BDK12 Data annotation and curation

Unit 2: Data preparation and planning and File and Directory Naming

Exercise:

In a hypothetical study, you are collecting survey data from participants in your community. You will collect structured survey data in a table from 100 participants by asking them questions over the phone. This data will include their names, addresses and birthdates.

1. While you are collecting the data, what type of file should you use?
   1. None, it should be handwritten X Handwritten tables are unstructured, not searchable, difficult to share, and potentially difficult to read.
   2. .docx (MS Word or equivalent) X Tables within documents are less structured and more difficult to share and analyze.
   3. .xlsx (MS Excel or equivalent) ✓ A spreadsheet is the easiest, most effective way to collect data.
   4. .mp3 (a voice recording) X Sound recordings are not structured and may not be able to capture correct spellings.
2. For long-term preservation and sharing, will you change the file type? Why? What will the new file type be?

Change to a .csv (comma separated value) file for sharing and preservation because it can be read by many different types of software.

1. To share survey questions or other documents, what file type(s) should you use for sharing and preservation?

The survey questions should be in either a .txt (plain text) or a .pdf. These can be read by most software. (While you are writing the questions, it is all right to use a .docx file)

1. Following the recommended practices for file naming, write examples of file names that you would use for your data files and survey documents.

CommunitySurvey\_Data\_2015-09-08.xlsx

CommunitySurvey\_Data\_2015-09-08.csv

CommunitySurvey\_Questionnaire\_2015-09-08.docx

CommunitySurvey\_Questionnaire\_2015-09-08.txt

CommunitySurvey\_Questionnaire\_2015-09-08.pdf

1. A data dictionary is used to describe data objects so others can understand them, and can enter the data properly and consistently. You can read more about data dictionaries here: https://en.wikipedia.org/wiki/Data\_dictionary

Create a data dictionary for your survey data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Name | Description | Format | Units | Values |
| Name | Name of respondent | Last, First | N/A | No restrictions |
| Address | Current mailing address of respondent (could also be address of residence) | Street name and number, city, state, ZIP. Each item separated by commas, no new lines. For example, 123 1st Ave, Portland, OR, 97201 | N/A | Must have current address be in Portland, OR (for example) |
| Birthdate | Date of birth of respondent | YYYY-MM-DD | N/A | Must be over 18 at time of survey(for example) |

Variable Name: how it appears in your data

Description: a brief explanation of the variable

Format: using best practices or your own judgement, decide how the data will be recorded. For example, for a date, will the format be YYYY-MM-DD, MM/DD/YY, etc.

Units: units of the variable. For example, if the variable were height, would you be using centimeters or feet/inches.

Values: any restricted values. For example, if you are only collected data about current residents of Oregon, they could not have a non-Oregon address.