README

****
About the NEWID
****

The original geographical identifiers in the DVN versions were scrambled in order to conform to public data dissemination protocol.

Please note that the "newid" for HH, KI, IT, and BAAC modules has slightly different composition. The scrambled "NEWID" still consistently represents unique geographical location (although not the real geo ID) of each case.

For non-mapping project, it is possible to track a given Village/ Tambon /Amphoe over time using the public data without the original codes:

For the Household Surveys 1997-2010, the NEWID is composed by:

1-2 digits = Changwat (province) code
3-4 digits = Amphoe (county) code
5-6 digits = Tambon (subcounty) code
7-8 digits = Village code
9-12 digits = Household code

Example...
072124510902
changwat: 07
ampho: 21
tambon: 24
village: 51
household: 0902

For the Key Informant Surveys 1997-2010, the NEWID is composed by:

1-2 digits = Changwat (province) code
3-4 digits = Amphoe (county) code
5-6 digits = Tambon (subcounty) code
7-8 digits = Village code

Example.
27324053
changwat: 27
ampho: 32
tambon: 40
village: 53

For the Institution Surveys 1997-2010, the NEWID is composed by:

1-2 digits = Changwat (province) code
3-4 digits = Amphoe (county) code
5-6 digits = Tambon (subcounty) code
7-8 digits = Village code
9-12 digits = Institution code

Example...
532938571812
changwat: 53
ampho: 29
tambon: 38
village: 57
institution: 1812

For the BAAC Surveys 1997 & 2000, the NEWID is composed by:

1-2 digits = Changwat (province) code
3-4 digits = Amphoe (county) code
5-6 digits = Tambon (subcounty) code
7-8 digits = Village code
9-11 digits = Group code

Example.

07212451008
changwat: 07
ampho: 21
tambon: 24
village: 51
group: 008

By extract the first 4 digits from the "NEWID", a user can obtain the village dummies.
In STATA, we can simply use the following commands to extract the substring:

```
gen villageid=substr(newid, 1, 4)
```

****

How to Use the Data Dictionary

Variable Name: Data item names are unique throughout the entire file. Most of variable names are sequential identifier such as LA1, LA2 indicating the question number on the questionnaire that the variable comes from. Some derived variables may be a mnemonic string of character such as CASEID or INITEM.

Label: The label is a short description of the data item.

Type: The data type indicates whether the data item is numeric (N) or character (C).

Position: The position of a variable in the Stata data file.

Survey/Derived: Most variables come directly from original data collection forms and instruments. There are also constructed variables that are derived after the data collection.

Question Num: The question number indicates the original question in the questionnaire, where the variable comes from.

Question Text: Text from the questionnaire is provided if the variable was obtained directly.

Frequencies & **Means:** For most categorical variables in the codebook, the counts and percentages of each variable in the data file are shown. For continuous variables, we provide descriptive statistics on minimum, maximum, mean, and standard deviation. The number of valid cases and missing cases are also shown for all variables.

Unique Values: A list of all of the possible non-missing values for the variable and the description of the values.

Unique missing data codes: There are five types of missing values:

For Numeric Variable	For String Variable		Missing Value Type
.a	**RF**	=	refuse to answer
.b	**DK**	=	don't know the answer
.c	**NA**	=	not applicable
.x	**IV**	=	invalid value
.y	**MI**	=	missing value

.a or **RF**: The subject explicitly refused to answer the question or did not answer the question when he or she should have.
.b or **DK**: The subject was unable to answer the question, either because he or she had no opinion or because the required information was not available.
.c or **NA**: The subject was never asked the question for one reason or another. Usually this results from "skip patterns" that occur.
.x or **IV**: A response is inconsistent with related responses or is incompatible with response categories. The circumstances could be: interviewers/data-entry worker incorrectly record a response when recording or keying in the data.
.y or **MI**: Items should be filled out but have no data entry found. This is enumerator's own mistake and nothing to do with the respondent. The circumstances can be interviewers failing to ask a question or forgetting to record a response.