CREATING A CODEBOOK

One of the most important documents generated during a research project is the codebook. This document provides details regarding the variable construction and coding, database generation, and other factors associated with data quality. The codebook is relied upon during the analysis process and mistakes can jeopardize the project. In addition, codebooks are often consulted many years after the data were collected and as such, it needs to be developed carefully.

Codebooks must contain:
1. **Variables.** Every measure used to collect the data must be written out in detail (verbatim). For example, if a survey was used, every question must be listed along with the coding for each item. The "name" of each item must also be included. In first the example provided below (Figure 2), the name of the variables are in blue text while the coding is in red. Please note that open-ended items have a string notation. A codebook must also contain instructions for how to code missing data.

Figure 2. Segment of a Survey Instrument

FAMILY HISTORY

FH_1 Did you live with both parents until you were 18 years old? 1= Yes 0 = No ______

FH_2 Before 18 years old, were you taken to live with another family member or foster care? 1= Yes 0 = No ______

FH_3 How often did you move before age 18? ______
1 = more than once a year 3 = once every 3 to 6 years 5 = never moved as a child
2 = once every 1 to 2 years 4 = once 7 or more years

FH_4a Did you miss days of school on a regular basis? 1= Yes 0 = No ______

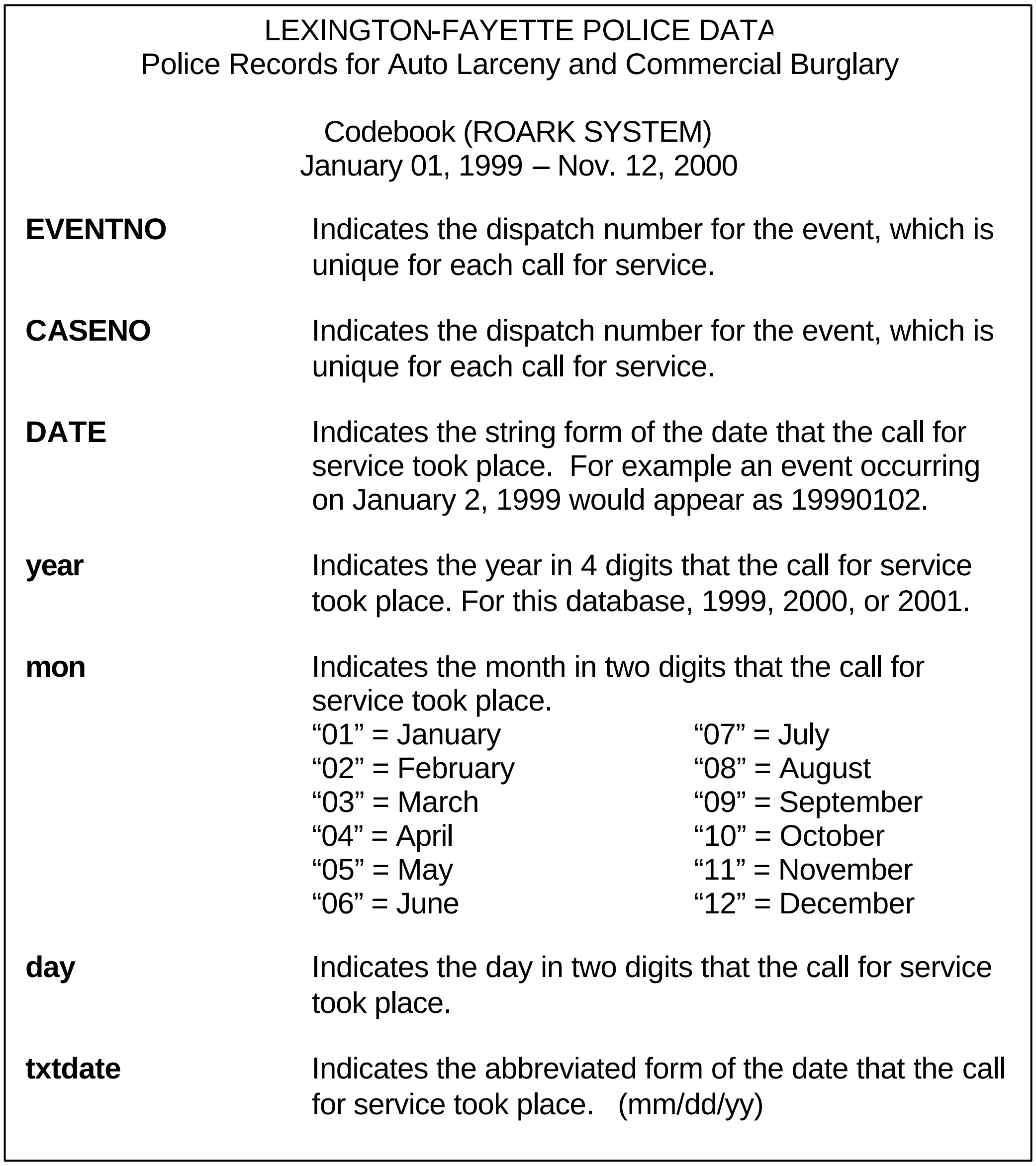
FH_4b If so, how many days did you miss in a row (consecutively)? ______
1 = one day/week 2 = 2 – 4 days/week 3 = 1 – 3 weeks 4 = more than 3 weeks

FH_4c On average, what was the main reason for these absences? _____string (actual answer given in text)____

FH_5 Were you involved in sports teams or school clubs? 1= Yes 0 = No ______

The Figure 3 shows a partial codebook for a dataset extracted from secondary sources, in this case it is from calls for service (911 calls to a police department). You will note that in this example there is a sentence describing what the variable is and when necessary, there is a coding structure. Here, the name of the variable is in bold; all capitals denotes the variable name recorded in the original data as forwarded by the agency and the lowercase names reflect new variables that were generated or recoded by staff at our Center. *Year*, *mon*, *day*, and *textdate* were developed from data contained in the *DATE* variable.

Figure 3. Secondary Data Codebook



LEXINGTON-FAYETTE POLICE DATA
Police Records for Auto Larceny and Commercial Burglary

Codebook (ROARK SYSTEM)
January 01, 1999 – Nov. 12, 2000

EVENTNO Indicates the dispatch number for the event, which is unique for each call for service.

CASENO Indicates the dispatch number for the event, which is unique for each call for service.

DATE Indicates the string form of the date that the call for service took place. For example an event occurring on January 2, 1999 would appear as 19990102.

year Indicates the year in 4 digits that the call for service took place. For this database, 1999, 2000, or 2001.

mon Indicates the month in two digits that the call for service took place.

“01” = January
“02” = February
“03” = March
“04” = April
“05” = May
“06” = June
“07” = July
“08” = August
“09” = September
“10” = October
“11” = November
“12” = December

day Indicates the day in two digits that the call for service took place.

txtdate Indicates the abbreviated form of the date that the call for service took place. (mm/dd/yy)

2. **Metadata.** Codebooks also contain information about the data collection procedure. This may include a description of the problems or issues that arose during data collection that impact on the quality or coding of the data. For example, during the second year of a multiyear project, one of the variables may have been dropped. All subsequent entries would be coded as missing data. Thus, any analysis of this variable must only include the cases up until the time data collection was dropped. Metadata must include:

- status: date the dataset was cleaned
- source: what was used to gather the data
- geographic application: list of areas associated with the data
- data collection methods: one paragraph or so outlining how the data were collected and subjects chosen and the time period of collection

- name of staff that worked on data collection
- subject selection criteria
- number of cases
- who created the final dataset
- who entered or coded the data
- name of data file (and shapefile or coverage where appropriate)
- date the data were last updated
- geocoding results: frequency and percent matched, partial matches, and no matches.

The metadata is essential as this information provides important methodological information that is used to write the reports. Some of the information listed above is also used to ascertain study limitations. Codebooks must be updated when new variables are created or whenever limitations or research issues arise.