World Bank

Sharing Household Survey Data Practice & Tools

Matthew Welch, Senior Statistician, World Bank

Seminar on Poverty Measurement, 12-13 July 2016, Geneva

Outline

- Why disseminate, document microdata
- Documentation
- Dissemination
- Anonymization
- Conclusion
- Discussion

Why?
Disseminate

- Replicability, transparency
- Visibility
- Improved quality through user feedback
- Knowledge generation (if disseminate microdata) > increase and demonstrate the value of data > more funding
- Satisfy a legal\Donor requirement
- Participate in Open Data / Data Liberation movement

- Also some obstacles: privacy protection; exposure to criticism; others

Why?
Documentation, or metadata, helps the researcher to:

- Find the data they are interested in.
- Understand what the data are measuring and how the data have been created.
- Assess the quality of the data.
- Rich metadata **reduces the burden on the data producer** to answer queries

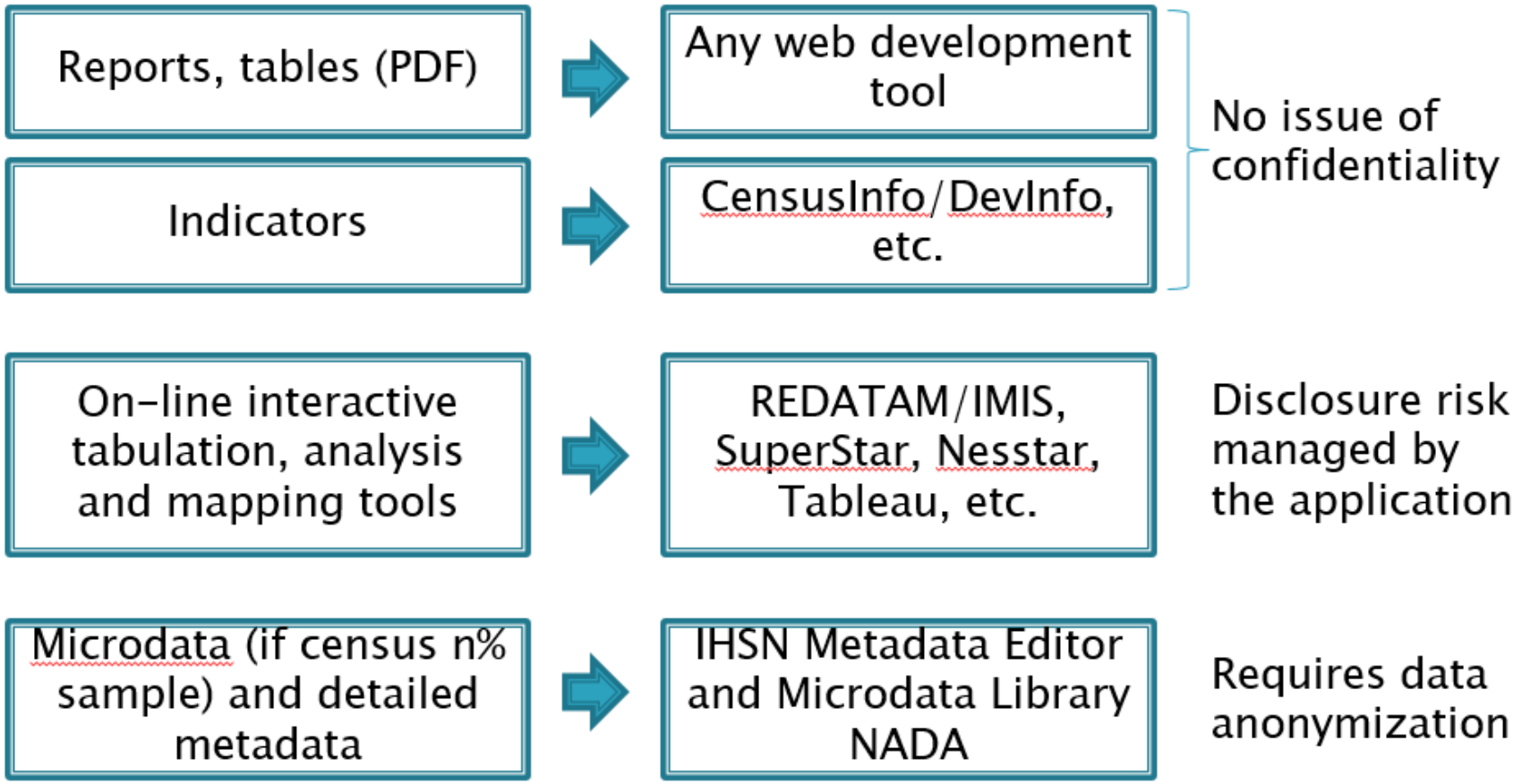
Documentation
Tools available:

- International metadata standards (in particular the DDI) and specialized software are available to help document and catalog microdata.
- Nesstar Metadata Editor
  - Freeware – full documentation
  - Multi-Language compliant
- Compliant with the DDI/DCMI (XML) standards (Data Documentation Initiative and Dublin Core)
- World Bank are developing a new multi platform open source editor

Documentation
International Standard: DDI

- Recommendation: use the DDI-Codebook metadata standard
  - Good for survey, census, administrative micro-datasets
  - Standard checklist of what you need to know about a survey and its dataset(s)
  - Documents the full survey life-cycle
- Free tools and support available (used in 75+ countries)
- Cost of documenting: a small fraction of the survey cost
- Required even for data not intended for dissemination (preserve institutional memory; disseminate metadata for transparency on methods, etc.)

Dissemination Channels



Dissemination
Enabling environment

- NSO should disseminate microdata in accordance with good practice
  - With detailed metadata to ensure discoverability and usability of data
  - Under an enabling legislation
  - According to a formal policy and transparent protocols
    - What data can be published?
    - To whom?
    - Under what terms of use?
    - In what form (anonymized)?

Dissemination
Policy and protocols

- NSO must have a formal (and publicly accessible) policy with protocols to define who can get access to what, under what conditions
  - Not a “one-size-fits-all” policy; various access levels depending on content of data (open data; public use files, licensed datasets; data accessible in enclave; data not accessible)
  - Can have multiple versions of a dataset from the same survey
  - For Example: MOU’s between the World Bank and Countries allow internal use for official business.
- Anonymization of microdata may be required
- NSO must have staff, equipment and system in place to offer proper support

Dissemination
World Bank Software

- National Data Archive (NADA) software
- An open source package (PHP)
- Discover, Browse, Search, Download > Metadata > Data
- Used by the World Bank and many countries (IHSN, NSOs, ILO, WHO, UIS, MCC ,etc.)
- Development plan based on feedback / requests from users
  - Originally for survey/census microdata (DDI standard)
  - Soon for any kind of data / metadata standards: geospatial, time series, etc.

Dissemination
Policy & Protocol Resources

- (IHSN) Data Dissemination policy guidelines
- Dissemination of Microdata Files - Principles, Procedures and Practices

Disclosure Control
Anonymization

http://dilbert.com/strip/2010-11-24

Disclosure Control
Anonymization (or "SDC")

- Required to comply with legal and ethical rules
- No international standard to define what to do and how to do it
- Tools and methods are available, but require "arbitrary" decisions to be made
  - Acceptable risk level?
  - Nature of the risk? (Worse case scenario?)
- Can be relatively simple … or made very complex
- Need a good understanding of what anonymization (or "statistical disclosure control" – SDC) means

Disclosure Control
Anonymization: identifiers

- Direct identifiers (names, tel. numbers, addresses, etc.) are excluded from datasets
- But (the combinations of) other variables – indirect identifiers or "key variables" – can lead to re-identification of respondents
  - E.g., combination of date of birth, age, geographic location and profession can be highly identifiable
- SDC is a process of treating identifiers in order to reduce the risk of disclosure
  - Identifies and treats unique or rare combinations of indirect identifiers

Disclosure Control
Anonymization: disclosure scenarios

- First step is to understand what we are protecting against
  - Must define disclosure scenarios
  - Need to assess what intruders know/have
- Mostly, we protect data against the risk of re-identification through data matching (not against nosy-neighbor)
- Assume that the intruder has access to a database with names + variables common to the NSO datasets, allowing matching records
  - Country-specific: what other datasets are available to intruders?

Disclosure Control
Anonymization: Measure Risk

- Based on modelled estimate of re-identification risk at the individual, household, and global levels
  - Provides probability of being re-identified at record-level (taking sample rates into account)
  - Provides expected number of re-identification
  - → can establish a threshold as a target
- Based on frequency of combination in dataset (k-anonymity)
  - E.g., 3-anonymity imposes that each combination of key variables would have a frequency of 3 or >

Disclosure Control
Anonymization: Methods

- Global recoding (e.g., replace date of birth with age or age group; group districts into regions, etc.)
- Top/bottom coding (e.g., all incomes >1,000,000 recoded as “1,000,000 and higher”)
- Local suppression (selective suppression of rare values)
- Micro-aggregation
- Post-randomization
- Rank swapping
- Shuffling
- Noise addition
- Displacement of GIS Co-ordinates

Disclosure Control
Anonymization: Utility

- SDC will reduce the disclosure risk, not eliminate it
- All SDC methods result in some information loss (reduction of utility)
- In general the greater the level of protection the higher the information loss and loss of utility
- The goal is to produce a dataset with acceptable risk of identification while keeping utility as high as possible
  - The level of acceptability of disclosure risk is at the discretion of the data producer and may be guided by legislation
  - Level of acceptable risk also depends on access type (e.g., licensed datasets may be less anonymized than public use data)

Disclosure Control
Anonymization: Support & Guidelines

- Introduction to Statistical Disclosure Control (SDC) by Matthias Templ, Bernard Meindl and Alexander Kowarik
- IHSN website http://ihsn.org/home/node/118

Disclosure Control
Anonymization: Tools

- Free open source R packages sdcMicro and sdcMicroGUI developed by, Templ, Meindl and Kowarik with IHSN\World Bank support. https://cran.r-project.org/web/packages/sdcMicro/index.html

- Requires knowledge of R
  - developing a Graphic Interface to lower this barrier
- Implements all the major anonymization routines.
  - Quantifying and measuring risk, recoding, suppression, randomization, noise addition, swapping.
- Requires the input of the subject specialist
- Requires computational power for some routines

Conclusion

- Documentation of data is a must (DDI standard recommended); cheap and “easy” – training and tools are available
- Publishing a clear dissemination policy is needed if NSO is to publish microdata
- Data should be (reasonably) anonymized; level of anonymization depends on the terms of use, type of data, legal environment
- Training/technical support may be available, keeping in mind that:
  - Anonymization can be a complex issue; requires strong analytical skills
  - A regional approach?
  - Focus on household surveys, population censuses and some administrative datasets (anonymizing enterprise surveys is a major challenge)

Thank you and discussion

Matthew Welch

mwelch@worldbank.org