Food and Agriculture Organization
of the United Nations

PROVIDING ACCESS TO AGRICULTURE MICRODATA

GUIDELINES

Publication prepared in the framework of the Global Strategy to improve Agricultural and Rural Statistics

Providing Access to Agriculture Microdata: A Guide

August 2014

Table of Contents

Preface

The development of these guidelines falls under the framework of the *Global Strategy to Improve Agricultural and Rural Statistics* and builds on the *International Household Survey Network* methods and practices.The Global Strategy provides the framework essential to meet current and emerging data requirements, and the demands of policy makers and other data users. Its goal is to contribute to greater food security, reduced food price volatility, higher incomes and greater well-being for rural populations, through evidence-based policies. The Global Action Plan of the Global Strategy is centred on 3 pillars: (1) establishing a minimum set of core data; (2) integrating agriculture into the National Statistical System (NSS); and (3) fostering sustainability of the statistical system, through governance and statistical capacity building.

The second pillar mentioned above (integrating agriculture into the NSS) recommends that countries establish a strategy for data dissemination, to ensure that the data is accessible. One important advancement made over the last few decades is the regular creation of user-accessible microdata files for in-depth analysis. The indications provided in this Guide are intended to help producers of agricultural data to navigate the process of providing researchers with access to microdata, while at the same time respecting the statistical and privacy requirements binding them.

This Guide presents a set of operational tools, methods and good practices that are the result of a long process, taking advantage of knowledge from country experiences and existing material developed by the World Bank and PARIS21 on household survey microdata, within the International Household Survey Network. Access to agricultural microdata is still not common, but the few practices available have been included. In designing their microdata policies, countries will be able to refer to existing tools. This Guide will be updated regularly, thanks to countries' feedback and experiences in implementing agricultural microdata dissemination.

Acknowledgments

These Guidelines were developed by Ernie Boyko, former Director of Agriculture, Corporate Planning, Electronic Publishing, and Operations at Statistics Canada, with the help of Nancy Chin and François Fonteneau, Statisticians, FAO.

This work draws heavily upon an earlier guide prepared by Olivier Dupriez from the International Household Survey Network (IHSN) and Ernie Boyko in 2010, which is accessible at http://www.surveynetwork.org/home/sites/default/files/resources/IHSN-WP005.pdf.

Comments from Wendy Watkins from the Carleton University Data Centre, and from Jean-Louis Tambay from Statistics Canada's Methodology Directorate, were greatly appreciated by the author. Proofreading was carried out by Lorna Boyko. The preparation of this publication was supported by the Trust Fund of the Global Strategy to Improve Agricultural and Rural Statistics, funded by the Department for International Development (DFID) of the United Kingdom, and the Bill and Melinda Gates Foundation (BMGF).

Acronyms

ABS	Australian Bureau of Statistics
ABSDL	Australian Bureau of Statistics Data Laboratory
ADP/PARIS21	Accelerated Data Program/Partnership in Statistics for Development in the 21st Century
CDER	Canadian Centre for Data Development and Economic Research
CES	Center for Economic Studies
CSPro	Census and Survey Processing System (Software)
DDI	Data Documentation Initiative
DLI	Data Liberation Initiative
GPS	Global Positioning System
IHSN	International Household Survey Network
NADA	Microdata Cataloguing Tool
NESSTAR	Networked Social Science Tools and Resources (Software)
NBS	Nigerian Bureau of Statistics
NISR	Rwandan National Institute of Statistics
NSO	National Statistics Office
PUF	Public Use File
RAF	Remote Access Facility
RADL	Remote Access Data Laboratory
RDC	Research Data Centre
RTRA	Real Time Remote Access
SAS	Statistical Analysis System (Software)
SDC	Statistical Disclosure Control
SDCMicro	Statistical Disclosure Control for Microdata files (Software)
SO	Statistical Organization
SPSS	Statistical Package for the Social Sciences (Software)
STATA	Data Analysis and Statistical Software (Software)
XML	Extensible Markup Language

Introduction

The nature and role of information in society

Information plays a vital role in the agricultural sector. It provides us with a picture of the status and contributions of the sector, in terms of the production of food, clothing, shelter, income, and employment. Decisions concerning agriculture and the sector's major players are often based on data collected from agricultural operators. A single survey or census can support decision making in a number of ways, after the basic data have been collected. Figure 1 below depicts the different stages involved in the interrelated processes of using measures in the real world to make decisions, which may in turn result in changes to the real world itself.

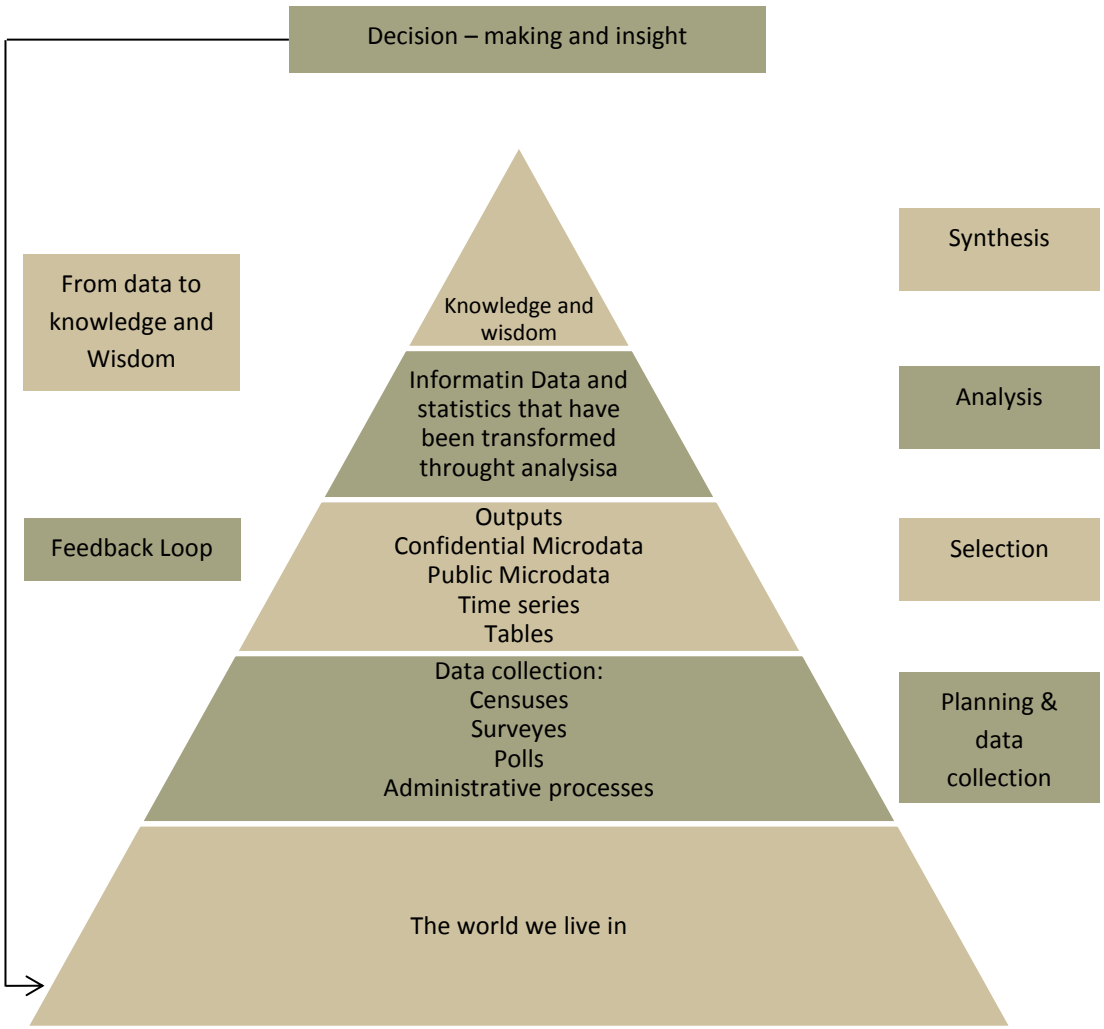


FIGURE 1. Evidence-Based Decision Making
(Public Microdata: Microdata)

There are several sources for such data, including censuses, surveys, administrative data and even remotely sensed data from satellites. The Outputs section of Figure 1 above refers to confidential microdata, public microdata, time series and tables. Users of agricultural data are accustomed to accessing aggregate statistics, such as time series and tables, to paint a picture of the sector. While these aggregate statistics may have been prepared after consultation with the user community, they represent but a small fraction of the possible tabulations that can be produced from a given source. Users who require additional tabulations are left with two choices:

- Return to the data producers (the holders of the microdata files) and request additional outputs, or
- Obtain access to a (public) version of the microdata files and perform their own data manipulations.

The purpose of this Guide is to explore the options available to data producers for providing researchers and other users with better access to census and survey data. More specifically, the Guide will explore the creation of publicly accessible microdata.

This document is structured in the following Sections:

1. What are microdata?
2. Data producers' rationales for providing access to microdata
3. Alternative models for providing access to microdata
4. Preparing data files for user access
5. Providing access to agricultural microdata
6. Legal and policy frameworks for providing access to agricultural microdata
7. Technical infrastructure and institutional requirements for providing access to microdata
8. Promoting the use of microdata
9. The Open Data Agenda
10. Concluding observations
11. Bibliography
12. Appendices

Notes to readers:

- The term "providing access to agricultural microdata" is used to indicate all forms of dissemination of microdata. This includes files that can be transferred to users as well as those that users can access remotely and download only specified outputs thereof. Accessing data on-site in a data enclave is also included.
- The term "Statistical Organization" (SO) is used to refer to the agency or organization responsible for producing agricultural statistics, regardless of whether this is the National Statistics Office (NSO) or the Ministry of Agriculture.

1. What Are Microdata?

1.1 Defining Microdata

Figure 1 above shows microdata as the output of a survey or census. Microdata refer to the information recorded by or from the respondent, when a survey or census is conducted. The following definitions for microdata were taken from the OECD website[1]. The additions within square brackets [ ] were made by the author of this Guide.

*"**Definition: [of microdata]***

…observation data collected on an individual object - statistical unit.

[For agriculture, the statistical unit could be the household, the family, the farm or the community.]

Context:

Microdata is data on the characteristics of units of a population, such as individuals, households, or establishments, collected by a census, survey, or experiment. (Survey Design and Statistical Methodology Metadata, Software and Standards Management Branch, Systems Support Division, United States Bureau of the Census, Washington D.C., August 1998, Section 3.4.4, p. 39).

[In an agricultural survey, the characteristic measured could be the number of heads of livestock for a particular respondent, on a specific (anonymized) unit.]"

1.2 How are Microdata Organized?

Microdata are the basic building blocks of statistical information. They can be aggregated, to form tables that can be published for use by researchers and others. Microdata are stored in computer files as records.

The screenshots below were taken from the *Comprehensive Food Security and Vulnerability Analysis and Nutrition Survey 2012*, conducted by the Rwanda National Institute of Statistics.

[1]Organisation for Economic Cooperation and Development. 2005. "Statistical Microdata." Glossary of Statistical Terms. Available at http://stats.oecd.org/glossary/detail.asp?ID=1656. In particular, the source refers to Economic Commission for Europe of the United Nations (UNECE), "Terminology on Statistical Metadata", Conference of European Statisticians Statistical Standards and Studies, No. 53, Geneva, 2000. Accessed on 1 January 2014.

FIGURE 2. Catalogue entry for the Rwanda survey *on Comprehensive Food Security and Vulnerability Analysis and Nutrition Survey 2012,* from the IHSN NADA catalogue[2].

It must be noted that the entry provides all the information that researchers need to access and utilize the data, including

- The study website
- A description of the study
- A data dictionary
- A link to the data files and
- Related materials, including the questionnaires used.

Clicking on the 'Get Microdata' link leads to a page that requires users to identify themselves (the reason for this will be explained below). The files can be downloaded in a compressed format and will appear as shown in Figure 3 below.

[2] Available at http://catalog.ihsn.org/index.php/catalog/4149.

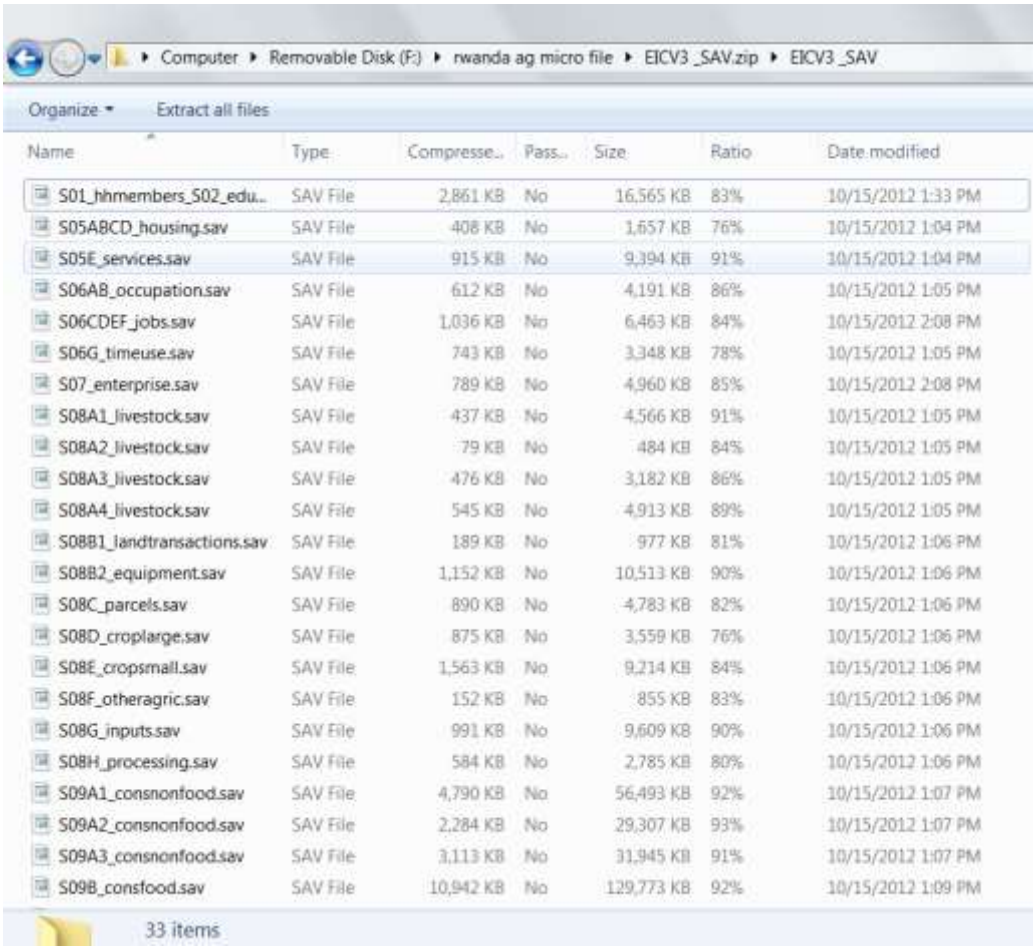
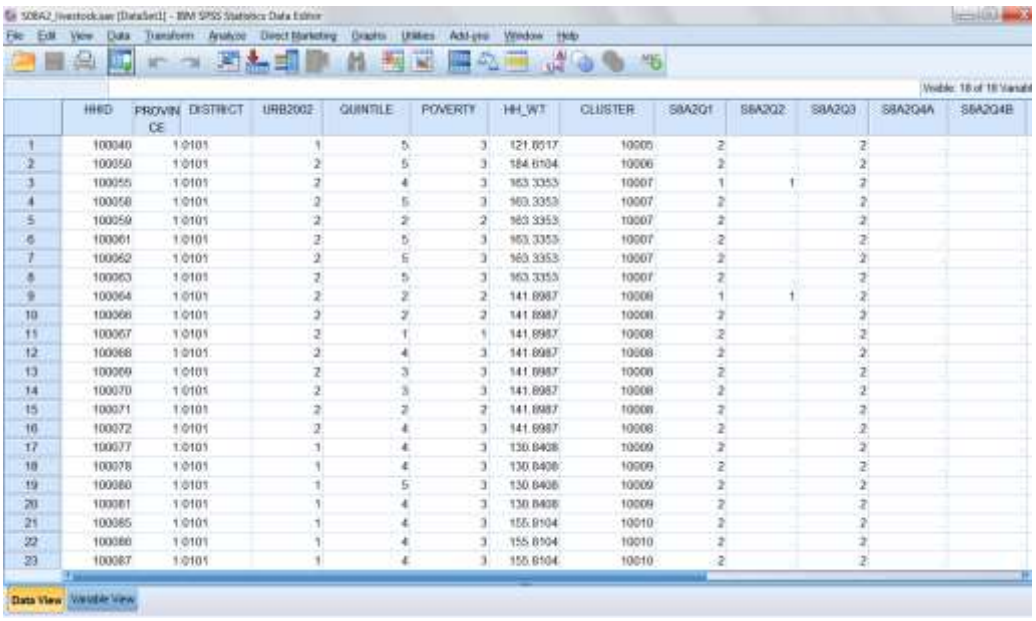


FIGURE 3. Thirty-three separate files compose the microdata for the *Comprehensive Food Security and Vulnerability Analysis and Nutrition Survey 2012*[3]. (file S08A2: a view of this file, opened, is provided in Figure 4)

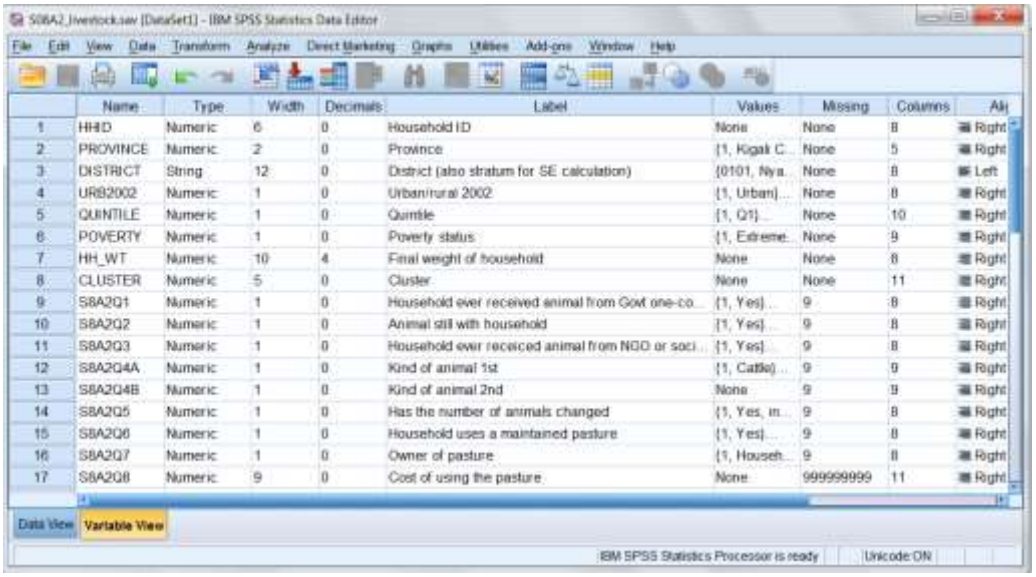
File "S08A2_Livestock" is opened below. Two views of the file are shown. "Data View" shows the file organized with the variables (the questions) in the columns and the individual responses in the rows. For ease of processing, the variable names are mnemonics. Figure 5 below shows the same file in "Variable View", which displays the variable labels for ease of human comprehension.

[3] Rwandan National Institute of Statistics (NISR), Ministry of Agriculture and Animal Resources (MINAGRI) and World Food Programme. *Rwanda Comprehensive Food Security and Vulnerability Analysis and Nutrition Survey 2012*. Ref. RWA_2012_CFSVA_v01_M. Data set downloaded from http://nada.vam.wfp.org/index.php/catalog on 6 March 2014.



	HHID	PROVINCE	DISTRICT	URB2002	QUINTILE	POVERTY	HH_WT	CLUSTER	S8A2Q1	S8A2Q2	S8A2Q3	S8A2Q4A	S8A2Q4B
1	100040	1	0101	1	5	3	121.6517	10005	2		2		
2	100050	1	0101	2	5	3	184.6104	10006	2		2		
3	100055	1	0101	2	4	3	163.3353	10007	1	1	2		
4	100058	1	0101	2	5	3	163.3353	10007	2		2		
5	100059	1	0101	2	2	2	163.3353	10007	2		2		
6	100061	1	0101	2	5	3	163.3353	10007	2		2		
7	100062	1	0101	2	5	3	163.3353	10007	2		2		
8	100063	1	0101	2	5	3	163.3353	10007	2		2		
9	100064	1	0101	2	2	2	141.8987	10008	1	1	2		
10	100066	1	0101	2	2	2	141.8987	10008	2		2		
11	100067	1	0101	2	1	1	141.8987	10008	2		2		
12	100068	1	0101	2	4	3	141.8987	10008	2		2		
13	100069	1	0101	2	3	3	141.8987	10008	2		2		
14	100070	1	0101	2	3	3	141.8987	10008	2		2		
15	100071	1	0101	2	2	2	141.8987	10008	2		2		
16	100072	1	0101	2	4	3	141.8987	10008	2		2		
17	100077	1	0101	1	4	3	130.8408	10009	2		2		
18	100078	1	0101	1	4	3	130.8408	10009	2		2		
19	100080	1	0101	1	5	3	130.8408	10009	2		2		
20	100081	1	0101	1	4	3	130.8408	10009	2		2		
21	100085	1	0101	1	4	3	155.6104	10010	2		2		
22	100086	1	0101	1	4	3	155.6104	10010	2		2		
23	100087	1	0101	1	4	3	155.6104	10010	2		2		

FIGURE 4. File "S08A2_livestock" opened in SPSS and shown in "Data View"



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align
1	HHID	Numeric	6	0	Household ID	None	None	8	Right
2	PROVINCE	Numeric	2	0	Province	{1, Kigali C...	None	5	Right
3	DISTRICT	String	12	0	District (also stratum for SE calculation)	{0101, Nya...	None	8	Left
4	URB2002	Numeric	1	0	Urban/rural 2002	{1, Urban}...	None	8	Right
5	QUINTILE	Numeric	1	0	Quintile	{1, Q1}...	None	10	Right
6	POVERTY	Numeric	1	0	Poverty status	{1, Extreme...	None	9	Right
7	HH_WT	Numeric	10	4	Final weight of household	None	None	8	Right
8	CLUSTER	Numeric	5	0	Cluster	None	None	11	Right
9	S8A2Q1	Numeric	1	0	Household ever received animal from Govt one-co...	{1, Yes}...	9	8	Right
10	S8A2Q2	Numeric	1	0	Animal still with household	{1, Yes}...	9	8	Right
11	S8A2Q3	Numeric	1	0	Household ever received animal from NGO or soci...	{1, Yes}...	9	8	Right
12	S8A2Q4A	Numeric	1	0	Kind of animal 1st	{1, Cattle}...	9	9	Right
13	S8A2Q4B	Numeric	1	0	Kind of animal 2nd	None	9	9	Right
14	S8A2Q5	Numeric	1	0	Has the number of animals changed	{1, Yes, in...	9	8	Right
15	S8A2Q6	Numeric	1	0	Household uses a maintained pasture	{1, Yes}...	9	8	Right
16	S8A2Q7	Numeric	1	0	Owner of pasture	{1, Househ...	9	8	Right
17	S8A2Q8	Numeric	9	0	Cost of using the pasture	None	999999999	11	Right

FIGURE 5. File "S08A2_livestock" opened in SPSS and shown in "Variable View"

Another way to conceive of the organization of microdata is to consider a questionnaire with several questions. Each row (as in Figure 4 above) would represent an individual questionnaire; each column, the respondent's coded answers to the questions.

In addition to the variable labels, researchers need metadata that can help them understand the codes, definitions and concepts underpinning the data collected. The concept of metadata will be described later in this Guide.

1.3 The Power of Microdata

The best way to appreciate the power of microdata in relation to aggregate tables is to consider the following analogy.

[Micro]data are unlike other tools of the research endeavour. They provide the raw material from which information and knowledge can be created. By their nature, data allow for exploration of topics of interest to the researcher. Unlike printed tables which, like a postcard, provide a picture of one view of a larger phenomenon, data can act as a camera, allowing the researcher to manipulate the background, change the foreground and more fully investigate the object under study.[4]

The utility and power of microdata files can be demonstrated by considering File "S08A2_livestock". The survey from which this file was taken has hundreds of variables. The National Institute of Statistics Rwanda (NISR) and the World Food Programme (WFP) have issued reports which draw upon the data from this survey. These reports contain summary tables and analyses relating to topics that were deemed important for readers and decision makers.

If researchers wish to use the survey data to analyse questions that were not addressed during the initial analysis, they can make use of the microdata file. For example (with reference to Figure 6 below), if more information on the impact of certain policies (e.g. the "one cow" policy) with regard to poverty were required, researchers could choose Variable 1 (Poverty Status) and Variable 2 (whether the respondent ever received one cow from the government). Researchers could then perform a simple analysis, such as a cross-tabulation, to show the poverty status of households to households receiving animals under the one-cow policy (Table 1 below). The results of the cross-tabulation can be used to describe the policy's performance, with a further analysis, e.g. a regression analysis, carried out to generalize the quantitative impact of the policy measure, including for particular sub-groups such as gender, geography etc.

[4] Watkins, W. & Boyko, E. 1996. Data Liberation and Academic Freedom. *Government Information in Canada/Information gouvernementale au Canada,* 3(2). Available at: http://www.usask.ca/library/gic/v3n2/watkins2/watkins2.html. Accessed on 23 May 2014

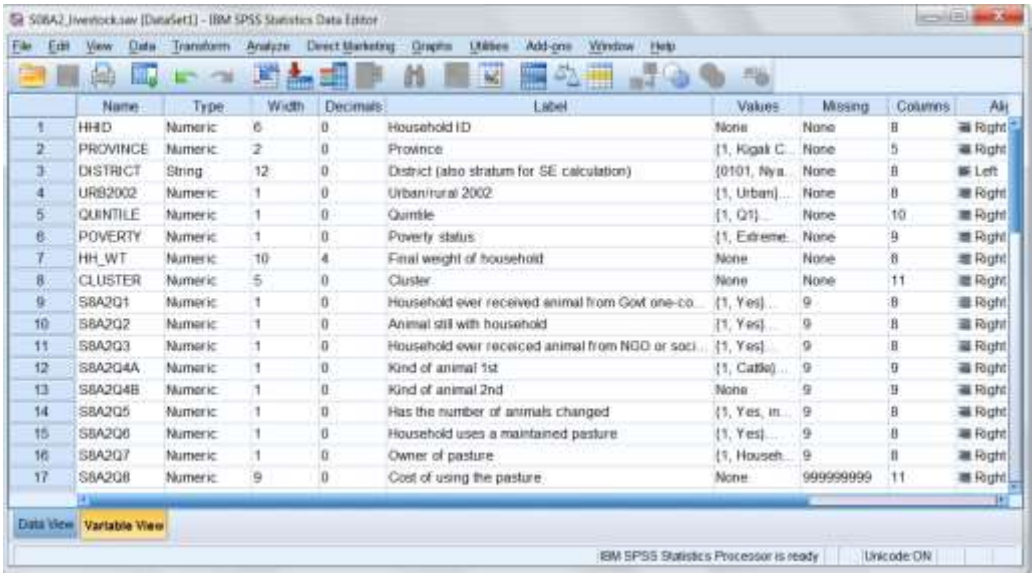


FIGURE 6. Selecting variables to analyse policy
(Data view 2: variable 1. Data view 5: variable 2)

	Households ever receiving animals from one-cow policy	
Poverty Status	Yes	No
Extremely Poor		
Poor		
Non Poor		
Total		

TABLE 1. Structure of cross-tabulation of "Households' poverty status" by "Households ever receiving animals from the government one-cow policy"

Access to microdata is essential to perform this type of analysis. Producing a microdata file enhances the value of the survey, compared to the limited secondary analysis that can be performed on static tables.
Thus, microdata is clearly a powerful tool for researchers, enabling them to leverage the utility of a survey file. However, before microdata can be made available to researchers (typically researchers at universities, research departments in ministries and research institutes), steps must be taken to ensure that the respondents' privacy and the data producers' confidentiality requirements are respected. It may not always be possible to create a public file that can be moved to the researcher's premises, as was the case with the NISR survey. In some cases, data producers may have to provide researchers with access to the microdata via a mediated service. If this cannot

be done, then custom tabulations may be researchers' only option. The rest of this Guide will explore the rationale, mechanisms and statistical issues involved in enabling researchers' adequate access to microdata from agricultural surveys and censuses, while at the same time not compromising respondents' privacy and confidentiality.

Throughout this Guide, reference will be made to policies and procedures for providing access to microdata, along with examples from various countries. Many of the examples are drawn from the work of developed countries, as these have been distributing microdata for several years. However, it is encouraging to note that developing countries are also pursuing this path. Some examples of the latter are included in this Guide.

1.4 Microdata Files as Part of the Survey Cycle

To facilitate comprehension of the options available to data producers for providing access to microdata, Figure 7 below illustrates the stages of the survey life cycle. Stage F shows the final file produced before commencement of the dissemination process. There are two dissemination streams: Stages G and I result in tables and updates of time series that are typically published on the Internet or in a printed publication. The file from Stage F can also be used to create a microdata file for researcher access. Section 1.5 below will describe the various access options.

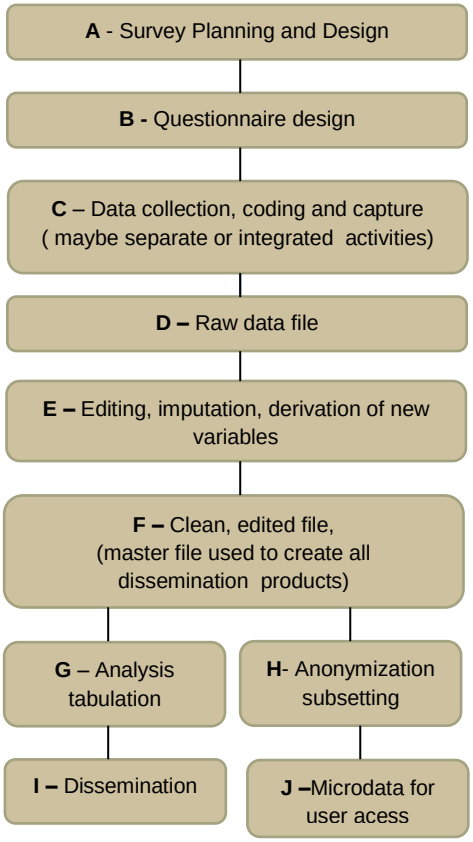


Figure 7. Survey Life Cycle

1.5 Different Types of Microdata Files

Several considerations must be made when choosing the best option for providing access to the microdata produced in Stage J. Some options (Public Use Files and Licensed Files) involve creating a file that has been subjected to statistical disclosure control (anonymization) and can be transferred to the user's premises. Others (e.g. remote access facilities) require files to be accessed on the data producers' computers, through remote job submission for batch or interactive processing. Data enclaves require the user to visit a location controlled by the data producer. Finally, researchers can also be invited to join the statistical organization (SO) as deemed employees, when the research they conduct is of direct interest to the SO or its mother organization. A brief description of each is presented in Sections 3 and 5 of this Guide.

1.6 A Note on Agricultural Microdata

Microdata for agriculture can be derived from agricultural censuses, surveys, polls or administrative data. For censuses and surveys, each question within the questionnaire to which operators respond constitutes a record. A questionnaire is generally completed for each holding, whether households or larger commercial farms, although the questionnaire may contain more than one section. The respondents to the questionnaire may be the owners or operators of the household farms, while a manager or contract worker will generally complete the questionnaires for commercial or corporate farms.

The agriculture industry's profile in terms of size of farming area varies widely across countries. Examples of the number of farms by size for Uruguay and Ethiopia are shown below.

Number and area of holdings by size		
	Number of holdings	**Area (ha)**
Total	57 131	16 419 683
1 - 4 ha	6 260	16 516
5 - 9	7 086	47 611
10 - 19	7 118	97 841
20 - 49	8 934	285 254
50 - 99	6 647	472 928
100 - 199	6 382	910 286
200 - 499	6 783	2 162 836
500 - 999	3 887	2 725 637
1000 - 2500	2 912	4 441 627
2500 - 5000	838	2 837 134
5000 - 10000	228	1 504 482
10000 >	56	917 531

TABLE 2. Uruguay Agriculture Census 2000[5]

Legal status		
	Number of holdings	**Area (ha)**
Total	57 131	16 419 683
Civil person	49 302	10 159 084
Corporation	7 336	6 103 333
Government	395	75 543
Other	98	81 723

TABLE 2. Uruguay Agriculture Census 2000

These two countries illustrate the extremes of agricultural organization. In Uruguay, there are 57,131 holdings, each having an average size of 284 hectares. Perhaps more strikingly, there are 56 holdings expanding over 10,000 hectares, which account for over 5% of the agricultural land area. In addition, 7,336 of the holdings in Uruguay operate as corporations. In Ethiopia, there are 10,758,597 holdings, each having an average size of approximately 1 hectare. Of these, 10,333 expand over 10 or more hectares. This group accounts for slightly over 1% of agricultural land.

Number and area of holdings by size		
	Number of holdings	**Area (ha)**
Total	10 758 597	11 047 249
< 0.1 ha	819 394	38 418
0.1– 0.5	3 175 027	933 428
0.5 – 1	2 767 746	2 021 798
1 - 2	2 612 288	3 682 947
2 - 5	1 276 773	3 605 515
5 - 10	97 037	612 070
10 >	10 333	153 072

TABLE 3. Ethiopia Agricultural Census 2001/02[6]

This Guide seeks to discuss models that can be used to provide access to agriculture microdata. As will be shown below, meeting researchers' needs, while ensuring the greatest possible protection for the privacy of respondents, are the pre-eminent considerations when choosing a microdata access system. One method involves creating files that have undergone a process known as anonymization. Uruguay and Ethiopia are compared to demonstrate the diversity of the populations that disseminators must handle. Without further information, it appears that anonymizing farm holdings for Uruguay would be much more difficult than for Ethiopia.

[5] FAO. 2000. 2000 *World Census of Agriculture: Main Results and Metadata by Country*, FAO

[6] Ibid. The legal status of Ethiopian holdings is not available from this Table. The information available on the FAO census site varies by country.

BOX 1

Note: This Guide seeks to discuss models that can be used to provide access to agriculture microdata. As will be shown below, meeting researchers' needs, while ensuring the greatest possible protection for the privacy of respondents, are the pre-eminent considerations when choosing a microdata access system. One method involves creating files that have undergone a process known as anonymization. Uruguay and Ethiopia are compared to demonstrate the diversity of the populations that disseminators must handle. Without further information, it appears that anonymizing farm holdings for Uruguay would be much more difficult than for Ethiopia.

If a PUF or Licensed File must be made on the basis of an agricultural census, a two-step process should be adopted. Step 1 would be the drawing of a sample from the master file; Step 2 would be to anonymize the sample file to be released.

2. Data Producers' Rationales for Providing Access to Microdata

Providing access to microdata requires SOs to balance the demands of the research community with their legal obligations to maintain the confidentiality of the information collected from respondents. If SOs fail to do so, they risk undermining the confidence of respondents and thus losing their support. At the same time, they are under pressure to accede to researchers' demands as fully as possible, because the research that could be supported may lead to significant benefits for the sector and for the country.

The research community includes those working in government research departments, academic institutions, and researchers working in non-governmental organizations and international agencies. In this context, the research conducted often involves working within government-funded agencies and institutions, or with a grant from an international agency.

The demand for microdata access is driven by the research and policy agendas within the country and by the research and development programs of international institutes and organizations. The research community is an important player in the processes of offering policy analysis, stimulating debate and helping to assess policies and programs. These activities can lead to revised or new programs to further national goals and directly benefit the sector.

To perform this work, researchers require access to good quality statistical data. If statistical organizations possess such data, they should strive to satisfy researchers' demands. Otherwise, researchers may attempt to collect their own data through their own studies and surveys. Since these parties may not have the same resources and capabilities as statistical offices, the resulting data may be of an inferior quality.

Below are some of the factors that statistical offices should consider in deciding whether to provide access to microdata.

- Providing access to microdata supports research and thus helps statistical offices to fulfill their mandate. Statistical offices are usually established to serve the informational needs of the country or community. Therefore, supporting researchers should be seen as part of the offices' overall mandate.
- If researchers can meet their own needs by accessing microdata from statistical offices, they are less likely to collect their own data, thus avoiding a greater response burden and duplication.
- Creating additional uses for collected data will increase the returns on the investments made in data collection.
- Satisfying a wider user demand can bring greater recognition to statistical offices' work and enhance their credibility. This may also improve the chances of receiving future funding, as there will be a broader base of supporters for new programs.
- Feedback from researchers who have analysed the data may lead to quality improvements in survey methodology.
- In an age of "open government" and open data, permitting broader access to data may lead to innovation and better tools for using data.
- Identifying ways to provide access to microdata may be less expensive than attempting to fill information gaps by providing further custom tabulations.
- Finally, statistical offices may be obliged to permit access to microdata pursuant to the legal contract under which the survey was funded.

The rationale for providing access to microdata must be tempered by considering three points that often raise concerns among statistical organizations:

- The need to adhere to the SO's confidentiality policies
    - The legislation under which SOs operate varies considerably across countries. In some cases, the release of microdata may be prohibited (although this is decreasingly common as statistical legislation is revised), while in other examples, microdata can be released only under certain circumstances. The interpretation of these provisions is generally left to SOs' professionals and experts.

- Data quality
    - In addition to undertaking special surveys such as those for the World Food Programme, SOs are responsible for publishing official statistics (often, a time series). In these cases, they must take a variety of indicators into account. Problem encountered during the execution of an annual survey may lead agencies to publish estimates that are beyond the survey coefficients of variation. This may lead to a reluctance to release surveys as microdata files.

- The cost of providing microdata access options

    - The cost of releasing microdata is an issue mainly if the SO's personnel is not familiar with the techniques required for anonymizing data files. If the files are under the control of the relevant Ministry of Agriculture (to which the NSO delegated the power to conduct the survey), while the expertise required for preparing the file lies with the NSO, then the parties would have to formalize an agreement. The main expertise required is generally a combination of subject matter knowledge (which would be found in the Ministry of Agriculture) and mathematical statistics (generally available in the NSO or through an international consultant).

The existing trade-offs cannot be described in more detail here. It may be advisable to consult the examples provided in various international documents, such as:

CENEX-SDC. 2007. "Handbook on Statistical Disclosure Control". Available at http://neon.vb.cbs.nl/casc/handbook.htm.

Eurostat. 2009."Work Session on Statistical Data Confidentiality. Manchester 17-19 December 2007",in Methodologies and Working Papers. Available at http://www.unece.org/stats/publications/Proceedings_statistical_data_confidentiality.pdf.

Eurostat. Monographs of official statistics – Work session on statistical data confidentiality. Geneva, 9-11 November 2005. Available at http://ec.europa.eu/eurostat/ramon/statmanuals/files/KS-73-05-623-EN.pdf.

Templ, M., Meindl, B. and Kowarik, A. 2014. Introduction to Statistical Disclosure Control, Paper prepared for The International Household Survey Network, Vienna. Available at http://www.ihsn.org/home/software/disclosure-control-toolbox.

United Nations. 2007. Managing Statistical Confidentiality & Microdata Access. Available at http://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf.

United Nations Statistical Commission. 2007. Principles and Guidelines for Managing Statistical Confidentiality and Microdata Access, Thirty-eighth Session. Available in English only for dates between and including 27 February - 2 March 2007. Accessible at http://unstats.un.org/unsd/statcom/doc07/BG-Microdata-E.pdf.

3. Alternative Models for Providing Access to Microdata

Figure 7 above shows the different stages of a survey leading to Stage F. During this Stage, a clean file is produced, on the basis of which dissemination can take place. This file is also known as the master file, as it contains the maximum amount of information that can be used for dissemination purposes. It is maintained by the data producer in a secure location, and is used to derive either aggregate or microdata products. This file is used to create versions that can be accessed by users, both directly and indirectly. Below is a brief description of the types of products and services that can be provided.

Public Use Files (PUFs): These files (which may be obtained from a survey or a sample of census records) undergo a rigorous statistical disclosure control (SDC) process, so that the chance of re-identifying respondents is minimal (the various SDC options are described in Section 5 below). However, it should be recognized that there is no such thing as a completely "safe" file. Indeed, most data producers require researchers to agree to certain conditions. For PUFs, this involves agreeing to a set of conditions published online, often referred to as a "click-through" agreement. An example of these conditions is provided in Appendix A to this Guide.

PUFs present a minimal amount of geographic detail below the national or larger subnational areas. In addition, all direct and indirect identifiers are removed. Certain records and variables may be suppressed, regrouped and recoded.

Pros: These files can be distributed broadly, to a wide range of users, who can access the data on their own premises with a minimal risk of disclosure. They are widely used to teach data analysis skills and to provide the bases for initial analyses of a topic. Researchers may use PUFs before attempting access to more detailed work.

Cons: These files lack geographic detail and certain other variables deemed too sensitive or revealing. Their production generally takes longer than for tables (although a statistical disclosure control must be performed on tabular data as well). Statistical and subject matter specialists must work together to create a file that can be considered safe enough for release.

Some disseminators of public use files allow users to freely download files from their website, while others require users to identify themselves by providing contact information. This requirement serves two purposes:

- it enables the disseminator to build a profile of their user community, which can be asked to provide important feedback when e.g. negotiating continued funding of the program; and
- it establishes a contact list that can be used to advise users of any changes made to the files.

BOX 2

Examples of Public Use Files

1) The Comprehensive Food Security and Vulnerability Analysis and Nutrition Survey 2012 from Rwanda is a good example of a households survey that focuses on agriculture and food. The survey is available at http://catalog.ihsn.org/index.php/catalog/4149. Users are required to register; then, they may proceed to download and analyse the data.

2) **The Nigerian General Household Survey – Panel 2010-2011 (Post-Planting), First Round (Wave 1)** is another good example of a survey with agriculture-related questions. Users must register to download the file. See http://www.nigerianstat.gov.ng/nada/index.php/catalog/31.

3) The United States has a long history of distributing public use files, beginning with the 1961 population census. This was distributed on computer punch cards and UNIVAC tapes. The current US approach for the **American Community Survey** (which replaced the census' long form) is to make it freely available for download. Users must access the website http://www.census.gov/acs/www/; the **Data and Documentation** tab lists the relevant files and documentation. Note: in the US, Public Use Files are called Public Use Microdata files (PUMS).

4) Similarly, **the British Household Panel Survey** is available for download after users register with the website, http://discover.ukdataservice.ac.uk/series/?sn=200005. Registration is a common condition for providing access to microdata.

5) The **Afrobarometers** are available at http://www.afrobarometer.org/ where SPSS files can be downloaded without registration. The access and usage policy is published on the website: http://www.afrobarometer.org/data/data-usage-policy.

6) The **German National Social Science Infrastructure Service** (GESIS) makes a variety of data sets available once users register. The Eurobarometers may be accessed at http://www.gesis.org/en/eurobarometer/data-access/. The registration form is available at https://dbk.gesis.org/register/register.asp?db=E.

7) The **Multiple Indicator Cluster Survey** (http://www.childinfo.org/mics2_datasets.html) provides access to authenticated users. Authentication is not the same as licensing, but requires the user to respect certain conditions.

8) The **Demographic and Health Survey** (http://www.measuredhs.com/data/Access-Instructions.cfm) requires users to register.

Requesting contact information is not the same as requiring the users to sign a license.

BOX 4

Example of conditions users must accept before using microdata files
[Department: The Department of Census and Statistics (DCS) for Sri Lanka
Source: http://www.statistics.gov.lk/databases/data%20dissemination/DataDissaPolicy_2007Oct26.pdf.]

Confidentiality
Under the Statistical ordinance, microdata cannot be released with identifications [of respondents] for public use. Procedures are in place to ensure that information relating to any particular individual person, household or undertaking will be kept strictly confidential and will not be divulged to external parties. Information on individual or individual Household/establishment will not be divulged or published in such a form that will facilitate the identification of any particular person or establishment as the data have been collected under the Census/Statistical ordinance, according to which the information at individual level cannot be divulged and such information is strictly
confidential.

Access conditions
The data set has been anonymized and is available as a Public Use Dataset. It is accessible to all for statistical and research purposes only, under the following terms and conditions:

1. The data and other materials will not be redistributed or sold to other individuals, institutions, or organizations without written agreement.

2. The data will be used for statistical and scientific research purposes only. They will be used solely for reporting of aggregated information, and not for investigation of specific individuals or organizations.

3. No attempt will be made to re-identify respondents, and no use will be made of the identity of any person or establishment discovered inadvertently.

4. No attempt will be made to produce links among data sets provided by the Department or among data from the Department and other data sets that could identify individuals or organizations.

5. Any books, articles, conference papers, theses, dissertations, reports, or other publications that employ data obtained from the Department will cite the source of data in accordance with the Citation Requirement provided with each data set.

6. An electronic copy of all reports and publications based on the requested data will be sent to the Department.

BOX 4

The following rules apply to microdata released by the Department of Census and Statistics.

• Only the requests of Government Institutions, Recognized Universities, Students, and selected international agencies are entertained. However, the Data users are required to strictly adhere to the terms stipulated in the agreement form.

• All the data requests should be made to Director General (DG) of the DCS as the sole authority of releasing data is vested with the DG of the DCS. The DCS of Sri Lanka reserves sole right to approve or reject any data request made depending on the confidential nature of the data set and intended purpose of the study or analysis.

• Requests for microdata should be made through the agreement form designed by DCS for this purpose (Form D.R.1). The agreement form should be filled in triplicate and the Study/project proposal should accompany the filled agreement form. If requests are made for the microdata of more than one survey, a separate agreement should be signed.

• If the data request is from a student, a letter from the respective Dept. Head/Dean/Supervisor, recommending the issue of data, should also be accompanied.

• If the request is approved only 25% of the data file is released at the first stage. The release of the total data file is considered only after reviewing the draft report prepared on the basis of the 25% sample data file.

• The released Data file should be used only for the specific Study/Analysis mentioned in the agreement form and shall not be used for any other purpose without the prior approval of the Director General of the DCS. Moreover, Copies of the microdata file, obtained from the DCS, shall not be given to anyone else without the prior written approval of the Director General of the DCS.

• The draft report of the Study/Analysis should be submitted to the DCS and the concurrence of the DG of the DCS, should be obtained before publishing it. Once published, a copy of the final report should be submitted to the DCS.

Licensed Files: Licensed files are also anonymized, but fewer SDC procedures may be applied, depending upon the nature of the file and the producer's policies. Thus, more detail may be provided. The data producers ask researchers to identify themselves and provide explicit details on their research. The researchers are required to sign a license identifying who can access the file and the conditions of its use. The researchers' organizations may also impose certain undertakings.

"Licensing agreements permit a researcher to use confidential data offsite, but under highly restricted conditions as spelled out in a legally binding agreement. Arrangements that place restrictions on who has access, at what locations, and for what purposes access is allowed, normally require written agreements between

agency and users. These agreements usually subject the user to fines, being denied access in the future and/or other penalties for improper disclosure of individual information and other violations of the agreed conditions of use. Users may be subject to external audits conducted by the agency to assure terms of the agreement are being followed.

Users in violation may be required to pay fines or be subject to other legal penalties."[7]

Pros: Since researchers agree to conditions that restrict manipulation of the files, licensed files **may** contain more detailed information than PUFs. Unless the country's specific laws provide otherwise, the only sanction that can be imposed upon those who breach the license conditions is the removal of access privileges from users and their organizations. Data producers generally feel that there is less risk associated with a licensed file than with a PUF, as there is greater interaction with the user and opportunities to ensure that the importance of the conditions is understood. Requiring researchers' organizations to sign an undertaking provides greater security for the data producers, as most organizations do not want to risk sanctions.

Cons: The cons of licensed files are similar to those regarding PUFs. From the users' perspective, obtaining access requires more time because of the licensing process. These files will still lack geographic detail and some variables that are deemed sensitive or revealing in relation to the respondents. As with PUFs, these files generally take longer to produce than tables. The skills required to produce the files are similar to those for PUFs. In addition, staff members must be allocated to managing the licensing process.

[7] United States of America. 2005. US Federal Committee on *Statistical Methodology. Statistical Policy Working Paper 22 (Revised 2005) – Report on Statistical Disclosure Limitation Methodology*. Available at: http://www.fcsm.gov/working-papers/SPWP22_rev.pdf. Accessed on 06 January 2014.

BOX 5

Examples of Licensed Files:

The examples below illustrate a variety of approaches to providing access to microdata.

1) The **USDA's Agricultural Resource Management Survey (ARMS).** See http://www.ers.usda.gov/data-products/arms-farm-financial-and-crop-production-practices.aspx#.UvpaffldWh0**.** This is an example of an agricultural survey conducted in a country with a well-developed agricultural sector, and that has found a way to release microdata files covering the entire sector, including large commercial farms. Access is restricted to government agencies and university researchers in the United States. Users are required to enter into a Memorandum of Understanding with the USDA. See http://www.ers.usda.gov/data-products/arms-farm-financial-and-crop-production-practices/contact-us.aspx#Raw and navigate to **Procedures for Requesting Access to Raw ARMS Data**.

2) **IPUMS:** International (Integrated Public Use Microdata Series, International) is the world's largest collection of publicly available individual-level census data. The data are samples from population censuses from around the world, since 1960. Users are required to register and may use the data for scholarly and educational purposes, including policy analysis. See https://international.ipums.org/international-action/faq#ques6.

3) Statistics Netherlands: "For its social sample surveys Statistics Netherlands (*Centraal Bureau voor de Statistiek*, or CBS) releases about ten standard microdata files each year. The microdata are protected against disclosure but not to the last detail. The remaining risk is dealt with by a contract (or license). The microdata are available to legitimate researchers. They are released on tape or on disk, usually in the SPSS format." This statement is available at http://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf; see p. 41 of this file: Annex 1.6. Case Study Release of Licensed Microdata Files – Netherlands.

4) Statistics Canada uses a similar approach to that adopted in the Netherlands. The 75 universities that participate in the Data Liberation Initiative (DLI) can sign one license, which enables access for all students to all publicly available files. Individual users outside the DLI must sign a license for each file. The files are free, but the license ensures that users will abide by the conditions. See http://www.statcan.gc.ca/dli-idd/caselaw-jurisprudence/license-licence-eng.htm. Recent changes to data policy pursuant to an Open Data initiative enables the commercial use of these files.

5) The Australian Bureau of Statistics provides various avenues of access to its microdata files. Confidentialised Unit Record Files (CURFs) are files containing the responses given to ABS surveys; they contain highly detailed data, while also protecting individual respondents' confidentiality. Users must register to be licensed. See http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Microdata+Entry+Page and http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Registration+Centre.

6) Accessing Eurostat microdata. Access to Eurostat data is given to accredited researchers for scientific purposes, if their institution is a recognized research entity: see http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/introduction.

Remote Access Facilities (RAFs): RAFs feature a service window, provided by the data producers, which allows researchers to supply the algorithm that they will use in their analysis. Researchers are provided with a synthetic file that replicates the structure and content of the actual data sets. Researchers can then develop programs and procedures using tools such as SAS, SPSS, STATA or R. The programs can be transmitted to the data producer's staff, which then runs the job against the actual data set and vets the results for disclosure before returning the output to the users.

There are two types of RAF. The first involves remote execution: researchers submit a program and receive the vetted output over the Internet. The second type involves the submission of data requests into an interactive system, which can produce tabulations and vet the data "on the fly". These data requests require staff intervention only if the access software detects disclosure issues.

Pros: Both types of RAF effectively ensure fulfilment of the data producer's confidentiality requirements. The use of synthetic files informs researchers of the full structure and content of the data set.

Cons: This type of service requires the data producers' staff to be available for submitting jobs and vetting results. This is an expensive process, for which the costs must likely be recovered, and users may find it slow. Interactive services require resources to develop the systems. These systems require less staff time. Many applications of RAFs can only produce tabulations, while others may only enable analytic outputs.
Users should be able to specify models, using synthetic files to be run against the detailed files. However, running models on the data set often requires a great deal of iteration, which would be impossible to perform in an RAF environment because this could lead to residual disclosure. It should be noted that a risk of residual disclosure arises when the comparison of successive retrievals can isolate individual respondents.

BOX 6

Examples of Remote Access:

1) Australian Bureau of Statistics – the Remote Access Data Laboratory (RADL). The RADL is a secure online data query service that approved clients may access via the ABS website. Within the RADL, users submit queries in SAS, SPSS or Stata analytical languages against Expanded CURFs maintained within the ABS environment. The query results are automatically checked, and then made available to users. Since the CURFs are kept within the ABS environment, the ABS can release more detailed CURFs in the RADL than may be available on CD-ROM. (Note: CURFs are Confidentialised Unit Record Files, the name given to PUFs in Australia). See http://abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+Remote+Access+Data+Laboratory+(RADL).

2) Statistics Canada – the Real Time Remote Access (RTRA) system. The RTRA system is an online remote access facility that enables users to run SAS programs, in real time, against microdata sets located in a central and secure location. Researchers using the RTRA system do not gain direct access to the microdata and cannot view the content of the microdata file. Instead, users submit SAS programs to extract results in the form of frequency tables. As RTRA researchers cannot view the microdata, becoming a deemed employee of Statistics Canada is no longer necessary. This relationship is the basis that enables the RTRA to service its clients rapidly. See http://www.statcan.gc.ca/rdc-cdr/rtra-adtr/inf-eng.htm.

Data enclaves: A data enclave consists of a facility within the premises of the statistical organization that researchers can frequent to perform their research on detailed files. These files are the most detailed files available to researchers, other than the master files themselves. A data enclave is equipped with computers that are not linked to the Internet or to an external network. Information cannot be downloaded using USB ports or written onto a CD/DVD drive. A data enclave could contain a complete agricultural census or an agricultural survey data set. Users must identify the part of the data set that they wish to consult; only that data subset is made available to them. The results produced by the researcher must be vetted by a statistical organization staff member before they can be removed from the premises. Researchers must state specific goals before they are allowed to perform research in the data enclave. Commonly, statistical organizations require researchers to identify their objectives and to demonstrate a legitimate need for access to these data. They also require the proposed research to be consistent with the organization's objectives[8].

Pros: A major advantage of data enclaves is the amount of detail that can be provided to researchers. Compared to RAFs, data enclaves enable researchers to conduct complex analyses.

Cons: The major disadvantages of data enclaves are that they are the most expensive form of access for statistical organizations; also, they are inconvenient for researchers, who must travel to a central location, which may be in a different city.
Data enclaves are generally operated under a cost-recovery basis unless the SO receives a grant financing their operation.

Deemed Employee: A final model for consideration is swearing in the researcher to work with the agency as a temporary staff member. In these cases, researchers would be subject to the same secrecy and ethical provisions as regular staff. This employment can also take the forms of research fellowships and post-doctoral programs. Deemed employment is generally limited to projects that assist the data producer to meet organizational goals, and for which it does not possess the necessary skills. There is little documented evidence as to which statistical organizations use this approach, but there is anecdotal evidence that this occurs in a number of countries.

Pros: this model has the advantage of providing researchers with access to the master file in its full detail. The data producer's staff is available to answer questions on the data. This model protects the data's confidentiality and furthers the organization's goals.

[8] Statistics Canada distinguishes between academic and government researchers. An outline of the application requirements are available at "Application process and guidelines." Statistics Canada. Available at: http://www.statcan.gc.ca/rdc-cdr/process-eng.htm. Accessed on 23 May 2014.

BOX 7

Examples of data enclaves (also known as Research Data Centres – RDCs)

1) Australian Bureau of Statistics. The ABS Data Laboratory (ABSDL) is the data analysis solution for high-end data users who wish to extract the full value from ABS microdata. ABSDL enables analysis of Basic, Expanded or Specialist (customized) Confidentialised Unit Record Files (CURFs). ABSDL provides a more responsive and interactive environment in which to analyse CURFs than that offered by the Remote Access Data Laboratory (RADL). See http://abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+About+the+ABS+Data+Laboratory+(ABSDL).

2) Statistics Canada. The Research Data Centres Program provides researchers with access to microdata from population and household surveys, in a secure university setting. The RDCs are staffed by Statistics Canada employees and are operated under the provisions of the Statistics Act, in accordance with all relevant confidentiality rules; they are accessible only to researchers with approved projects who have been sworn in under the Statistics Act as "deemed employees". RDCs are located throughout the country; therefore, researchers need not travel to Ottawa to access Statistics Canada's microdata. See http://www.statcan.gc.ca/rdc-cdr/index-eng.htm.

3) Statistics Canada. The Canadian Centre for Data Development and Economic Research (CDER) is the repository for business and economic microdata files that contain sufficient detail for researchers to undertake meaningful and complex analyses. At the same time, the confidentiality of the results and the privacy of business respondents and workers are safeguarded (see http://www.statcan.gc.ca/cder-cdre/). The Statistics Canada CDER project is used by the Canadian Ministry of Agriculture and Food to access the files relating to 7 quinquennial agricultural censuses (1986 through 2011). Since the farms can be linked over time, the file is longitudinal in nature and can track the evolution of farms over time. The files are placed in a database held and maintained at the Centre, and a synthetic file is created with perturbed data from the subset of the variables of interest to the researcher. The data visible to researchers are stripped of identifying information such as names, contact information, business numbers, detailed geography and industry. The synthetic file is used to develop the analytic routines which are then brought into the Centre for running against the full file, under the supervision of Centre staff. The main outputs are analytic rather than tabular.

4) United States. The Center for Economic Studies (CES) Economic data refer to the Economic Census of establishments and various surveys and data for establishments and firms. With very few exceptions, the publicly available versions of these files contain only data presented in aggregate form. A list displays the establishment- and firm-based data available, the survey period, the frequency of data collection, the years of data available at the CES, and the sponsoring federal agency. Access to these data is only granted to qualified researchers, for approved projects, with an authorization to use specific data sets. All researcher access to restricted–use data can take place only at one of the secure Census Research Data Centers. http://www.census.gov/ces/dataproducts/economicdata.html. The main outputs are analytic rather than tabular.

Note: The latter two enclaves are examples of bodies that allow users to access establishment data.

Figure 8 below illustrates the relationship between the different models of access. The clean edited master file (Box A) derives from Stage F of the survey cycle (see Figure 7 above). Box B represents the procedures used to anonymize microdata files. Section 5 describes this process in further detail.

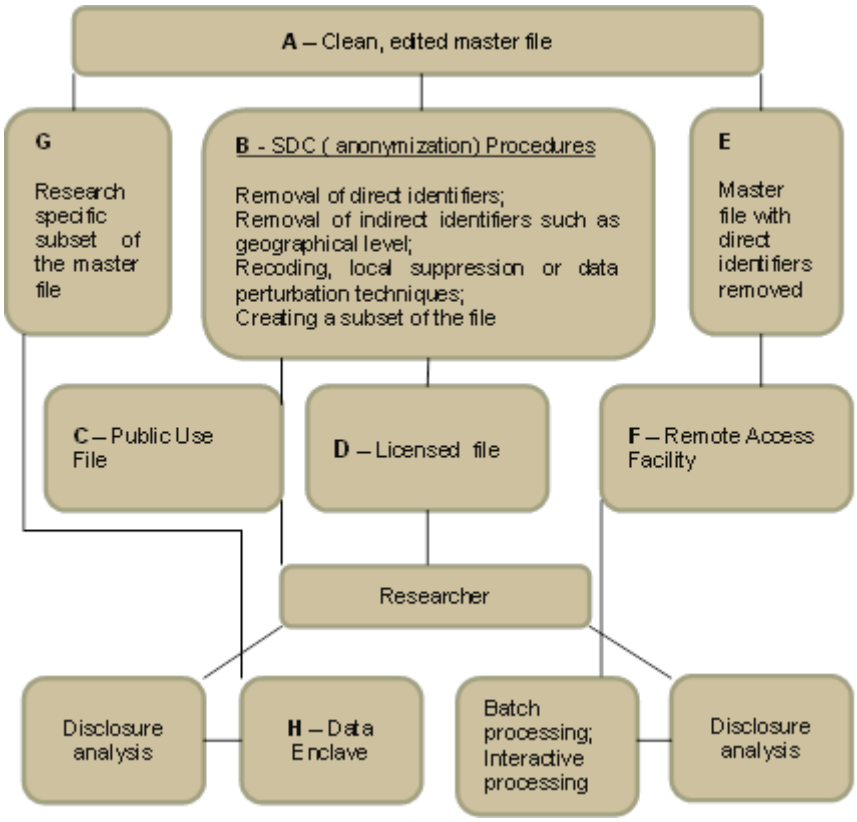


Figure 8: Creating Microdata Files for Researcher Access

(SDC(many techniques can be used, SOs need not apply any more than those necessary to mask the data. The file must be analysed at each stage, to determine whether unique records can be found. If so ,it must be determined whether such records can result in the identifiability of a given respondent. Additional SDC techniques can be applied as required)anonymization) Procedures: Anonymization is the key to releasing a microdata file. While

After an anonymized file has been produced, it can be made available to researchers as either a Public Use File or as a Licensed File (Boxes C and D). To support research performed using a Remote Access Facility (Box F), a special version of the master file is produced. This is cleared of direct identifiers, to avoid spontaneous recognition of a respondent. After the researchers have completed their analysis and produced tabular or analytic output, this is subjected to a disclosure analysis before it can be accessed by the researchers.

In a data enclave (Box H of Figure 8 above), researchers will be provided with a subset of the master file (again cleared of direct identifiers), which will serve the specific research needs outlined in their application to perform research in the enclave. Researchers sworn in as employees could work directly with the master file.

A good overview of the different methods of providing access to microdata can be found on the website of the Inter-university Consortium for Social and Political Research (ICPSR). The ICPSR manages several thousands of files and provides access to these records to researchers around the world.
See http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter5.html.

Comparison of different methods of accessing microdata						
Type of access	**Effort required to create file**	**Level of detail**	**Method of Distribution**	**Type of analysis**	**Turnaround time**	**Convenience for researchers**
Public Use File	High. The SDC must be implemented.	Direct identifiers are removed; some data are masked; variables may be removed.	CD, DVD, SO website or NADA catalogue at the IHSN	All types of analysis are possible. Analysis may be limited by suppression of variables or cases.	Real time, on the users' computer.	High
Licensed File	High. The SDC must be implemented.	Direct identifiers are removed; some data are masked; variables may be removed.	CD, DVD, SO website or NADA catalogue at the IHSN	All types of analysis are possible.	Real time, on the users' computer.	**High.** Signing the license requires more time than accessing a PUF from a website.
Remote job submission	File takes less time to prepare than for PUFs; a remote access server must be set up.	Direct identifiers are removed; sensitive variables may be removed.	Jobs must be submitted to the SO server; results are received after a confidentiality review.	Most systems only permit tabulations.	Review of the output requires time, in terms of hours or days, depending upon the level of service offered.	**Medium**. Access is from one's own office; turnaround may be an issue; analysis may be limited to tabulations.
Interactive remote	File preparation takes less time than for PUFs; a remote access server must be set up; a process for screening output on the fly must be developed.	Direct identifiers are removed; sensitive variables may be removed.	Jobs must be submitted to the SO server.	Some organizations only offer tabulations, while others permit regression analysis.	Output is reviewed on the fly.	**High.** Access is from one's own office, and will meet all research needs only if all forms of analysis are permitted.
Enclave	File preparation time is minimal. The physical space, the workstation, the software and the security required must all be in place. The output leaving the enclave must be screened.	Direct identifiers removed; sensitive variables may be removed.	Access is on the SO's premises.	Mainly analytic models, with a minimum possibility to produce tabulations.	Output is reviewed by a disclosure analyst before it can be removed from the premises.	**Low**, because it requires the user to travel to the SO's location.

Type of access	Effort required to create file	Level of detail	Method of Distribution	Type of analysis	Turnaround time	Convenience for researchers
Deemed employees	Full access to master file; no additional preparations is necessary; al the producers of the SO must be followed	Full access to the master file	Acess is on the SO's premises (including regional offices(	All types of the analysis; users must abide by all the rules of the SO	All analysis done in real time	Low, because it requires the user to be located on the SO's premises

TABLE 4: Comparison of the Different Methods of Accessing Microdata

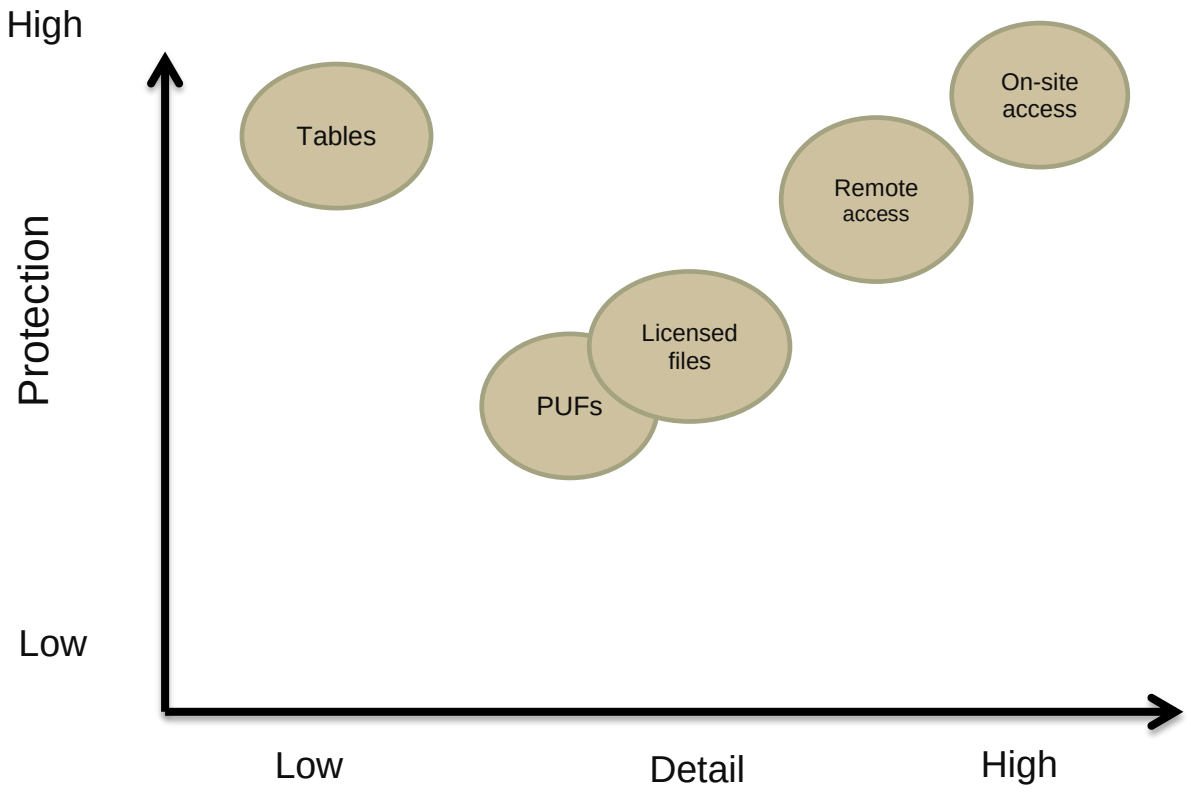


FIGURE 9. Trade-off between Levels of Protection and Detail in Microdata Files

(PUFs – Licensed Files: there may or may not be a difference in detail between a PUF and a

Licensed File, depending on the SO. Some may take advantage of the fact that a licensed file provides more protection by placing a gr a greater onuson the user to respect the license conditions)

A diagrammatic representation of the trade-off between the level of detail contained in a microdata file and the degree of protection (ensuring that individual data cannot be re-identified). It should be noted that confidentiality is a concern for published tables as well as for microdata.

4. Preparing Data Files for User Access

Statistical organizations must carry out two major tasks before microdata files can be used by researchers. The first task is to prepare the appropriate documentation (metadata) to ensure that the data can be understood. This is a necessary step even if the SO does not release microdata files, a part of good data management that ensures that the data can be preserved and reused in the future. For more information on data preservation, readers are referred to IHSN Working Paper No. 003, entitled *Principles and Good Practice for Preserving Data*[9].

The second task is to anonymize the data by carrying out various Statistical Disclosure procedures.

4.1 Metadata

4.1.1 What Are Metadata?

The definition of metadata was extensively addressed in Dupriez and Boyko, 2010[10]. The definitions below are drawn from this work.

"Metadata are usually defined as 'data about data'. The previous chapter mentions the importance of providing users with a proper data dictionary describing the content of all variables included in a dataset. But good metadata contains much more than a data dictionary.

Metadata are intended to help researchers understand what the data are measuring and how they have been created. Without a proper description of a survey's design and the methods used when collecting and processing the data, there is a significant risk that the user will misunderstand and even misuse them. Good documentation also

[9] Inter-University Consortium for Political and Social Research (ICPSR). 2009. *Principles and Good Practice for Preserving Data"*, IHSN Working Paper No. 003, International Household Survey Network. Available at http://www.ihsn.org/home/sites/default/files/resources/IHSN-WP003.pdf. Accessed on 17 July 2014.

[10] Dupriez, O. & Boyko, E. 2010. *Dissemination of Microdata Files. Formulating Policies and Procedures*, International Household Survey Network, IHSN Working Paper No. 005. See, in particular, Section 2 of this Report. Available at: http://www.surveynetwork.org/home/sites/default/files/resources/IHSN-WP005.pdf . Accessed on 17 July 2014.

reduces the amount of user support statistical staff must offer external users of their microdata.

Metadata are also intended to help users assess the quality of data. Knowledge of data collection standards – as well as of any deviations from the planned standards – is important to any researchers who wish to know whether particular data are useful to them.

Lastly, metadata are needed to develop data discovery tools, such as survey catalogues that help researchers locate datasets of interest. Note [that] data not intended for dissemination must also be fully documented. Producing good metadata helps build the institutional memory of data collection, and can assist in training new staff and improving data consistency over time."

Metadata are best applied throughout the survey cycle, as a means of ensuring data quality and the preservation and reuse of data. Without metadata, users may not be able to understand the meaning of the data. While metadata are important to support the dissemination of all products, they are especially important for microdata files. Without codebooks, users cannot use the microdata files.

4.1.2 Producing and Applying Metadata

When constructing survey level metadata, it is best to consider standards and procedures that are capable of supporting a number of different objectives:

- Catalogues, to help users find the appropriate data set on the SO website and to manage current and historical survey files. Systems for catalogues are less important if there is only a limited number of files for users to choose from; however, this list will undoubtedly grow over time.
- Machine-readable codebooks, which facilitate use of the data set with popular statistical software.
- Software-independent standards for long-term preservation.

These are the objectives considered in establishing different metadata standards for household and establishment surveys. Several guides and best practices exist on the creation of metadata and the application of the various standards[11]. For the purposes of this Guide, one standard and its application will be examined in detail. The Data Documentation Initiative (DDI) embodies all the characteristics necessary to achieve the above objectives.

A brief description of the five sections of the DDI (Version 2.4) is presented below.

Section 1.0: Document Description
section includes information (metadata) not only on the study itself, but also on the documentation process. The documentation may be compiled by a different

[11] Some of these are outlined in Dupriez and Boyko, 2010, *ibid.*

organization from that which carried out the survey.

Section 2.0: Study Description
The Study Description consists of an overview of the survey or study

Section 3.0: Data File Description
This section describes the content of each data file, e.g. record and variable counts, version, producer, *etc*.

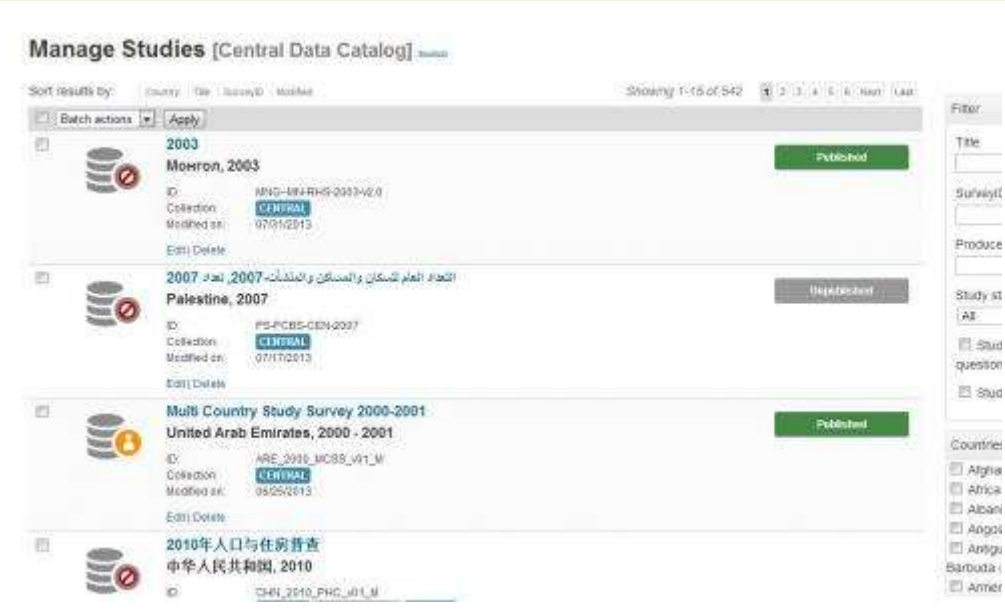
Section 4.0: Variable Description
This section presents details of each variable, including literal question text, universe, variable and value labels, derivation and imputation methods, *etc.*

Section 5.0: Other Material
This section allows for a description of other material related to the study (e.g. questionnaires, coding information, technical and analytical reports, interviewers' manuals, data processing and analytical programs, photos and maps).

Box 8

Example of a Database with Rich Metadata NADA: International Household Survey Network, Microdata Cataloguing Tool

NADA is a web-based cataloguing system that allows users to browse, search, compare, apply for access to, and download relevant census or survey information. It uses the Data Documentation Initiative (DDI), an XML-based international metadata standard. The system enables publishing of DDI documents on the Internet, creation of a rich web interface for managing DDI-based codebooks, performance of searches using rich metadata and the distribution of external resources and microdata.



Box 9

Examples of Metadata Records

Canadian Census of Agriculture
http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3438
United Kingdom Data Archive
http://www.data-archive.ac.uk/create-manage/document/metadata
International Household Survey Catalog for the Third Integrated Household Survey 2010-2011, Malawi
http://catalog.ihsn.org/index.php/catalog/2301

4.2 Preparing the data files

Preparation of the data (as opposed to the metadata) files consists of the application of Statistical Disclosure Control processes. These processes must be appropriate for the method of microdata access chosen for the project.

4.1.1 Statistical disclosure

A **disclosure** occurs when an individual using a microdata file recognizes or learns something not previously known to him/her about a survey respondent. This can happen in two ways. If a **direct identifier** is left in the file (e.g. a name, telephone number or address), from which the respondent's identity can be learned, **identity disclosure** occurs. Disclosure can also take place if an attribute (e.g. a large farm) can be directly associated with a particular respondent; this is known as **attribute disclosure**.

For example, if a record in a survey file contains agricultural information, and the head of the household is a young widow who owns a large land area, then individuals with some knowledge of the region would probably be able to identify that person. This would contrast with the requirements of statistics confidentiality legislation, and thus the record would have to be modified.

Residual disclosure is yet another form of risk that must be considered. This occurs when successive retrievals from a file can be compared (subtracted) to isolate a respondent's value. For example, if the first retrieval contains a grouping from 1-100, and a subsequent retrieval by the same user is for the group 1-99 and the two retrievals are compared, the individual value can be identified by subtraction. This may also occur when one data retrieval is compared to a table published previously.

4.2.2 What does Statistical Disclosure Control (SDC) mean?

SDC refers to the process of ensuring that the confidentiality requirements governing the SO's work are met, and that the risk of revealing information about the respondent is minimal. This is also referred to as anonymization. In previous pages of this Guide, it was mentioned that completely "safe" data is impossible to obtain. Thus, the risk of disclosure must be weighed against the benefits of access. In this evaluation, many factors must be considered:

- The sensitivity of the data
- The existence of external sources of information that can be used to attempt re-identification of respondents
- Whether the microdata file is a sample (e.g. a subsample of a census file)

Naturally, this will vary according to the country under examination.

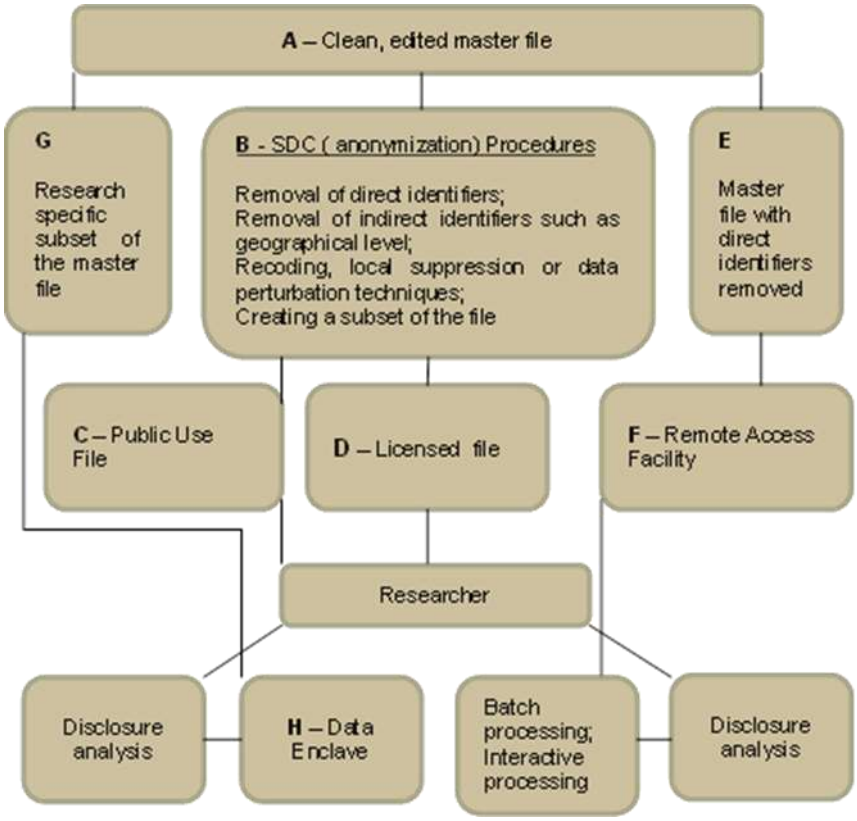
Sometimes disclosure can occur based on the released data alone; other times disclosure may result from combining the released data with publicly available information; and sometimes disclosure is possible only through combining the released data with detailed external data sources that may or may not be available to the general public[12].

4.2.3 SDC Techniques

This subject is treated extensively in Dupriez and Boyko, 2010. Further detailed information may be found in Chapter 7 thereof.

SDC will be discussed at a general level, with reference to Figure 8 of Section 3 of this Guide, reproduced below for readers' convenience.

[12] United States. US Federal Committee on Statistical Methodology. 2005. *Statistical Policy Working Paper 22 (Revised 2005) – Report on Statistical Disclosure Limitation Methodology*. Available at: http://www.fcsm.gov/working-papers/SPWP22_rev.pdf. Accessed on 06 January 2014.



Reproduction of Figure 8, Section 3, on Creating Microdata Files for Researcher Access

Preparing a file for dissemination is a five-step process:

1. At the most basic level, direct identifiers[13] should be removed from ALL types of microdata files prepared for use by others. This also applies to data enclaves (H) and remote access facilities (F) where the research results are examined by an analyst or by a software routine, to prevent researchers from spontaneously recognizing respondents.

2. After direct identifiers have been removed, indirect identifiers must be addressed. These can lead to disclosure when there are respondent attributes that can identify the respondent, e.g. farm units with very large areas, or rare crops or animals. This is generally done by conducting a large number of frequencies and cross-tabulations to identify unique records and variables.
3. The next step is to evaluate the unique records and values found in Step 2 and decide the relevant action to take in their regard. Not all unique values can lead to re-identification; therefore, subject specialists will have to provide advice on this matter.

[13] I.e. names, addresses, telephone numbers, detailed location of agricultural units.

4. The penultimate step is to apply an anonymization technique to the data. Care must be taken to ensure that the technique chosen is appropriate to the type of variable under consideration, e.g. discrete, continuous, categorical, etc.

5. The final step is to evaluate the file for utility and information loss. While an anonymized file cannot have the same structure as the original file, the file should be compared to benchmarks derived from the original file for key statistics.

A summary of the approaches that can be used to anonymize files is provided by the American Statistical Association, and presented below:

- **"Aggregation:** coarsen categorical data, e.g., do not release geographic units under 100,000 people
- **Top and bottom coding:** report values (e.g., incomes, ages) above thresholds only as "above the threshold" or creating catch-all categories for small values
- **[Local] suppression:** make sensitive values missing in the released file
- **Data swapping:** switch one record's values on key variables with another record's values
- **Noise addition:** fuzz data values by adding randomly generated values to sensitive real data values
- **Micro-aggregation:** cluster numerical data (e.g., incomes) in groups of at least three records, and replace each cluster member's value with the average value in its cluster
- **Multiple imputation** for disclosure limitation (also called synthetic data): replace sensitive values with simulated values drawn from statistical models"[14]

Just as there are no "safe"[15] data sets, there is also no best way to anonymize a file[16]. Section 5 of this Guide will focus on the challenges encountered when providing access to agricultural microdata.

[14] American Statistical Association. Committee on Privacy and Confidentiality. N.d. *Overviews of Statistical Disclosure Protection Methods*. Available at: http://community.amstat.org/CPC/Methods//MethodsB_Overviews. Accessed on 6 January 2014.

[15] Safe data refers to microdata files that have been anonymized to the degree that there is very little risk of re-identification.

[16] A good description of anonymization methods is available in Section 3 of the Guide to the SDCMicro tool: Matthias, T., Meindl, B. & Kowarik, A. 2014. *Introduction to Disclosure Control*. Data-Analysis OG Publication: Vienna. Available at: http://cran.r-project.org/web/packages/sdcMicro/vignettes/sdc_guidelines.pdf. Accessed on 15 April 2014.

4.2.4 Managing the Trade-off Between Statistical Disclosure Control and Information Loss

It is important to assess the impact that SDC techniques may have on the file in terms of information loss. Suppressing records and altering the structure of the file results in less risk, but at the expense of less information being provided, which is detrimental to the research conducted upon the file. This is particularly critical if researchers cannot replicate the indicators published by SOs in their official reports[17].

BOX 10

Guidelines from ICPSR
http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/index.html

Removing Identifiers
Two kinds of variables often found in social science datasets present problems that could endanger research subjects' confidentiality. Some variables point explicitly to particular individuals or units. Examples of direct identifiers include:

- Names
- Addresses, including ZIP codes
- Telephone numbers, including area codes
- Social Security numbers
- Other linkable numbers such as driver license numbers, certification numbers, *etc*.

All variables directly identifying research subjects must be removed or masked prior to deposit. Variables that can also be problematic are the **indirect identifiers** that may be used in conjunction with other information to identify individual respondents. Examples of indirect identifiers include:

1. Detailed geographic information (e.g., state, county, or census tract of residence)
2. Organizations (to which the respondent belongs)
3. Educational institutions (from which the respondent graduated and year of graduation)
4. Exact occupations
5. Place where respondent grew up
6. Exact dates of events (birth, death, marriage, divorce)
7. Detailed income
8. Offices or posts held by respondent

ICPSR may recode data to remove the threat of disclosure. Recoding can include converting dates to time intervals, exact dates of birth to age groups, state of residence to regional codes, and income-to-income ranges or categories.

ICPSR staff work closely with data depositors to resolve confidentiality issues. ICPSR strongly recommends that data producers remove all respondent identifiers before they deposit their data in the archive. For more information, see Phase 5: Preparing Data for Sharing in the ICPSR Guide to Social Science Data Preparation and Archiving, 5th Edition.

If removing all identifiers will unacceptably reduce the analytic utility of the data, depositors should contact ICPSR about releasing a restricted-use dataset. Restricted-use datasets retain confidential information so investigators must meet stringent requirements to access them.

Box 11

Surveys with Geospatial Coordinates

Some surveys may include GPS coordinates as part of the recorded information relating to area sample segments or plot locations. If the precise GPS coordinates can be associated with a specific farm unit, then they must be removed. If the area segment is used to identify farm operators in an effort to compensate for an incomplete list frame, then the farm identified should be treated in the same way as those identified from the list, but without any GPS coordinates attached. If the GPS coordinates are used to identify plot locations for yield or ground cover surveys, the files that contain them are unlikely to be used as microdata files.

Box 12

SDCMicro

The International Household Survey Network has recently released a software tool called Statistical Disclosure Control (SDCMicro). "*SDCMicro is a free, R-based open-source software for the generation of protected microdata for researchers and public use. The package provides multiple options for reducing the statistical disclosure risk in categorical or continuous variables. SDCMicro can be used from the R command line interface or by using the application's graphical user interface (GUI). The package can also be used in batch-mode from other software."* This tool will help statisticians to anonymize their files and, at the same time, preserve the maximum amount of information for researchers. It is envisaged that technical support to assist SOs in using this tool will be offered from 2014.

See http://www.ihsn.org/home/software/disclosure-control-toolbox.

Box 13

Eurostat Handbook

Eurostat has prepared an extensive handbook on file anonymization.
The Manual on Disclosure Control Methods is widely used as a reference tool by statisticians who work on file anonymization.

See
http://ec.europa.eu/eurostat/ramon/statmanuals/files/manual_on_disclosure_control_methods_1996.pdf

5. Providing Access to Agricultural Microdata

5.1 Data Access Framework for Agriculture

So far, this guide has addressed the following topics:

- What are microdata?
- The rationale for producing microdata
- Alternative models for providing access to microdata
- Preparing microdata files for access

Most of this material is general and is based on the experience gained from providing access to household survey data. However, an important question remains: which microdata systems are most appropriate for agriculture? Agriculture presents many characteristics of a household survey and of a business (establishment) survey. This makes it more challenging to access agricultural microdata.

A producer of agricultural data must evaluate the following approaches, when considering whether to release microdata for access by researchers:

6. Public Use File (PUF)
7. A licensed microdata file
8. Remote access (batch)
9. Remote access (interactive)
10. Data enclaves
11. Temporary employees

In the first two approaches, a file is released to the researcher, generally with conditions of use attached to its release. In approaches 3-5, the SO retains control over the file and much greater control over the results pulled from the data. Researchers using data enclaves and acting as temporary employees must be sworn in under the relevant statistics legislation, and take the same oath as regular employees.

Each of the access methods described in Section 3 will be discussed below, along with the implications of their use in the agricultural sector.

Access Method	Implementation	Assessment
1) Public Use Files	Microdata files (entire surveys or samples of census files) are subjected to rigorous disclosure procedures, possibly using a number of SDC procedures. A simple **"click-through" user agreement is placed on the website and users must "agree" before using the data. If the size distribution is skewed, the larger values must be omitted, top-coded or aggregated.**	This method is feasible only if the farm units are relatively homogeneous and the universe contains a large number of them. Since agricultural populations often consist of a commercial (or estate) sector and a household sector, this method can usually be applied only to the household sector. The commercial sector would have to be analysed separately from the microdata, as these units cannot be anonymized safely without impairing the usefulness of the file. In certain cases, it may be possible to obtain the permission of the respondents to share their data for specific studies. There may be some risk of disclosure in virtually all public use files.
2) Licensed microdata files	Microdata files (entire surveys or samples of census files) are also subjected to rigorous disclosure, possibly using a number of SDC procedures. A formal agreement (license or Memorandum of Understanding – MOU) must be entered into with the organization of *bona fide* researchers, who must complete an application specifying, *inter alia*, the purpose for which the data will be used and who will have access to it. Possible sanctions will be specified.	In practice, in terms of implementation, licensed files are very similar to public use files. The main difference is the existence of a license or an MOU. As with PUFs, there may be some risk of disclosure due to identifiable attributes or residual disclosure. The terms of the license agreement forbid users from disclosing respondents' information. In certain cases, it may be possible to obtain the permission of the respondents to share their data for specific studies. Data producers generally feel more assured when they have concluded agreements with users. An example is the US ARMS Farm Financial and Crop Production Survey. This is a unique example of a licensed microdata file that is made available to government departments and university researchers in the United States. This is available at http://www.ers.usda.gov/data-products/arms-farm-financial-and-crop-production-practices/contact-us.aspx#Raw.

Access Method	Implementation	Assessment
3) Remote Access (batch processing)	Users are given a synthetic data set that replicates the structure and content of the actual data sets. This enables researchers to develop programs using tools such as SAS, SPSS or Stata. The programs are then transmitted to the data producer's staff, who run the job against the actual data set. Finally, the tabular results are vetted for disclosure and returned to the user. An example of a user undertaking is the Australian Bureau of Statistics' Remote Access Data Laboratory Access. Further information is available at: http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/1406.0.55.003main+features70Sep%202009.	Agriculture files could be accessed by this method if identifiable values are properly masked. This process offers data confidentiality greater protection than the two methods seen above. There will usually be a limit to the procedures that can be run on the data. Some remote access systems may be used only for tabulations. Producing tabulations prompts the need to review the tables for disclosure risk. If repeated tabulations are produced, residual disclosure is also a risk, including possible comparisons with previously published data. The cost of supporting job submission services can be high; historically, users have found job turn-around from this process to be slower than desired. Cost recovery is generally applied. An example of this method is the Australian Bureau of Statistics' Remote Access Data Laboratory Access (see http://abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+Remote+Access+Data+Laboratory+(RADL).
4) Remote Access (interactive processing)	Users have access to web-based data tabulation and analysis software (or can use SAS, STATA, SPSS etc.), which is constrained so that users cannot download data sets or generate tables that would reveal identifying details of individuals (i.e., identifiable records cannot be extracted). Disclosure analysis is performed by software routines, but SO staff may have to intervene. Commercial software applications are often used (e.g. Nesstar, Beyond 20/20, SuperCross, Redatam). Some advanced data centres develop their own applications.	Agriculture files could be accessed by this method if identifiable values are properly masked. This process protects data confidentiality better than PUFs or licensed files. There will usually be a limit to the procedures that can be run on the data; very often, it can only be used for tabulations. If more complex procedures are run, there will be limits to ensure that individual records cannot be isolated. Residual disclosure may be a risk. These systems take less staff time but are more expensive to develop. Cost recovery is generally applied. An example is Statistics Canada's Real Time Remote Access System. Further details are available at: http://www.statcan.gc.ca/rdc-cdr/rtra-adtr/rtra-adtr-eng.htm.

Access Method	Implementation	Assessment
Data Enclave	Enclaves are part of the SO premises or are under complete SO control. They are a facility equipped with computers that are not linked to the Internet or to an external network, and from which no information can be downloaded via USB ports, CDs/DVDs or other drives. The data files are detailed, but direct identifiers have been removed. Users interested in accessing a data enclave must become deemed employees (see below) and will not necessarily have access to the full data set, but rather only to the particular data subset that they require. They will be asked to complete an application form demonstrating a legitimate need to access these data to fulfil a stated statistical or research purpose. Before release, the outputs generated must be scrutinized in a full disclosure review.	Agriculture-related files could be placed in such a facility. Confidentiality is protected. A full of range analytical procedures can be applied. Operating a data enclave is expensive, although it can be scaled to meet the expected demand. Data enclaves may require special premises and computer equipment. They also demand staff with the skills and time to review outputs before their removal from the data enclave, to ensure that there is no risk of disclosure. Such staff must be familiar with the content of the surveys and with data analysis, and be capable of reviewing request processes and managing file servers. Cost recovery is generally applied. This service is aimed at " high end" users. Examples are Statistics Canada's Research Data Centres program (see http://www.statcan.gc.ca/rdc-cdr/index-eng.htm) and Australia's Data Laboratory (http://abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+About+the+ABS+Data+Laboratory+(ABSDL). **Note:** Some SOs (e.g. Statistics Canada) do not put agriculture-related files or establishment survey data into their regular RDC program; they are currently working on a new approach for business data.

Access Method	Implementation	Assessment
On site Deemed Employee (temporary employees)	Researchers join the SO as temporary employees and have access to the files required, to perform research previously agreed with the SO. Any research material removed from the SO must be reviewed for confidentiality.	Confidentiality is maintained and access to the full data set is possible; however, researchers must be able to work at the SO premises, in their headquarters (or regional offices, if they exist).

5.2. Options for Accessing Agriculture Microdata – an Analysis

Providing access to agricultural microdata involves balancing the use of approaches that:

- can be feasibly implemented by the SO;
- minimize the risk of disclosure;
- provide sufficient data access to achieve research objectives; and
- do not impose undue constraints upon researchers.

From a researcher's point of view, the best outcome is a PUF or a licensed file with a minimal loss of data. The best approaches for protecting confidentiality are those which enable the SO to retain full control over the access to the files, and to vet the results before they leave the SO offices. This applies to the deemed employee, data enclave and remote access approaches. While disclosure analysts can vet the files for attribute and identity disclosure, they cannot monitor residual disclosure with similar ease. This is why some organizations place limits on the number of procedures and retrievals that can be performed. The data enclave and temporary employee approach provide the best access to the data, but may be less convenient for researchers, if these do not reside in the same location as the SO's headquarters or as one of their regional offices. Remote access is more convenient for researchers, but access to data is limited to the procedures that can be implemented in the remote access service. Tabulations are less powerful than regression and modelling techniques.

The use of licensed and Public Use Files offers the broadest access to data, as the geographical location of the researchers is not a barrier. However, the files will be less complete, due to SDC; indeed, from the SO's point of view, the risk of disclosure is greater. Confidentiality management depends on the SO's ability to evaluate requests for licenses, the effectiveness of the license agreements and the feasibility of sanctions as deterrents to attempting the re-identification of respondents.

Creating microdata files for a complete agricultural survey file may not be feasible when the distribution of farms is such that large (commercial) operators cannot be masked[18]. However, the feasibility of creating a microdata file for the household sector (smaller holdings) should be explored. The utility of this type of file depends on the research objectives pursued. For example, this approach may be useful if the research focuses on development strategies for the household sector. On the other hand, if the research objectives relate to the sector as a whole, it is less effective.
The question of what constitutes disclosure for an agricultural operation is also to be considered. Many of the characteristics of farms are readily visible to passers-by (livestock, crops, *etc*.), but some other characteristics remain hidden (e.g. number of holdings, income, expenditures). More research in this area is required.

In assessing which approach is feasible, SOs must consider a number of factors:

- Do they possess the statistical expertise to analyse and implement SDC procedures on the files? In countries where the agricultural SO operates

[18] Anonymization depends on there being a significant number of units in the population that ar

separately from the NSO, are these skills available in the National Statistics Offices (NSOs)? What are the legal requirements? (Legal requirements will be addressed in Section 6, below).

- How are the policy research needs of the relevant Ministry of Agriculture and its national and international partners best satisfied?
- Do they have the technical skills and infrastructure required to implement a remote access service?
- Are physical premises and staff available to operate a data enclave?
- Are there other ways to meet the policy research objectives?

Naturally, there is no one best approach to meet research needs. SOs may wish to start with one approach (that which best meets their circumstances) and then evolve to others as required and appropriate. This is the trajectory followed by many developed countries. Public Use and licensed microdata files have been produced by NSOs for over 60 years[19], whereas remote access and data centres only came into prominence during the past decade.

5.3 Agriculture Microdata – Past Practices

Interestingly, the practice of producing microdata files for agriculture is only in the early stages of establishment. In developed countries, or in countries with a skewed size distribution, it is very difficult to create "safe" files. Agricultural sectors present several facets (in terms of size, commodities, and location), which makes disclosure control difficult. However, important research and policy decisions for agriculture have been made. Did the researchers then involved rely strictly on published tables and custom tabulation, or did they have other means of access? Several explanations are possible:

- The researchers relied on standard and custom tabulations. Much research in agriculture relies on small area data (e.g. crop yields). While the level of detail in standard tabulations may not meet this need, most SOs are willing to provide more detailed tabulations on a cost-recovery basis.
- The researchers from Ministries of Agriculture could easily be brought under the umbrella of the SO, which was often within the same organization as the policy analysts. This provided an avenue for supporting research using microdata.
- The researchers were sworn in as deemed employees, or given access to files pursuant a contract with the relevant Ministry. This theory is supported by anecdotal evidence exclusively.
- The literature presents substantial evidence supporting the possibility that the researchers collected their own data. This is the case especially when the SOs did not collect the type of data sought by the researchers, e.g. detailed cost of production data.
- It may have been a question of resources and priorities. For most agricultural SOs, the main priorities are the production of food supply estimates, and the

[19] The 1960 US Population Census created a Public Use Microdata file, which was distributed on 13 UNIVAC tapes or 18,000 punch cards. See http://researchmatters.blogs.census.gov/2012/08/02/steven-ruggles-census-data-processing-part-2/ (accessed on 26 January 2014).

provision of inputs for estimates of national economic aggregates. Unfortunately, many surveys were often not completed on a timely basis due to a lack of resources. Thus, the estimates were often based on subjective information. Production of microdata files were a lower priority compared to the two tasks mentioned above.

- There may have been concerns that the production of microdata files could reveal quality issues in the survey data. It is possible that the SOs feared that the methodology they used would not compare well to more rigorous statistical standards.
- The use of aerial photography and remotely sensed satellite techniques was sufficient to provide detailed land use/land cover information.

5.4 A Note on Anonymizing Agriculture Data Files

Which approach is "best" depends on several factors, such as the nature of the survey (households only, or households and commercial units), the structure of the sample (skewed or not skewed) and whether independent sources exist against which the files can be matched. These questions must be assessed on a country-by-country basis. Generally, the fewer the number of SDC procedures that must be applied, the better. A generic process for preparing a file could be the following:

1. Removal of all respondent information from the file (names, addresses, plot numbers, telephone and other contact information, administrative information, *etc*.).
2. Reduction of the file's geographic levels to a broad level, such as the provincial or other subnational level.
3. Performance of a disclosure analysis using techniques such as frequency distributions, multi-way cross-tabulations and regression analysis to identify unique respondents. The unique values must be examined to determine if they represent a disclosure risk. Generally, for agriculture, disclosure risks are high for units with large land holdings or high incomes (if income is asked), or if units present a combination of characteristics that may relate to age, marital status, off-farm work, *etc*.
4. If this analysis shows that a risk of disclosure exists, an additional procedure should be applied, such as[20]:

a. Micro-aggregations
b. Top coding
c. Bottom coding
d. Deleting variables
e. Deleting cases
f. Perturbation

[20] For a more detailed discussion of the techniques for anonymizing data, see Dupriez, O. and Boyko, E. 2010. Dissemination of Microdata Files. op. cit. See, in particular, Chapter 7.

Box 14: Some Notes on Survey Anonymization

A combination of variables in a microdata record that can be applied to re-identify a respondent is referred to as a 'key'. Re-identification can occur when a respondent is (a) rare in the population, with respect to a certain key value; and (b) this key can be used to match a microdata file to other data files that might contain direct or other identifiers such as voter lists, land registers or school records (or even publicly accessible Internet search engines).

In most developing countries, the risk of disclosure from matching a household survey microdata file to other data files is currently limited, either because they do not exist or because they are not disseminated. In practice, the risk of disclosure from disseminating household survey microdata files will be sufficiently minimized by simply stripping records of direct identifiers, and removing geographical identifiers below the stratum level. However, the disclosure risk should be assessed for each microdata file, as important exceptions to the aforementioned rule of thumb will exist that warrant additional control measures. Survey microdata files containing records from small-target populations – for example, enterprise and business records – are considerably more difficult to anonymize.

Taken from Dupriez, O. and Boyko, E. 2010. *Dissemination of Microdata Files, op. cit.,* Chapter 7.

5.5 Examples of Anonymized Surveys with Agriculture and Food Questions

As mentioned in Section 1 of this Guide, most of the examples of microdata access have been taken from developed countries. However, examples from developing countries are emerging. A summary of two examples from Rwanda and Nigeria is given below.

Rwanda – Comprehensive Food Security and Vulnerability Analysis and Nutrition Survey 2012 (CFSVA)[21]

The CFSVA process generates a document describing the food security status of various segments of a population, over various parts of a country or region; analyses the underlying causes of vulnerability; and recommends appropriate interventions to address problems. CFSVAs are undertaken in all crisis-prone food-insecure countries. The shelf life of CFSVAs is determined by the indicators collected and reported. In most situations, CFSVA findings are valid for three to five years unless drastic food security changes arise[22].

Anonymization

The team from Rwanda's National Institute of Statistics consisted of statisticians who worked on the survey, staff in charge of the survey and census documentation, data processing staff with IT skills, and an advisor from ADP/PARIS21 to provide general and technical assistance in the first round of survey and census documentation and creation of microdata files. This work usually takes between 3 to 6 months.

Tools
Data entry and data processing are usually performed in CSPro, and then converted into SPSS and STATA for analysis. The microdata files are created using mostly SPSS, and occasionally STATA, based on demand from data users.

Procedures

Creation of Public Use Files: (1) Removal of respondents' names; (2) Removal of all addresses from the village level (lowest administrative entity) up to the district level (since the survey estimates are representative up to the district level), to avoid computation of any estimates at a level lower than the district – these will not be reliable due to the sample size; (3) Collapsing of all geographic information (similar to Step 2 above). The disclosure techniques applied were micro-aggregations, top coding, and bottom coding.

***Creation of Licensed and Enclave Files**:* The procedure for public use files is applied.

[21] Information obtained by email from D. Habimana, Director, Statistical Methods, Research and Publication Unit National Institute of Statistics of Rwanda, 17 April 2014.

[22] Rwandan National Institute of Statistics (NISR), Ministry of Agriculture and Animal Resources (MINAGRI) and World Food Programme. Rwanda Comprehensive Food Security and Vulnerability Analysis and Nutrition Survey 2012. Ref. RWA_2012_CFSVA_v01_M. Data set downloaded from http://nada.vam.wfp.org/index.php/catalog on 6 March 2014.

The disclosure techniques applied were micro-aggregations, top coding, and bottom coding.

Review/approval process: The data release committee reviewed the microdata files before they were released online. This committee checks whether files are created properly. It also advises the Director General upon approval of the created files and authorization of their release; in practice, this power is delegated to the Director of Statistical Methods, Research and Publication.

Publication: The catalogue record for this survey is available at http://catalog.ihsn.org/index.php/catalog/4149.
This catalogue is based on the NADA software developed and is housed on the IHSN website.
The survey can be downloaded as a Public Use File by accessing the "**Get data"** tab on the catalogue record, and completing a registration form.

Nigeria: General Household Survey-Panel 2010-2011 (Post-Planting)[23]

The GHS-Panel survey is a sub-sample of the annual GHS cross-section survey conducted by NBS. The GHS - Panel survey is the first stage of a long-term project aiming to collect panel data on households and their characteristics, welfare and agricultural activities. The GHS-Panel is to be conducted every two years.

This first stage consists of two visits to the household. The first visit was the post-planting visit (August-October 2010), which occurred directly after the planting season to collect detailed information on household characteristics, including preparation of plots, the inputs used, the labour used for planting, and other issues relating to the planting season. The second visit, the post-harvest visit (February-April 2011), occurred after the harvest season and collected additional information on household characteristics, along with information on the crops harvested, the labour used for cultivating and harvest activities, and other issues relating to the harvest cycle[24].

The team consisted of staff from the subject matter and data processing area, with systems and computing expertise. The team was familiar with the surveys, as they also perform the survey documentation for all NBS surveys.

Tools and procedure: After a clean file was produced, the team removed all direct identifiers (similarly to the Rwandan approach) and then used the IHSN's SDC Micro tool to identify the risks of carrying out the suppression.

Publication: The catalogue record for this survey is available at:
http://nigerianstat.gov.ng/nada/index.php/catalog/31/overview

This catalogue is based on the NADA software (freely distributed by the IHSN) developed and housed on the Nigerian Bureau of Statistics website. The survey can be downloaded as a Public Use File by accessing the "Get data" tab on the catalogue record and completing a registration form.

[23] Information obtained from a conference call between Nancy Chin (FAO) and Biyi Fafunmi (Data Access), National Bureau of Statistics, Nigeria, 7 April 2014.
[24] National Bureau of Statistics, *General Household Survey-Panel 2010-2011 (Post-Harvest) version 1.0*, accessed from http://nigerianstat.gov.ng/nada/index.php/catalog/31/overview on 4 April 2014.

5.6 Coordination between Agriculture and the National Statistics Office

In many countries, NSOs have delegated responsibility for agricultural statistics to the Ministry of Agriculture. In this case, production of a microdata file requires coordination and agreement between the two offices. In all likelihood, the agricultural body will be identified as the owners of the file, while the NSO will bear the responsibility to ensure observance of the provisions of the relevant statistics legislation. To the extent that the statistical skills required to implement the SDC procedures lie only with the NSO (and not with the Ministry of Agriculture), an agreement on resource sharing will also be necessary. The microdata release policy will have to be a joint one (i.e. will apply to both partners), and must also specify the approval process. Since the majority of users' questions will probably concern the files' content, user support would be best overseen by the Ministry of Agriculture.

Box 14: Some Notes on Survey Anonymization

A combination of variables in a microdata record that can be applied to re-identify a respondent is referred to as a 'key'. Re-identification can occur when a respondent is (a) rare in the population, with respect to a certain key value; and (b) this key can be used to match a microdata file to other data files that might contain direct or other identifiers such as voter lists, land registers or school records (or even publicly accessible Internet search engines).

In most developing countries, the risk of disclosure from matching a household survey microdata file to other data files is currently limited, either because they do not exist or because they are not disseminated. In practice, the risk of disclosure from disseminating household survey microdata files will be sufficiently minimized by simply stripping records of direct identifiers, and removing geographical identifiers below the stratum level. However, the disclosure risk should be assessed for each microdata file, as important exceptions to the aforementioned rule of thumb will exist that warrant additional control measures. Survey microdata files containing records from small-target populations – for example, enterprise and business records – are considerably more difficult to anonymize.

Taken from Dupriez, O. and Boyko, E. 2010. *Dissemination of Microdata Files, op. cit.,* Chapter 7.

6. Legal and Policy Frameworks for Providing Access to Agricultural Microdata

Before Statistical Organizations (SOs) embark on a mission to provide access to microdata, they must ensure that they are operating within the limits of their enabling legislation and of the charters under which they function. To do otherwise could undermine the agency's credibility and the confidence of respondents. It is essential to maintain respondent support if the agency is to succeed. At the same time, it is recognized that
methods exist to ensure confidentiality while at the same time making microdata available for statistical purposes.
It is important that clear policies be established on the actions that the agency can take, and that this information be transparent and available to the public. This Section outlines the legal and policy considerations for disseminating agricultural microdata.

6.1 Legal Considerations

The first factor to consider is the national legal framework under which the SO operates. This usually derives from legislation defining the organization's role and mandate. In some countries, there is a single National Statistics Office (NSO) with the mandate to supply statistics for all sectors of society and the economy. In other cases, agricultural statistics may be a responsibility that is delegated to the Ministry of Agriculture. Nonetheless, the principles under which these bodies operate usually resemble those outlined by the United Nations Statistics Division:

Individual data collected about natural persons and legal entities, or about small aggregates that are subject to national confidentially rules, are to be kept strictly confidential and are to be used exclusively for statistical purposes or for purpose mandated by legislation

Good practices include:
Putting measures in place to prevent the direct or indirect disclosure of data on persons, households, businesses and other individual respondents
Developing a framework describing methods and procedures to provide sets of anonymous micro-data for further analysis by bona fide researchers, maintaining the requirements of confidentiality).[25]

This indicates that while it is essential that microdata be managed in accordance with national confidentiality rules, it is possible to implement measures to prevent disclosure. This applies both to microdata files, to which access may be given, and to the task of ensuring that no cells in a table of aggregate data reveal information about respondents.

[25] United Nations Statistical Office, Principles Governing International Statistical Activities. Available at http://unstats.un.org/unsd/methods/statorg/Principles_stat_activities/principles_stat_activities.htm. Accessed on 14 March 2014.

The legal frameworks under which SOs operate vary by country. It is up to each individual SO to interpret the legislation and provide services accordingly. In some cases, it may be necessary to amend the legislation under which the organization works before microdata can be disseminated[26]. Once the legal framework has been verified, the next step is to ensure that the agency has an explicit policy governing the dissemination of microdata.

Box 15

Notes on Statistical Legislation

Statistical laws vary among countries, as does the extent and the entities to which SOs permit access to microdata. Today, probably no statistical laws exist that wholly prohibit the release of microdata. Rather, legislation generally indicates that information without identifiers *may* be released to certain users (at the discretion of the head of the agency). Most countries refer to Principle 6 of the United Nations Principles Governing International Statistical Activities, available at: http://unstats.un.org/unsd/methods/statorg/Principles_stat_activities/principles_stat_activities.asp

Example of restrictive legislation: Microdata may be released free of charge or for a fee. Only requests from Government Institutions, Recognized Universities, Students, and selected international agencies are entertained. However, data users are strictly required to adhere to the terms stipulated in the agreement form. All data requests should be made to the Director General (DG), as the sole authority to release data is vested with the DG. The organization reserves the sole right to approve or reject any data request, depending on the confidential nature of the data set and the intended purpose of the study or analysis.

Example of a less restrictive policy: The data collected by the national statistical system's services through surveys or any other collection method are protected by statistical confidentiality. Statistical confidentiality implies that the dissemination of such data, as well as of the statistical data that can be calculated from them, is to be executed only in such a way that those who provided the data cannot be directly or indirectly identified.

Example of less restrictive guidelines: These would seek to ensure that arrangements for protecting confidentiality are sufficient to guard the privacy of individual information, but would not be so restrictive as to unduly limit the practical utility of official statistics. The obligation to establish disclosure control methods that are sufficient yet not overly restrictive makes the selection of the disclosure control method important: while any method can be adopted to achieve confidentiality protection, some methods achieve a better utility, for the same level of protection. There are no restrictions as to who can receive the data, but a prior agreement must be in place.

Coordination: In view of the fact that a Ministry of Agriculture may be collecting and dissemination data under delegation from the NSO, it is important to clarify whether and under which conditions a microdata file can be prepared and released by the Ministry of Agriculture. These projects will most likely be a joint activity.

[26] The Canadian Statistics Act was revised in 1970, to ensure that the legislation accommodated the release of a Public Use Microdata File (PUMF) from the 1971 Population and Housing Census. Section 17(1)(b) prohibits the disclosure of respondent information. See http://www.statcan.gc.ca/about-apercu/act-loi-eng.htm.

6.1 Policy Framework

The dissemination policies of SOs constitute the framework governing dissemination. These policies are based on the legislative mandate entrusted to SOs within their respective countries. As mentioned above, this depends on how statistical activities are organized within individual countries (i.e. whether the system adopted is centralized or decentralized). This must be ascertained on a case-by-case basis.

Policy frameworks generally contain the following types of information:

- **Agency name:** The National Statistics Office (NSO) for [Name of country]

- **Policy objectives, rationale and definitions:** The objective of this policy is to outline the principles for user access to the microdata produced by the NSO. This policy objective adheres to the principle that all data collected by the NSO should be publicly available, in line with governments' policies of supporting an informed citizenry while conforming to international best practices for statistics, including ensuring the confidentiality and anonymity of data providers.

- **Types of microdata files and access processes:** This defines the range and nature of the available data files, such as Public Use Files with conditions to accept, licensed files, remote access, on-site access, *etc*.

- **Classes of use and users:** This identifies the potential users, who can range from all researchers to only researchers from certain organizations, and the type of agreement that must be established with them.
-
- **Responsibilities of players within the NSO**

- **Responsibilities of users**

- **Timing and release policy**

- **Cost recovery policy**

A generic draft of a microdata dissemination policy may be found in Appendix C below. This draft is quite detailed; a simpler policy could certainly be used, but this choice is up to the competent parties within the country.

6.2 Licensing and Agreements

Statistical agencies recognize that it is virtually impossible to produce a file that is both anonymized to the extent that re-identification is impossible, and is useful to researchers. Accordingly, they take several precautions when releasing files to researchers. Countries generally use one of three different approaches, briefly described below.

- For **Public Use Files**, users may not be required to sign a contract, but may be asked to identify themselves and to provide **contact information**. This provides the SO with information on their user community as well as the ability to contact them, if changes must be made to the file. Similarly, users can contact the SO if they detect errors in the file. Appendix A below shows conditions that users may typically be asked to accept. If users do not respect these conditions, they will be denied future access to data files and will lose credibility.

- For **licensed files,** Appendix B below outlines the conditions that may apply to researchers who wish to use licensed files. Ideally, the researchers' organizations will also sign the agreement. In some cases, this is a necessity, as it adds an incentive to ensure that the data are used and managed properly. Indeed, the organization would not wish to suffer any sanctions or losing future access to the data, nor damage to their reputation.

- For **on-site/data enclave access**, users must enter into an **agreement** with the SO. The operation of a data enclave is best understood by referring to some examples and webpages from different countries.

Australia

The ABS [Australian Bureau of Statistics] Data Laboratory (ABSDL) is the data analysis solution for high-end data users who want to extract full value from ABS microdata. ABSDL provides an interactive environment, enabling the analysis of Basic, Expanded or Specialist (customised) Confidentialised Unit Record Files (CURFs). ABSDL provides a more responsive and interactive environment in which to analyse CURFs than that offered by the Remote Access Data Laboratory (RADL).[27]

With this service, the ABS offers a more advanced remote access facility within its premises. The initial remote access facility (the Remote Access Data Laboratory) was available from outside the ABS, but had limited functionality (tabulations); the ABSDL, instead, has no such restrictions. The ABSDL's major limitation is the fact that users must travel to the ABS offices. The next generation of microdata access in Australia will focus on providing online remote access through the REEM project – Remote

[27] Australian Bureau of Statistics, Remote Access Data Laboratory. See http://abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+About+the+ABS+Data+Laboratory+%28ABSDL%29.

Execution Environment for Microdata. This will feature an expanded functionality, going beyond tabulations28.

Brazil

In response to increasing demands from the research community, the Brazilian Institute of Geography and Statistics (IBGE) has also established an on-site facility[29].

The following describes the administrative and technical measures to regulate the access of restricted microdata and to ensure that the output is released with an adequate level of protection so that individual data cannot be disclosed. The procedures cover the following steps:

(1) application

The researcher submits the research project to be evaluated if it is for public or academic interest, for statistical purposes and also whether it is feasible.

(2) evaluation of the project

A Committee of Assessment of Restricted Data Access evaluates the project, based on submissions of the thematic area responsible for the survey microdata. The Committee authorizes (or not) the access to internal data files under the appropriate conditions.
The Committee is chaired by the Deputy Director for Surveys and composed of senior staff members dealing with business, methodology and dissemination coordination.

(3) formal agreements to access

Once a project has been authorized, formal agreements between the researcher and the agency are established. These agreements involve a written contract (contractual arrangement), and an agreement form outlining the conditions of access and setting out fees for the proposed work.

(4) on-site access

The databases are installed in the room with special computers for the researchers. The security features of the computers include a blockade to external networks to prevent transfer of data. Furthermore, the external disk drives and serial parallel ports are disabled. The identification of the enterprises is recoded in the databases from businesses surveys of IBGE or from external sources.

[28] Thorne, R., Nicholls, P. and Boettcher, K. Microdata Dissemination Architectures and Systems, Australian Bureau of Statistics, Australia. See http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2011/wp.7.e.pdf, accessed on 14 March 2014.

[29] Instituto Brasileiro de Geografia e Estatística, Data laboratory microdata access – Brazil. For more information. see United Nations Economic Commission Europe, http://www1.unece.org/stat/platform/display/confid/Annex+1.15.+Data+laboratory+microdata+access+-+Brazil, accessed on 17 July 2014.

The researchers do the work and save the output in the hard disk of the special computer and then prepare a report document. A CD-Rom with this information is prepared by IBGE staff, to be analysed by the thematic survey area.

(5) evaluation of output

Canada

Canada adopts different application procedures depending on whether the requests to access data are from government departments, or academic researchers; in the latter case, the procedure is more stringent.
For government-sponsored research, the process is as follows:

As of August 2005, an additional mechanism is offered to facilitate the conduct of research projects focusing on statistical support for policy development. All requests are assessed by a provincial or territorial representative on the Federal-Provincial-Territorial Consultative Council on Statistical Policy. The request for access is submitted to the Program Manager of Statistics Canada's Research Data Centre Program who coordinates a review of the proposal by Statistics Canada subject matter experts. The review is completed within 10 working days. If it is determined that Statistics Canada can quickly and efficiently carry out the work, the Departmental representative will be informed of this and of the associated cost to complete the work. However, if Statistics Canada does not have the resources to complete the work quickly and efficiently, the provincial/territorial employee identified is eligible to become a "deemed" Statistics Canada employee, under Section 10 of the Statistics Act, for purposes of completing the work. The work conducted in the RDC is subject to standard operational procedures. More details on this process can be obtained in the Data access procedures - statistical support for policy development.[30]

Academic requests must be sent to a peer-review committee, which will approve the project based on the following criteria:

- scientific merit and viability of the proposed research;
- relevance of the methods to be applied - the data to be analysed;
- demonstrated need for access to detailed microdata; and
- expertise and ability of the researchers to carry out the proposed research as illustrated in the CVs and list of contributions.[31]

The statistical output must be analysed before its release to the researcher to ensure the technical assessment of disclosure risks and confidentiality requirements. The analysis is undertaken by the thematic area responsible for the survey microdata, the same that gave submissions for the committee decision.[32]

The key reason for establishing data enclaves is to ensure maximum protection for confidentiality while at the same time providing researchers with the widest possible

[30] Statistics Canada, Research Data Centres. For more information, see http://www.statcan.gc.ca/rdc-cdr/index-eng.htm.
[31] As above, note 33.

access to the detailed files. If the statistical legislation does not permit the creation of microdata access channels which are not fully under SO control, data enclaves may be the only legal way to access microdata. Since the researchers using data enclaves must effectively become deemed employees, their work must be relevant to the SO in some manner. Once an agreement has been signed, the researchers can proceed.

If the agreement is not honored, access will be terminated and possible sanctions under statistic legislation may be taken.

6.4 Conclusion

SOs usually take rigorous precautions before providing access to microdata, simply because of the great stakes involved. SOs must operate within the law, and yet attempt to meet the increasing demands from researchers. Licensed files are preferred over PUFs. Remote access and data enclaves provide SOs with a much greater degree of control, but are usually costly and present disadvantages for researchers. The best approach to adopt requires judgment on the part of SOs. This depends on the requestor and his or her objectives.

7. Technical Infrastructure and Institutional Requirements for Providing Access to Microdata

Within an SO, the main requirements to sustain a process for preparing and releasing files for access include the following:

- Technical infrastructure
- Metadata tools
- Human skills
- Review and approval process

7.1 Technical Infrastructure

The computing infrastructure required to anonymize files is no greater than that required for processing a survey. Modern microcomputers with sufficient storage and CPUs are adequate for the tasks of analysing the files and making them available for access. Analysing files for possible unique records involves running a number of 2- and 3-way cross-tabulations and examining the cells for unique records. Generally, software such as CSPro SPSS, SAS, STATA, and R is used.

The introduction of the SDCMicro toolbox is a great step forward for anonymization and risk assessment. However, it cannot solve all problems.

SDCMicro is a free, R-based[33] *open-source software for the generation of protected microdata for researchers and public use. The package provides multiple options for reducing the statistical disclosure risk in categorical or continuous variables. SDCMicro can be used from the R command line* interface or by using the application's graphical *user interface (GUI). The package can also be used in batch-mode from other software.34*

The release of the SDCMicro software and the possible forthcoming offer of technical support to countries wishing to use this tool could be a major benefit, for those who wish to produce anonymized files.

Providing the actual access to such a file may require additional infrastructure, depending on the method of access chosen. For example, providing access to a PUF or a licensed file requires a means of transfer such as a CD or DVD, or electronic file transfer via Internet. Many countries are adopting the NADA catalogue system developed by the IHSN. This can be done either on the IHSN website, or by installing and rebranding it on the SO's website.

Creating a remote access facility requires a network server and communication infrastructure, as well as software to perform the remote access data manipulations. Interactive remote access requires manipulation software, as well as routines to check for possible disclosure.

7.2 Metadata Tools and Knowledge

As noted in Chapter 4, good metadata are essential if the data files are to be useable. The Data Documentation Initiative (DDI) standard is highly recommended, as it is widely used for household and other types of surveys. To assist survey managers in implementing the DDI, a metadata editor such as that from NESSTAR is commonly used. This can be freely downloaded from the IHSN or the NESSTAR websites[35].

If researchers may choose from a large number of files, then there must be a catalogue to enable them to browse and identify relevant data. Researchers prefer to search files at the variable level, as opposed to simply looking at higher-level descriptions of surveys or studies. For this purpose, the NADA tool is ideal.

NADA is a web-based cataloging system that serves as a portal for researchers to browse, search, compare, apply for access, and download relevant census or survey information. It was originally developed to support the establishment of national survey data.[36]

[33] R is a free programming environment/language for statistical computing. See http://www.r-project.org/, accessed on 21 April 2014.
[34] SDCMicro anonymization software. See http://www.ihsn.org/home/software/disclosure-control-toolbox, accessed on 21 April 2014.
[35] NESSTAR, Norwegian Social Science Data Services, DDI Metadata Editor. See http://www.ihsn.org/home/software/ddi-metadata-editor or access the NESSTAR website at http://www.nesstar.com/software/publisher.html. Accessed on 12 March 2014.
[36] NADA was developed by the IHSN to help countries make their microdata files available to users. See http://www.ihsn.org/home/software. Accessed on 21 March 2014.

7.3 Human Skills

A variety of skills are required. Statistical and software knowledge is required to perform the analysis and determine which disclosure techniques to adopt in masking the data. The computing skills necessary are commensurate with those required to process the survey. Analysing the files for disclosure issues requires subject matter knowledge on the survey's content, to decide which information would be considered sensitive and what is the best way to mask it.

Producing the metadata with a tool such as the NESSTAR Editor is a greatly simplified process, but requires a thorough knowledge of the content and concepts behind the survey. If NESSTAR is not chosen, it is possible to prepare the metadata using an XML (Extensible Markup Language) editor; the DDI standard is expressed in XML[37]. Expressing the metadata in XML and saving the data as an ASCII file affords users the greatest flexibility in choosing software and computing platforms, and provides a preservation format for future research.

It is unfortunate that these skills are in short supply, even in developed countries. Hiring staff with the right background and training and making use of international expertise are the keys to overcoming this challenge.

7.4 Review and Approval Process

Microdata dissemination policies generally identify the head of the agency (e.g. the Chief Statistician or the Director General) as the official who gives the final approval for releasing a file. He or she will normally rely on a Microdata Release Committee for advice on whether to release the data file. This committee generally consists of subject matter specialists, statistical experts, computer specialists and dissemination staff. This committee can provide leadership in the process of preparing microdata for release, by drafting guidelines and procedures. The outcome of the review process can be the approval of the file for release, or recommendations on changes to be made to the file before it is released. Files for surveys that are repeated on a regular basis (e.g. a Labour Force Survey, conducted on a monthly basis) would only have to be reviewed once, as long as the content and methodology have not changed. Of course, files may not receive the committee's approval for release. A discussion with the committee prior to beginning work on a file is advisable, to avoid wasting efforts for files that simply cannot be anonymized to a degree satisfactory for release.

[37] Extensible Markup Language

TABLE 6. Summary of Infrastructure Requirements for Disseminating Microdata

	Public Use File	Licensed File	Remote Access[38]	Enclave
Human Skills	Knowledge of statistical techniques for analysing files, to identify potential disclosure issues; implementing disclosure prevention techniques; subject matter knowledge to identify sensitive variables and to prepare the metadata; user support.	Same as for a public use file, plus the need to oversee the license.	Subject matter knowledge to identify sensitive variables to be removed; metadata; vetting of files if output is not automated; systems development, if building a remote access real-time access system; user registration; user support.	Subject matter knowledge to identify sensitive variables to be removed; vetting of files; security; may require the establishment of a special physical location; administration; user support.
Software and Computing Infrastructure	A computer workstation and software such as SAS,SPSS, STATA or R.	A computer workstation and software such as SAS, SPSS, STATA or R; a metadata editing tool.	An Internet server with software developed to serve the data; on-the-fly disclosure analysis requires additional programming.	Secure room with standalone workstations loaded with data and software, or a local server with data and software.
Dissemination	Files can be distributed on CDs or DVDs, on the SO website or through the IHSN NADA catalogue and server. See e.g. http://catalog.ihsn.org/index.php/catalog.	Same as for Public Use Files.	Internet server with the software required to support access.	Users can remove material after it has been cleared by the disclosure analyst.
Approval and review	Survey program managers, review committee, head of the SO.	Survey program managers, legal services, review committee, head of the SO.	Survey program managers, review committee, head of the SO.	Survey program managers, review committee, head of the SO.

[38] Developing remote access infrastructure is generally done in stages. The cost of a fully-functioning remote access system is in the range of USD 500,000 – 1,000,00

8. Promoting the Microdata Access Program

The experience of several countries has been that, when it comes to microdata*, "if you build it, they will come"*. This has been true every time that microdata was made available where no access previously existed. While SOs produce files in response to demands from their respective user communities, it was found that the potential community can be significantly larger than the researchers most active in lobbying for access. This is positive as, having expended the time and effort required to produce these files, it is beneficial if they are used as broadly as possible. These new research projects can be promoted on the SO website and at seminars or workshops for the user community is invited; also, notices could be sent to the key ministries that could be interested in the files.

While research using microdata requires a different set of skills from those needed to analyse aggregate data, the increased sophistication of the analysis makes it appropriate to provide quantitative research training to researchers. SOs may wish to work with universities or leading government agencies for this purpose.

The ability to increase Canadian researchers' policy-analysis capacity was one of the objectives of Canada's establishment of the Data Liberation Initiative[39] (DLI). This program provides university students and researchers with access to all publicly-available data files from Statistics Canada, including approximately 350 PUF titles. Access to these data files is affordable for faculty, staff and students, as the 75 universities involved are required to pay a small fee to belong to the Initiative. Membership of the DLI means that, unlike all other users, who must sign a license each time an individual accesses a file, the institutions are covered by a single license. Training in the use of microdata and other files is a cornerstone of the Initiative. All 18 years of training material from the DLI is freely available on the Internet[40]. This partnership between Statistics Canada and Canadian universities has greatly increased the research community's capacity to perform the complex analysis enabled by microdata files.

With reference to Canada, experience has shown that when researchers are only able to access PUFs, they eventually request more detailed files. The best recommendation for the program will certainly be made by the researchers who have successfully used the files for their research. Over time, it should be possible to construct a network of users that includes government departments, universities and international organizations.

[39] Statistics Canada, Data Liberation Initiative. See http://www.statcan.gc.ca/dli-idd/dli-idd-eng.htm. Accessed on 20 January 2014.
[40] Data Liberation Initiative, Training Repository. See http://cudo.carleton.ca/. Accessed on 20 January 2014.

9. The Open Data Agenda

9.1 Introduction

It may be wondered whether opening microdata files for dissemination and access is part of the Open Data agenda. Open data is one of the latest movements to appear on the global scene, together with ideas such as open government[41], open source, open access, and open science, to name only a few. The premise of open data is that governments at all levels (from the national to the local) hold vast stores of data. When these data are made freely available through portals, metadata, and search tools for reuse by governments, citizens, NGOs, academia, and the private sector, they can be used in new and unanticipated ways. The aim is to encourage and facilitate innovation, and to improve government. Most of the user community activity in this area may be attributed to developers who seek to create applications for "fun or profit". Gatherings of developers and technically-skilled individuals in metropolitan centres looking to promote and exploit open data are referred to as "hackathons", and are attended by individuals searching for ways to start a business or to contribute to the public good.

The types of data released do not reveal citizens' personal information, and are often in formats that are not easily used by the general public. An example of a useful application in one city involves the retrieval of bus route information from the local bus company and its combination with GPS feeds from individual buses; the smartphone application developed enables bus riders to know when the next bus will reach their stop. Users must only send a message to the central server with the number of their bus stop, and the server returns the requested information within seconds. Governments often hold extremely detailed geo-spatial information, which can be the basis for public service and mapping applications. Suffice it to consider the benefits of applications such as Google Maps, which can be combined with local information to improve access to various services.

9.2 How does open data affect the work of Statistical Organizations?

The open data agenda is compatible with the agendas and mandates of statistical agencies in many ways. SOs collect, compile, analyse and disseminate information gleaned from surveys, censuses and administrative sources. Data are analysed and often interpreted to enable a general audience to understand the underlying message. However, the open data movement is not concerned with the creation of Public Use Microdata files; this task is left to statistical agencies' specialists. The open data movement encourages SOs to release more detail and data (of a non-confidential nature) in a generic format, to permit easier further processing. However, the task of retrieving data from all the different branches of government, including SOs, and building a single portal at the national level, is an enormous and complex task. This work is often hampered by a lack of metadata standards. Thus, the real progress of the open data movement is slow.

The open data movement has brought the following benefits to SOs and other holders of data:

[41] The Open Government Partnership started with the G8 countries and now has 63 members. See http://www.opengovpartnership.org/. Accessed on 16 March 2014.

- The recognition that data collected with public funds should be made more freely available to the public, including developers and commercial entities;
- The broad adoption of the concept of an open license, attached to the data disseminated. For example, Statistics Canada once used a commercial and "legalistic" license for its data; today, it has adopted an open license agreement[42];
- The use of open data formats to permit easier, wider access to data, by removing the barriers caused by proprietary formats;
- A move towards using common metadata standards.

The open research data agenda is a close ally of open data. It stresses the need to preserve and share data with other researchers, to facilitate data re-use. To comprehend the benefits of these actions, suffice it to consider the success of the Genome Project, in which scientists deposit research findings on genetic sequencing into an open repository that other researchers may access. A similar institution in the statistical world is the Accelerated Data Program's Initiative[43] that encourages data producers to document their data for depositing into the International Household Survey Network's[44] Catalog/NADA or a replicate of that system on its own website.

Open data, the pressures towards a greater availability of anonymized data files, and the ability to publish detailed databases that enable users to help themselves, all indicate the desire to use data as evidence, for making informed decisions and for creating a better-informed citizenry.

[42] Statistics Canada Open Data Agreement. See http://www.statcan.gc.ca/eng/reference/licence-eng. Accessed on 15 March 2014.
[43] Accelerated Data Program. See http://adp.ihsn.org/node/203. Accessed on 15 March 2014.
[44] International Household Survey Network. See http://www.ihsn.org/home/. Accessed on 15 March 2014.

10. Concluding Observations

A statistical organization's goals are to serve the needs of its government, businesses, institutions and individuals, while respecting its obligations to respondents to maintain the confidentiality of their information. As the needs and demands of the user community grow, and the demand for access to microdata increases, SOs must find innovative ways to meet these demands, or their relevance will fade. There is growing recognition that access to microdata is important; methods providing safe access must be found. The international statistical community has devoted considerable effort the challenges of providing access to microdata. This has resulted in a wide range of online resources that may assist agencies in planning appropriate approaches. SOs are pressured to find ways to balance researcher needs with their capacity for action and risk management.

Providing access to agricultural microdata requires balancing a number of variables, to determine a course of action that meets researchers' needs and protects respondent confidentiality. The decision depends on:

- The legislation under which the statistics are collected
- The organization of the statistical system (centralized or decentralized)
- The value of the data and the potential for enhancing its usefulness by providing access to the microdata
- The ease of user access to the data
- SOs' capacity to prepare and support any given access approach and
- SOs' capacity to manage or tolerate risk

It is also necessary to ask whether alternatives to accessing microdata exist. Can the needs of researchers be satisfied by providing a database of aggregate statistics? The view of several SOs who deal with agriculture is that agricultural data should be treated in the same way as business/establishment data. Generally, SOs provide access to such data through custom tabulations and on-site data enclaves, where researchers are deemed employees. However, rather than accepting this as a general rule, SOs should consider the structure of their agricultural industry, and the possibility of providing access to at least some microdata. Encouragingly, a growing number of countries have made available the results of their surveys with agricultural and food-related questions, as PUFs and licensed files.

Some countries have relatively small overall numbers of agricultural units, and a preponderance of larger, more specialized units, which are difficult to anonymize. In this context, Uruguay was cited as an example. Countries with such an agricultural profile are unlikely to attempt creation of access to agricultural microdata in forms other than a data enclave. On the other hand, countries having a large number of smaller units dispose of more options (see the example of Figure 7 as applied to Ethiopia, or to countries with a similar structure, in Table 3 above). It should be feasible to create an anonymized file of the subset of smaller units (generally referred to as household units) which can be licensed to users. It should be noted that this is indeed feasible only if the needs of the researchers can be met by using this subset. In this connection, further research and testing would assist comprehension of the risks of disclosure.

Because it is crucial for the SO to observe their country's statistical legislation, the policies for microdata access enabled by the legislation must be clear and transparent to the broader community. Otherwise, the basic principles under which SOs operate

may be violated[45].

Communication between researchers and SOs is important. Researchers must communicate the data's potential value and the research that they are conducting, to meet national or international objectives. The widespread availability of the technology necessary to exploit large data sets increases the demand that SOs will face from the research community. Thus, the onus will be on SOs to meet this demand.

Assessing the risk of statistical disclosure requires several factors to be considered.

- The nature of the user community and the extent to which it has been educated by the SO to understand the issues and risks involved as a result of disclosure. Are they willing to accept the SO's values and ethics, and adhere to the agency's security practices? Are they willing to sign formal agreements and to accept the consequences of sanctions, if a breach is committed? Unfortunately, the impact of a breach may have lasting impact for the SO.
- The existence of external databases against which agricultural microdata can be matched to individual units, thus enabling their identification. Matching files may be forbidden by the licenses or contracts in place, but if it is a technical possibility, then it remains a potential risk. For example, a company that supplies inputs to farmers or buys their products may obtain information about their clients from a database. If such a party could access agriculture microdata and had a malicious intent, they could build an expanded profile of these units for business purposes. Most likely, the user will not engage in such activities, but it remains a practice that SOs seek to prevent. Linking databases should be forbidden by the terms of the license, and sanctions should include prohibiting all future access to users who violate the agreement.
- The structure of the population and the sample to which they wish to provide access.

In developing approaches to provide access to microdata, the location and the capacity of the relevant community should be taken into account. If the researchers are within easy geographic access to the SO, then it may be best to operate a data enclave or to use deemed employees. To satisfy the research needs of international organizations, creating a licensed file and signing an MOU may be the best approach, because these researchers cannot easily access the SO's location. A second alternative, for both types of users, would be to establish some form of remote access. The related infrastructure would have to be developed; however, once developed, it could be shared among countries.

The skills required to produce microdata files for dissemination must be in place before an SO launches a microdata project. The experience of the International Household Survey Network (IHSN) and the Accelerated Data Program (ADP) indicates that countries are currently performing much better in managing and preserving their

[45] United Nations Statistics Division, Principles Governing International Statistical Activities. See http://unstats.un.org/unsd/methods/statorg/Principles_stat_activities/principles_stat_activities.htm. Accessed on 25 January 2014.

survey census files[46]. The next step, of anonymizing survey files to create licensed files, adds a further layer of complexity to the project. This applies to surveys in the household sector, which are generally regarded as being less problematic than agriculture surveys.

Concerns may arise as to the survey's quality, as well as to a lack of expertise to manage the statistical disclosure process. If a program to support microdata access is to be launched in several regions of the world, training programs to support this must be established. The recent release of the SDCMicro tool is a promising step in file anonymization, which is often the last hurdle in releasing microdata files. The IHSN is one of the many international organizations that recognize the growing importance of access to microdata.

[46] Data catalogues are available from the IHSN website. See http://www.ihsn.org/home/survey-catalogs. Accessed on 26 January 2014.

References

Afrobarometer. n.d. *Afrobarometer*. Website. Available at: www.afrobarometer.org. Accessed on 23 May 2014

American Statistical Association. Committee on Privacy and Confidentiality. N.d. *Overviews of Statistical Disclosure Protection Methods*. Available at: http://community.amstat.org/CPC/Methods//MethodsB_Overviews. Accessed on 6 January 2014.

Australian Bureau of Statistics. 2012. *About the ABS Data Laboratory (ABSDL)*. Website. http://abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+About+the+ABS+Data+Laboratory+%28ABSDL%29. Accessed on on 23 May 2014

Australian Bureau of Statistics. 2013. *Microdata Entry Page*. Website. Available at http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Microdata+Entry+Page. Accessed on 23 May 2014

Australian Bureau of Statistics. 2013. *Remote Access Data Laboratory (RADL)*. Available at http://abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+Remote+Access+Data+Laboratory+(RADL). Accessed on 23 May 2014

Canada. *Statistics Act*. Last amended in 2005. Available at: http://laws-lois.justice.gc.ca/eng/acts/S-19/FullText.html. Accessed on on 23 May 2014
Childinfo. *Multiple Indicator Cluster Surveys/MIC2 – National datasets.* Website. Available at: http://www.childinfo.org/mics2_datasets.html. Accessed on 23 May 2014

DDI Alliance. 2014. *Data Documentation Initiative*. Website. Available at: http://www.ddialliance.org/. Accessed on 23 May 2014

The Department of Census and Statistics (DCS) for Sri Lanka. 2007. *Micro-data dissemination policy of the Deparmtnet of Census and Statistics (DCS)*. Available at: http://www.statistics.gov.lk/databases/data%20dissemination/DataDissaPolicy_2007Oct26.pdf. Accessed on 23 May 2014

The DHS Program. N.d. *Data – Access Instructions*. Website. Available at: http://dhsprogram.com/data/Access-Instructions.cfm. Accessed on 23 May 2014

Dupriez, O., and Boyko, E. 2010. *Dissemination of Microdata Files: Principles, Procedures and Practices*. IHSN Working Paper No. 005. International Household Survey Network. Available at: http://www.surveynetwork.org/home/sites/default/files/resources/IHSN-WP005.pdf . Accessed on 17 July 2014.

Eurostat. 2013. *Access to microdata.* Website. Available at: see http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/introduction. Accessed on 23 May 2014

Eurostat. 1996. *Manual on Disclosure Control Methods*. Eurostat Publication:

Bruxelles, Luxembourg. Available at: http://ec.europa.eu/eurostat/ramon/statmanuals/files/manual_on_disclosure_control_methods_1996.pdf. Accessed on 23 May 2014

FAO. 2000. *2000 World Census of Agriculture: Main Results and Metadata by Country*. FAO Publication: Rome. Available at: http://www.fao.org/economic/ess/ess-wca/wca90-country0/en/. Accessed on 12 January 2014.

Federal Government of Nigeria. 2011*. Nigerian General Household Survey – Panel 2010-2011 (Post-Planting), First Round (Wave 1)*. Available at: http://www.nigerianstat.gov.ng/nada/index.php/catalog/31. Accessed on 4 April 2014.

Gardner, T. "Steven Ruggles, Census Data Processing, Part 2." *Research Matters.* United States Census Bureau. Website. Accessed on 26 January 2014.

Habimana, D. Director, Statistical Methods, Research and Publication Unit National Institute of Statistics of Rwanda. 17 April 2014. E-mail message.

International Household Survey Network. 2014. *Central Data Catalog*. Available at http://catalog.ihsn.org/index.php/catalog/4149. Accessed on 23 May 2014

International Household Survey Network. n.d. *Accelerated Data Program.* Available at: http://adp.ihsn.org/node/203. Accessed on 15 March 2014.

International Household Survey Network. n.d. *Survey Catalogs*. Available at: See http://www.ihsn.org/home/survey-catalogs. Accessed on 26 January 2014.

Inter-university Consortium for Political and Social Research (ICPSR). n.d. *Guide to Social Science Data Preparation and Archiving Phase 5: Preparing data for Sharing*. Website. Available at: http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter5.html. Accessed on 23 May 2014

Inter-university Consortium for Political and Social Research (ICPSR). n.d. "Confidentiality." *Data Management & Curation*. Website. Available at: http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/index.html. Accessed on 23 May 2014

Inter-university Consortium for Political and Social Research (ICPSR). 2009. *Principles and Good Practice for Preserving Data", IHSN Working Paper No. 003.* International Household Survey Network. Available at http://www.ihsn.org/home/sites/default/files/resources/IHSN-WP003.pdf. Accessed 17 July 2014.

Leibniz Institute for the Social Sciences. 2014. *Access to Eurobarometer primary data*. Website. Available at: http://www.gesis.org/en/eurobarometer/data-access/. Accessed on 23 May 2014

Matthias, T., Meindl, B. & Kowarik, A. 2014. *Introduction to Disclosure Control*. Data-Analysis OG Publication: Vienna. Available at: http://cran.r-project.org/web/packages/sdcMicro/vignettes/sdc_guidelines.pdf. Accessed on 15 April 2014.

Minnesota Population Center, University of Minnesota. n.d. "How do I get access

to IPUMS-International data?" Frequently Asked Questions (FAQ). *Integrated Public Use Microdata Series, International.* Available at https://international.ipums.org/international-action/faq#ques6. Accessed on 23 May 2014

Rwandan National Institute of Statistics (NISR), Ministry of Agriculture and Animal Resources (MINAGRI) and World Food Programme. 2012. *Rwanda Comprehensive Food Security and Vulnerability Analysis and Nutrition Survey 2012*. Ref. RWA_2012_CFSVA_v01_M. Data set downloaded from http://nada.vam.wfp.org/index.php/catalog, on 6 March 2014.

Statistical Microdata. 2005. In Organisation for Economic Cooperation and Development, Glossary of Statistical Terms. Available at http://stats.oecd.org/glossary/detail.asp?ID=1656. Accessed on 1 January 2014.

Statistics Canada. n.d. *The Research Data Centres (RDC) – Application process and guidelines*. Available at: http://www.statcan.gc.ca/rdc-cdr/process-eng.htm. Accessed on 23 May 2014

Statistics Canada. n.d. *The Research Data Centres (RDC) Program*. Available at: http://www.statcan.gc.ca/rdc-cdr/index-eng.htm. Accessed on on 23 May 2014

Statistics Canada. n.d. *Data Liberation Initiative*. Available at: http://www.statcan.gc.ca/dli-idd/dli-idd-eng.htm. Accessed on 20 January 2014.

Statistics Canada. n.d. *Open Data Agreement*. Available at: http://www.statcan.gc.ca/eng/reference/licence-eng. Accessed on 15 March 2014.

Statistics Canada. 2011. *Census of Agriculture*. Available at: http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3438. Accessed on 23 May 2014

Statistics Canada. 2012. "The RTRA system". *Statistics Canada*. Website. Available at http://www.statcan.gc.ca/rdc-cdr/rtra-adtr/inf-eng.htm. Accessed on 23 May 2014

Statistics Canada. "The Canadian Centre for Data Development and Economic Research (CDER)." *Statistics Canada*. Website. Available at: http://www.statcan.gc.ca/cder-cdre/. Accessed on 23 May 2014

Thorne, R., Nicholls, P. and Boettcher, K. Microdata Dissemination Architectures and Systems. *Australian Bureau of Statistics*. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2011/wp.7.e.pdf. Accessed on 14 March 2014.

UK Data Archive. "Create & Manage Data: Documenting Your Data." *UK Data Archive*. Website. Available at: http://www.data-archive.ac.uk/create-manage/document/metadata. Accessed on 23 May 2014

United Kingdom Data Service. n.d. *British Household Panel Survey – Series*. Website. Available at: http://discover.ukdataservice.ac.uk/series/?sn=200005. Accessed on 23 May 2014

United Nations Economic Commission for Europe – Conference of European

Statisticians. 2007. *Managing Statistical Confidentiality & Microdata Access: Principles and Guidelines of Good Practice.* United Nations Publication: New York and Geneva. Available at http://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf. Accessed on 23 May 2014

United Nations Statistical Office. 2006. *Principles Governing International Statistical Activities*. Available at http://unstats.un.org/unsd/methods/statorg/Principles_stat_activities/principles_stat_activities.htm. Accessed on 14 March 2014.

United Nations Statistics Division. 2006. *Principles Governing International Statistical Activities*. Available at: http://unstats.un.org/unsd/methods/statorg/Principles_stat_activities/principles_stat_activities.htm. Accessed on 25 January 2014.

United States Census Bureau. n.d. *American Community Survey*. Website. Available at: http://www.census.gov/acs/www/. Accessed on 23 May 2014

United States Census Bureau. 2014. "Center for Economic Studies (CES)." *Census.gov.* Website. Available at: http://www.census.gov/ces/dataproducts/economicdata.html. Accessed on 23 May 2014

United States Department of Agriculture – Economic Research Service. n.d. *Agricultural Resource Management Survey (ARMS) Farm Financial and Crop Production Practices***.** Available at: http://www.ers.usda.gov/data-products/arms-farm-financial-and-crop-production-practices.aspx#.UvpaffldWh0. Accessed on 23 May 2014

United States of America. 2005. *US Federal Committee on Statistical Methodology. Statistical Policy Working Paper 22 (Revised 2005) – Report on Statistical Disclosure Limitation Methodology*. Available at: http://www.fcsm.gov/working-papers/SPWP22_rev.pdf. Accessed on 06 January 2014.

Watkins, W. & Boyko, E. 1996. Data Liberation and Academic Freedom. Government Information in Canada/Information gouvernementale au Canada, 3(2). Available at: http://www.usask.ca/library/gic/v3n2/watkins2/watkins2.html. Accessed on 23 May 2014

Willis-Núñes, F. 2013. *Annex 1.15. Data laboratory microdata access – Brazil*. UNECE Statistics Wikis. Website. Accessed on 23 May 2014

Appendix A: Terms and conditions of use of public data files

1. The data and other materials provided by the [SO] will not be redistributed or sold to other individuals, institutions, or organizations without the written agreement of the [SO].

2. The data will be used for statistical and scientific research purposes only. They will be used solely for reporting aggregated information, and not for investigation of specific individuals or organizations.

3. No attempt will be made to re-identify respondents, and no use will be made of the identity of any person or establishment discovered inadvertently. Any such discovery would be reported to the [SO] immediately.

4. No attempt will be made to produce links among data sets provided by the [SO] or among data from the [SO] and other data sets that could identify individuals or organizations.

5. Any books, articles, conference papers, theses, dissertations, reports, or other publications that employ data obtained from the National Data Archive will cite the source of data, in accordance with the Citation Requirement provided with each data set.

6. An electronic copy of all reports and publications based on the requested data will be sent to the [SO].

7. The original collector of the data, the [SO], and the relevant funding agencies, bear no responsibility for the use made of the data or for interpretations or inferences based upon such uses.

Appendix B: Form for access to licensed file

Application for Access to a Licensed Data set

Title and reference number of the data set(s) you are requesting (use the exact title, year and reference number as listed in our survey catalogue):

Instructions

This form is to be mailed or faxed to the ABCBOS, with a cover letter printed on the sponsoring agency letterhead.

Mail to: [address]
[address]
[address]
Fax to: [fax number, with country and area code]
E-mail scanned copy to: [email address]

Access to licensed data sets is only granted when there is a legally registered sponsoring agency (government ministry, university, research centre, national or international organization, *etc*.).

Requests are reviewed by our data release committee. If approved, you will be provided with the data and documentation on CD-ROM/DVD or through a secure FTP server.

The information you provide in this form will not be shared with others, unless a breach to the legal agreement is confirmed, in which case the ABCBOS may inform partner statistical agencies in other countries.

Terms

In this agreement,

1. 'Primary Data Investigator' refers to the investigator who serves as the main point of contact for all communications involving this agreement. The Primary Data Investigator assumes all responsibility for compliance with all terms of this Data Access Agreement by employees of the receiving organization.

2. 'Other Investigators' refers to individuals other than the Principal Investigator, including research assistants, who will have access to the restricted data.

3. 'Receiving Organization' refers to the organization/university/establishment which employs the Primary Data Investigator.

4. 'Representative of the Receiving Organization' refers to an individual who has the authority to represent the Receiving Organization in agreements of this type.

Section A. **Primary Data Investigator**
First name
Last name
Title

Prof/ Dr/ Mr/ Mrs/ Ms
Organization
Position in organization
Postal address
Telephone (with country code)
Fax (with country code)
E-mail

Section B. Other Investigators

Provide names, titles, and affiliations of any other members of the research team who will have access to the restricted data.
Name (last / first), Position Affiliation

Section C. Receiving Organization

Organization name
Type of organization

- Line ministry / public administration
- University
- Research centre
- Private company
- International organization
- Non-governmental agency (national)
- Non-governmental agency (international)
- other (specify)

Organization website (URL)
Postal address

Section D. Representative of the Receiving Organization

First name
Last name
Title
Prof/ Dr/ Mr/ Mrs/ Ms
Position in organization
Postal Address
Telephone (with country code)
Fax (with country code)
E-mail

Section E. Description of intended use of the data

Please provide a description of your research project (research question, objectives, methods, expected outputs, partners). If the information given is insufficient, your request may be rejected or additional information will be requested. You may provide such information in an attached appendix to this request.

List of expected output(s) and dissemination policy

Section F. Identification of data files and variables needed

The ABCBOS provides detailed metadata on its website, including a description of data files and variables for each data set. Researchers who do not require access to the entire data set may indicate which subset of variables or cases are of interest. As this reduces the risk of disclosure, providing us with such information may increase the probability that the data will be provided.

This request is submitted to access:

- The entire data set (all files, all cases)
- A subset of variables and/or cases, as described below (note that variables such as sample weighting coefficients and records identifiers will always be included in subsets).

Section G. Data access agreement

The Primary Data Investigator, the Other Investigators, and the Representative of the Receiving Organization agree to comply with the following:

1. Access to the restricted data will be limited to the Primary Data Investigator and Other Investigators.

2. Copies of the restricted data or any data created on the basis of the original data will not be copied or made available to any party other than those mentioned in this Data Access Agreement, unless formally authorized by the ABCBOS.

3. The data will be processed only for the stated statistical purpose. They will be used for solely for reporting aggregated information, and not for investigation of specific individuals or organizations. Data will not be used for any administrative, proprietary or law enforcement purposes, in any way.

4. The Primary Data Investigator undertakes that no attempt will be made to identify any individual person, family, business, enterprise or organization. If such a unique disclosure is made inadvertently, no use will be made of the identity of any person or establishment discovered, and full details of the discovery will be reported to the ABCBOS. The identification will not be revealed to any other party not included in the Data Access Agreement.

5. The Primary Data Investigator will implement security measures to prevent unauthorised access to licensed microdata acquired from the ABCBOS. Upon completion of this research, the microdata must be destroyed, unless the ABCBOS obtains satisfactory guarantee that the data can be secured and provides written authorization as to their retention to the Receiving Organization. Destruction of the microdata will be confirmed in writing to the ABCBOS by the Primary Data Investigator.

6. Any books, articles, conference papers, theses, dissertations, reports, or other publications that employ data obtained from the National Data Archive will cite the

source of the data, in accordance with the citation requirement provided with the data set.

7. An electronic copy of all reports and publications based on the requested data will be sent to the ABCBOS.
8. The original collector of the data, the ABCBOS, and the relevant funding agencies bear no responsibility for the use made of the data or for interpretations or inferences based upon such uses.

9. This agreement will come into force on the date that approval is given for access to the restricted data set and remain in force until the end date of the project or an earlier date if the project is completed ahead of time.

10. If there are any changes to the project specification, security arrangements, personnel or organization detailed in this application form, it is the responsibility of the Primary Data Investigator to seek the agreement of the ABCBOS to these changes. Where there is a change to the employer organization of the Primary Data Investigator, this will involve a fresh application being made, and termination of the original project.

11. Breaches of this agreement will be taken seriously and the ABCBOS will take action against those responsible for the lapse if wilful or accidental. Failure to comply with the directions of the ABCBOS will be deemed to be a major breach of the agreement and may involve recourse to legal proceedings. The ABCBOS will maintain and share with partner data archives a register of those individuals and organizations responsible for breaching the terms of the Data Access Agreement, and will impose sanctions on releasing future data to these parties.

Signatories

The following signatories have read and agree with the Data Access Agreement as presented in Section G above:

The Principal Data Investigator

Name _____ Signature _____ Date _____

The Representative of the Receiving Organization

Name _____ Signature _____ Date _____

Appendix C: Generic Microdata Dissemination Policy

1.1 Introduction

In its quest to maximize the use of official statistics by users, and in line with Section xx of the Statistics Act, the NSO makes available anonymised micro-data to users for research purposes. Microdata, which are organised in electronic data files, refer to the information about each unit of observation. This policy addresses the conditions and the manner in which anonymised micro-data files may be released.

1.2 Policy Objective

The objective of this policy is to define the nature of the anonymised microdata files that will be released, the intended use of these files and the conditions under which these files will be released.

1.3 Rationale

The NSO is committed to achieving excellence in the provision of timely, reliable and affordable official statistics for informed decision making, to maximize the welfare of the general population. This policy aims to support the needs of specialised researchers, students and other users by providing anonymised microdata files, to be used strictly for research purposes.

1.4 Policy Statement

The NSO will release microdata files for use by potential users for research purposes, when:

- The Director of Statistics is satisfied that all reasonable steps have been taken to prevent the identification of individual respondents;
- The release of the data will substantially enhance the analytic value of the data that have been collected;
- The users have disclosed the nature and objectives of their intended research;
- The users have signed an appropriate undertaking and
- The applicable legal provisions have been observed.

1.5 Definitions

1. For the purposes of this policy, microdata are defined as files of records pertaining to individual respondent units. Microdata files for dissemination purposes are those in which all direct and indirect identifiers have been removed, through various anonymization procedures.
2. Anonymization refers to the process of removing direct and indirect identifiers from a survey, census or administrative data file, to conceal the characteristics and the identity of individual respondents.
3. Direct identifiers include information such as names, addresses or other direct personal identifiers, which must be removed from all files made available to users.

4. Indirect identifiers refer to characteristics that are shared with several other respondents and that, when combined with other information, may compromise an individual respondent's identity.
5. For the purposes of this policy, dissemination refers to the act of making microdata files and supporting metadata available for access and use with the data.

1.6 Types of Microdata files

The following two types of microdata files are covered by this policy:

1. **Public Use Files (PUFs):** Microdata files that are disseminated for general public use. They have been highly anonymized by removing the names and addresses of respondents, and by collapsing geographic and respondent characteristic details to ensure that identification of individuals is highly unlikely. These files may be downloaded from the [NSO Name]'s website to individuals who identify themselves by name, provide their email addresses and agree to abide by the terms and conditions appropriate for a PUF (See Appendix A).

2. Licensed files: These files require that there be a signed agreement between [NSO Name] and major users, to permit them to access data files that are less highly anonymized and/or more sensitive than PUFs. For these files, all individual identifiers have been removed and some characteristic details may be collapsed or removed. Licensing agreements are only entered into with bona fide users, working for registered organizations. The primary and secondary researchers must be identified by name, and a responsible officer of the organization must co-sign the license agreement.

1.7 Classes of users

The NSO recognizes that microdata files are intended for specialized users with advanced quantitative skills. This policy distinguishes the following classes of users:

- Policymakers and researchers employed by line-ministries and planning departments engaged in the development of regional and national strategies and programs, including the monitoring and evaluation of these programs;
- International agencies involved in the conduct of special studies aimed at identifying development and support opportunities and the development programs and infrastructures within the NSO's country;
- Research and academic institutions involved in social and economic research;
- Students and professors engaging mainly in educational activities, and

Other users involved in conducting scientific research (to be approved on a case-by-case basis).

1.8 Notes on the degree of anonymization

The terms of this policy cover two types of anonymized files, distinguished mainly by the levels of geographic and characteristic detail. The types of files are:

1. Files having less geographic and less characteristic detail (i.e. the geography has been collapsed and the variable detail reduced), and
2. Files that have less geographic detail and more variable detail.
   If all variables are available in a file and the geography is detailed, then the risk of identifying a respondent is greater. Accordingly, users requiring the maximum level of geographic detail may have to be prepared to work with less characteristic (variable) detail.
   Files that have had both the variable detail and the geographic codes reduced are invaluable for students and professors for teaching and learning purposes so long as they are easily available to them. Such files can be made available as PUFs from the NSO web site (see below).
   All files will be reviewed by a Microdata Release Committee prior to release.

1.9 Notes on Access Methods

Public Use Files are made freely available, by having users complete a form in which they provide their contact information.
Licensed files are made available after the user completes and returns an access form, together with an appropriate fee (see Section 1.12 below).

1.10 Responsibilities

Officers in charge of surveys/censuses will be responsible for:

- identifying the needs of key stakeholders and ensuring the creation of an anonymized file that meets the needs of the user community, to the extent possible under the Statistics Act;
- preparing an initial screening of the microdata file, identifying any potential problem areas that must be resolved and preparing a submission to the Microdata Release Committee.

The Microdata Release Committee

This Committee will be responsible for:

- reviewing all requests for the release of anonymized microdata files, using established criteria;
- approving all files for release or providing guidance to Officers in charge of surveys/censuses on how to improve the file before it may be released;
- overseeing the licensing process and resolving issues dealing with possible breaches;
- revising the guidelines used by Officers in charge of surveys/censuses to create anonymized microdata files as necessary.

The Officer in charge for dissemination

This Officer will be responsible for:

- providing access to the data files for users who have been cleared by the Microdata Release Committee; and
- responding to users' requests for support and additional information.

The Director of Statistics

The Director of Statistics will approve all releases of anonymized microdata files to users, based on the advice and recommendations of the Microdata Release Committee.

1.11 Timing and Release Policy

Recognizing the importance of the needs of users and of timely data, the NSO will strive to release microdata files within 6-12 months after the first release of data from a survey/census.

1.12 Cost Recovery Policy

It is the policy of the NSO to encourage broad use of its products by making them affordable for users. Accordingly, the NSO attempts to ensure that the costs of creating anonymized microdata files are built into the office budget. However, the NSO attempts to recover costs associated with the provision of special services that benefit only a specific group.
The following cost recovery principles apply to the dissemination of microdata files:

- Public use files will be made available from our website to all users, without charge.
- Licensed data files will be made available to specialized users as per the NSO's Pricing Policy.

1.13 Policy Update

This policy was adopted on [... date ...] and became effective on [... date ...].

1.14 Further Information

For further information on access to microdata, please contact: []

Prepared on [day month year]