ensure your data is preservation ready.

We recommend anonymizing the sensitive data (i.e. student performance data) by removing direct identifiers (i.e.names, addresses, postal codes, telephone numbers or pictures). Indirect identifiers (i.e. workplace, occupation, or exceptional values of characteristics like salary or age) also need to be removed because when they are placed together with "publicly available information sources" (Corti, Van den Eynden, Bishop, & Woolard, 2014), people can be identified. Additionally, using plain text or CSV machine readable format is better for long term preservation because Microsoft updates Excel regularly and older versions cannot be accessed anymore.

Is it necessary to preserve all the data?

We recommend thinking critically about the data sets you have acquired. It is not necessary to preserve all of the data. You have collected many datasets, and it is important to only retain what you actually use and what you can reasonably see yourself working with in the future. Preservation is costly and it also leaves more potential for disorganization.

Sharing and Reuse

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

When you're finished working with your datasets, we recommend sharing data to Dalhousie University's dataverse. Dataverse is an open source web application to share, preserve, cite, explore, and analyze research data.

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

We recommend sharing your data in the form of raw anonymized datasets.

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

MIT license. This is a proprietary free software license.

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?

We recommend only publishing the anonymized data, and using encryption during the analytical phases so only people with the passwords can access the data.

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

Sensitive data will not be published. If a researcher feels they require access to the sensitive data, they will be required to submit a request. Requests will be judged on a case by case basis.

RCASE 4: Professor Chartreuse

Admin Details

Project Name: Science of Science Research
Principal Investigator / Researcher: Professor Chartreuse
Institution: Dalhousie University

Data Collection

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

The data types are largely in tabular format. In the future, interview and survey data is expected to be collected (audio/textual).

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?

We recommend utilizing Comma-Separated Values (.csv) for your tabular data. For your interview data we recommend storing your audio recordings as Free Lossless Audio Codec (.flac) files, and transcribing your interviews into Plain Text Data (.txt) files. These open, standard, interchangeable, long lasting formats are important to use because they allow others to reuse the data in the future. They also allow for long-term access to the data because they are sustainable digital file formats.

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?

We recommend creating a standard file naming system that everyone involved in the project will use (ie: YYYY-MM-DD-time-NameOfResearcher). Using software such as Git is useful in version control, making it possible to go back to an older version of the file in the event of mistakes. You mentioned a past issue where a dataset was lost and it was immensely difficult to retrieve. Using a system with version control, such as Git, will solve this problem for you. With such a wide range of data formats, it is important to practice maintaining organized data.

Documentation and Metadata

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

As you mentioned, the large amount of documents you store has led to confusion about what data is stored where and how old it is. As a solution, we recommend using a metadata standard. Using a metadata standard is good practice and very important. Metadata standards help sum up datasets by providing data about them (ie: title, date, size, etc.) We recommend using the Data Documentation Initiative (DDI) standard, as it is a metadata standard commonly used for the social sciences.

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

We recommend that everyone involved in the project be aware of the Data Documentation Initiative guidelines and all follow them. Maintaining consistency in documentation allows for easy searchability later. We recommend having a meeting before beginning the project to discuss these documentation standards and the importance of following them.

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?

You mentioned your data was currently 20GB and that you plan to continue your research indefinitely. As such, we recommend for your current and future research needs that you have an unlimited data cloud storage plan.

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

With the data you plan to analyze later, we recommend storing at least 3 copies of the data, in at least 2 locations, with 1 location being offsite. For the offsite storage, we recommend using a cloud storage system. Your institution provides cloud storage through OneDrive. For onsite storage we recommend using encrypted external hard drives.

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

The research team and collaborators will be able to access, modify, and contribute data through their institution's access to SharePoint.

Preservation

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

We recommend storing your data on your institution's Dataverse for long term-preservation and access at the end of your research. Dataverse is an open source web application to share, preserve, cite, explore, and analyze research data.

Indicate how you will ensure your data is preservation ready.

In the future research involving interviews and surveys, we recommend anonymizing all sensitive data by removing direct identifiers (i.e.names, addresses, postal codes, telephone numbers or pictures). Indirect identifiers (i.e. workplace, occupation, or exceptional values of characteristics like salary or age) also need to be removed because people can be identified when they are placed together with your information from public domain resources. The audio files and transcripts should be especially encrypted and password protected as they contain more obvious identifiers for the participants. Additionally, using plain text or CSV machine readable format is better for long term preservation because other softwares update regularly and older versions cannot be accessed anymore.

Sharing and Reuse

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

We recommend sharing all analyzed data that has been anonymized. You need to share your data so that people can reuse it. Reuse of data is important because it can help provide historical context in the future, be used for comparative research, restudy or follow up and secondary analysis. Additionally, it can be used for replication or validation, research design and methodological advancement, and for teaching and learning purposes (Corti, Van den Eynden, Bishop, & Woolard, 2014).

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

MIT license. This is a proprietary free software license.

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?

We recommend only publishing the anonymized data, and using encryption during the analytical phases so only people with the passwords can access the data.

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

Sensitive information will not be published. If a researcher feels they require access to the sensitive data, they will be required to submit a request. Requests will be judged on a case by case basis.

References

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2014). *Managing and sharing research data: A guide to good practice*. London, UK: Sage Publications.

Olshan, J. (2013). 88% of spreadsheets have errors. *Marketwatch*. Retrieved from https://www.marketwatch.com/story/88-of-spreadsheets-have-errors-2013-04-17.