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



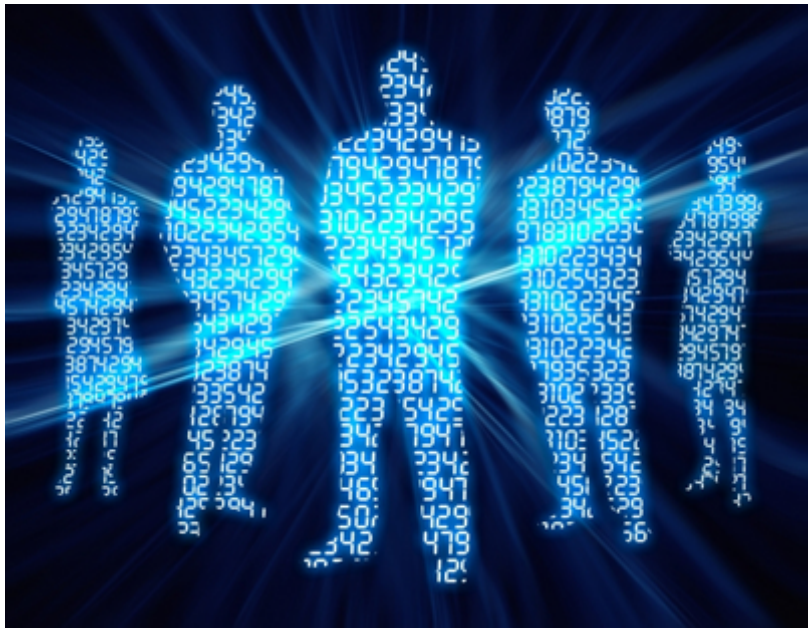
The big questions

. . . 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

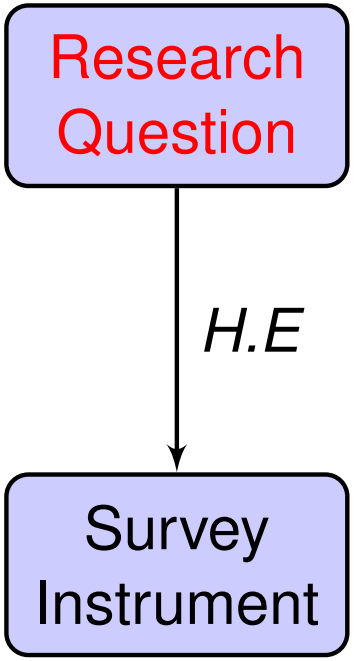
Research Question

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

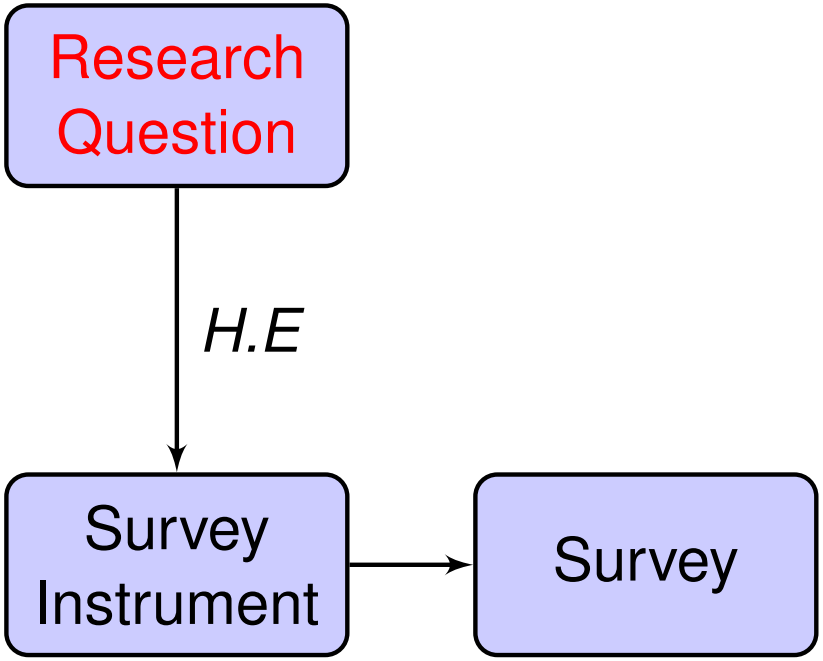


Survey Instrument

> *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

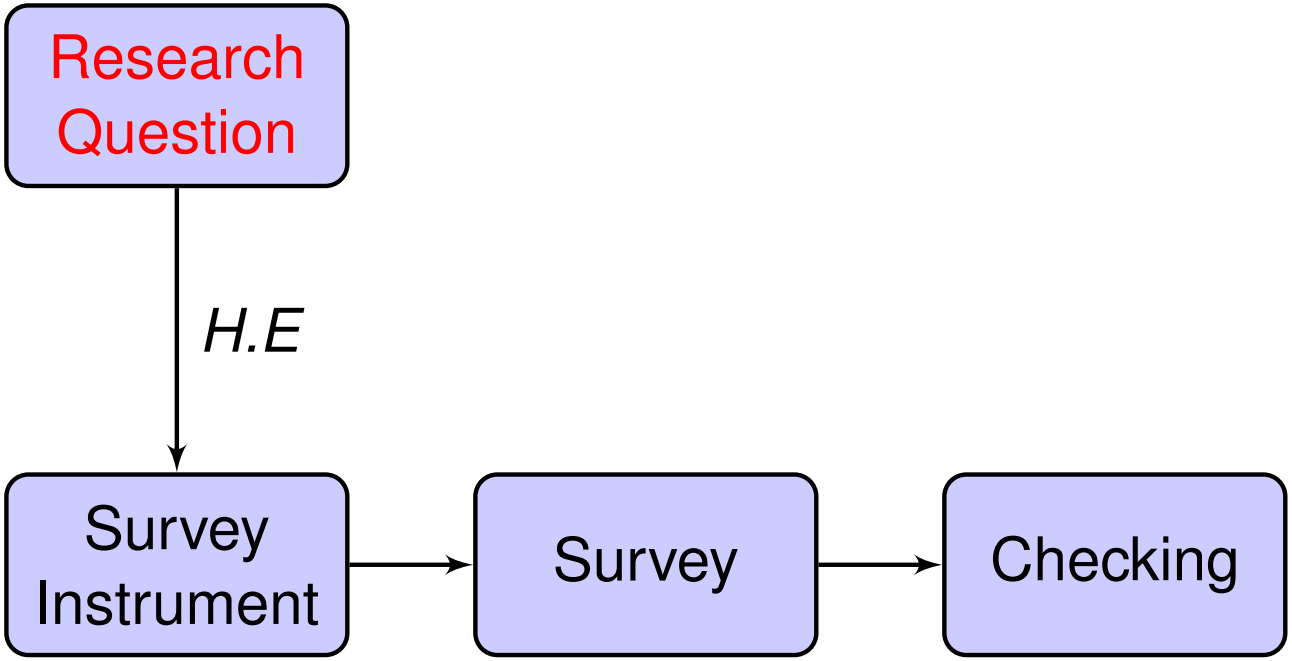


The Survey

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

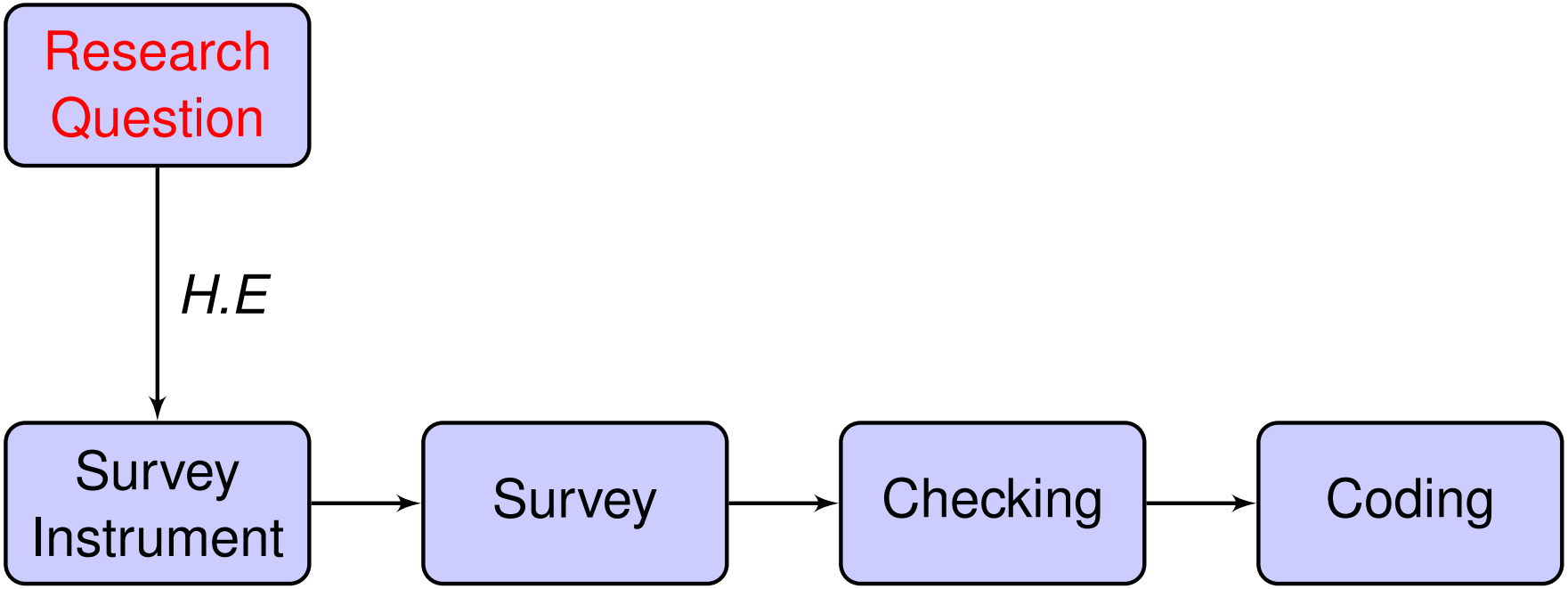


Checking

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

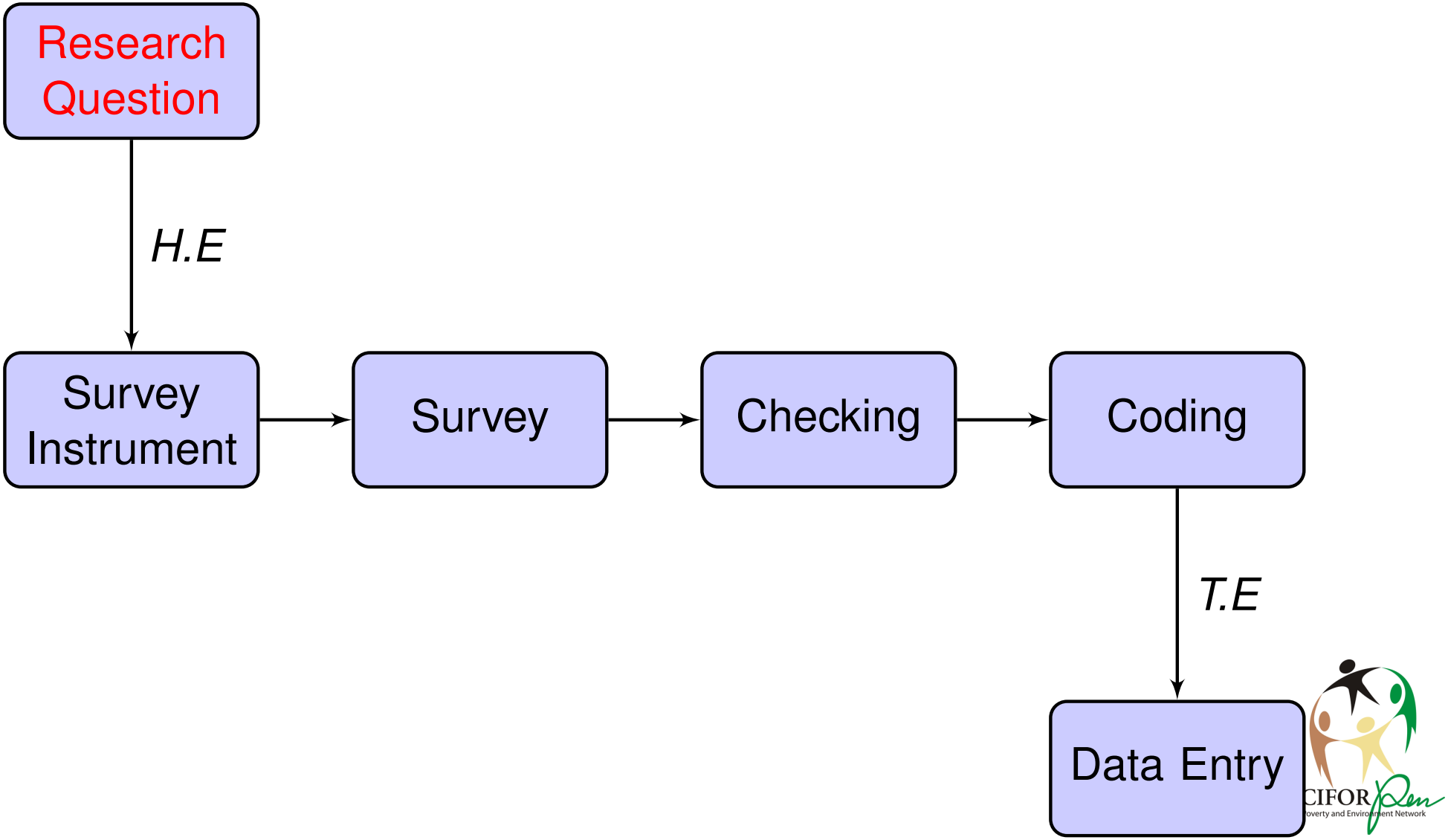
Good Quality Data Workflow



Coding

- Is a 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



Data Entry

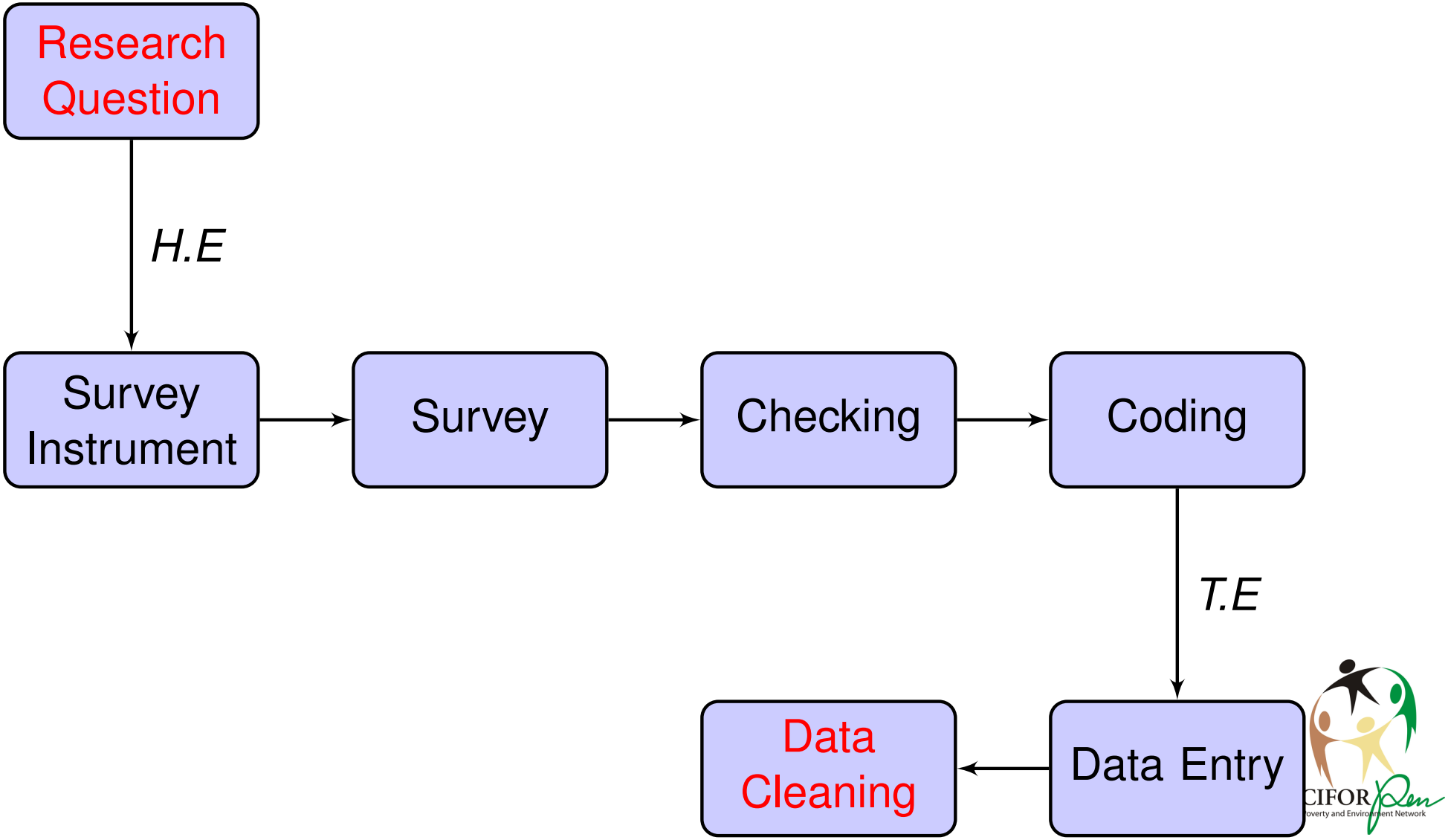
Good practice

- 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



Data Cleaning

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))

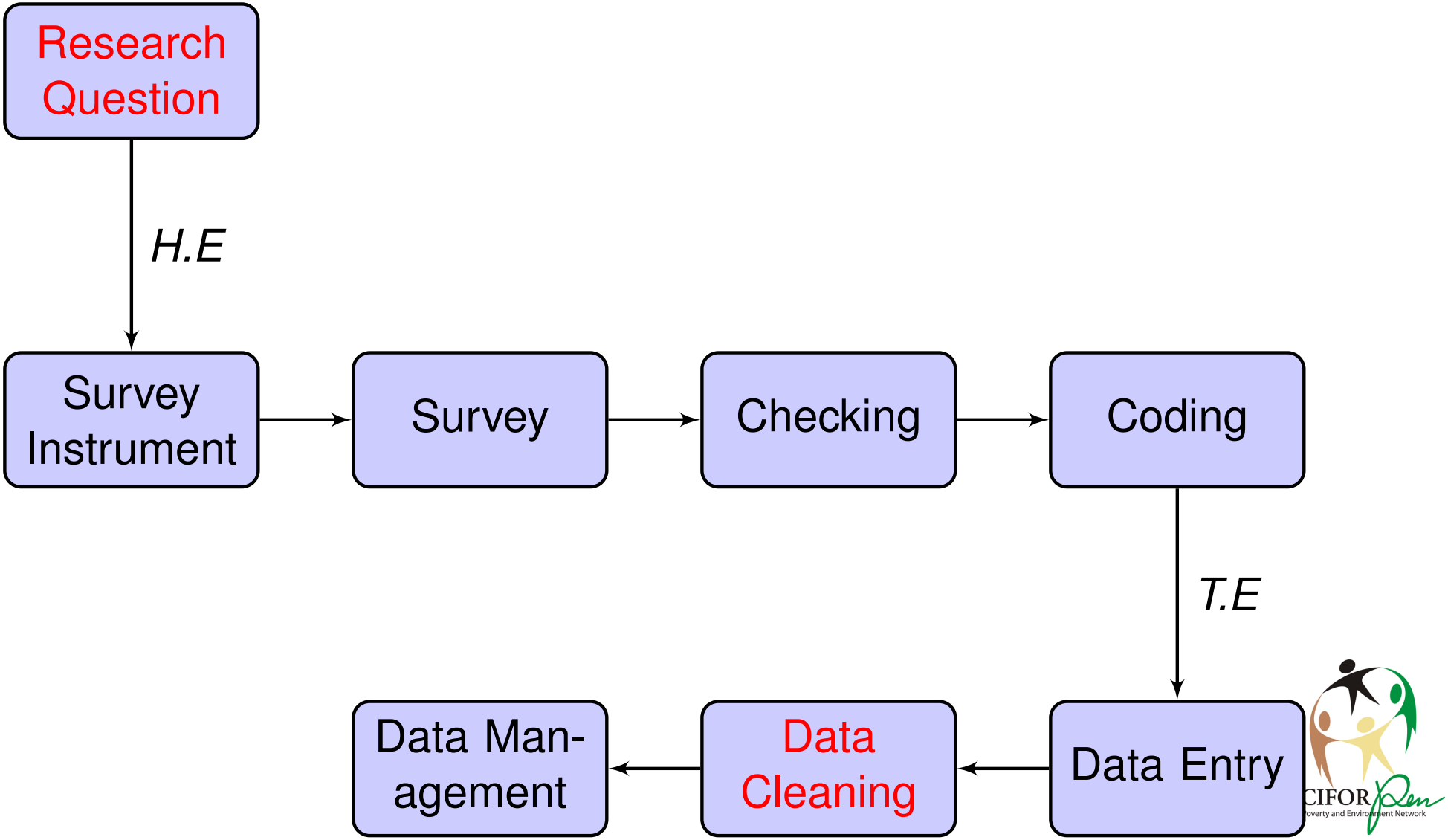
What is it

- 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

Good Quality Data Workflow



Data Management

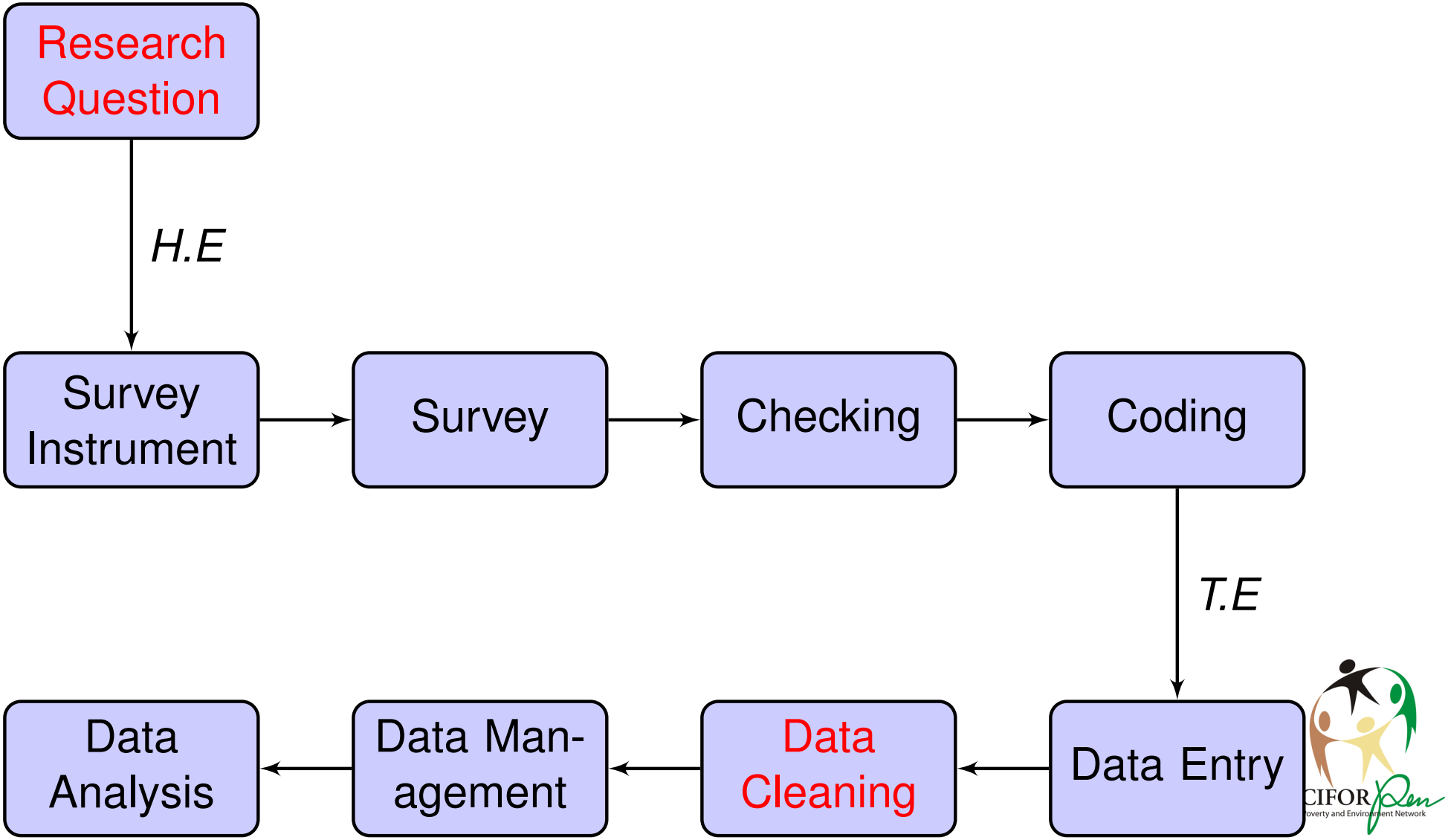
What is it

- 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.

Good Quality Data Workflow



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.

Columb, M.O, P Haji-Michael, and P Nightingale. 2003. Data collection in the emergency setting. Emergency Medicine Journal 20 (5):459-463.

Part II

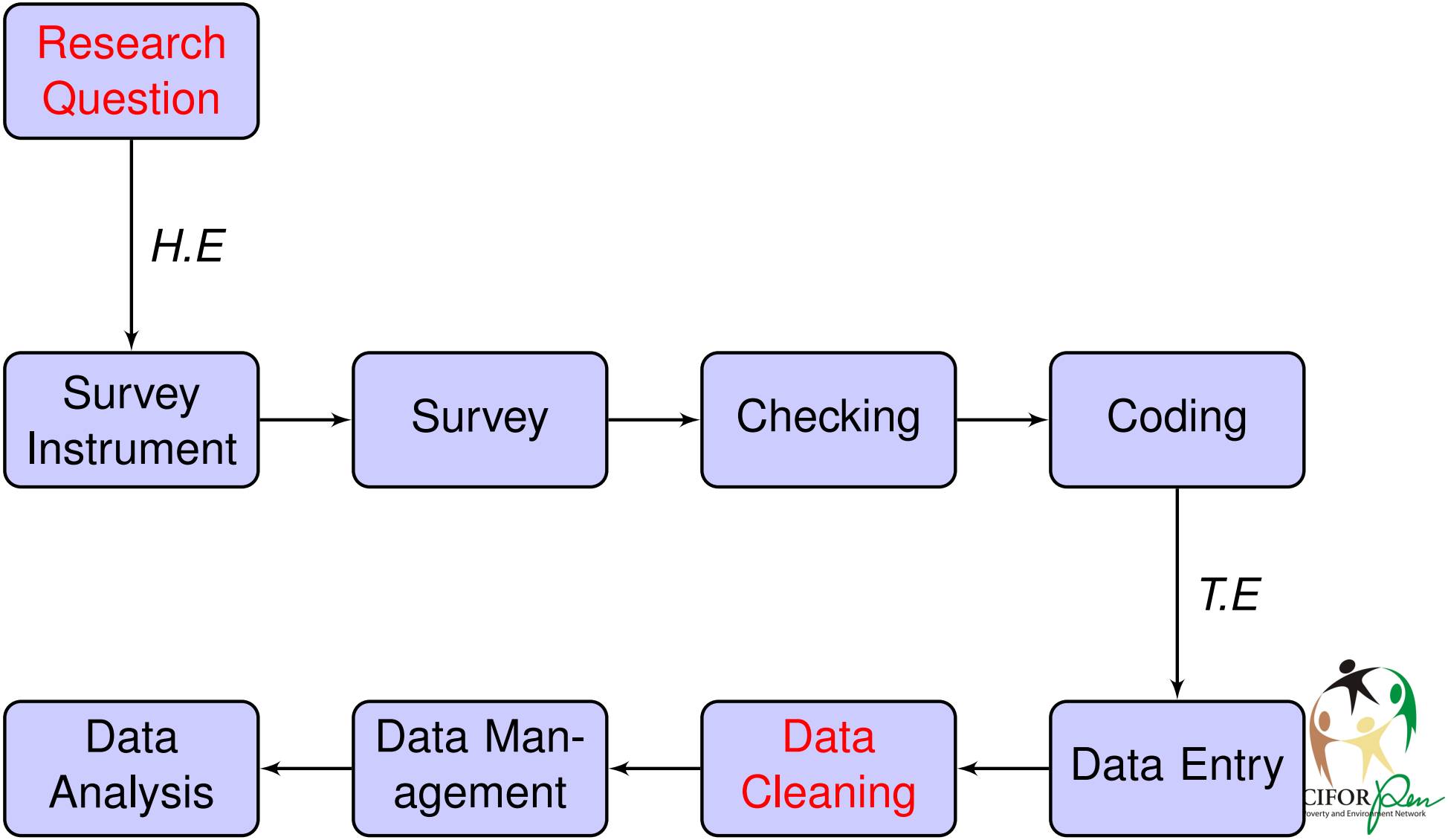
PEN's Approach to DQ



PEN in a Second

30 + researchers, scientists and partners to answer a big question about the the importance of forests to the livelihoods of millions of poor people world wide.

PEN in the Workflow



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

Data entry

- 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 [1] have designed and implemented MS Access database modules for data entry

What

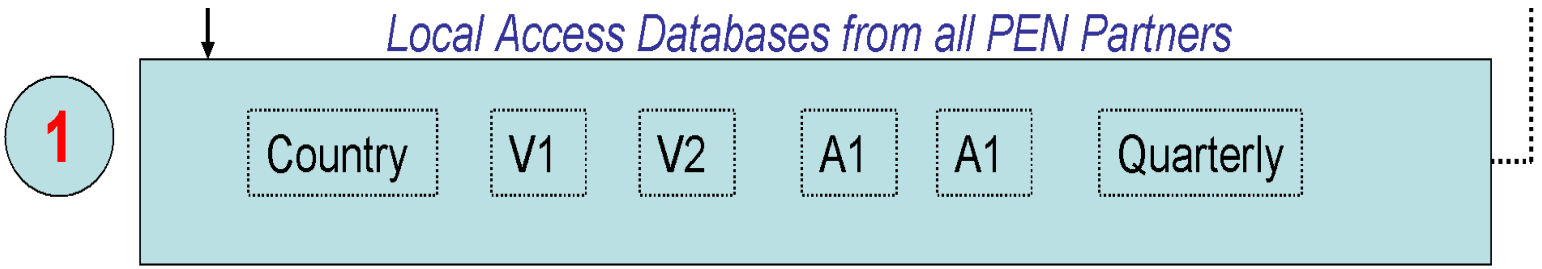
- Each survey is a database
- Within each database, “each” page of the questionnaire is a table

Why

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

[1] The database development team is Arild Betty and Ronnie

A Modular approach



Intuitive and user friendly GUI

Questionnaire

Village Survey 1 (V1)

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

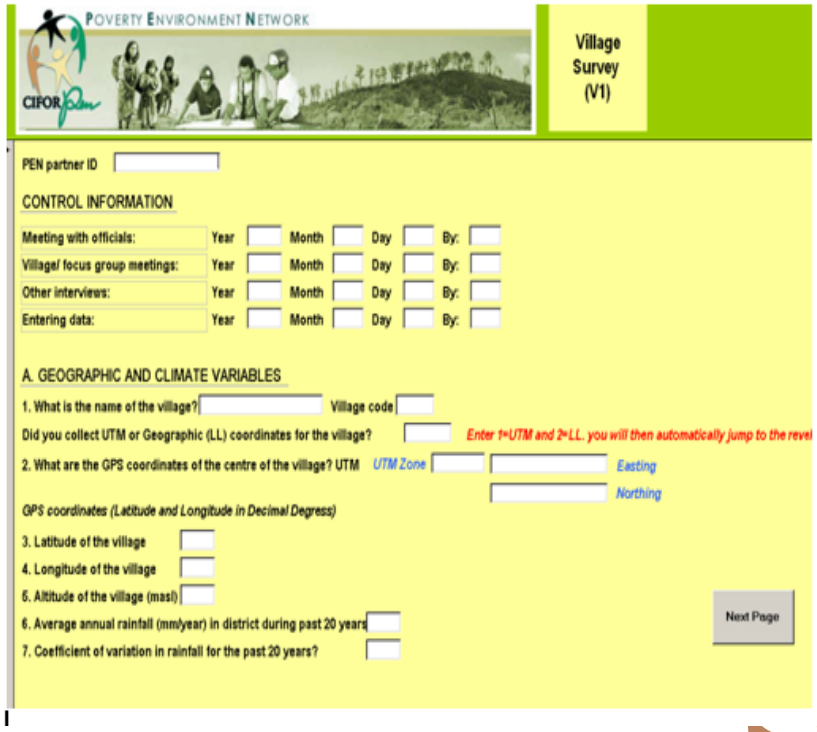
Control information

Task	Date(s)	By who?	Status OK? If not, give comments
Meeting with officials			
Village/focus group meetings			
Other interviews			
Checking questionnaire			
Coding questionnaire			
Entering data			
Checking & approving data entry			

A. Geographic and climate variables

1.	What is the name of the village?	*1. (name)*	*2. (village code)*
2.	What are the GPS coordinates of the centre of the village? (UTM format)		
3.	What is the latitude of the village?		*degrees*
4.	What is the longitude of the village?		*degrees*
5.	What is the altitude (masl) of the village?		*masl*
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.)*		

Database Form



CIFOR Poverty and Environment Network

Controls

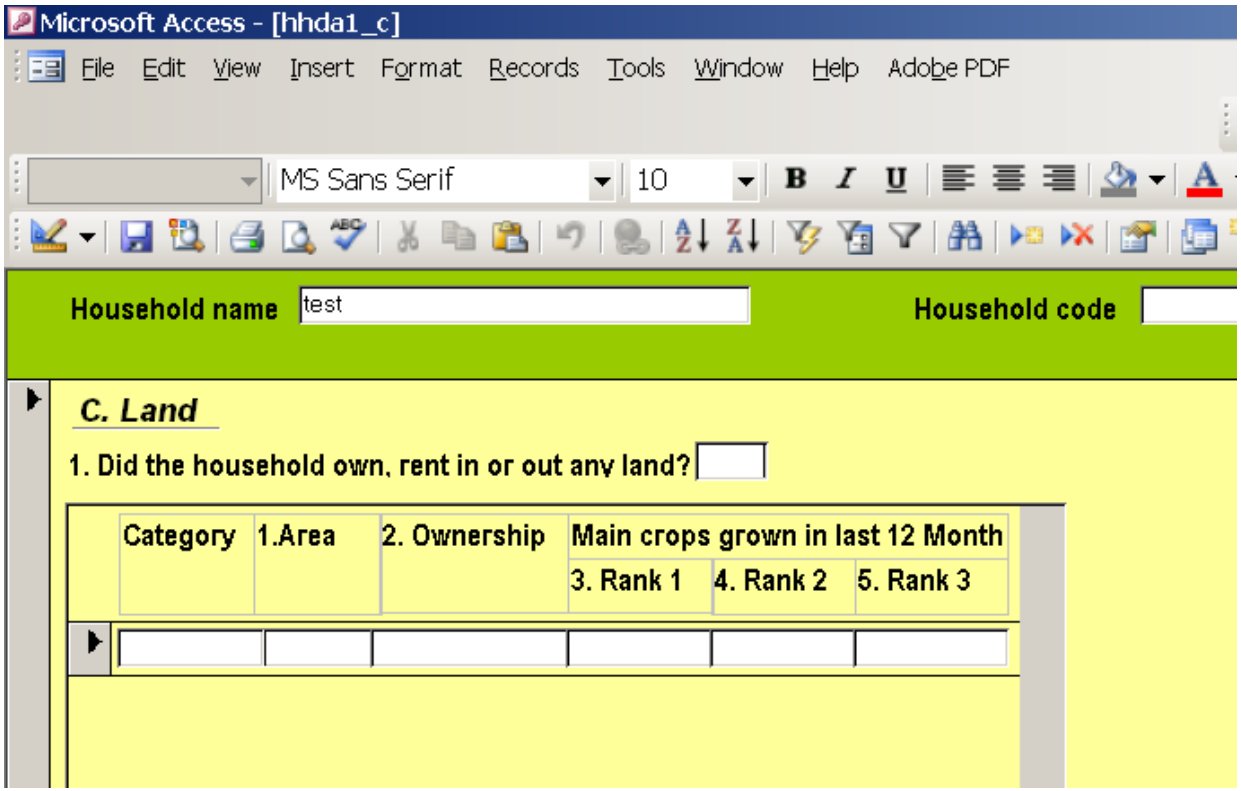
C. Land

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.

Category	**1. Area** *(ha)*	**2. Ownership** *(code-tenure)*	**Main products grown/harvested in the past 12 months** Max 3 *(code-product)*		
			3. Rank1	**4. Rank2**	**5. Rank3**
Forest:					
1. Natural forest					
2. Managed forests					
3. Plantations					
Agricultural land:					
4. Cropland					
5. Pasture (natural or planted)					
6. Agroforestry					
7. Silvipasture					
8. Fallow					
9. Other vegetation types/land uses (residential, bush, grassland, wetland, etc.)					
10. Total land owned (1+2+3+…+9)					
11. Land rented out (**included** in 1-9)					
12. Land rented in (**not included** in 1-9)					

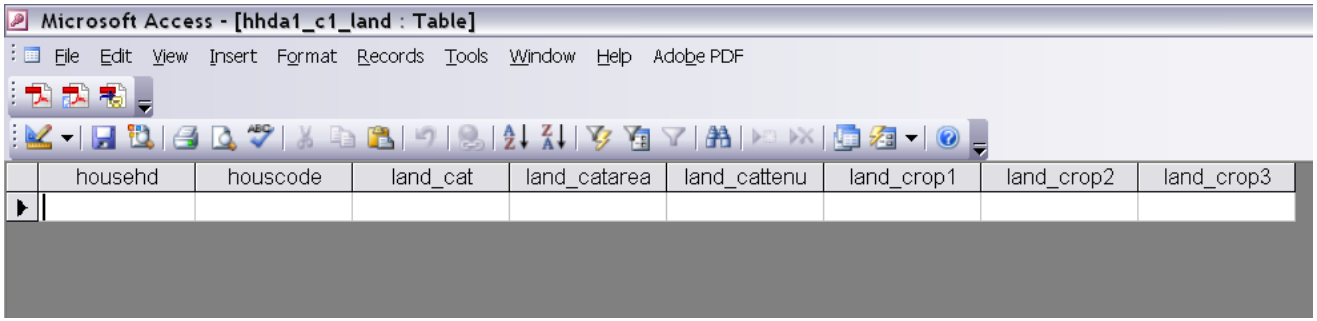
Questionnaire

Controls



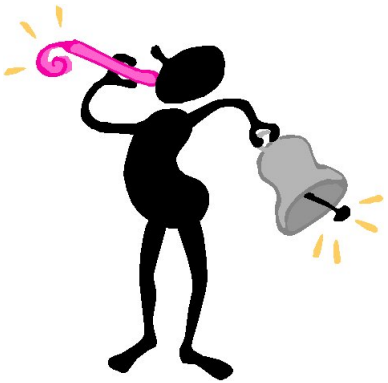
Eye Candy

Controls

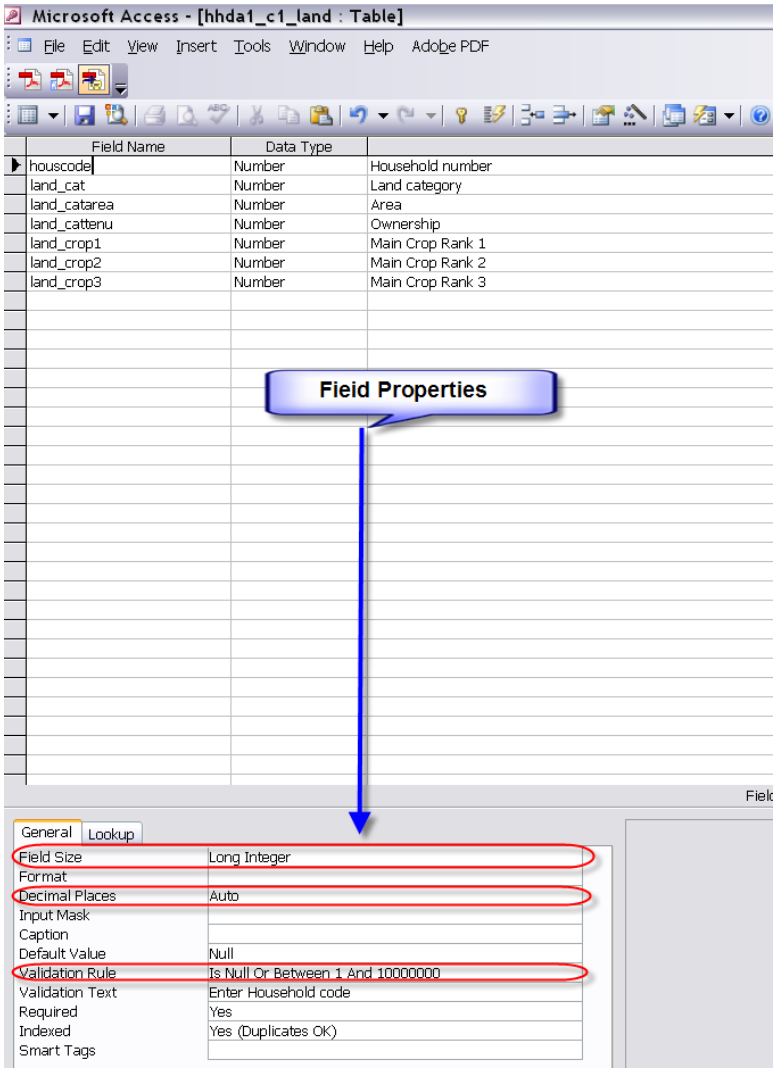


Deceptively Simple

Controls



Bells and Whistles



Flexible

B. Household composition

1. Who are the members of the household?

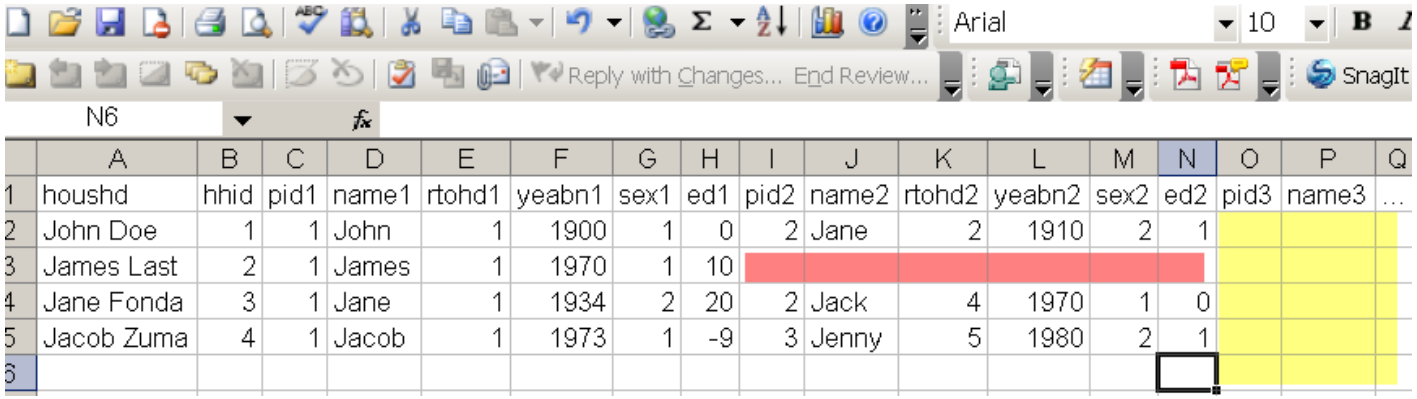
Note: Recall the definition of households in the Technical Guidelines.

1. Personal Identification number (PID)	* Name of household member	2. Relation to household head[1)]	3. Year born[2)] *(yyyy)*	4. Sex *(0=male 1=female)*	5. Education *(number of years completed)*
1		Household head			
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					

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=uncle/aunt; 10=nephew/niece; 11=step/foster child; 12=other family; 13=not

Household Data

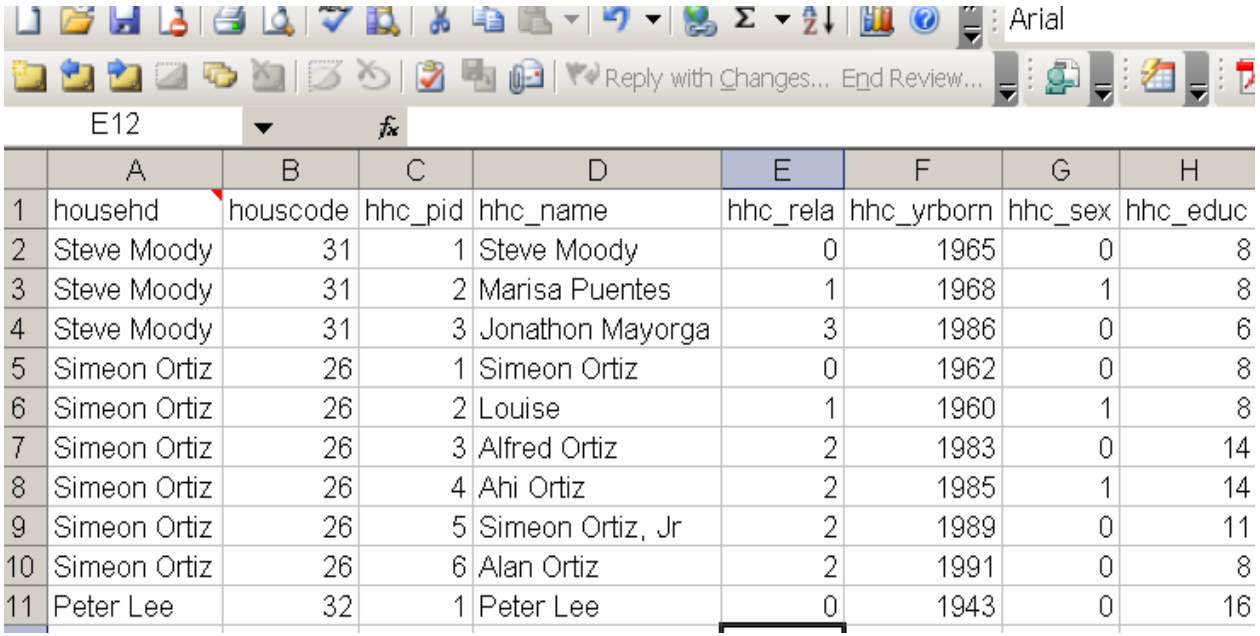
Flexible



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	houshd	hhid	pid1	name1	rtohd1	yeabn1	sex1	ed1	pid2	name2	rtohd2	yeabn2	sex2	ed2	pid3	name3	…
2	John Doe	1	1	John	1	1900	1	0	2	Jane	2	1910	2	1			
3	James Last	2	1	James	1	1970	1	10									
4	Jane Fonda	3	1	Jane	1	1934	2	20	2	Jack	4	1970	1	0			
5	Jacob Zuma	4	1	Jacob	1	1973	1	-9	3	Jenny	5	1980	2	1			
6																	

Common Solutions

Flexible



E12

	A	B	C	D	E	F	G	H
1	househd	houscode	hhc_pid	hhc_name	hhc_rela	hhc_yrborn	hhc_sex	hhc_educ
2	Steve Moody	31	1	Steve Moody	0	1965	0	8
3	Steve Moody	31	2	Marisa Puentes	1	1968	1	8
4	Steve Moody	31	3	Jonathon Mayorga	3	1986	0	6
5	Simeon Ortiz	26	1	Simeon Ortiz	0	1962	0	8
6	Simeon Ortiz	26	2	Louise	1	1960	1	8
7	Simeon Ortiz	26	3	Alfred Ortiz	2	1983	0	14
8	Simeon Ortiz	26	4	Ahi Ortiz	2	1985	1	14
9	Simeon Ortiz	26	5	Simeon Ortiz, Jr	2	1989	0	11
10	Simeon Ortiz	26	6	Alan Ortiz	2	1991	0	8
11	Peter Lee	32	1	Peter Lee	0	1943	0	16

Access.. Flexible format

The last work on Flexibility

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

Wide Format			
hhid	village	district	d_mkt
1001	Moss	Akershus	15
1002	Ås	Akershus	1
1003	Drøbak	Akershus	42

Long Format		
hhid	pid	sex
1001	1	1
1001	2	1
1001	3	0
1002	1	2
1002	2	1
1002	3	1
1002	3	2
1003	1	1
1003	1	2

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

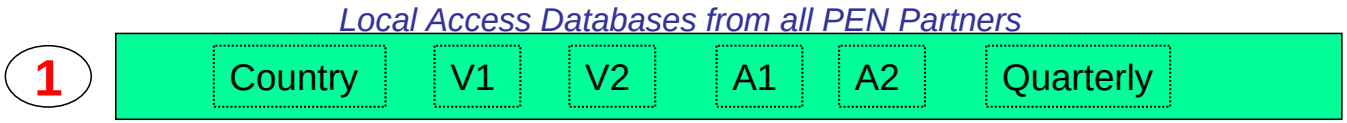

/* Some simple checks
1. Year is between 2005 and 2007,
2. Month is between 1 and 12
3. Day is between 1 and 31
*/

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 !=.)
```

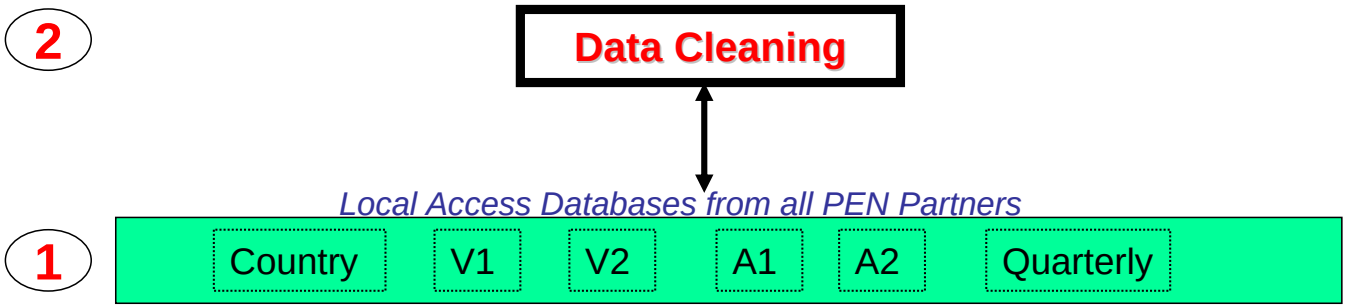
PENs data submission Workflow

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

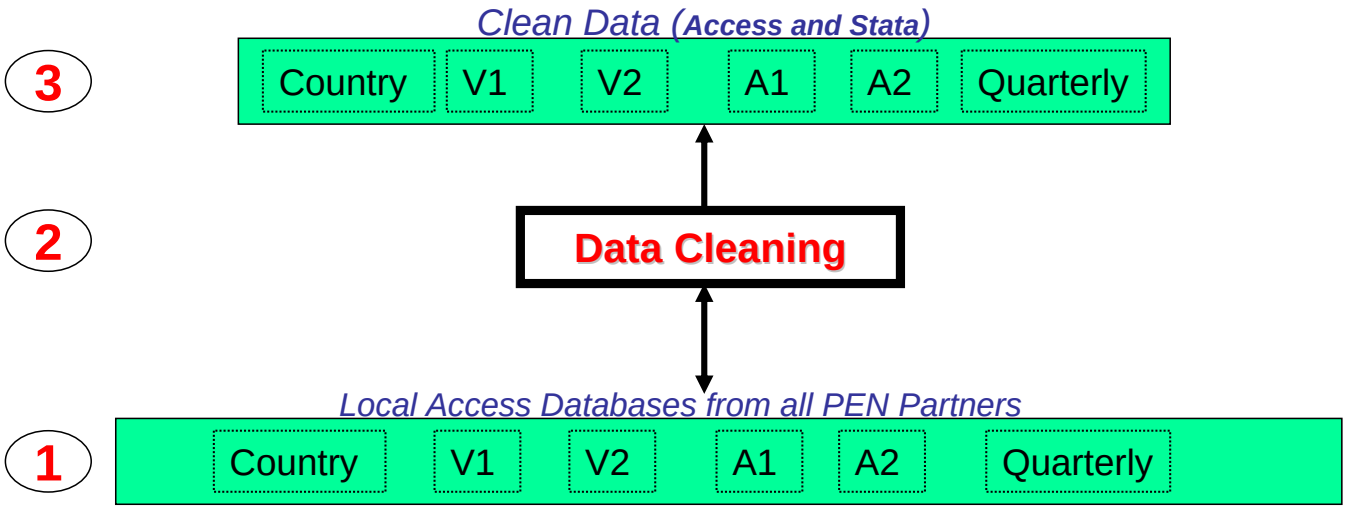
Putting it together



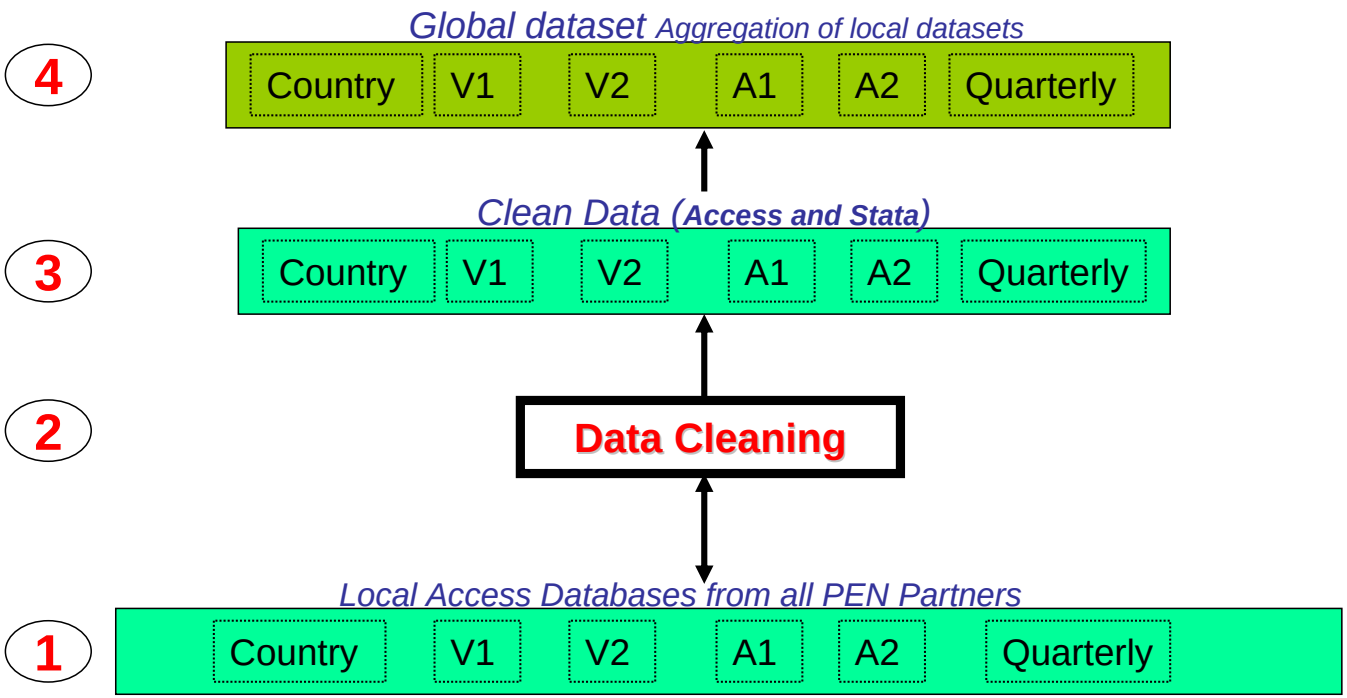
Putting it together



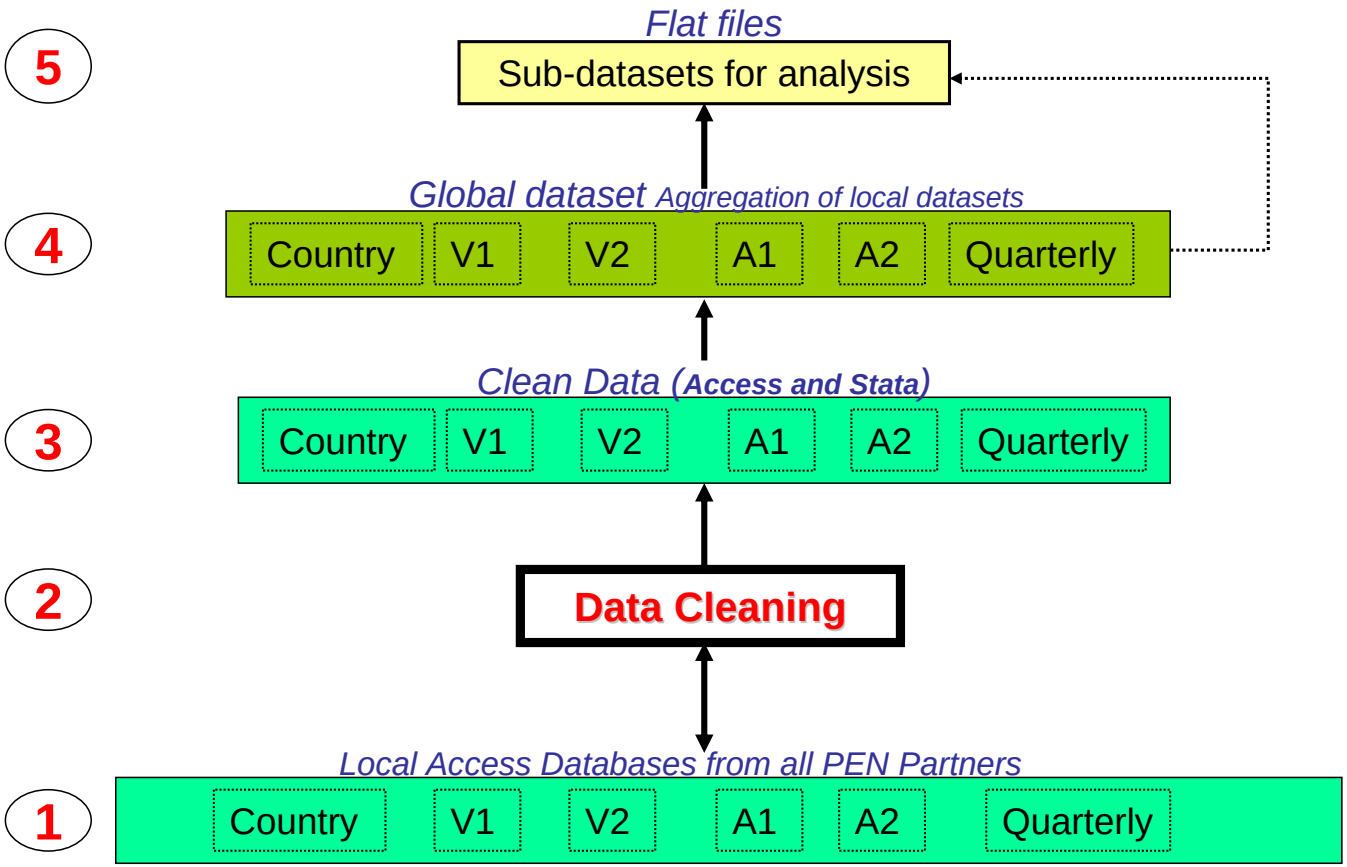
Putting it together



Putting it together



Putting it together



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***