Microdata confidentiality

Agenda

1 Bangladesh Bureau of Statistics

2 Disclosure control

3 Way forward

19.08.2018

Kabir Uddin Ahmed
Director, Industry and Labour Wing, BBS

Agenda

- Bangladesh Bureau of Statistics
- Microdata
- Disclosure control
- Proposal for microdata dissemination
- Way forward

Microdata confidentiality

Agenda

1 Bangladesh Bureau of Statistics

2 Disclosure control

3 Way forward

19.08.2018

Kabir Uddin Ahmed
Director, Industry and Labour Wing, BBS

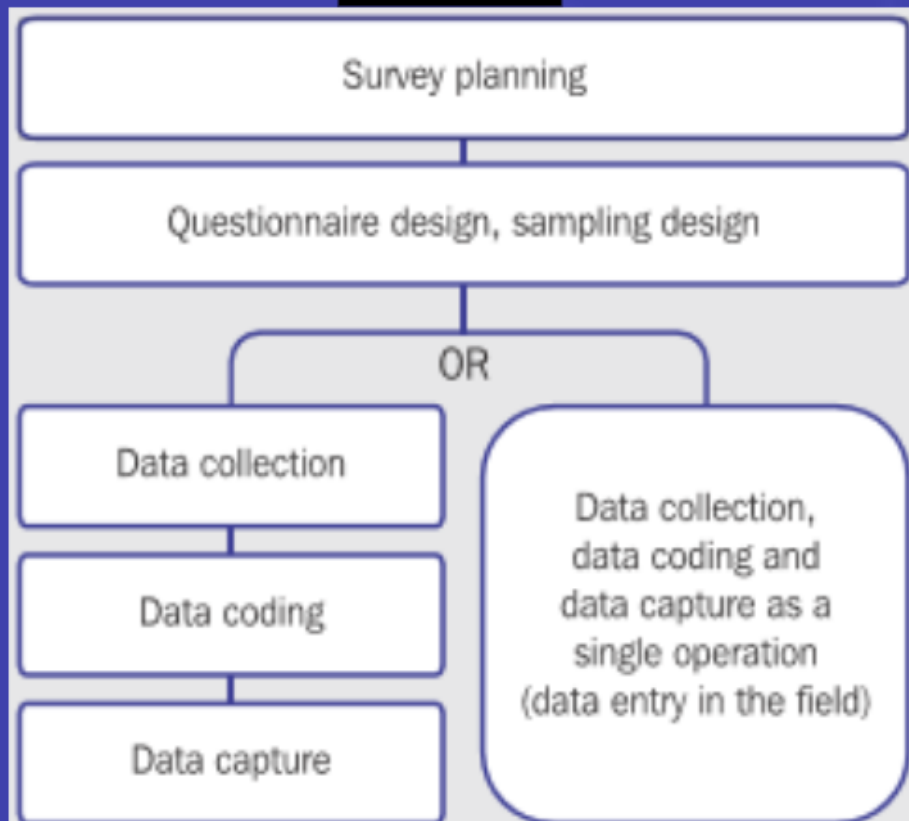1 Bangladesh Bureau of Statistics

Functions of BBS

Protocol

Data dissemination

The BBS

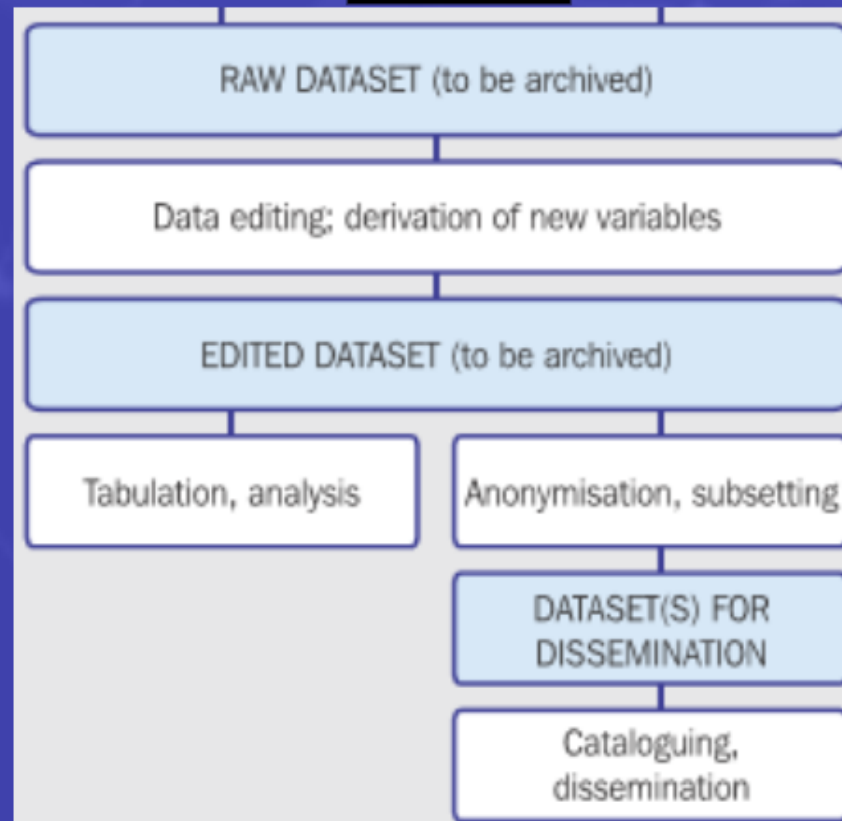- BBS, the NSO of Bangladesh, is to collect and disseminate data.
- Decades ago, this meant producing books and reports primarily consisting of tabular data – designed to answer pre-defined questions.
- The increasing complexity of 21st century society, however, has put increasing pressure on BBS to produce microdata – designed to allow policy analysts and researchers to pose and answer questions of their own choosing.

Typical Data production framework

1st part

- Survey planning
- Questionnaire design, sampling design

OR

- Data collection
- Data coding
- Data capture

OR

- Data collection, data coding and data capture as a single operation (data entry in the field)

2nd part

- RAW DATASET (to be archived)
- Data editing; derivation of new variables
- EDITED DATASET (to be archived)
- Tabulation, analysis
- Anonymisation, subsetting
- DATASET(S) FOR DISSEMINATION
- Cataloguing, dissemination

Data dissemination is the release to the users of information obtained through a statistical activity.

Forms of data dissemination

- Paper publication;
- Macrodata/aggregated data;
- Microdata;
- Presentation/interview;
- Press release;
- Web sites;
- Social media ...

Microdata confidentiality

1 Bangladesh Bureau of Statistics

2 Disclosure control

3 Way forward

Agenda

19.08.2018

Kabir Uddin Ahmed
Director, Industry and Labour Wing, BBS

2 Disclosure control

To protect statistical data in such a way that they can be released without giving confidential information that can be linked to specific individuals or entities.

UNFPOS

Microdata

Metadata

TIMELINE

TIMELINE

United Nations FPOS

Principle 6.

Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.

Microdata

- Microdata is the records at the level of individual respondents.
- **For instance**, a census might collect age, home address, educational level, employment status, and many other variables, recorded separately for every person who responds; this is microdata.

psu	spc	stratum	str_name	div	div_name	zila	zl_name	uz	uz_name	un	un_name	mz	mza_name	vill	vill_name	rmo
10	Rural	1	Barisal Rural	10	BARISAL	6	BARISAL	7	BAKERGANJ	33	DUDHAL	513	KABIRAJ	1	KABIRAJ	1

psu	hhold	member	s01a_q01	sex	rel	age
10	2	1	Mofajall gazi	1	1	54
10	2	2	paria begum	2	2	42
10	2	3	sadia	2	3	10
10	2	4	Aminul islam	1	3	14
10	2	5	Farjana	2	3	21
10	2	6	suraia	2	5	4
10	2	7	Taslima	2	4	23
10	2	8	sanjida	2	5	3

```
1 1 54 1 1 1   2 99 99 2 0 1 0 0   0 2 2 0 0 0 0 0 0 0 0 0 0 2 2 1 1   5 1 1 2 2   0 0     0 0     0 0
                                    2 2 0 0 0 0 0 0 0 0 0 0 2 2 1 1   5 1 1 2 2   0 0     0 0     0 0
                                    2 2 0 0 0 0 0 0 0 0 0 0 2 2 1 1   4 1 1 2 1   5 2     0 0     0 0
1 3 14 1 2 2   0 1   2 2 0 2 2 0   3 2 2 0 0 0 0 0 0 0 0 0 0 2 2 1 1   7 1 2 2 1   8 0     0 2     0 0
2 3 21 1 1 2                                                                               0 0     0 0     0 0
2 5  4 1 2 2                                                                               0 0     0 0     0 0
2 4 23 1 1 2                                                                               0 0     0 0     0 0
2 5  3 1 2 2                                                                               0 0     0 0     0 0
1 1 33 1 1 1                                                                               0 0     0 0     0 0
2 2 26 1 1 2                                                                               0 0     0 0     0 0
1 3  5 1 2 2                                                                               0 0     0 0     0 0
1 1 65 1 1 2                                                                               0 0     0 0     0 0
2 2 52 1 1 1                                                                               0 0     0 0     0 0
1 3 24 1 1 1                                                                               0 0     0 0     0 0
2 4 21 1 1 2                                                                               0 0     0 0     0 0
2 5  4 1 2 2                                                                               0 0     0 0     0 0
1 1 52 1 1 1   2 99 99 2 0 1 0 0   0 2 2 0 0 0 0 0 0 0 0 0 0 2 2 1 1   9 1 1 2 2   0 0     0 0     0 0
2 2 43 1 1 2   1 99 99 2 0 2 2 0   1 2 2 0 0 0 0 0 0 0 0 0 5 2 1 1   8 1 1 2 2   0 0     0 0     0 0
1 3  9 1 2 2   0 1   2 2 0 2 2 0   3 2 2 0 0 0 0 0 0 0 0 0 0 2 2 1 1   1 1 1 2 1   2 1 600 0     0 0
```

Columns 1-3: Variable *Household ID Number*
Column 4: Variable *Area* (code 2 = 'rural')
Columns 5-6: Variable *Person ID*
Columns 7-8: Variable *Relationship to Head of Hhld*
Column 9: Variable *Sex* (1 = 'Male', 2 = 'Female')
Columns 10-11: Variable *Age* (age in years)

Record 1 (information on 1ˢᵗ person) →
Record 2 (information on 2ⁿᵈ person) →
Etc

```
12  1  114021
12  2  223921
12  3  321711
12  4  321311
12  5  32 5
12  6  31 1
22  1  124711
22  2  321611
22  3  814311
22  4  629933
```

Microdata - Stata format

hhid		area	pid	relat	sex	age
1	1	Rural	1	Head	Male	40
2	1	Rural	2	Spouse	Female	39
3	1	Rural	3	Son/Daughter	Female	17
4	1	Rural	4	Son/Daughter	Female	13
5	1	Rural	5	Son/Daughter	Female	5
6	1	Rural	6	Son/Daughter	Male	1
7	2	Rural	1	Head	Female	47
8	2	Rural	2	Son/Daughter	Female	16
9	2	Rural	3	Brother/sister	Male	43
10	2	Rural	4	Parent	Female	99
11	3	Rural	1	Head	Male	35
12	3	Rural	2	Spouse	Female	39
13	3	Rural	3	Son/Daughter	Male	1
14	3	Rural	4	Not related	Female	16
15	3	Rural	5	Not related	Female	4
16	3	Rural	6	Not related	Male	4

Metadata

- Metadata means "data about data";
- Metadata is defined as the data providing information about one or more aspects of the data;

  Components:
  - Questionnaire
  - Sample design
  - Coding
  - Weighting
  - Post-stratification

Metadata

- Metadata describes the microdata.
- Standard metadata formats/protocols should be used.
- Metadata gives meaning to the microdata, and without metadata the microdata is just a collection of numbers with no meaning and is just a load of junk.
- The wrong metadata with microdata can be a disaster.

Paradata

Paradata are measurements related to the process of collecting survey data. Paradata can provide a means of additional control over or understanding of the quality of the primary data (the responses to the survey questions).

- Using paradata to monitor fieldwork activity in face–to–face, telephone, and web surveys;
- Guiding intervention decisions during data collection;
- Analysis of measurement, nonresponse, and coverage error via paradata.

Methods of Statistical Disclosure Control (SDC) of microdata

1. Non-perturbing-masking methods reduce the amount of information released by suppressing or aggregating data.

2. Perturbing-masking methods edit and modify the data before publication by introducing an element of error purposely for confdentiality reasons.

1. Non-perturbing-masking techniques

- Sampling (5 %), instead of disseminating the entire microdata population census data;
- Re-coding involves aggregation of the observed survey values into pre-defnedclasses in such a way that individual responses are not visible;
- Local suppression is a basic technique used when two variables taken together could lead to identifying a unique person;
- Collapsing lower level geographic variables.

1. Non-perturbing-masking techniques

- Removing variables from a microdata file for information is regarded as too sensitive to be released, for example ethnicity or religion.
- Removing records is sometimes necessary to protect the anonymity of respondents with a unique set of variables. When a record is removed entirely from a microdata file, it is necessary to compute and include adjusted weighting factors.

2. Perturbing techniques:

- Additive noise
- Data swapping
- Rank-swapping
- Micro-aggregation
- Rounding
- Re-sampling
- Post-randomisation

Microdata confidentiality

1 Bangladesh Bureau of Statistics

2 Disclosure control

3 Way forward

Agenda

19.08.2018

Kabir Uddin Ahmed
Director, Industry and Labour Wing, BBS

Annonymize

Proposal

3 Proposal for microdata dissemination

DL

Way forward

Way forward

1. Annonymize microdata and disseminate

- Strictly follow the SDC and annonymize microdata
- Develop a protocol for confidentiality
- Identify proper persons to disseminate microdata.

2. Remote Access Facilities (RAFs) - a service window

These are facilities that provide researchers with the ability to produce statistical outputs from microdata through computer networks without researchers actually 'seeing' the microdata. The microdata itself does not leave the National Statistical Office. Remote Access Facilities may be of two types.

(a) Remote execution where a researcher submits a programme and receives the output later by email.
(b) Remote facilities where the researcher performs the analysis and can immediately see the answer on the screen.

3. Data Laboratories (DL): For researchers

- On-site access to more identifiable microdata, typically with NSO supervision.
- The access to more detailed data creates some inconvenience to the researcher, because of the requirement of working at the NSO.

Material that constitute metadata:

1. Explanatory material

- Information about data collection methods;
- Information about dataset structure;
- Variables and values, coding and classification schemes;
- Information about derived variables;
- Weighting and post-stratification;
- Confidentiality and anonymisation.

2. Contextual information

- Objectives of the activity;
- Geographical coverage;
- Problems encountered during data collection;
- data entry, and data checking and cleaning;
- Dataset's life- cycle viz. Panel, or Time-series;
- Additional information describing, for example, changes in the question text, variable labeling; and
- Sampling procedures.

To whom should microdata be made available?

- Policymakers and researchers employed by line-ministries and planning departments;
- International agencies and other sponsoring agencies;
- Research and academic institutes involved in social and economic research;
- Academic staff and students; and
- Other users involved in scientific research.

Way forward

- Regular work of Computer Wing, BBS;
- Alternatively, from the New project of Computer wing, BBS.

Microdata confidentiality

Agenda

1 Bangladesh Bureau of Statistics

2 Disclosure control

3 Way forward

19.08.2018

Kabir Uddin Ahmed
Director, Industry and Labour Wing, BBS