Data Quality Step by Step

Ronnie Babigumira

PEN Workshop, 08/01/08
Outline of Part I

1. A generalized approach
   - Background
   - How can we get it right?
Outline of Part II

2. PEN’s DQ Procedures
   - The Dilemma
   - The Fixes
Part I

General Principles
The big questions

... need Data

... At one level "data" are the world that we want to explain .... At the other level, they are the source of all our troubles ... Griliches, Zvi. "Data and Econometricians-the Uneasy Alliance." American Economic Review 75, no. 2 (1985): 196-200.

- Anyone who has delved into data from the real world knows it can be messy, very messy
- And yet the quality of a data is of prime importance for accurate, reliable and valid results.
Where does it go wrong

Main Culprit
The Human Element

Accomplice
The Tech Element
Good Quality Data Workflow

Research Question
... It is better to use an approximate solution to the right question than an exact solution to a wrong question ...

- If you don’t ask the right question, you will likely correct the wrong data
- Clarity
- Optimal ignorance
Good Quality Data Workflow

Research Question

H.E

Survey Instrument
My heart sinks when someone produces their own questionnaire, consisting of questions that they have thought of on the bus [Dr Fisher’s Casebook: A shy nurse consults the good doctor. 2006. Significance 3 (3):122–122.]

- Hurriedly throwing a questionnaire together is at best a waste of time and at worst, a source of flawed data that could affect your work and a reputation.
- More details in section 6.4 of the PEN technical guidelines
Good Quality Data Workflow

Research Question → H.E → Survey Instrument → Survey
The Government are very keen on amassing statistics. They collect them, add them, raise them to the \( n^{th} \) power, take the cube root and prepare wonderful diagrams. But you must never forget that every one of these figures comes in the first instance from the village watchman, who just puts down what he damn pleases - Sir Josiah Stamp, Inland Revenue Department of England, 1896-1919

- Pre-testing.
- Hiring and training enumerators (probing.. respondents may mis-report)
- Sampling
- Jens will talk more about this
- Also see section 6.5 of the PEN technical guidelines
Good Quality Data Workflow

Research Question

Survey Instrument

Survey

Checking

H.E.
Visual inspection of hard copies of the questionnaires as they return. Best done while in the field. Look out for

- Gaps
- Legibility
- Consistency e.g. No livestock but you have livestock data
- A debriefing session recommended where enumerators share their experiences
Good Quality Data Workflow

- Research Question
- Survey Instrument
- Survey
- Checking
- Coding

H.E
Is a crucial part of the data preparation and should be treated as such.
“Close” all open questions before entering data
The creation of the “other” code is best done on the PC i.e.
Only code something as other as a last resort
All questionnaires should be coded before data entry
Good Quality Data Workflow

Research Question

Survey Instrument

Survey

Checking

Coding

Data Entry

H.E

T.E

Ronnie Babigumira

Data Quality Step by Step
Data Entry

Good practice

- Is about capturing & storing raw data as accurately as possible
- Not to be confused with data management or analysis
- Do it in a way that eases analysis. Preferably done in the field

What to use

- Word processors or text editors
- Spreadsheets ("To err is human. But to really foul things up, you need Excel")
- Statistical packages e.g SPSS. Base not good for entry, pay more and you get the data entry module. Epidata highly recommended (gratis)
- Database packages. Hard to set up but good rewards
- Use the wrong software & not only will it garble your data, it might eat you up as well.
Good Quality Data Workflow

Research Question

Survey Instrument

Survey

Checking

Coding

Data Entry

Data Cleaning

H.E

T.E

Ronnie Babigumira

Data Quality Step by Step
**Data Cleaning**

**What is it**
- A set of procedures aimed at ensuring the reliability and correctness of the data by detecting and correcting (or removing) corrupt or inaccurate records.
- Almost always necessary yet often ignored

**What to look for**
- Completeness
- Outliers (what to do about them?)
- Range
- Logic
- Arithmetic

_I found mothers younger than their children. I found men who claimed to have cervical smears_ ("Dr Fisher’s Casebook the Trouble with Data." Significance 4, (2007))
Good Quality Data Workflow

- Research Question
  - H.E
  - Survey Instrument
  - Survey
  - Checking
  - Coding
  - Data Entry
  - Data Cleaning
  - Data Management

How can we get it right?

Ronnie Babigumira

Data Quality Step by Step
Data Management

What is it

- Some say everything from data entry to analysis
- Our definition: All the pre-analysis processing of clean data. Includes
  - Data documentation (labeling)
  - Creating new variables / obs and changing existing ones
  - Managing files separating, combining, collapsing, reshaping

Good practice

- Raw data: Leave it alone, create flat files.
- Transparency transparency transparency: Leave a clear audit trail.
- Document: Some say, if data is not documented, run as fast as you can.
- Tech Element: Save your hair & sanity. Use a command drive program.

Ronnie Babigumira
Good Quality Data Workflow

Research Question

Survey Instrument

Survey

Checking

Coding

Data Entry

Data Cleaning

Data Management

Data Analysis

H.E

T.E
Data Analysis
This is what it is all about

It is all well and good to collect data at great length, expense, and effort but the most important aspect is often not the actual information collected but the interpretation and use it is put to.

Part II

PEN’s Approach to DQ
30 + researchers, scientists and partners to answer a big question about the importance of forests to the livelihoods of millions of poor people worldwide.
PEN in the Workflow

Research Question

Survey Instrument → Survey → Checking → Coding

Data Analysis

Data Management → Data Cleaning → Data Entry

H.E

T.E

Ronnie Babigumira

Data Quality Step by Step
PEN in the Workflow

- Research Question ✓
- Survey instrument ✓
- Survey ✓
  - 364 villages
  - 9,100 households
- Visual Inspection ✓
- Coding ✓
  - All responses have been coded
  - Coding centralized which has provided an unified understanding of products and no duplication
A monster is born
Desafío

Stock
- 294,150 questionnaire pages
- 700,000 + tables.
- Heterogeneity in data management skills of partners

How do we
1. Accurately capture the raw data from the questionnaire’s
2. Store it in an organized fashion
3. Prepare and make it ready for individual and global analysis
4. Analyze it

Ronnie Babigumira
Data Quality Step by Step
Excel was never an option. Why?
- Individual creativity a nightmare for the collective
- More serious excel issues such as mixed variable types, propagation of blanks, checks, and dangerous proximity to data.

A discussion on what to use, in the end, MS Access was chosen
The Database

We have designed and implemented MS Access database modules for data entry.

What

- Each survey is a database
- Within each database, “each” page of the questionnaire is a table

Why

- Uniformity and compatibility among users i.e. variable names, codes etc
- Intuitive and user friendly interface
- Data quality controls
- Flexible

The database development team is Arild Betty and Ronnie.
A Modular approach

Local Access Databases from all PEN Partners

1. Country  V1  V2  A1  A1  Quarterly
Intuitive and user friendly GUI

### Questionnaire

**Village Survey 1 (V1)**

Notes: See the Technical Guidelines for the appropriate source of information and respondents for the various questions in the village survey.

<table>
<thead>
<tr>
<th>Task</th>
<th>Date(s)</th>
<th>By who?</th>
<th>Status OK? If not, give comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meeting with officials</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Village focus group meetings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other interviews</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Checking questionnaire</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coding questionnaire</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entering data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Checking &amp; approving data entry</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Database Form

**A. Geographic and climate variables**

1. What is the name of the village? 

2. (name) (village code)

3. What is the latitude of the village? 

   degree

4. What is the longitude of the village? 

   degree

5. What is the altitude (m) of the village? 

   meter

6. What has been the average annual rainfall (mm/year) in the district during the past 20 years (or less, see guidelines)? 

   mm/year

7. What is the coefficient of variation in rainfall for the past 20 years? 

   (Note: To be filled in if data are readily available.)
C. Land

1. Please indicate the amount of land (in hectares) that you currently own and have rented in/out.
   Note: See definitions of land categories in the Technical Guidelines.

<table>
<thead>
<tr>
<th>Category</th>
<th>1. Area (ha)</th>
<th>2. Ownership (code-tenure)</th>
<th>Main products grown/harvested in the past 12 months Max 3 (code-product)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forest</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Natural forest</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Managed forests</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Plantations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agricultural land:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Cropland</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Pasture (natural or planted)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Agroforestry</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Silviculture</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Fallow</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Other vegetation types/land uses (residential, bush, grassland, wetland, etc.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Total land owned (1+2+3+...+9)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Land rented out (included in 1-9)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Land rented in (not included in 1-9)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Eye Candy
Controls

Deceptively Simple
Controls

Bells and Whistles
## B. Household composition

1. Who are the members of the household?

   Note: Recall the definition of households in the Technical Guidelines.

<table>
<thead>
<tr>
<th>1. Personal Identification number (PID)</th>
<th>2. Relation to household head*</th>
<th>3. Year born (YYYY)</th>
<th>4. Sex (0=male; 1=female)</th>
<th>5. Education (number of years completed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Household head</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1) Codes: 1 = spouse; 2 = son/daughter; 3 = son/daughter in law; 4 = grandchild; 5 = mother/father; 6 = mother/father in law; 7 = brother or sister; 8 = brother/sister in law; 9 = uncles/uncles; 10 = nephew/niece; 11 = step/foster child; 12 = other family; 13 = not
### Flexible

#### Common Solutions

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>houshd</td>
<td>hhid</td>
<td>pid1</td>
<td>name1</td>
<td>rtohd1</td>
<td>yeabn1</td>
<td>sex1</td>
<td>ed1</td>
<td>pid2</td>
<td>name2</td>
<td>rtohd2</td>
<td>yeabn2</td>
<td>sex2</td>
<td>ed2</td>
<td>pid3</td>
<td>name3</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>John Doe</td>
<td>1</td>
<td>1</td>
<td>John</td>
<td>1</td>
<td>1900</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>Jane</td>
<td>2</td>
<td>1910</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>James Last</td>
<td>2</td>
<td>1</td>
<td>James</td>
<td>1</td>
<td>1970</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>Jane</td>
<td>4</td>
<td>1900</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Jane Fonda</td>
<td>3</td>
<td>1</td>
<td>Jane</td>
<td>1</td>
<td>1934</td>
<td>2</td>
<td>20</td>
<td>2</td>
<td>Jack</td>
<td>4</td>
<td>1970</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Jacob Zuma</td>
<td>4</td>
<td>1</td>
<td>Jacob</td>
<td>1</td>
<td>1973</td>
<td>1</td>
<td>-9</td>
<td>3</td>
<td>Jenny</td>
<td>5</td>
<td>1980</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Access.. Flexible format

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>household</td>
<td>housecode</td>
<td>hhc_pid</td>
<td>hhc_name</td>
<td>hhc_rela</td>
<td>hhc_yrborn</td>
<td>hhc_sex</td>
</tr>
<tr>
<td>2</td>
<td>Steve Moody</td>
<td>31</td>
<td>1</td>
<td>Steve Moody</td>
<td>0</td>
<td>1965</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Steve Moody</td>
<td>31</td>
<td>2</td>
<td>Marisa Puentes</td>
<td>1</td>
<td>1968</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Steve Moody</td>
<td>31</td>
<td>3</td>
<td>Jonathon Mayorga</td>
<td>3</td>
<td>1966</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>Simeon Ortiz</td>
<td>26</td>
<td>1</td>
<td>Simeon Ortiz</td>
<td>0</td>
<td>1962</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>Simeon Ortiz</td>
<td>26</td>
<td>2</td>
<td>Louise</td>
<td>1</td>
<td>1960</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>Simeon Ortiz</td>
<td>26</td>
<td>3</td>
<td>Alfred Ortiz</td>
<td>2</td>
<td>1983</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>Simeon Ortiz</td>
<td>26</td>
<td>4</td>
<td>Ahi Ortiz</td>
<td>2</td>
<td>1985</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>Simeon Ortiz</td>
<td>26</td>
<td>5</td>
<td>Simeon Ortiz, Jr</td>
<td>2</td>
<td>1989</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>Simeon Ortiz</td>
<td>26</td>
<td>6</td>
<td>Alan Ortiz</td>
<td>2</td>
<td>1991</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>Peter Lee</td>
<td>32</td>
<td>1</td>
<td>Peter Lee</td>
<td>0</td>
<td>1943</td>
<td>0</td>
</tr>
</tbody>
</table>
The last work on Flexibility

We focus on ease of entry and deal with data management later

<table>
<thead>
<tr>
<th>Wide Format</th>
<th>Long Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>hhid</td>
<td>village</td>
</tr>
<tr>
<td>1001</td>
<td>Moss</td>
</tr>
<tr>
<td>1002</td>
<td>Ås</td>
</tr>
<tr>
<td>1003</td>
<td>Drøbak</td>
</tr>
</tbody>
</table>

Ronnie Babigumira
Data Quality Step by Step
Data Cleaning

Example of data cleaning in Stata

* *-----------------------------------------------------------------
* CIFOR PEN Poverty Environment Network
* A demo do-file to clean v2
* By: Ronnie Babigumira
* *-----------------------------------------------------------------

cd e:\pen\cleaning\miriam\s_data\v2
use v2_a_geo, clear

/* Some simple checks
1. Year is between 2005 and 2007,
2. Month is between 1 and 12
3. Day is between 1 and 31
*/

list village villcode intyear if intyear < 2005 | (intyear > 2007 & intyear !=.)
list village villcode intmon if intmon < 1 | (intmon > 12 & intmon !=.)
list village villcode intday if intday < 1 | (intday > 31 & intday !=.)
PENs data submission Workflow

1. Partner submits access data
2. Data transferred to Stata
3. Cleaning program run on data. Report of bugs sent back to partners
4. Partner addresses bugs, re-submits data.
5. Steps 1 to 4 repeated till all bugs are addressed or can be explained. Final report on data compiled.
6. Data sent to master database for archiving
7. Programs written to access master data set and create flat files for data analysis
Local Access Databases from all PEN Partners

Quarterly

Ronnie Babigumira
Data Quality Step by Step
Putting it together

Local Access Databases from all PEN Partners

1. Data Cleaning

<table>
<thead>
<tr>
<th>Country</th>
<th>V1</th>
<th>V2</th>
<th>A1</th>
<th>A2</th>
<th>Quarterly</th>
</tr>
</thead>
</table>
| Local Access Databases from all PEN Partners

2. Ronnie Babigumira

Data Quality Step by Step
Putting it together

1. Local Access Databases from all PEN Partners
2. Data Cleaning
3. Clean Data (Access and Stata)
## Putting it together

### 1. Local Access Databases from all PEN Partners

<table>
<thead>
<tr>
<th>Country</th>
<th>V1</th>
<th>V2</th>
<th>A1</th>
<th>A2</th>
<th>Quarterly</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 2. Data Cleaning

**Clean Data (Access and Stata)**

<table>
<thead>
<tr>
<th>Country</th>
<th>V1</th>
<th>V2</th>
<th>A1</th>
<th>A2</th>
<th>Quarterly</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 3. Aggregation of local datasets

**Global dataset**

<table>
<thead>
<tr>
<th>Country</th>
<th>V1</th>
<th>V2</th>
<th>A1</th>
<th>A2</th>
<th>Quarterly</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 4. Data Quality Step by Step

- **Clean Data (Access and Stata)**
- **Global dataset**
- **Local Access Databases from all PEN Partners**
Putting it together

1. Local Access Databases from all PEN Partners
2. Data Cleaning
3. Clean Data (Access and Stata)
4. Global dataset Aggregation of local datasets
5. Sub-datasets for analysis

Ronnie Babigumira  Data Quality Step by Step
Conclusion

What is a good data set

- Accurate
- Complete
- Clean
- Documented
- Easy to use.

But

- ... data preparation may occupy 90% of a project time line and is a major source of delay. Mikhail Golovnya, Salford Systems 2007.
- A number of Steps, a number of opportunities to mess up. Its all in your hands.
Minds think with ideas, not information. No amount of data, bandwidth, or processing power can substitute for inspired thought. **Clifford Stoll**