Taskforce for household survey microdata dissemination
May 2020

1. Rationale

Data collected through censuses, surveys, administrative systems, and other sources form the foundation of official statistics, and are an invaluable input to research and decision making. These data are commonly referred to as microdata, defined as unit-level information on people or entities (such as individuals, households, business enterprises, farms, or even geographic areas). They are aggregated and analyzed to generate national estimates by official statisticians and analyzed by researchers and policy analysts to gain scientific insights which can be translated into policy.

The power of microdata stems from its granularity. Because microdata contain unit-level information, they facilitate the investigation of the unique ways a certain phenomenon may affect different groups or sub-populations. Microdata can be costly to collect, and microdata files often contain many variables, more than any single organization can fully exploit. An effective way to maximize the return on investment and fully leverage microdata's analytical potential is to share them widely. The role of governments in promoting the wider use of microdata has become increasingly important, and many initiatives have been implemented with not only the goal of maximizing potential uses, but also in the interest of transparency and openness.

Over the past two decades, a consensus has formed that data, particularly data produced by public bodies, should be openly and freely available[1]. This is in keeping with principle 1 of the Fundamental Principles of Official Statistics (FPOS), which says, in part, "[O]fficial statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honour citizens' entitlement to public information." Several indexes have been created to measure the openness of government datasets[2] However, these datasets largely consist of aggregate indicators. Microdata, which often contain identifiable information of individuals or other entities, are governed by principle 6, which says, "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."[3] This places limitations on what the data can be used for and who can use it. The dissemination of microdata requires careful attention to the disclosure of confidential information.

The World Bank, the International Household Survey Network (IHSN), UNECE, and other organizations have published best practice guidelines, handbooks, and software tools to assist organizations in disseminating microdata. Additionally, a previous report on *Standards and best practices for survey data documentation* was produced for the CCSA[4] in 2014. These documents identify the key principles,

[1] For example, the International Open Data Charter (https://opendatacharter.net/principles/)

[2] For example, the Open Data Inventory (http://odin.opendatawatch.com/) and the Open Data Barometer (https://opendatabarometer.org/).

[3] https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx

[4] https://unstats.un.org/unsd/accsub-public/microdata.pdf

technical standards, and best practices for disseminating microdata. They have paved the way for international organizations and national statistical offices to develop microdata dissemination policies.

Despite the availability of these manuals and tools, many countries still do not disseminate microdata, and those that do don't always follow the recommendations. Furthermore, in recent years, the data ecosystem has changed dramatically. To improve timeliness and fill gaps between survey rounds, National statistical offices are increasingly looking to take advantage of new sources of data, such as remote sensing, devices that make up Internet of Things (IoT), and social networks. These new data sources are seen as possible cost-effective complements to sample surveys. New or higher frequency data may fill gaps which surveys are unable to fill quickly enough while at the same time surveys may provide ground truthing (and quality checks) for these new sources of data. From a privacy and quality perspective many of these new data sources provide additional challenges for microdata dissemination. For example, geo-referenced data retained in microdata are needed to link with remote sensed data, which may breach respondents' confidentiality. Additionally, as more and more new data sources become publicly available the ability to link data and identify individuals may increase substantially. It follows that new practices of integrating new data sources with microdata will require some changes in data management and dissemination practices.

In this context, the task force aims to provide an overview of the state of the art in terms of microdata dissemination in countries comparing the prevailing practices with the recommended practices. Such a review will be carried out in a number of countries that are selected to represent those in different levels of statistical development and in different geographical regions. Furthermore, due the changing data ecosystem, emerging challenges will be identified in order to set an agenda for future research and international efforts to promote microdata dissemination.

2. Proposed output(s)

The task force will result in a technical report consisting of four chapters. The report will follow the structure below:

Chapter 1: Introduction: Microdata dissemination in the era of Open Data – *This chapter will provide the reader the context and motivation of the report. An introduction on the use and value of microdata, a brief history of international efforts to promote microdata dissemination, and a description of open data principles will motivate Chapter 2. A description of the changing data ecosystem and potential impacts on microdata dissemination and use will motivate Chapter 3.*

Chapter 2: State of the art of microdata dissemination – *The objective of this chapter is to provide an overall picture on the availability of microdata, and compare the prevailing practices in a set of countries varying in geography and statistical capacity with recommended practices. The group will leverage information already obtained by the World Bank, Open Data Watch, and FAO to estimate the availability of microdata across countries and statistical domains. To provide examples of current practice, a group of countries will be selected, as much as possible, representing those in different levels of statistical* *development and in different geographical regions. Comparisons will be made of their current practices and best practices defined in international manuals across the following aspects:*

1. *Enabling legal framework, licensing, and terms of use*
2. *Quality of metadata, assessed by their accuracy, completeness, and adherence to recognized standards such as the Data Documentation Initiative (DDI)*
3. *An assessment of statistical disclosure control (SDC) policies developed by NSO's (including methods applied, if publicly disclosed) as well as well the presence of well composed legislation around SDC of microdata.*
4. *Availability and accessibility of microdata (How countries choose which datasets to disseminate, as well as IT, legal, and other constraints)*

Whenever applicable to any of the items listed above, good practices of international agencies in disseminating survey microdata will also be discussed.

Chapter 3: Microdata dissemination in an evolving data ecosystem – *The objective of this chapter is to identify new challenges and gaps in the existing manuals on microdata dissemination, and point to the way forward for the international statistical community to focus on. These can be organized based on the aspects analyzed in Chapter 2. For example:*

1. *Enabling legal frameworks: An analysis based on a review of a few key data privacy legislations. For example, using the GDPR as an example of how privacy legislation impacts microdata dissemination in Europe.*
2. *Quality of metadata: Enabling data interoperability through detailed standards-based metadata.*
3. *Statistical disclosure control: Additional SDC considerations presented by alternative data sources*
4. *Availability and accessibility: Exploring the feasibility of trusted repositories, remote access and virtual enclaves in low- and middle-income countries as a means of overcoming barriers to microdata dissemination. This includes an assessment of sustainable funding, technical requirements, legal requirements.*

Chapter 4: Conclusion (summary) – *The objective of this chapter is to summarize (a) current status of microdata dissemination, highlighting issues and challenges faced; (b) recommendations on various approaches and tools that work for countries at different capacities; and (c) areas for future work*