IHSN

INTERNATIONAL HOUSEHOLD SURVEY NETWORK

Dissemination of Microdata Files

Principles, Procedures and Practices

Olivier Dupriez and Ernie Boyko

IHSN Working Paper No 005
August 2010

www.ihsn.org

Dissemination of Microdata Files Principles, Procedures and Practices

Olivier Dupriez and Ernie Boyko

IHSN Working Paper No 005

August 2010

Abstract

In all countries, data producers are faced by expanding demand for microdata. Determining the best way to disseminate these data is a challenge. The challenge is technical, as data producers have to implement procedures for the documentation, cataloguing and dissemination of the data. The challenge is also legal and ethical. While data producers are often well aware of the power and importance of microdata, they have to balance this demand with the need to keep respondent information confidential. This is a requirement of a country's statistical and privacy legislation and is often an undertaking given to respondents when the information is collected. This imposes the establishment of formal policies and procedures defining the conditions of access to microdata. This guide provides an overview of such policies and procedures, and documents existing good practice.

About the Authors

Ernie Boyko is a former staff member of Statistics Canada where he held a number of Directorships, including Agriculture, Corporate Planning, Electronic Publishing, and Operations for the 1991 Census. His responsibilities included overseeing the work of the Canadian *Data Liberation Initiative*. He is an active member of the Canadian Association of Public Data Users and the International Association for Social Science Information Services and Technology.

Olivier Dupriez is a Senior Economist-Statistician in the World Bank Development Data Group and also the Manager of the International Household Survey Network. He coordinates technical assistance programmes for a large number of countries in areas related to microdata documentation and dissemination.

Acknowledgments

The document was developed for the International Household Survey Network (IHSN) with financial support from the World Bank Development Grant Facility, Grant No 4001009-06, administered by the PARIS21 Secretariat.

It was prepared by Ernie Boyko and Olivier Dupriez with contributions, in the form of input or comments and suggestions, from François Fonteneau (OECD-PARIS21), Julia Lane (National Opinion Research Center,, University of Chicago), Johan Mistiaen (World Bank), Dennis Trewin, and Wendy Watkins (Carleton University, Canada). Discussions with many colleagues in IHSN-member agencies and with official statisticians in a number of countries were another important input. John Wright edited the document, and typesetting was by Rhommell Rico.

The paper (or a revised copy) is available on the website of the International Household Survey Network at www.ihsn.org

Citation

Dupriez, Olivier and Ernie Boyko. 2010. "Dissemination of Microdata Files. Formulating Policies and Procedures", International Household Survey Network, IHSN Working Paper No 005.

Table of Contents

List of Appendices

List of Figures

List of Boxes

List of Acronyms

ABS	Australia Bureau of Statistics
ACS	American Community Survey
API	Application-programming interfaces
ASCII	American Standard Code for Information Interchange
CENEX	Centre of Excellence for Statistical Disclosure Control
CESSDA	Council of European Social Science Data Archives
CSO	Central Statistics Office
CURF	Confidentialised Unit Record Files
DCMI	Dublin Core Metadata Initiative
DDI	Data Documentation Initiative
DHS	Demographic and Health Survey
DLI	Data Liberation Initiative (Statistics Canada)
GPS	Global Positioning System
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
ICPSR	Inter-University Consortium for Political and Social Research
IHSN	International Household Survey Network
ISO/IEC	International Organization for Standardization/International Electrotechnical Commission
JSI	Job Submission Interface
LIS	Luxemburg Income Study
MCRDC	Michigan Census Research Data Center
MICS	Multiple Indicator Cluster Surveys
MIT	Massachusetts Institute of Technology
NCHS	US National Center for Health Statistics
NCSA	National Center for Supercomputing Applications
NDE	National Data Enclave
NGO	Non-governmental Organisation
NORC	National Opinion Research Center (University of Chicago)
NSD	Norwegian Social Science Data Services
NSO	National Statistics Office
NSS	National Statistical System
OAIS	Open Archival Information System
OCLC	Online Computer Library Center
OECD	Organisation for Economic Co-operation and Development
PDF	Portable Document Format
PSU	Primary sample unit
PUF	Public Use File
PUMA	Public Use Microdata Areas
PUMS	Public Use Microdata Sample
RDC	Research Data Centre (Canada)
SAS	Statistical Analysis System (software)
SDC	Statistical Disclosure Control
SNZ	Statistics New Zealand
SOAP	Simple Object Access Protocol
SQL	Structured Query Language
UKDA	United Kingdom Data Archive (University of Essex)
UN	United Nations
UNECE	United Nations Economic Commission for Europe
UNF	Universal Numeric Fingerprint
UNICEF	United Nations Children's Fund
UNSD	United Nations Statistics Division

URL	Uniform Resource Locator
US	United States of America
USB	Universal Serial Bus
XML	eXtensible Markup Language
XSL	Extensible Stylesheet Language

Introduction

Collection of statistical data to support a country's public and private decision-making is an enormous enterprise generally supported by public funds. Ensuring maximum return from this investment by promoting the use of these data is the responsibility of all publicly-funded data producers, researchers and sponsor organisations.

Socio-economic data, which are the focus of this document, are obtained typically from censuses, sample surveys and administrative recording systems. These activities produce *microdata* (data at the level of the individual respondent). Microdata can then be processed (edited, analysed and tabulated) before being made accessible to users. Traditionally, outputs have consisted of aggregated data presented in the form of tables, charts, briefs, descriptive reports and analytical papers. Content of these tables and reports is determined by their value to producers and sponsors. Most data collection activities are undertaken for a specific purpose: satisfying this purpose is the priority and often the only objective of the data producer or sponsor. But microdata collected for one purpose can often serve many others, including some that cannot be anticipated at the time of collection. In other words, microdata can be re-purposed. Unlocking them for access to the research community is a cost-effective and efficient way of multiplying and diversifying the analysis and use of existing information. If thoroughly exploited, these data offer almost inexhaustible opportunities to generate new knowledge.

Since the 1980s the increasing power of computers and software has made microdata all the more desirable to researchers. In all countries, data producers are faced by an expanding demand for access to the underlying microdata on which published statistics are based. Access to microdata not only enables new and more diverse research, it also allows the development of innovative ways of using, processing and displaying information, and the generation of new datasets by combining data from multiple sources.

However, determining the best way of disseminating microdata is a challenge for data producers. The challenge is both technical and organisational, as data producers must implement proper procedures for documenting, cataloguing and sharing their microdata. International standards and good practice have been developed by the data archive community to address such issues. But the challenge is also legal and ethical. While data producers are well aware of the power and importance of microdata sharing, they have to balance this with the need to keep respondents' information confidential. This is a requirement of a country's statistical and privacy legislation – and often an undertaking data producers give respondents when collecting the information. Statistical agencies and other data producers must ensure they maintain the trust and confidence of respondents. Without such trust, cooperation with surveys would diminish and the quality of statistics suffer. Therefore, microdata dissemination demands the establishment of policies and procedures to define formally the conditions of access to microdata.

The context within which this can be accomplished varies from country to country. But "whatever differences there may be between practices of, and policies on, data sharing, and whatever legitimate restrictions may be put on data access, practically all research could benefit from more systematic sharing." [17]

This guide is intended to help microdata producers and depositors develop their own policies and procedures for disseminating microdata files. It is important such policies and procedures are formal and transparent. Proper microdata dissemination involves not only the provision of data and related documentation, but also of the conditions attached to using the data. This information should be made visible and easily accessible, preferably via the Internet.

While most of this guide is generic, it is meant primarily for official data producers – national statistical offices and line ministries –– in developing countries. And when the guide mentions data, it refers typically to microdata obtained from sample surveys, censuses and administrative data collection systems.

The guide was produced under the auspices of the International Household Survey Network (IHSN). It draws heavily from work by the United Nations Economic Commission for Europe, by a Task Force on Managing Statistical Confidentiality and Microdata Access set up by the Conference of European Statisticians, and by Eurostat [5] [24] [25] [26]. It also benefits from the experience of statistical agencies in those parts of the world where the practice of providing

access to microdata files has been in existence for more than 40 years in some cases, and of various academic data centers.

The information in this guide answers twelve key questions to be addressed when formulating a policy to disseminate microdata files. These are:

1. What are microdata?
2. What are metadata?
3. Why should data producers disseminate microdata?
4. What are the costs and risks, and how can they be addressed?
5. To whom should microdata be made available?
6. Under what conditions should microdata be provided?
7. What is meant by microdata 'anonymisation'?
8. Should microdata be sold or provided free of charge?
9. When in the dissemination cycle should microdata be released?
10. What are the technical infrastructure requirements for disseminating microdata?
11. What are the institutional requirements for disseminating microdata?
12. How to promote use of microdata?

The guide mainly addresses policy aspects of microdata dissemination. Proper and safe microdata dissemination also requires appropriate technical solutions for documenting, making anonymous, cataloguing and preserving data and metadata. These challenges are addressed here briefly but covered in more detail in other guidelines by the IHSN or others.

1. What are microdata?

1.1 What are microdata?

When statistical agencies or other data producers conduct surveys or censuses or collect administrative data, they gather information from each unit of observation. Such a unit can be a household, a person, a firm or enterprise, an agricultural holding, a school, a health facility, or other. In the context of this guide, *microdata* are the electronic data files containing the information about each unit of observation. Microdata are thus opposed to *macrodata* or *aggregated data*, which provide a summarised version of this information in the form of means, ratios, frequencies or other summary statistics.

Typically, microdata are organised in data files in which each line (or *record*) contains information about one unit of observation. This information is stored in *variables*. Variables can be of different types (e.g. numeric or alphanumeric, discrete or continuous, etc). They can be obtained directly from the respondent via a questionnaire or by observation or measurement (e.g. by GPS positioning) or imputed or calculated.

Information in statistical microdata files is stored in the form of coded values. For example, the sex of the respondent may be stored in a variable named 'H01a': this would include values 1 or 2, where 1 is the code for male and 2 for female. Therefore, microdata must be accompanied by a *data dictionary* containing the list of variables, a description of their content and the meaning of each code used. Such is the minimum documentation or *metadata* that must always accompany the data. Chapter 2 shows that much more metadata are actually required.

Typically, a survey or census dataset comprises multiple data files, often resulting from multiple levels of observation in the same data collection operation. In most cases, household surveys and censuses collect data at a minimum of two levels: the household (with, for example, variables describing dwelling characteristics) and the individual (with, for example, information on age, marital status, education level and economic activity). The dataset may comprise one or multiple file(s) at each of these levels. The data files contain some variables named *key variables*; these allow users to link information from one file to that of another file. Datasets organised thus are named *hierarchical datasets*.

1.2 In what format are microdata files stored and disseminated?

Microdata files can be stored in different formats. Common formats include the non-proprietary ASCII format and proprietary formats like those generated by specialised statistical software such as SAS, SPSS and Stata. Microdata can also be stored in SQL or other database formats. However, this is less common and less convenient for survey and census data as database applications are not specifically designed for statistical tabulation and analysis.

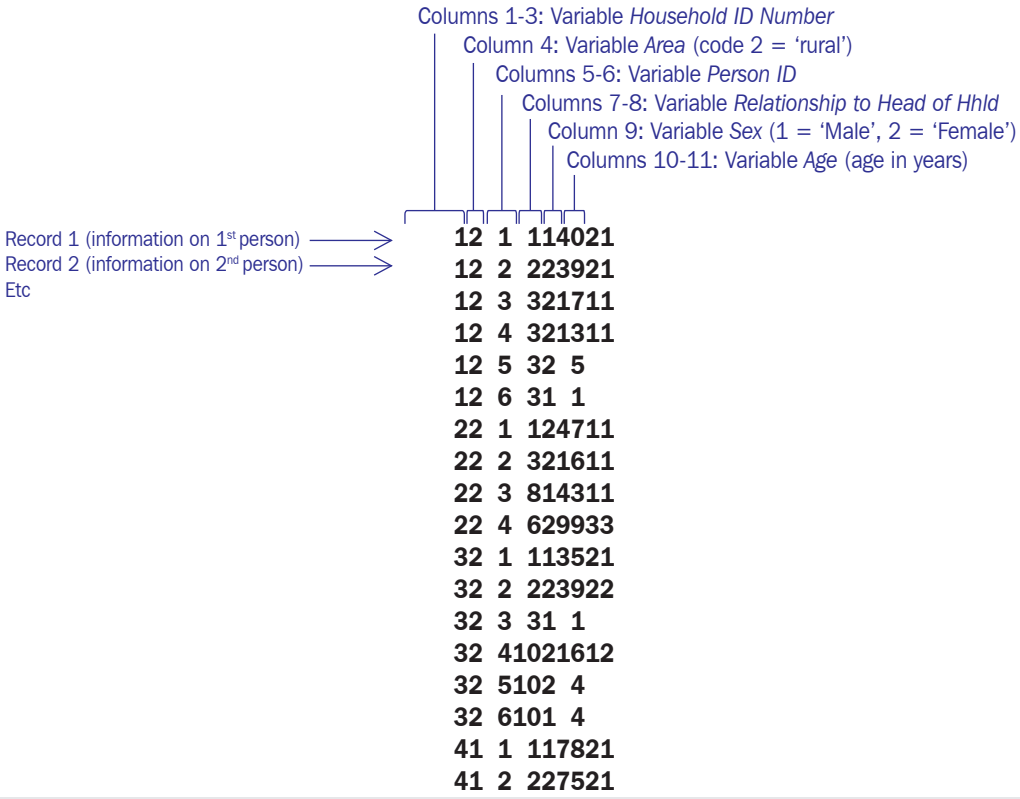
The ASCII file format is not specific to any particular software or platform. ASCII data files only contain data, readable by most software applications. As they are not associated with software liable to obsolescence, ASCII files are optimal to guaranteeing long-term data preservation. But ASCII files cannot be understood or used unless a data dictionary is provided as a separate file or document. Figure 1 presents a screenshot of a typical statistical data file in fixed ASCII format.[1]

To tabulate or analyse ASCII data, users must first import them into other software. All statistical and database software applications offer tools and commands for this purpose. Below is an example of a Stata script that would import the ASCII data shown in Figure 1 and add labels to the variables and codes to make the data file more user-friendly. Obviously, writing such scripts requires the user to be provided with a *data dictionary* giving information on content and structure of the ASCII data file.

Once imported in Stata by running this script, the ASCII file shown in Figure 1 will be displayed as in Figure 2. Proprietary file formats from SAS, SPSS, Stata or equivalent software include both the data and the variables and value labels.

1 ASCII files can be *fixed* or *character delimited*. In fixed ASCII files, data related to a variable always will be found in the same position (column). In ASCII delimited files, information related to each variable is separated by a special character (a semi-colon, a tab, a comma or other user-defined character). For example, in a *comma-delimited* ASCII file each variable would be separated by a comma.

Figure 1 **Screenshot of a Fraction of an ASCII Fixed Format Data File**



Example of Stata 'set-up' for importing ASCII data

```
* · Read the ASCII data found in file test.dat and import the values in new variables;
· · infix hhid 1-3 area 4 pid 5-6 relat 7-8 sex 9 age 10-11 using test.dat;

* · Add a label to describe each new variables;
· · label variable hhid "Household ID";
· · label variable area "Area";
· · label variable pid "Person ID";
· · label variable relat "Relationship to head of household";
· · label variable  sex "Sex";
· · label variable age "Age in completed years";

* · Add label to each code used by the variables;
· · label define relatcod 1 "Head"  2 "Spouse" 3 "Son/Daughter" 4 "Son/Daughter-in-law"
· · 5 "Grandchild" 6 "Parent" 7 "Parent in law" 8 "Brother/sister" 9 "Other relative"
· · 10 "Not related", add;
· · label values relat relatcod;
· · label define areacod 1 "Urban"  2 "Rural", add;
· · label values area areacod;
· · label define sexcod 1 "Male"  2 "Female", add;
· · label values sex sexcod;

* · Save the file as a Stata file;
· · save "test.dta", replace;
```

Figure 2 **Screenshot of a Fraction of a Stata Data File**

	hhid	area	pid	relat	sex	age
1	1	Rural	1	Head	Male	40
2	1	Rural	2	Spouse	Female	39
3	1	Rural	3	Son/Daughter	Female	17
4	1	Rural	4	Son/Daughter	Female	13
5	1	Rural	5	Son/Daughter	Female	5
6	1	Rural	6	Son/Daughter	Male	1
7	2	Rural	1	Head	Female	47
8	2	Rural	2	Son/Daughter	Female	16
9	2	Rural	3	Brother/sister	Male	43
10	2	Rural	4	Parent	Female	99
11	3	Rural	1	Head	Male	35
12	3	Rural	2	Spouse	Female	39
13	3	Rural	3	Son/Daughter	Male	1
14	3	Rural	4	Not related	Female	16
15	3	Rural	5	Not related	Female	4
16	3	Rural	6	Not related	Male	4
17	4	Urban	1	Head	Male	78
18	4	Urban	2	Spouse	Female	75

Survey datasets may contain hundreds of variables, even thousands. Writing scripts to import and document such data files from ASCII is time-consuming and may lead to errors. Therefore, for their users' convenience and to minimise the risk of errors, data providers should supply their datasets either in ASCII format but with accompanying pre-written SPSS, SAS and Stata scripts, or in the most common proprietary statistical formats. Specialised applications such as StatTransfer from Stata Corporation are available to convert data files automatically from one statistical package format to another.

1.3 Which version of the data files should be disseminated?

Data producers often create multiple versions of any given microdata file; these differ in the quality, content and number of records. They range from raw microdata files —containing all replies by each respondent obtained immediately after data entry— to cleaned and edited files for public use.

Figure 3 shows the typical life-cycle of a survey or census.

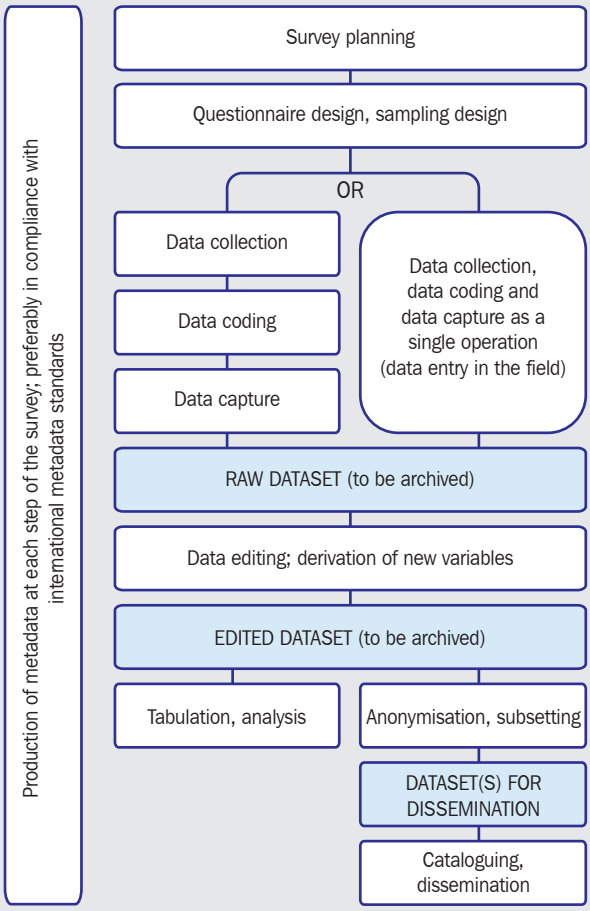
1.4 What is sensitive in microdata?

Data obtained from censuses and sample surveys are only to be used for statistical and research purposes. Information on the name or address of respondents is often collected in questionnaires for practical reasons but usually not captured in data files. Therefore, data files from sample surveys and censuses usually do not contain variables that are *direct identifiers*. The same is not true of administrative data files, which often include names, addresses, telephone numbers, social security numbers, etc.

However, most datasets include variables that are *indirect identifiers*. Detailed geographical information, composition of household by age and sex and detailed information on professional status, for example, could be used to attempt to identify respondents.

In addition to information considered sensitive because it may allow identification of units of observation, some variables in microdata sets can be sensitive due to the nature of the information contained in them. For example, this is so for information on health status, sexual behaviour, income, etc. Enterprise surveys are by nature sensitive as their information can be used by competitors.

Figure 3 **Life-cycle of a Survey**



1.5 What are the main types of microdata files for dissemination?

Microdata files for dissemination almost always differ from those strictly for use by staff of data producing agencies. Preparing raw microdata files for dissemination involves processes that may adjust the content and/or number of records. The *content of records* in microdata files for dissemination is edited by suppressing information from direct and indirect identifiers to protect the anonymity of respondents. But suppressing information does not necessarily mean removing variables. In some cases, re-coding variables into less detailed categories to make them less informative is sufficient. Sometimes this also requires truncating the *number of records* contained in a disseminated microdata file – especially in the case of population census data. Processes to safeguard respondents' identity are referred to collectively as *Statistical Disclosure Control* (SDC) or *anonymisation*.

Microdata files collected for official statistics should be disseminated if respondent confidentiality can be protected adequately. Consideration of three types of files is recommended when it comes to establishing dissemination policy: public use files, licensed files and data enclaves. These files differ in their level of accessibility to users and the extent to which they are anonymised.

"No individual (...) may claim entitlement to obtain or access identifiable data (...) by virtue of his or her employment. Access to identifiable data is not determined solely by employment status, organizational affiliation, or financial commitment. More important are the need for the identifiable data, the use to which the data will be put, and the requestor's role and responsibility with respect to the data collection activity. Since any access to identifiable data poses risk, access to such data will be carefully evaluated and tracked after access is granted." [14]

Public Use Files (PUF)

Public Use Files (PUFs) are available to anyone agreeing to respect a core set of easy-to-meet conditions. Such conditions relate to *what* can be done with the data (e.g. the data cannot be sold), not to access to the data. In some cases PUFs are disseminated with no conditions; often being made available on-line. These data are made easily accessible because the risk of identifying individual respondents is considered minimal. Minimising the risk of disclosure involves eliminating all content that can identify respondents directly—for instance, names, addresses and telephone numbers. In addition this requires purging relevant indirect identifiers from the microdata file. These vary across survey designs, but commonly-suppressed indirect identifiers include geographical information below the sub-national level at which the sample is representative. Occasionally, certain records may be suppressed also from PUFs, as might variables characterised by extremely skewed distribution or outliers. However, *in lieu* of deleting entire records or variables from microdata files, alternative SDC methods can minimise the risk of disclosure while maximising information content. Such methods include top-and-bottom coding, local suppression or using data perturbation techniques.[2] PUFs are typically generated from census data files – a

2 An overview and discussion of SDC methods to minimise the risk of disclosure is provided in Chapter 7.

sub-set of records rather than the entire file – and household surveys. While technically possible to create PUFs for business surveys, this presents a particular set of challenges that will be addressed separately.

PUFs should be as informative as possible. As stated by the US National Center for Health Statistics (NCHS,) "the objective is to make microdata available as widely and in the most detailed form possible, subject only to limits imposed by resources, data quality, technology, and the need to protect confidentiality." [14]

Licensed Files

Licensed Files – also called *Research Files* – are distinct from PUFs: their dissemination is restricted to users who have received authorisation to access them after submitting a documented application and signing an agreement governing the data's use. While typically licensed files are also anonymised to ensure the risk of identifying individuals is minimised when used in isolation, they may contain potentially identifiable data if linked with other data files.[3] Direct identifiers such as respondents' names must be removed from a licensed dataset. The data files may, however, still contain indirect variables that could identify respondents by matching them to other data files such as voter lists, land registers or school records.

When disseminating licensed files, the recommendation is to establish and sign an agreement between the data producer and external *bona fide* users – trustworthy users with legitimate need to access the data. Such an agreement should govern access and use of such microdata files. Sometimes, licensing agreements are only entered into with users affiliated to an appropriate sponsoring institution. i.e. research centres, universities or development partners. It is further recommended that, before entering into a data access and use agreement, the data producer asks potential users to complete an application form to demonstrate the need to use a licensed file (instead of the PUF version, if available) for a stated statistical or research purpose. Template licensed files' application forms and agreements are provided in Chapter 6 which discusses the conditions under which access to microdata files should be given.

Files accessible in data enclave

Some files may be offered to users under strict conditions in a *data enclave*. This is a facility equipped with computers not linked to the internet or an external network and from which no information can be downloaded via USB ports, CD-DVD or other drives. Data enclaves contain data that are particularly sensitive or allow direct or easy identification of respondents. Examples include complete population census datasets, enterprise surveys and certain health-related datasets containing highly-confidential information. Users interested in accessing a data enclave will not necessarily have access to the full dataset – only to the particular data subset 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 (an example of which is provided in Chapter 6). The outputs generated must be scrutinised by way of a full disclosure review before release.

Operating a data enclave may be expensive – it requires special premises and computer equipment. It also demands staff with the skills and time to review outputs before their removal from the data enclave in order to ensure there is no risk of disclosure. Such staff must be familiar with data analysis and be able to review the request process and manage file servers.

Because of the substantial operating costs and technical skills required, some statistical agencies or other official data producers opt to collaborate with academic institutions or research centres to establish and manage data enclaves. Examples of some data enclaves with informative websites include: the Michigan Census Research Data Center (MCRDC), a joint project of the US Census Bureau and the University of Michigan (www.isr.umich.edu/src/mcrdc/); the National Opinion Research Center (NORC) at the University of Chicago (www.norc.org/DataEnclave); the Research Data Centres (RDC) program of Statistics Canada (www.statcan.gc.ca/rdc-cdr/index-eng.htm); and the US NCHS Research Data Center (http://www.cdc.gov/nchs).

3 More detailed information on definitions and distinctions between Public Use Files and Licensed Files is provided in the work of the Economic Commission for Europe, Conference of European Statisticians [24].

Box 1 **One Survey, Multiple Products**

Data producers may decide to create more than one kind of product from a single census or survey dataset. For a population census it might include a PUF with a small sample or with a subset of variables, allowing its distribution widely to create an awareness of microdata products without the danger of respondent disclosure. They may also have an extended version with a bigger sample and which is licensed. And, finally, the full file (with or without identifiers) may be available in a data enclave.

The US Census Bureau, for example, produced two different public use files from the Census 2000 dataset with sampling rates of 1% and 5%.

> *Because of the rapid advances in computer technology and the increased accessibility of census data to the user community, the Census Bureau has had to adopt more stringent measures to protect the confidentiality of public use microdata through disclosure-limitation techniques. At the same time, the Census Bureau recognizes the needs of data users for greater characteristic detail and greater geographic specificity. Hence, two sets of files will be produced: one that provides a fuller range of detailed characteristics (the 1 percent national characteristics file) and one that provides greater geographic detail but less characteristic detail (the 5 percent state files).*
>
> Source: http://www.census.gov/population/www/cen2000/pums/index.html accessed on April 8, 2010.

The full microdata are available in various data enclaves in the US, such as the Michigan Census Research Data Center at the University of Michigan.

> *The Michigan Census Research Data Center (MCRDC) enables qualified researchers with projects approved by the United States Census Bureau to conduct research using unpublished data from the Census Bureau's economic and demographic programs and from the National Center for Health Statistics. All MCRDC research is conducted within its secure laboratory facility located in the Institute for Social Research at the University of Michigan in Ann Arbor.*
>
> Source: http://www.icpsr.umich.edu/icpsrweb/ICPSR/access/restricted/index.jsp accessed on May 7, 2010.

1.6 Are there alternatives to sharing microdata files?

In the situations described above, the user is offered access to microdata files. Other forms of access to microdata include remote job submission and remote access to microdata where users have no direct access to them. Policy requirements for such access are not covered by our guidelines but merely summarised here with references to enable the reader to learn more. Note that these approaches are typically costly and technically demanding.

Job Submission

One approach to making it possible for users to conduct analyses of confidential data is creating a process that enables them to submit data processing and analysis programs remotely to the data depositor. The user is given a synthetic dataset that replicates the structure and content of the actual datasets. This enables the researcher to develop programs using tools such as SAS, SPSS or Stata. The programs are then transmitted to the data depositor staff, who run the job against the actual dataset. The results are then vetted for disclosure and returned to the user.

One example of this approach is the Luxembourg Income Study. This provides access to its micro-databases through an automated job submission system called LISSY (Box 2).

While this protects data confidentiality, the cost of supporting job submission services can be high. And, if sufficient resources are not allocated, users may find it a slow process.

Remote Data Access

This consists of providing users with access to web-based data tabulation and analysis software, with no possibility of users downloading datasets or generating tables that would reveal individual or small numbers of records.

Various, mostly commercial, software applications are available (Nesstar, Beyond 20/20, SuperCross, Redatam, PcAxis and others). Some advanced data centres develop their own applications. For example, the UK Data Archive (UKDA) is piloting a Secure Data Service (SDS) "intended to promote excellence in research by enabling safe and secure remote access by bona fide researchers to data hitherto deemed too sensitive, detailed, confidential or potentially disclosive to be made available under standard licensing and dissemination arrangements." [12]

Box 2 **Luxemburg Income Study - LISSY**

LISSY, a fully-automated remote execution system running 24 hours a day, seven days a week, allows researchers to submit statistical batch programs (written in SAS, SPSS or Stata) from their own locations. LISSY automatically processes the jobs and returns aggregated results on average within a couple of minutes.

The micro-databases cannot be downloaded and no direct access to the data is permitted. Only results from statistical requests are returned to the users.

Registration is required

While the LIS Key Figures are available to the general public, the Luxembourg Income Study grants access to the micro-databases to registered users only and for a limited period of one year, renewable annually.

Two ways to submit jobs

LISSY provides secure remote access to the micro-data via two submission paths:

- a Job Submission Interface (JSI)
- email software such as Outlook, Thunderbird, etc.

Although these two paths generate identical results, the Luxembourg Income Study strongly recommends users to access LISSY through the Job Submission Interface (JSI). This is more user-friendly and provides additional features such as access to the user's job history.

Job submission instruction

Regardless of the way jobs are submitted, there are a few exceptions to the standard statistical programming syntax needed for LISSY to process user requests properly.

The statistical packages currently available in the LISSY system are SAS 9.2, Stata 11.0, and SPSS 11.5.

User support

All queries about use and content of the LIS databases are to be addressed to LIS User Support, rather than individual LIS staff members. This allows LIS staff to maintain a coordinated record of all queries.

Source: The content of this box was extracted (with minor editing) from http://www.lisproject.org/data access/data access.html accessed on April 9, 2010.

This approach is satisfactory for tabulations – in particular for population and housing censuses – but not for advanced analysis.

Engaging a researcher as a temporary staff member

Some data producers provide researchers with access to microdata by engaging them as temporary staff. This makes them subject to the same secrecy provisions as permanent staff of the data producing agency. Such arrangements should be limited to those where the researcher is actually assisting the work of the data producer; otherwise the arrangement could be seen as a sham. [24]

2. What are metadata?

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 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 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.

What constitute good metadata?

The description of good metadata provided below is extracted from the UK Data Archive's "Good Practices in Data Documentation." [20] Another source of useful information is the International Household Survey Network's website (www.ihsn.org) and their Quick reference Guide for Data Archivists. [4]

"A crucial part of creating a good dataset with long-lasting usability is ensuring that the data are easy to understand and analyze. This requires accompanying data description and documentation that is user-friendly, clear and detailed, yet comprehensive." (http://www.data archive.ac.uk)

There are three main types of material that constitute ideal documentation for a dataset:

1. Explanatory material

This is the minimum required to ensure the long-term viability and function of a dataset – without it, there cannot be full understanding of the dataset and its contents.

Information about data collection methods

This section describes the data collection process – whether a survey, collection of administrative information or transcription of a document source. It should cover the instruments used, and methods employed and how these were developed. If applicable, details of sampling design and frameworks should be included. It is also extremely useful to include information on any monitoring of data collection, as well as details of quality control.

Information about dataset structure

Central to this is a detailed document describing the dataset structure, including information on relationships between individual files or records within the study. It should include, for example, key variables required for unique identification of subjects across files. It should also include the number of cases and variables in each file, and the number of files in the dataset. For relational models, a information on the structure and relationship between records should be provided.

Technical information

This information relates to the technical framework and should include:

- the computer system used to generate the files;
- the software packages used to create the files;
- the medium in which the data were stored; and
- a complete list of all data files in the dataset.

Variables and values, coding and classification schemes

The documentation should contain a full list describing all variables (or fields) in the dataset, including a

complete explanation and full details of the coding and classifications used in the information allocated to those fields. It is especially important to have blank and missing fields explained and accounted for. It is also helpful to identify variables to which standard coding classifications apply, and to record the version of the classification scheme used – preferably with a bibliographical reference to that code.

Information about derived variables

Many data producers derive new variables from original data. This may be as simple as grouping raw age data (age in years) according to groups of years appropriate to the needs of the survey. Or it may be much more complex, requiring the use of sophisticated algorithms. When grouped or derived variables are created, it is important to make the logic for them clear. Simple grouping, such as for age, can be included within the data dictionary. More complex derivations require other means of recording. The best method of describing these is by using flow charts or accurate Boolean statements. It is recommended that sufficient supporting information be provided to allow an easy link between the core variables used and the resultant variables. It is further recommended that the computer algorithms used to create the derivations be saved, together with information on the software.

Weighting and grossing

Weighting and grossing variables need to be fully documented, with an explanation of the variables' construction and a clear indication of the circumstances in which they should be used. The latter is particularly important when different weights need to be applied for different purposes.

Data source

Details of the source from which the data are derived should be included in some detail. For example, when the data source is composed of responses to survey questionnaires, each question should be recorded carefully in the documentation. Ideally, the text will include reference to the generated variable(s). It is also useful to explain the conditions under which a respondent would ask a question, including, if possible, the cases to which it applies, and, ideally, a summary of response statistics.

Confidentiality and anonymisation

It is important to note if the data contain any confidential information on individuals, households, organisations or institutions. Whenever this occurs, it is recommended that such information be recorded, together with any agreement on how the data are to be used – for example, with survey respondents. Confidentiality issues may restrict the analyses to be undertaken or the results published, particularly if the data are to be made available for secondary use. If the data were to be made anonymous to prevent subject identification, it would be wise to record the anonymisation procedure and its impact on the data. Such modification may restrict subsequent analysis and some indication of it would be useful.

2. Contextual information

This provides users with material about the context of the collection of the data, and how they were used. This type of information adds richness and depth to the documentation. It enables the secondary user to understand fully the background to and processes behind the data collection exercise. This also forms a vital historical record for future researchers.

Description of the originating project

Details should be provided of the project's history, or the process that gave rise to the dataset. This should offer information on the intellectual and substantive framework. For example, the description could cover topics such as:

- why the data collection was felt necessary;
- aims and objectives of the project;
- who or what was being studied;
- geographical and temporal coverage;
- publications or policy development to which it contributed to or that arose in response; and
- any other relevant information.

Provenance of the dataset

This information relates to such aspects as the history of the data collection process, changes and developments that occurred in the data themselves and the methodology, or any adjustments made. In addition, he following can be provided:

- details of data errors;
- problems encountered during data collection, data entry, and data checking and cleaning;

- conversion to a different software or operating system;
- bibliographical references to reports or publications stemming from the study; and
- any other useful information on the dataset's life-cycle.

Serial and time-series datasets, new editions

For repeated cross-section, panel or time-series datasets, it is extremely helpful to obtain additional information describing, for example, changes in the question text, variable labelling or sampling procedures.

3. Cataloguing material

This material serves two purposes. Firstly, it is a bibliographical record of the dataset. This allows for the dataset to be acknowledged properly and cited in publications. The material also acts as a formal record for long-term preservation purposes. Secondly, it is the basic instrument used for resource discovery. This enables the dataset to be identified uniquely within the collection by providing appropriate information to help secondary users identify the study as useful to their purpose.

Without names, abstracts, keywords and other important metadata elements, it might be difficult for researchers to locate the datasets and variables to meet their requirements. Any cataloguing and resource-location systems, manual or digital, are based on metadata.

"Producing good data documentation is easiest when planned from the start of a project and considered throughout the course of research (during the data lifecycle). Advanced planning can significantly reduce the time and money needed to prepare documentation." (http://www.data archive.ac.uk). See also [21].

Metadata standards and good practice

"Technological and semantic interoperability is a key consideration in enabling and promoting international and interdisciplinary access to and use of research data. Access arrangements must pay due attention to the relevant international data documentation standards."[17]

Eager to facilitate data communication between organisations and software systems and improve the quality of statistical documentation provided to users of data, the data archive community has developed a set of metadata standards. These provide a structured framework for organising and disseminating information on content and structure of statistical information.

ISO 11179 - Information Technology - Metadata registries (MDR)

The International Standard ISO/IEC 11179-1 was developed by the Joint Technical Committee ISO/IEC JTC 1, Information Technology, Sub-Committee SC 32, Data Management Services. "ISO/IEC 11179 describes the standardizing and registering of data elements to make data understandable and shareable. Data element standardization and registration as described in ISO/IEC 11179 allow the creation of a shared data environment in much less time and with much less effort than it takes for conventional data management methodologies." [9]

ISO11179 is used by some data organisations to design and databases of concepts and definitions, but it is important to note that it does not provide a handy documentation and dissemination tool, contrary to the XML-defined standards described below.

The Data Documentation Initiative (DDI)

Traditionally, data producers wrote text-based codebooks. To take full advantage of web technology, most standards are now defined in XML language. The Data Documentation Initiative specification (or DDI) is a standard dedicated to microdata documentation. [13] [4]

The DDI developed standards that provide a structured framework for organising the content, presentation, transfer and preservation of metadata in the social and behavioural sciences. It enables documenting even the most complex microdata files in a way simultaneously flexible and rigorous.

The DDI seeks to establish an international XML-based standard for microdata documentation. Its aim is to provide a straightforward means of recording and communicating to others all

4 Description of DDI standard taken from http://www.ddialliance.org.

the salient characteristics of micro-datasets. The DDI specification is a major transformation of the once-familiar electronic 'codebook': it retains the same set of capabilities but greatly increases the scope and rigour of the information contained therein. The DDI metadata specification originated in the Inter-university Consortium for Political and Social Research (ICPSR), a membership-based organisation with more than 500 member colleges and universities worldwide. It is now the project of an alliance of institutions in North America and Europe. Member institutions comprise many of the largest data producers and data archives in the world.

The DDI specification is designed to encompass fully the kinds of data resulting from surveys, censuses, administrative records, experiments, direct observation and other systematic methodology for generating empirical measurements. In other words, the unit of analysis could be individual persons, households, families, business establishments, transactions, countries or other subjects of scientific interest. Similarly, observations may consist of measurements at a single point in time in a single setting – such as a sample of people in one country during one week. Or they may comprise repeated observations in multiple settings – including longitudinal and repeated cross-sectional data from many countries, as well as time-series of aggregated data. The DDI specification also provides for full descriptions of the study's methodology (mode of data collection, sampling methods if applicable, universe, geographical areas of study, responsible organisation and persons, and so on).

Structure

The DDI specification permits all aspects of a survey to be described in detail: the methodology, responsibilities, files and variables. It provides a structured and comprehensive list of hundreds of elements and attributes that may be used to document a dataset, although it is unlikely that any one study would use all of them. However, some elements, such as 'Title', are mandatory (and must be unique). Other elements are optional and can be repeated – for example 'Authoring Entity/Primary Investigator', since it includes information on the person(s) and/or organisation(s) responsible for the survey.

Box 3 **Who Uses the DDI ?**

The DDI metadata standard is used by a large community of data archivists, including data librarians from academia, data managers in national statistical agencies and other official data producing agencies, and international organisations.

Academic users include, for example:

- DataFirst at the University of Cape Town (www.datafirst.uct.ac.za)
- UK Data Archive at the University of Essex (www.data archive.ac.uk/)
- Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan (www.icpsr.umich.edu)
- Canadian universities participating in the Research Data Centres' programme (http://www.statcan.gc.ca/rdc-cdr/network-reseau-eng.htm)
- DataVerse Network at Harvard-MIT Data Center and Harvard University Library (http://thedata.org/)
- Agency members of the Council of European Social Science Data Archives (CESSDA) (http://www.cessda.org)

Official data producers in more than 50 countries, including:

- Statistics Canada with the Data Liberation Initiative (DLI) (http://www.statcan.gc.ca/dli-ild/dli-idd-eng.htm)
- National Institute of Statistics of Bolivia (http://www.ine.gov.bo/anda/)
- National Statistics Office, Bureau of Agricultural Statistics and Bureau of Labour and Employment Statistics of the Philippines (www.census.gov.ph www.bas.gov.ph www.bles.dole.gov.ph)
- Department of Census and Surveys of Sri Lanka (http://statistics.sltidc.lk/)
- Central Statistical Agency of Ethiopia (www.csa.gov.et)
- And many more (see www.ihsn.org/adp)

International organisations:

- UNICEF, for the Multiple Indicator Cluster Surveys (MICS) (http://www.childinfo.org/mics3_surveys.html)
- The World Bank (http://data.worldbank.org/)
- The Global Fund (http://www.theglobalfund.org/html/5YEdata/)

Adoption of the DDI metadata standard is greatly facilitated by availability of user-friendly software, such as the DDI Metadata Editor and other DDI-compliant cataloguing tools provided by the IHSN (see www.ihsn.org/toolkit and www.ihsn.org/nada).

DDI (version 2.4) elements are organised in five sections:

Section 1.0: Document Description

A study (survey, census or other) is not always documented and disseminated by the same agency as the one that produced the data. Therefore, it is important to provide information (metadata) not only on the study itself, but also on the documentation process. The *Document Description* consists of an overview describing the DDI-compliant XML document, or, in other words, 'metadata about metadata'.

Section 2.0: Study Description

The *Study Description* consists of an overview of the study. This section includes information on how the study should be cited; who collected, compiled and distributed the data; a summary (abstract) of the data content; details of data collection methods and processing; and so on

Section 3.0: Data File Description

This section describes each data file's content, record and variable counts, version, producer, and so on

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, and so on.

Section 5.0: Other Material

This section allows for a description of other material related to the study. This can include resources such as documents (questionnaires, coding information, technical and analytical reports, interviewers' manuals, and so on), data processing and analytical programs, photos and maps.

The Dublin Core Metadata Standard (DCMI)

The content of this section was taken from the DCMI website (http://dublincore.org)

The DCMI Metadata Element Set (ISO standard 15836), also known as the Dublin Core metadata standard, is a simple set of elements for describing digital resources. This standard is particularly useful to describe resources related to microdata such as questionnaires, reports, manuals, data processing scripts and programs, etc. It was initiated in 1995 by the Online Computer Library Center (OCLC) and the National Center for Supercomputing Applications (NCSA) at a workshop in Dublin, Ohio. Over the years it has become the most widely used standard for describing digital resources on the Web and was approved as an ISO standard in 2003. The standard is maintained and further developed by the Dublin Core Metadata Initiative - an international organization dedicated to the promotion of interoperable metadata standards.

A major reason behind the success of the Dublin Core metadata standard is its simplicity. From the outset it has been the goal of the designers to keep the element set as small and simple as possible to allow the standard to be used by non-specialists. The purpose of the standard is to make it easy and inexpensive to create simple descriptive records for information resources, while providing for effective retrieval of those resources on the Web or in any similar networked environment. In its simplest form the Dublin Core consists of 15 metadata elements, all of which are optional and repeatable. The 15 elements are:

- Title
- Subject
- Description
- Type
- Source
- Relation
- Coverage
- Creator
- Publisher
- Contributor
- Rights
- Date
- Format
- Identifier
- Language

Box 4 **Who uses the Dublin Core?**

The Dublin Core is a highly-flexible and simple metadata standard. It does not provide the same level of detail as the DDI (for example, the DC does not contain elements to document data files and variables), but it can be used to document the general characteristics of datasets. It provides a useful option to document datasets when resources available do not allow detailed documentation. An adapted version of the Dublin Core is used for example by the US open data initiative (www.Data.gov).

Box 5 **The XML Language**

XML stands for e**X**tensible **M**arkup **L**anguage. It is a way of tagging text for meaning instead of appearance. In other words, XML can be used to organise text by tagging with meaningful information. Although the 'tags' are conceptually the same as the 'fields' in a database in terms of organisation, the major difference between XML files and database files is the former are regular text files that can be viewed and edited using any standard text editor. The file can be searched and queried like a regular database, using appropriate tools. Just as the content of a database can be converted into a report, XML documents can be read and transformed by other software applications into user-friendly formats such as spreadsheets, PDF files or web pages.

The example below shows how textual information about a survey could be presented in XML:

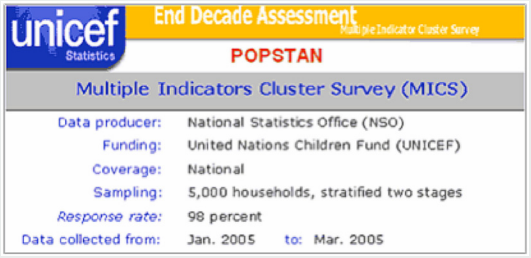
Assume the following information: *From January to March 2005, the National Statistics Office (NSO) of Popstan conducted the Multiple Indicators Cluster Survey (MICS) with the financial support of UNICEF. 5,000 households, representing the overall population of the country, were randomly selected to participate in the survey, following a two-stage stratified sampling methodology. 4,900 of these households provided information.*

The same information converted into XML using DDI tags would look like this:

<titl>**Multiple Indicator Cluster Survey 2005**</titl>
<altTitl>**MICS**</altTitl>
<AuthEnty>**National Statistics Office (NSO)**</AuthEnty>
<fundAg abbr="**UNICEF**">**United Nations Children Fund**</fundAg>
<collDate date="**2005-01**" event="start"/>
<collDate date="**2005-03**" event="end"/>
<nation>**Popstan**</nation>
<geogCover>**National**</geogCover>
<sampProc>**5,000 households, stratified two stages**</sampProc>
<respRate>**98 percent**</respRate>

Use of tags is particularly powerful when a community of users agrees on a common set of tags (such as the DDI or Dublin Core standards). Adoption of a common set of XML tags offers major advantages in documenting microdata including:

- Creation of a comprehensive checklist of useful metadata elements;
- Potential to assess the content of a file by determining whether particular tags are or are not within that file;
- Creation of a dataset catalogue that can be queried for key metadata elements;
- Potential to transform the file into more user-friendly formats. XML files can be converted into HTML, PDF or other types of documents using XSL Transformations. They can also be exchanged across networks or the Internet using web services or SOAP (an XML-based protocol allowing applications to exchange information over HTTP). An example of the application of XSL Transformation to the XML file above is the following HTML web page:



3. Why should data producers disseminate microdata?

Dissemination is one of the key responsibilities of a statistical agency. This chapter summarises the multiple benefits of microdata dissemination. A review of data producers' mission statements, dissemination policies and experience around the world underlines the importance of and reasons for giving access to microdata files.

3.1 Supporting research

The main reason – and often the only explicit reason – why data producers share their microdata is to support research work. After conducting a survey, the agencies that collect data normally produce a wide range of tabular output to give users the highlights and a broad overview of the results. They can hardly be expected – neither are they funded – to identify all the research questions that could be addressed using these data. Microdata files offer researchers considerable flexibility to identify relationships and interaction between phenomena covered in a survey, thereby fostering the diversity and quality of research work.[5]

Below are examples of how some national statistical agencies describe the goal of their data dissemination policies.

> "Access to microdata assists and encourages informed decision making through enabling wider use of ABS [Australian Bureau of Statistics] data for social and economic research and analysis. The ABS has been making microdata available, under certain conditions, for statistical purposes in the form of Confidentialised Unit Record Files (CURFs) since 1985. Today there is high user demand for the ABS to provide access to more detailed unit record data in a more flexible way, across a wider array of datasets (such as business data and longitudinal linked datasets). An inability to meet these demands will increasingly become a disadvantage to ABS core business, the relevance of the ABS and ultimately to the coherence of the NSS. This, along with a number of other drivers for change including the growing risk of identification, has led the ABS to propose a new strategy for accessing ABS microdata into the future." (Australian Bureau of Statistics - http://www.abs.gov.au/)

> "The primary objective of the (Central Statistics Office) CSO in providing access to microdata is to support the research community and to ensure that maximum usage is made of the data collected by the CSO. This approach supports the move towards evidence-based policy-making, has the potential to reduce the cost of research and also helps to avoid duplicate data collections." (Central Statistics Office, Ireland - http://www.cso.ie/)

> "The mission of the National Microdata Archive of Sri Lanka is to satisfy the data needs of the National and International research community who are striving hard to find answers to the socio-economic problems across the world. Backed by the Department of Census and Statistics – the central statistical agency in Sri Lanka, *LankaDatta* disseminates relevant, reliable and up-to-date statistical information being produced by the Agencies in the National Statistical System of Sri Lanka while preserving strict confidentiality of the respondents." (Department of Census and Statistics, Sri Lanka http://statistics.sltidc.lk/)

> "The DLI [Data Liberation Initiative] represents a major application of Canada's information highway technology. It allows post secondary institutions, for the first time, to offer a full range of data services to students and faculty alike. There is also growing evidence that the Initiative is making important contributions to Canadian teaching and research. (…) The DLI has been at the forefront of the creation of a data culture in Canada." (Statistics Canada http://www.statcan.gc.ca/dli-ild/about-apropos-eng.htm; see also Watkins [31])

3.2 Enhancing the credibility of official statistics

Broader access to microdata demonstrates producers' confidence in the data by making possible their replication or correction by independent parties.

3.3 Improving the reliability and relevance of data

Other benefits can accrue through a closer relationship between data providers and knowledgeable users. And

5 A good example of extended use of microdata files by researchers is found in the work of Hamilton and Humphrey [6]. They demonstrate that hundreds of research projects were carried out using the National Population Health Survey in Canada after it was released as a public-use microdata file. See also [30].

is often by the use of data that insights to possible improvement can be identified, e.g. to survey design and microdata dissemination. Feedback to a national statistical office can be built into the microdata dissemination process. For example, the US Census Bureau has formalised the process of feedback from researchers. Thus can user-feedback result in survey improvement over time.

3.4 Reducing duplication in data collection

Making microdata files available to users often discourages them from collecting data they require on their own.. This reduces the burden on respondents and minimises the risk of inconsistent studies of the same topic.

3.5 Increasing return on investment

"Sharing and open access to publicly funded research data not only helps to maximize the research potential of new digital technologies and networks, but provides greater returns from the public investment in research. (...)

"Continuously growing quantities of data are collected by publicly funded researchers and research institutions. This rapidly expanding body of research data represents both a massive investment of public funds and a potential source of the knowledge needed to address the myriad challenges facing humanity.

"To promote improved scientific and social return on the public investments in research data, OECD member countries [for example] have established a variety of laws, policies and practices concerning access to research data at the national level. In this context, international guidelines would be an important contribution to fostering the global exchange and use of research data." [17]

3.6 Leveraging funding for statistics

The more data files are disseminated and used, the more valuable they will appear to funding bodies. This can be used to encourage sponsor agencies to finance data collection. Indeed, in some cases, evidence of use is a requirement of financial sponsors.

Better use of data means a better return for survey sponsors, who thus will be more inclined to support data collection activities. Increasingly, funding of surveys by international sponsors is subordinated to dissemination of the resulting datasets.

3.7 Reducing the cost of data dissemination

A final benefit can accrue to data collectors by way of improved efficiency, in that they can possibly reduce the number of pre-defined tables they produce and devote more time to providing high-level analytical results. Highlighting this – for example via the media and the education sector – can appeal to a broader audience and encourage support for the work of data collectors, be they national statistical offices or others. However, when researchers probe more deeply, such tables will not be sufficient. On a final note, efficiency should be considered in the context of the costs of producing and disseminating microdata files. These issues are reviewed in the next section.

3.8 Complying with a contractual or legal obligation

In some countries, public agencies have an obligation to disseminate some of their microdata. Data collection is often funded by the taxpayer and thus considered a public good. In other cases, data collection is funded by sponsoring organizations that impose that the resulting data be made accessible to researchers. See example in Box 6. This obligation to disseminate microdata does not conflict with the obligation to maintain confidentiality and privacy. The responsibility of deciding on content of microdata to be published and procedures to be implemented to generate public-use files lies with the organisation's chief statistician or with a data release committee established by the organisation.

3.9 Promoting development of new tools for using data

A new movement, often referred to as the 'open-data' movement, has gained much ground in recent years. At its heart is the concept that data collected using public funds or under the auspices of a public agency are a 'public good'. Several governments have adhered to it – see, for example the *Open Government Initiative* in the US (www.data.gov) or its UK equivalent (http://*data.gov.uk)*. Such a movement challenges the public to add value to existing data. By providing unrestricted access to microdata, such initiatives are promoting the development of new software applications, especially innovative applications of web 2 technology.

These innovative web applications that make use of open data are often referred to as *mash-ups*:-

"In web development, a *mash-up* is a web page or application that uses or combines data or functionality

from two or many more external sources to create a new service. It implies easy, fast integration, frequently using open APIs and data sources to produce enriching results that were not necessarily the original reason for producing the raw source data.

"To be able to permanently access the data of other services, mash-ups are generally client applications or hosted online. (...) Mash-ups can be considered to have an active role in the evolution of social software and Web 2.0." (Source: http://en.wikipedia.org/wiki/Mashup accessed 5 April 2010].)

To facilitate their development of mash-ups, data providers not only disseminate data but often also provide software applications called *application-programming interfaces (APIs)*; these enable developers to use data more easily.

"An *application programming interface (API)* is an interface implemented by a software program to enable interaction with other software, similar to the way a user interface facilitates interaction between humans and computers. APIs are implemented by applications, libraries and operating systems to determine the vocabulary and calling conventions the programmer should employ to use their services. It may include specifications for routines, data structures, object classes, and protocols used to communicate between the consumer and implementer of the API." (Source: http://en.wikipedia.org/wiki/API accessed on April 5, 2010; see Box 7).

Box 6 **Legal Obligation to Disseminate Microdata: Example from the US National Center for Health Statistics**

An example of legislation governing microdata dissemination is provided by the National Center for Health Statistics (NCHS), attached to the US Centers for Disease Control.

"As a federal statistical agency NCHS must demonstrate that it has done all it can feasibly do to maximize data availability, including minimizing the time from data collection to dissemination, to maximize quality of data, and to minimize the risk of disclosure. NCHS' authorizing legislation mandates that data be made as widely available as practicable (...).

"However, the mandate to make data available must be guided by NCHS' role as a federal statistical agency and be balanced against the need to protect respondent confidentiality and to assure data quality. (...)

"The same law that requires that NCHS disseminate data also requires that NCHS safeguard the identity of individuals or establishments included in its data systems." [14]

Box 7 **Promoting Development of Mash-ups by Disseminating Open Data and APIs**

United States: data.gov – "Discover, Participate, Engage"

An underlying goal of the Open Government Initiative is to change the culture of information dissemination, institutionalizing a preference for making Federal data more widely available in more accessible formats. As one of the flagships of the Open Government Initiative, Data.gov is designed to facilitate access to Federal datasets that increase public understanding of Federal agencies and their operations, advance the missions of Federal agencies, create economic opportunity, and increase transparency, accountability, and responsiveness across the Federal Government – i.e., "high value" datasets.

Developers Corner: "Come on, brag a little!"

Since the launch of Data.gov we have been amazed by the number of dataset downloads and innovative applications popping up that use government data. One of the major reasons for creating Data.gov was to empower the community to innovate. And our developer community has been remarkable in this regard. We encourage developers and programmers to explore the datasets listed on Data.gov and to jump in – by actively participating in this vibrant and growing community.

Source: http://www.data.gov/open accessed on April 3, 2010.

United Kingdom: data.gov.uk – "Unlocking Innovation" – *"Show us a better way ..."*

We're very aware that there are more people like you outside of government who have the skills and abilities to make wonderful things out of public data. These are our first steps in building a collaborative relationship with you.

Source: http://data.gov.uk/ accessed on April 3, 2010.

Open Data is about accessibility. With the launch of data.worldbank.org we step up the effort to open the World Bank's data catalogues to direct and easy access via the web. During 2010, we will roll out two waves of web functionality that will serve as a platform for this push. (...)

Phase 2 will focus on promoting the use of the World Bank's data through the web API. This will require improvements of the actual API but equally important a place for communicating around the API and a lower threshold of entry especially for non-developers.

This phase will mark the departure of a developers-only web API to one that is geared towards researchers, decision makers and developers alike.

Source: http://data.worldbank.org/developers accessed on May 7, 2010.

4. What are the costs and risks, and how can they be addressed?

There are a number of issues for NSOs and other providers and collectors of data to consider as they formulate and implement microdata dissemination policies and programmes. These include the costs and expertise involved; questions of data quality, potential misuse and misunderstanding of data by users; legal and ethical matters; and maintaining the trust and support of respondents.

4.1 Ethical issues and maintaining respondents' trust

When collecting data from individuals, facilities or establishments, statistical agencies and other data producers usually give respondents assurances that the information they provide will be used only for statistical purposes. This is a moral or ethical and legal obligation.

To be successful "NSOs must maintain the trust of respondents if they are to continue to cooperate in their data collections. Confidentiality protection is the key element of that trust. If respondents believe or perceive that a NSO will not protect the confidentiality of their data, they are less likely to cooperate or provide accurate data. One incident, particularly if it receives strong media attention, could have a significant impact on respondent cooperation and therefore on the quality of official statistics. This is the dominant issue from the point of view of NSOs but there are other concerns. A key one is whether they have sufficient authority to support researcher access to microdata, either through a legal mandate or some other form of authorisation." [24]

In order to maximise microdata use, NSOs and other data providers need to balance carefully the need for confidentiality with provision for access. This may be achieved by using different types of microdata, as previously discussed. Another option, although of limited applicability, is to obtain formal consent from each respondent to share the collected data.

Obtaining consent from individuals

"Consent may be obtained by means of a signature or by construction. In the first instance, the respondent may be provided with a written description of the intended treatment of information he/she is asked to provide and asked to sign his/her name, thereby indicating permission to use that information as described. In some cases, however, he/she is given this information either in writing or verbally. If the respondent then supplies the requested data, [the investigator] 'construes' that the respondent agrees to those intended uses and sharing of data with parties he has read or been told about. The [investigator] can then make such uses of the data as have been described to the respondent, but no other uses of the data may be made." [15]

Obtaining consent from an establishment

"In the case of establishments, the approach depends partly upon whether the request for information is made in a personal interview or by mail.

A. If the request for information is made in person by a staff member or agent of [the investigator], the contact person first inquires as to who is authorized to provide the requested data on behalf of the establishment. When such authorized person is informed of the uses to be made of the data, and he/she then supplies the data, [the investigator agency's] staff construes that the establishment has given consent to the uses of data as specified.

B. When data are sought from an establishment by mail, the request may be addressed to the establishment itself, to the manager of the establishment, or to some other person who, as [the investigating agency] has previously ascertained, is authorized to provide requested data on behalf of the establishment. The letter transmitting the request explains the uses to be made of the data. When [the investigating agency] staff then receives the requested data from the establishment, it is construed that the establishment has consented to those uses of which it has been informed." [15]

4.2 Legal issues

Is it legal for a data producer to disseminate microdata files? There is no single answer. Legislation under

which a data producing agency works is specific to each country and programme framework (Box 8). As mentioned before, disseminating microdata is, in some cases, a legal obligation. But, in most cases, the legislation will formulate restrictions not obligations. Thus microdata dissemination policy for a country will be shaped by its legislative framework. It is crucial for data producers to "ensure there is a sound legal and ethical base (as well as the technical and methodological tools) for protecting confidentiality. This legal and ethical base requires a balanced assessment between the public good of confidentiality protection on the one hand, and the public benefits of research on the other. A decision on whether or not to provide access might depend on the merits of specific research proposals and the credibility of the researcher, and there should be some allowance for this in the legal arrangements." [24]

"Data access arrangements should respect the legal rights and legitimate interests of all stakeholders in the public research enterprise. Access to, and use of, certain research data will necessarily be limited by various types of legal requirements, which may include restrictions for reasons of:

- National security: data pertaining to intelligence, military activities, or political decision making may be classified and therefore subject to restricted access.

- Privacy and confidentiality: data on human subjects and other personal data are subject to restricted access under national laws and policies to protect confidentiality and privacy. However, anonymisation or confidentiality procedures that ensure a satisfactory level of confidentiality should be considered by custodians of such data to preserve as much data utility as possible for researchers.

- Trade secrets and intellectual property rights: data on, or from, businesses or other parties that contain confidential information may not be accessible for research. (...)" [15]

"Subscribing to professional codes of conduct may facilitate meeting legal requirements." [17]

United Nations Fundamental Principles of Official Statistics

Since many countries referred to the United Nations Fundamental Principle**s** of Official Statistics when setting up their legislation, it is useful to review these principles as they pertain to statistical confidentiality.

The sixth principle governing International Statistical Activities states: "Individual data collected by statistical agencies for statistical compilation, whether or not they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes." [27]

Any principles for microdata access must be consistent with this recommended principle or the principles contained in the NSOs' enabling legislation. The following points should be considered for managing the confidentiality of microdata.

Principle 1: Appropriate use of microdata

"It is appropriate for microdata collected for official statistical purposes to be used for statistical analysis to support research as long as confidentiality is protected. (...)

Making available microdata for research is not in contradiction with the sixth UN Fundamental Principle as long as it is not possible to identify data referring to an individual. Principle 1 does not constitute an obligation to provide microdata. The National Statistical Office should be the one to decide whether to provide microdata or not. There may be other concerns (for example, quality) that make it inappropriate to provide access to microdata. Or there may be specific persons or institutions to whom it would be inappropriate to provide microdata." [24]

Principle 2: Microdata should only be made available for statistical purposes

"For Principle 2, a distinction has to be made between statistical or analytical uses and administrative uses. In the case of statistical or analytical use, the aim is to derive statistics that refer to a group (be it of persons or legal entities). In the case of administrative use, the aim is to derive information about a particular person or legal entity to make a decision that may bring benefit or harm to the individual. For example, some requests for data may be legal (a court order) but inconsistent with this principle. It is in the interest of public confidence in the official statistical system that these requests are refused. If the use of the microdata is incompatible with statistical or analytical purposes, then microdata access should not be provided. Ethics committees or a similar arrangement may assist in situations where there is uncertainty whether to provide access or not.

Researchers are accessing microdata for research purposes but to support this research they may need to compile statistical aggregations of various forms, compile statistical distributions, fit statistical models, or analyze statistical differences between sub-populations. These uses would be consistent with statistical purposes. To the extent that this is how the microdata are being used, it could also be said to support research purposes." [24]

Principle 3: Provision of microdata should be consistent with legal and other necessary arrangements that ensure that confidentiality of the released microdata is protected

"With respect to Principle 3, legal arrangements to protect confidentiality should be in place before any microdata are released. However, the legal arrangements have to be complemented with administrative and technical measures to regulate the access to microdata and to ensure that individual data cannot be disclosed. The existence and visibility of such arrangements (whether in law or supplementary regulations, ordinances, etc) are necessary to increase public confidence that microdata will be used appropriately. Legal arrangements are clearly preferable but in some countries this may not be possible and some other form of administrative arrangements should be put in place. The legal (or other arrangements) should also be cleared with the privacy authorities of countries where they exist before they are established by law. If such authorities do not exist, there may be NGOs who have a "watchdog" role on privacy matters. It would be sensible to get their support for any legal or other arrangements, or at least to address any serious concerns they might have.

In some countries, authorizing legislation does not exist. At a minimum, release of microdata should be supported by some form of authority. However, an authorizing legislation is a preferable approach." [24]

Principle 4: The procedures for researcher access to microdata, as well as the uses and users of microdata should be transparent, and publicly available

"Principle 4 is important to increase public confidence that microdata are being used appropriately and to show that decisions about microdata release are taken on an objective basis. It is up to the NSO to decide whether, how and to whom microdata can be released. But their decisions should be transparent. The NSO web site is an effective way of ensuring compliance and also for providing information on how to access research reports based on released microdata." [24]

4.3 Exposure to criticism and contradiction

"Some NSOs are concerned that the quality of their microdata may not be good enough for further dissemination. Whilst quality may be sufficiently accurate to support aggregate statistics, this may not be the case for very detailed analysis. In some cases, adjustments are made to aggregate statistics at the output editing stage without amendment to the microdata. Consequently, there may be inconsistencies between research results based on microdata and published aggregate data." [24] If some parts of datasets are considered too unreliable, these may be removed before dissemination. Data-producers should be open and transparent about quality.

Another concern is that providing microdata to researchers opens up the possibility of their publishing results that could contradict data producer estimates. When the data producer is an official statistical agency, this may result in conflicting – official v non-official – estimates, and lead to questioning of the data, with possible political implications. There may be various reasons for differences. Firstly, there may be errors in official estimates, in which case outside scrutiny is of benefit. Secondly, differences may arise from use of different versions of the data (the full master file v an anonymised/reduced public version, further editing by researcher, etc). These differences should be marginal, and can easily be explained.

Thirdly, this may be the consequence of different methodologies used. This is often a more challenging issue for data producers, as the public will not always be able to understand highly technical explanations. It is important for data producers to be able to defend their own estimates. This means that the collection, processing and analysis of the data must be fully documented, and that this information be preserved for easy access. In some cases, published results may have been produced by or with the assistance of external experts who are no longer available to answer questions. Data producers can protect themselves against this risk by adopting and enforcing strict practices of documentation and preservation in compliance with the *replication standard*. Succinctly, the replication standard is defined as follows: "(...) the only way to understand and evaluate an empirical analysis fully is to know the exact process by which the

Box 8 **Examples of Legislation on Confidentiality**

Below are examples of statistical legislation and the way microdata dissemination is dealt with:

1. The US Bureau of the Census operates under Title 13-Census of the US Code. *"Title 13, U.S.C., Section 9 prohibits the publication or release of any information that would permit identification of any particular establishment, individual, or household. Disclosure assurance involves the steps taken to ensure that Title 13 data prepared for public release will not result in wrongful disclosure. This includes both the use of disclosure limitation methods and the review process to ensure that the disclosure limitation techniques used provide adequate protection to the information."*

   Source: http://www.census.gov/srd/sdc/wendy.drb.faq.pdf

2. In Canada the Statistics Act states that *"no person who has been sworn under section 6 shall disclose or knowingly cause to be disclosed, by any means, any information obtained under this Act in such a manner that it is possible from the disclosure to relate the particulars obtained from any individual return to any identifiable individual person, business or organization."*

   Source: http://www.statcan.ca/english/about/statact.htm

   Statistics Canada does release microdata files. Its microdata release policy states that Statistics Canada will authorise the release of microdata files for public use when:

   (a) the release substantially enhances the analytical value of the data collected; and
   (b) the Agency is satisfied all reasonable steps have been taken to prevent the identification of particular survey units.

3. In Thailand the Act states: *"Personal information obtained under this act shall be strictly considered confidential. A person who performs his or her duty hereunder or a person who has the duty of maintaining such information cannot disclose it to anyone who doesn't have a duty hereunder except in the case that:*

   (1) Such disclosure is for the purpose of any investigation or legal proceedings in a case relating to an offense hereunder.
   (2) Such disclosure is for the use of agencies in the preparation, analysis or research of statistics provided that such disclosure does not cause damage to the information owner and does not identify or disclose the data owner."

   Source: http://web.nso.go.th/eng/en/about/about0.htm section 15.

data were generated and the analysis produced. (...) The replication standard holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party could replicate the results without any additional information from the author". [10] See also [11].

4.4 Cost

"NSOs may also be concerned about costs. These include not only the costs of creating and documenting microdata files, but the costs of creating access tools and safeguards, and of supporting and authorising enquiries made by the research community; new users of data files need help to navigate complex file structures and variable definitions. Although the costs are borne by the NSOs, they are usually not provided with budget supplementation to do the additional work. And on the whole, researchers do not have the funding to contribute substantially to these costs." [24] Thus, whenever possible, such costs should be built into the survey budget as a means of ensuring that maximum use can be made of the survey results. It is in the public interest that insights from the data be made available to inform decision-makers and the public. Furthermore, if survey data are used more extensively in this way, they provide an extra level of protection against budget reductions to statistical programmes. Surveys that offer limited knowledge in support of policy-making are more vulnerable to elimination.

4.5 Loss of exclusivity

When disseminating microdata, data owners lose their exclusive right to discovery. This is more of an issue for academic researchers than official producers, although official data producers (or some of their staff members) sometimes take advantage of a monopolistic access to data to offer consulting services. Increasingly, survey sponsors define a legitimate and "reasonable" period of exclusive access to the data by the producer, after which data have to be made accessible to other users.

4.6 Technical capacity

A certain technical capacity is required to support dissemination of microdata files. The files need to be well-documented (preferably using the DDI metadata standard) and preserved. In addition, the files must be reviewed to identify the risk of disclosure of individual information, and the risk reduced using various techniques. The technical requirements for disseminating microdata files are explored in more detail in Chapter 10.

5. To whom should microdata be made available?

Microdata files are intended for specialised users with advanced quantitative skills. This includes typically

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

A key principle of microdata dissemination is *equitableness*. Microdata from publicly-funded data collection, when they can be disseminated legally, should be made available to all potential users openly and with comprehensive metadata. "Openness means access on equal terms for the international research community at the lowest possible cost, preferably at no more than the marginal cost of dissemination. Open access to research data from public funding should be easy, timely, user-friendly and preferably Internet-based." [17]

NSOs generally provide different products aimed at different audiences. High-level summaries (tables, graphs, analyses) are intended generally for a wide audience and made available via publications and agency websites. Microdata files generally are aimed at researchers from various institutions, including government agencies and ministries, non-governmental organisations, research institutes, academia and international bodies. These are known generically as the 'research community'.

"What is the research community? It includes those working in academic institutions, of course. It also includes researchers working in non-government organizations and international agencies. Furthermore, some researchers requiring access to microdata will work within government-funded agencies and institutions." [24]

Many statistics acts refer to making data available to users for legitimate statistical and research purposes. Legal and commercial use generally is not regarded as falling into this category. Stressing the potential use of microdata is a key aspect of adhering to the acts under which NSOs operate.

Eurostat, the statistical organisation of the European Communities, defines its access to microdata as follows:

> "Microdata are mainly accessible to the scientific community, institutes or universities for research purposes only, and on the submission of a research project. In some cases it is extended to students and to a larger non-scientific public. The highly protected microdata sets available for the general public are public use files (PUF). Although they are widely and freely made available by statistical institutes, their value for policy relevant research is limited. Generally, private enterprises are not allowed to work with microdata."[6]

While microdata dissemination generally is aimed at the research community, other qualifications are often attached. Reference is often made to providing such files to '*bona fide*' users, meaning genuine researchers acting in good faith.

'Ownership' of data files is not actually transferred to researchers: they are licensed to them for their own use. This gives the NSO an opportunity to discover the intended purpose. It is good practice for an NSO to spell out its access conditions in the form of a policy.

The policy should be generic and inform all types of user. Specific protocols may distinguish users based on their nationality or agency. Some considerations are as follows:

National users:

- Officials specifically covered by statistical legislation could have easier access to official microdata as they may operate under the same rules as the NSO. They can be sworn in, and enforcing penalties is easier – this serves the government's objectives.

- Other national users covered by statistical legislation still might be able to use data files, providing they sign a suitable licensing agreement.

6 Eurostat provides access to microdata via their website. The link below is part of privacy statement for the Eurostat Internet website; see http://epp.eurostat.ec.europa.eu/portal/page?_pageid=1493,58764075&_dad=portal&_schema=PORTAL accessed on September 23, 2008.

International users:

- There is often an obligation to share data with international organisations. This might result from membership of an international group or from agreements that provide funding for major projects, or stem from a country's participation in international development projects.

- Researchers in foreign universities or research centres are a trickier licensing challenge. It is more difficult to enforce licences in their case. However, there could be an incentive to share data with them – they often form a rich pool of expertise. The risk can be minimised, e.g. by licensing datasets to the university rather than the individual. This is discussed further below.

Technical assistants:

Domestic or international consultants are often involved in helping NSO staff to complete survey data processing and/or analysis. As such usage is compatible with NSO objectives, it should fall within the scope of this policy – provided that consultants sign the same undertaking required of other researchers. This would require them not to release data without prior authorisation from the data producer. An affidavit of confidentiality should be requested to ensure the data file is 'safe'.

It might be possible for NSOs to transfer some of the risk to researchers, as follows:

(i) "asking them to prove their bona fides as researchers and to demonstrate the public benefits of their research and that the microdata are necessary for this research;

(ii) making them sign a legally binding undertaking with similar penalties to those operating for NSO staff if they breach confidentiality provisions;

(iii) explaining the reasons NSOs are cautious. Ensuring researchers are fully aware of their obligations through appropriate education. Follow up with effective audit and monitoring procedures. It may be useful to establish a Code of Conduct in collaboration with the research community;

(iv) where offences occur, withdrawing all current and future services from the researcher and possibly their institution for a period of time (e.g. until the institution has undertaken appropriate disciplinary action against the offender). Make them realise that the future release of microdata to any researcher may be at risk if there is strong public criticism. Undertaking legal action where appropriate." [24]

Where a request for data access is made by a researcher on behalf of an organisation, it is advisable to oblige him or her to apply behalf of that organisation – or agency, employer, etc – rather than as an individual. Such bodies need to maintain their reputation and are in a better position to ensure undertakings are respected. These agreements are difficult to enforce in the case of cross-border data requests. One way to handle these is to cooperate with data archives or NSOs in the applicant's country.

Box 9 **Affidavit of Confidentiality – An Example**

I agree:

1. To make no copies of any files or portions of files to which I am granted access except those authorised by National Data Enclave (NDE) staff. No confidential data or information viewed or otherwise obtained while I am a researcher in NDE will be removed from NDE.
2. To return to NDE staff all NDE restricted material with which I may be provided during the conduct of my research at NDE, and other material as requested.
3. Not to use any technique in an attempt to learn the identity of any person, establishment or sampling unit not identified in public-use data files.
4. To keep in strictest confidence identification of any establishment or individual that may be inadvertently revealed in any documents, discussion or analysis. Such inadvertent identification revealed in my analysis will be immediately brought to the attention of NDE staff.
5. Not to remove any print-outs, electronic files, documents or media until they have been scanned for disclosure risk by NDE staff. I understand that NDE will perform a disclosure review and give me approval before I remove any data from NDE, whether they be in electronic or paper form.
6. Not to remove from NDE any written notes pertaining to the identification of any establishment, individual or geographical area that may be revealed in the conduct of my research at NDE.
7. To comport myself in a manner consistent with the principles and standards appropriate to a scientific research establishment.

I understand that deliberate violation of any of these conditions may result in cancellation of the data access agreement. I further agree that in such event I may be barred from any future use of NDE following a review and determination by the Director of NDE that finds such action is necessary to protect the integrity and confidentiality of NDE.

_____ _____
Consultant's name and signature Date

_____ _____
Witness's name and signature Date

This affidavit of confidentiality is an adapted version of the 2008 version of the *Agreement Regarding Conditions of Access to Confidential Data in the Research Data Center of the US National Center for Health Statistics* [16]

6. Under what conditions should microdata be provided?

Most publicly-funded data producing agencies have a mandate to make the data they collect as broadly available as practicable. However, all these agencies operate under legal and ethical obligations that place restrictions on the manner in which the data can be disseminated. "Making useful (...) statistics available to as many people as possible is very important, but it is equally important to do so in a manner that will not in any way harm the providers of these statistics." [15] "A combination of legal, administrative and technical measures (are thus) necessary to ensure public confidence in the arrangements." [24]

The conditions under which microdata are disseminated must be formal and transparent. They should be defined in a policy and in accompanying procedures or protocols. This chapter provides information for the formulation of such policies and procedures. It is important, however, to note that this information does not constitute a recommended standard. Each data producer has to define its policy and procedures based on technical, legal and ethical considerations.

Typically, a dissemination policy will be a relatively general statement offering:

- The objective of the policy, which is typically "to define the nature of microdata files to be released, their intended use and the conditions of their release".
- A statement describing briefly why the agency finds it important to disseminate microdata. Chapter 3 above provides various possible reasons.
- The legal basis under which microdata are disseminated (see 6.1 below).
- The overall principles applied to preserve privacy and confidentiality of the data.
- A statement related to the timing of data releases. This could, for example, state that "recognising the importance of user needs and timely data, the NSO will strive to release microdata files within six to 12 months from first release of survey data. Surveys may need to be released in phases to ensure their data have been reviewed and analysed by NSO staff prior to their being anonymised for use outside the agency."
- The key roles and responsibilities for the definition and implementation of the policy and accompanying procedures. These could for example include:

Survey managers, responsible for:

- Identifying the needs of key stakeholders and ensuring the creation of an anonymised file that meets the needs of the user community to the extent possible under the Statistics' Act; and
- Preparing an initial screening of the microdata file, identifying any potential problems to be resolved and drafting a submission to the Microdata Release Committee.

A Microdata Release Committee responsible for:

- Reviewing all requests by survey managers for the release of anonymised microdata files, using established criteria;
- Approving all files for release or providing guidance to the survey manger on how to improve the file beforehand;
- Overseeing the licensing process and resolving issues of possible breaches; and
- Revising as necessary the guidelines used by survey managers to create anonymised microdata files.

A Dissemination Group responsible for:

- Reviewing requests by researchers for access to licensed microdata files;
- Providing access to data files to users so approved; and
- Responding to users' requests for support and additional information.

The Agency's Director-General to approve all release to users of anonymised microdata files, based on the advice and recommendation of the Microdata Release Committee.

- An overall statement on the pricing policy. Ideally, data producers should have a policy that encourages wide use of their products by making them affordable. Accordingly, they should attempt to ensure that costs of creating anonymised microdata files are built into the survey budget. At the same time, the NSO may legitimately attempt to recover the cost of providing special services that benefit only a specific group.

- An overall statement on the pricing policy.

More detailed information on the procedures and conditions attached to the release of microdata files are specified in protocols or procedures that will define:

- How users can request access to the data (on-line requests, request forms to be used, etc.)

- The permissions and restrictions attached to the various types of datasets (see 6.2 below)

- Who is responsible for decisions to grant access, and other practical information on the review process

- What type of statistical disclosure control methods are in place

- What information is required from the researchers, and what can be done with this information

- The detailed pricing policy

- How much and what type of technical support is available to users

- And other practical information.

The UK Statistics Authority has published a Code of Practice for Official Statistics [22] that, without being specific to the release of microdata files, provides a good example of clearly-formulated principles and protocols.

6.1 Enabling legislation

An enabling legislation is essential for several reasons:

"(i) to provide public confidence in the arrangements – that there are legal constraints that determine what can and cannot be done;

(ii) to provide mutual understanding between NSOs and researchers on the arrangements;

(iii) to provide for greater consistency in the way research proposals are treated; and

(iv) to provide a basis for dealing with breaches." [17]

Some NSOs will have no provision in their statistics' acts for release of microdata files. In some cases this may be explicitly forbidden; in others the act may not deal explicitly with the matter, so could be subject to interpretation. This can arise if legislation dates back to a time when such a provision was not considered feasible. In such cases, legislation may need to be revised before microdata files can be distributed. For example, Canada's Statistics' Act has existed since the early 1900s but revised in 1971 – among other things, to permit the NSO to distribute microdata files.

But "the legislation need not exist in primary legislation or law. The detail may be better suited to regulations, ordinances, etc. that still have some legal impact. If legislation is not available, some other form of authorisation is essential. The reputation of the NSO is at risk if there is not some form of authority to enable the release of microdata even when anonymised. It is important that the legislation (or authorisation) covers the following aspects:

(i) what can and cannot be done and for what purposes;

(ii) the conditions of release; and

(iii) the consequences if these conditions are breached." [17]

6.2 Conditions for Public-Use Files (PUFs)

Generally, data regarded as **public** are open to anyone with access to an NSO website. It is, however, normally good practice to include statements defining suitable uses for and precautions to be adopted in using the data. While these may not be legally binding, they serve to sensitise the user. Prohibitions such as attempts to link the data to other sources can be part of the 'use statement' to which the user must agree, on-line, before the data can be downloaded.

An example of a PUF can be found at Statistics Canada. This refers to the microdata file produced from the Joint Canada-United States Survey of Health.[7] The European Social Survey provides another example.[8] Microdata files can be accessed simply by registering as a user. This is done to enable users to be notified about changes to files and availability of new ones.

Dissemination of microdata files necessarily involves the application of rules or principles. Box 10 below shows basic principles normally applying to PUFs.

6.3 Conditions applying to licensed files

For **licensed microdata files,** terms and conditions must include the basic common principles plus some additional ones applying to the researcher's organisation. There are two options: firstly, data are provided to a researcher or a team for a specific purpose; secondly, data are provided to an organisation under a blanket agreement for internal use, e.g. to an international body or research agency. In both cases, the researcher's organisation must be identified, as must suitable representatives to sign the licence.

Access to a researcher or research team for a specific purpose

If data are provided for an individual research project, the research team must be identified. This is covered by requiring interested users to complete a formal request to access the data (a model of such a request for is provided in Appendix 1). The conditions to obtain the data (see example in Box 12) will specify that the files will not be shared outside the organisation and that data will be stored securely. To the possible extent, the intended use of the data – including a list of expected outputs and the organisation's dissemination policy – must be identified. Access to licensed datasets is only granted when there is a legally-registered sponsoring agency, e.g. government ministry, university, research centre or national or international organisation.

Blanket agreement to an organization

In the case of a blanket agreement, where it is agreed the data can be used widely but securely within the receiving organisation, the licence should ensure compliance, with a named individual formally assuming responsibility for this. Each additional user must be made aware of the terms and conditions that apply to data files: this can be achieved by having to sign an affidavit. Where such an agreement exists, with security in place, it is not necessary for users to destroy the data after use. Box 13 provides an example of the formulation of such an agreement.

6.4 Conditions specific to data enclaves

Data enclaves are used for particularly sensitive data or for more detailed data for which sufficient anonymisation to release them outside the NSO premises is not possible. These can be referred to also

Box 10 **Conditions for Accessing and Using PUFs**

1. Data and other material provided by the NSO will not be redistributed or sold to other individuals, institutions or organisations without the NSO's written agreement.
2. Data will be used for statistical and scientific research purposes only. They will be employed solely for reporting aggregated information, including modelling, and not for investigating specific individuals or organisations.
3. No attempt will be made to re-identify respondents, and there will be no use of the identity of any person or establishment discovered inadvertently. Any such discovery will be reported immediately to the NSO.
4. No attempt will be made to produce links between datasets provided by the NSO or between NSO data and other datasets that could identify individuals or organisations.
5. Any books, articles, conference papers, theses, dissertations, reports or other publications employing data obtained from the NSO will cite the source, in line with the citation requirement provided with the dataset.
6. An electronic copy of all publications based on the requested data will be sent to the NSO.
7. The original collector of the data, the NSO, and the relevant funding agencies bear no responsibility for the data's use or interpretation or inferences based upon it.

Notes:

- *Items 3 and 6 in the list require that users be provided with an easy way to communicate with the data provider. It is good practice to provide a contact number, an email address, and possibly an on-line "feedback provision" system.*
- *For item 5, see Box 11.*

7 Statistics Canada; see http://www.statcan.ca/english/freepub/82M0022XIE/2003001/pumf.htm accessed on August 22, 2007.

8 European Social Survey; see http://www.europeansocialsurvey.org/index.php?option=com_content&task=view&id=78&Itemid=190 accessed on August 22, 2007.

as data laboratories or research data centres. A data enclave may be located at the NSO headquarters or in major centres such as universities close to the research community. They are used to give researchers access to complete data files but without the risk of releasing confidential data. In a typical data enclave, NSO staff supervise access and use of the data; the computers must not be able to communicate outside the enclave; and the results obtained by the researchers must be screened for confidentiality by an NSO analyst before taken outside. A model of a data enclave access policy is provided in Appendix 2, and a model of a data enclave access request form is in Appendix 3.

Data enclaves have the advantage of providing access to detailed microdata but the disadvantage of requiring researcher to work at a different location. And they are expensive to set up and operate. It is, however, quite likely that many countries have used **on-site researchers** as a way of providing access to microdata. These researchers are sworn in under the statistics' acts in the same way as regular NSO employees. This approach tends to favour researchers who live near NSO headquarters.

6.5 Managing breaches by researchers

Experience in countries with established microdata dissemination practices shows that breaches of data file confidentiality are very limited indeed. They are not in the researchers' interests. Their reputation would be at risk as would that of their organisation. Nevertheless, whether intentional or accidental, they may occur.

Box 11 **Citing Electronic Data Files**

No universal standards exist for citing microdata sets. The minimum information should contain the identification of the data producer, the dataset title and reference year (followed by an indication that the reference is to a microdata set), the dataset reference number (which should identify the version number), the identification of the distributor of the data, and the data when the data files were obtained (see also [18] for more information on citation of statistical products).

Example 1

U.S. Dept. of Commerce, Bureau of the Census. AMERICAN COMMUNITY SURVEY (ACS): PUBLIC USE MICRODATA SAMPLE (PUMS), 2005 [Computer file]. ICPSR04587-v1. Washington, DC: U.S. Dept. of Commerce, Bureau of the Census [producer], 2005. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2007-08-08.

Example 2

Survey of labour and income dynamics public use microdata file wave 8, 2000: economic families [computer file]./ Canada, Statistics Canada, Household Surveys Division, Version 2, Ottawa, Ont.: Statistics Canada [producer]; Statistics Canada. Data Liberation Initiative [distributor], 2003/08/28.

Such citations, although acceptable, will not solve all issues and may not be fully satisfactory for academic data centers. "Sometimes URLs are given, but they often do not persist. (…) Modified versions of data are routinely distributed under the same name, without any standard for versioning. Copyeditors have no fixed rules, and often no rules whatsoever. Data are sometimes listed in the bibliography, sometimes in the text, sometimes not at all, and rarely with enough information to guarantee future access to the identical data set." [1]

To solve these (and other) issues, M. Altman and G. King from the Harvard-MIT Data Center propose that "citations to numerical data include, at a minimum, six required components. The first three components are traditional, directly paralleling print documents. They include the author(s) of the data set, the date the data set was published or otherwise made public, and the data set title. These are meant to be formatted in the style of the article or book in which the citation appears. The author, date, and title are useful for quickly understanding the nature of the data being cited, and when searching for the data. However, these attributes alone do not unambiguously identify a particular data set, nor can they be used for reliable location, retrieval, or verification of the study. Thus, we add three components using modern technology, each of which is designed to persist even when the technology inevitably changes. They are also designed to take advantage of the digital form of quantitative data.

The fourth component is a unique global identifier, which is a short name or character string guaranteed to be unique among all such names, that permanently identifies the data set independent of its location. (…) Unique global identifiers thus guarantee persistence of the link from the citation to the object, but we also need to guarantee and independently verify that the object does not change in any meaningful way even when data storage formats change. Thus, we add as the next component a Universal Numeric Fingerprint or UNF. The UNF is a short, fixed-length string of numbers and characters that summarize all the content in the data set, such that a change in any part of the data would produce a completely different UNF. (…) (They) add as a final component of the citation standard a bridge service. (…) An example of a complete citation, using this minimal version of the proposed standards, is as follows:

Sidney Verba. 1998. "U.S. and Russian Social and Political Participation Data." hdl:1902.4/00754 ; UNF:4:ZNQRI14053UZq389x0Bffg?== ; http://id.thedata.org/hdl%3A1902.4%2F00754." [1]

Box 12 **Conditions for Accessing and Using Licensed Data Files**

Note: Items 1 to 8 below are similar to the conditions for use of public use files. Items 9 and 10 would have to be adapted in the case of a blanket agreement.

1. Data and other material provided by the NSO will not be redistributed or sold to other individuals, institutions or organisations without the NSO's written agreement.

2. Data will be used for statistical and scientific research purposes only. They will be employed solely for reporting aggregated information, including modelling, and not for investigating specific individuals or organisations.

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

4. No attempt will be made to produce links between datasets provided by the NSO or between NSO data and other datasets that could identify individuals or organisations.

5. Any books, articles, conference papers, theses, dissertations, reports or other publications employing data obtained from the NSO will cite the source, in line with the citation requirement provided with the dataset.

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

7. The NSO and the relevant funding agencies bear no responsibility the data's use or for interpretation or inferences based upon it.

8. An electronic copy of all publications based on the requested data will be sent to the NSO.

9. The researcher's organisation must be identified, as must the principal and other researchers involved in using the data must be identified. The principal researcher must sign the licence on behalf of the organization. If the principal researcher is not authorized to sign on behalf of the receiving organization, a suitable representative must be identified.

10. The intended use of the data, including a list of expected outputs and the organisation's dissemination policy must be identified.

(Conditions 9 to 11 may be waved in the case of educational institutions)

In view of the importance of maintaining a sustainable policy of microdata dissemination, NSOs need to consider enforcement procedures, such as:

- If a breach is identified, it should be dealt with forthwith. This is an important part of maintaining confidence in the agency and respondent trust.

- If a legal offence has occurred, legal action should be considered.

- If researchers violate their undertaking, the NSO should consider suspending their access rights and those of their organisation.

- If the undertaking is made by an organisation on a researcher's behalf, the organisation, rather than the NSO, may wish to consider the sanctions it should take towards one of its own. Loss of access by an individual could lead to a similar penalty for the whole organisation.

- If necessary, the NSO should take steps to ensure further breaches do not occur; and,

- if the breach is minor, a warning should be considered as the only action necessary.

Below is an additional note on managing breaches-of use-agreements.

> "It is good practice for such an undertaking to have some legal standing, for example by having provision for such undertakings within enabling legislation. This would allow legal actions to be taken in respect of breaches of the conditions of the undertaking. This does not preclude other actions that might be taken in respect of breaches such as not providing any further services to the researcher and/or possibly the researcher's institution." [24]

If the NSO desires feedback from users, it should consider regular follow-up of researchers. This is also an opportunity to remind them of their requirement to provide feedback on their outputs, and to solicit their views on how improve the survey programme.

Box 13 **Blanket Agreement**

Agreement between [*providing agency*] and [*receiving agency*] regarding the deposit and use of microdata

A. **This agreement relates to the following microdatasets:**

1. _____
2. _____
3. _____
4. _____
5. _____

B. **Terms of the agreement:**

As the owner of the copyright in the materials listed in section A, or as duly authorized by the owner of the copyright in the materials, the representative of [providing agency] grants the [receiving agency] permission for the datasets listed in section A to be used by [receiving agency] employees, subject to the following conditions:

1. Microdata (including subsets of the datasets) and copyrighted materials provided by the [providing agency] will not be redistributed or sold to other individuals, institutions or organisations without the [providing agency]'s written agreement. Non-copyrighted materials which do not contain microdata (such as survey questionnaires, manuals, codebooks, or data dictionaries) may be distributed without further authorization. The ownership of all materials provided by the [providing agency] remains with the [providing agency].
2. Data will be used for statistical and scientific research purposes only. They will be employed solely for reporting aggregated information, including modeling, and not for investigating specific individuals or organisations.
3. No attempt will be made to re-identify respondents, and there will be no use of the identity of any person or establishment discovered inadvertently. Any such discovery will be reported immediately to the [providing agency].
4. No attempt will be made to produce links between datasets provided by the [providing agency] or between [providing agency] data and other datasets that could identify individuals or organisations.
5. Any books, articles, conference papers, theses, dissertations, reports or other publications employing data obtained from the [providing agency] will cite the source, in line with the citation requirement provided with the dataset.
6. An electronic copy of all publications based on the requested data will be sent to the [providing agency].
7. The [providing agency] and the relevant funding agencies bear no responsibility the data's use or for interpretation or inferences based upon it.
8. An electronic copy of all publications based on the requested data will be sent to the [providing agency].
9. Data will be stored in a secure environment, with adequate access restrictions. The [providing agency] may at any time request information on the storage and dissemination facilities in place.
10. The [recipient agency] will provide an annual report on uses and users of the listed microdatasets to the [providing agency], with information on the number of researchers having accessed each dataset, and on the output of this research.
11. This access is granted for a period of [provide information on this period, or state that the agreement is open ended].

C. **Communications:**

The [receiving organisation] will appoint a contact person who will act as unique focal person for this agreement. Should the focal person be replaced, the [recipient agency] will immediately communicate the name and coordinates of the new contact person to the [providing agency]. Communications for administrative and procedural purposes may be made by email, fax or letter as follows:

Communications made by [providing agency] to [recipient agency] will be directed to:

Name of contact person: _____
Title of contact person: _____
Address of the recipient agency: _____
_____
Email: _____
Tel: _____
Fax: _____

Communications made by [recipient agency] to [depositor agency] will be directed to:

Name of contact person: _____
Title of contact person: _____
Address of the recipient agency: _____
_____
Email: _____
Tel: _____
Fax: _____

D. **Signatories**

The following signatories have read and agree with the Agreement as presented above:

Representative of the [providing agency]

Name _____

Signature _____ Date _____

Representative of the [recipient agency]

Name _____

Signature _____ Date _____

7. What is meant by microdata anonymisation?

This chapter draws extensively on the CENEX Handbook on Statistical Disclosure Control [3]. This can be freely downloaded from http://neon.vb.cbs.nl/cenex/CENEX-SDC_Handbook.pdf.[9]

When disseminating microdata files to the public, researchers or other agencies, the NSO faces a conflicting mission. On one hand, it aims to release microdata files supporting a wide range of statistical analyses; on the other, it must safeguard the confidentiality of respondents' identities. Processes aimed at the latter are referred to collectively as Statistical Disclosure Control (SDC) or anonymisation.

"Since confidentiality can never be absolutely assured, the risks and benefits of providing access must be weighed against the disclosure risks and sensitivity of the data." [14] Data protection is a fully legitimate concern for statistical agencies and other data producers, and there is an abundant literature on the topic. But literature on confidentiality management also contains numerous references to what Peter Madsen [13] has termed 'the Privacy Paradox'. At a 2003 workshop on Confidentiality Research, sponsored by the National Science Foundation in the USA, he argued that "the rush to ensure complete levels of privacy in the research context paradoxically results in less social benefit rather than more".[10]

7.1 Statistical Disclosure Control (SDC) concepts

A disclosure occurs when a person or organisation recognises or learns via released data something they did not know about another person or organisation. There are two types of disclosure risk: ***identity disclosure*** and ***attribute disclosure***.[11] The former occurs when a respondent's identity is directly associated with a disseminated data record. This can occur easily when the data record includes variables unambiguously identifying the respondent – for instance, the respondent's name, passport/identification number or telephone number. It is essential that such identifying variables be removed from any microdata files before dissemination. Attribute disclosure occurs when attribute values (or estimates thereof) in the disseminated data are associated with a particular respondent.

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 when (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 publically-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 because either they do not exist or are not generally disseminated. In practice, the risk of disclosure from disseminating household survey microdata files will be minimised sufficiently simply by stripping records of direct identifiers and removing geographical identifiers below the stratum level. However, disclosure risk should be assessed for each microdata file as there will be important exceptions to the aforementioned rule of thumb that warrant additional control measures. Survey microdata files containing records from small-target populations – for instance enterprise and business records – are considerably more difficult to anonymise.

A final important consideration for NSOs is the need to safeguard details of their sampling frames –particularly inter-census master sampling frames used to design several different household surveys. Dissemination of detailed sampling frames undermines attempts to safeguard respondents' identities as they provide an obvious data file that could be used for matching – for instance, geographical details (such as the name or location of primary sample units) could be identified by matching sampling weights (which much be released in microdata files, otherwise users cannot draw correct statistical inferences). A particular reason for safeguarding master sampling frames is to limit the likelihood of respondents in these PSUs being selected to participate in surveys conducted independently of the NSO. In countries where the master sample frame PSUs are updated only every ten years, this could result in higher rates of non-response – due to respondent fatigue – and introduce potential bias.

9 The American Statistical Society also maintains a regularly-updated website with information and material on Statistical Disclosure Control. See www.amstat.org/committees/cmtepc/index.cfm?fuseaction=main.

10 A summary of results of this workshop can be found at http://www.nsf.gov/sbe/ses/mms/nsfworkshop_summary1.pdf.

11 This terminology was introduced by D. Lambert [12].

7.2 Disclosure scenarios

First step in the process of anonymising a microdata file for dissemination is to identify those parts of the file at risk of disclosure. Thus, before applying SDC measures, it is useful to consider the types of scenario that could lead to a respondent's identification by someone using a microdata file. There are two situations that can lead to disclosure if a user endeavours to intrude into the file:

- The '***nosy-neighbour' scenario*** is when a user has sufficient information about the attributes of one or more records that stems from personal knowledge. In other words, this is a scenario in which the user belongs to the circle of acquaintances in a statistical unit. It is most likely when the sample target population is small – e.g. in enterprise, business and or facility surveys or household surveys conducted in small countries or sub-national areas with relatively few inhabitants.

- The ***external archive scenario*** refers to cases where a user links disseminated microdata file records to those in another available dataset (or register) that contains direct identifiers, even though this is specifically forbidden in the data use agreement. The intruder does this by using identifying variables available in both datasets as merging keys (data matching).

Prior to anonymising data, it is useful to define a disclosure scenario that describes which information potentially is available to a user, and how the latter could use it to identify an individual. Conservative assumptions are often made to define a worst-case scenario. More than one scenario may be necessary because different sources of information might be available to the user alternatively or simultaneously.

7.3 Assessing disclosure risk

Confidentiality is breached when a survey respondent is re-identified and an intruder – a user who is unauthorised or breaches the conditions set out in the data access and use agreement – can observe sensitive variables about the respondent.

When releasing microdata, NSOs must evaluate the data "to determine whether a public release would put the identity of individuals or establishments at risk. This evaluation takes into consideration issues such as 1) the level of detail for which data would be released (particularly as regards geographic specificity, and variables known to be held in common with outside data sources that serve as matching keys to increase the risk of identification); 2) certain variables or combinations of variables that render respondents unique within the sample and which might facilitate their recognition to outsiders; and 3) other linkable data already available outside [NSO], such as those already released from the same or a related survey or information held by others from the same respondent." [14]

"Determining the risk of disclosing identifiable data is a complex task that involves both empirical statistical analysis and judgment." [14] There are various approaches, each one with its own merits. However, currently none of these is considered 'the best' method. While accepting that disclosure risk cannot be eliminated entirely, there are measures that can offset reducing such risks against loss of utility. These involve adopting a threshold rule to establish whether release of a dataset is safe or not. There are two principal mathematical measurements of re-identification risk:

- ***Individual measurements*** assess the risk for each record. Usually these are expressed either as the probability of correctly re-identifying a respondent or by a measurement of uniqueness and rarity in the population sample.

- ***Global measurements*** assess the risk for the entire file. These are quantified as the expected number of correct re-identifications and can be derived by aggregating individual measures.

The advantage of using individual risk measurements is that only those records appearing unsafe for a given risk threshold need to be protected locally. This minimises loss of information and utility. Using only a global measurement of re-identification risk might prompt the application of SDC measures to each record in the microdata file. This might, therefore, result in a greater loss of information and potentially diminished utility for statistical analysis.

Re-identification-risk measurements also can be distinguished by the way they use the keys. Measurements based on keys in the sample identify unique or rare combinations of categorical variables (keys) in the sample. A unit is at risk if its combination of scores in the identifying variables is below a given threshold. For example, a 30-year-old male doctor with four female children might be unique in a survey

sample. In this case his record would be considered at risk. But there might be many people with these same characteristics in the total population. A unit's risk can be determined also by a combination of scores in the identifying categorical variables within the population or its probability of re-identification. Because the frequency in the population is generally unknown, these probabilities have to be estimated by modelling. When identifying variables are continuous, it is not possible to exploit the concept of rareness of the keys, since most, if not all, the keys would be unique. The disclosure risk via continuous variables is assessed by estimating the probability of re-identification by matching variables from two datasets based on the 'proximity' of their value.

There are various ways of identifying cases and variables in a file that can lead to disclosure. A commonly-used procedure is to generate a number of frequency distributions and multi-way tabulations to identify cells with low counts. Detailed geographical information is one of the main characteristics that can lead to identification of individuals, especially by users living in the same area who might know some of the respondents' characteristics.

Box 14 **Checklist to Help Assess Microdata Disclosure Scenarios and Risks**

In reviewing microdata files for dissemination, when such files are based on surveys and censuses for which the US Census Bureau is required to protect the confidentiality of respondents, one tool that facilitates this process is the *'Checklist on Disclosure Potential of Proposed Data Releases'* (www.census.gov/srd/sdc/). This is a series of questions designed to assist the reviewers in determining the suitability of releasing either public-use microdata files or tables. Section 3 of the checklist pertains to microdata files. The checklist focuses on major areas such as geographical information, variables presenting unusual risk of individual disclosure, contextual or ecological variables, potential links or matching with external data, and possible cross-tabulations that might identify a unique combination of attributes.

7.4 Statistical Disclosure Control (SDC) techniques for microdata files

First key step in SDC of a microdata file is to remove all direct identifiers – variables that unambiguously identify the respondent. Thereafter, a microdata file can be anonymised further either by applying ***masking methods*** or by constructing ***synthetic microdata***. A synthetic microdata file is generated randomly using a process that preserves certain statistical or internal relationships of the original raw microdata file. Masking methods, on the other hand, are techniques used to generate modified version of the original raw microdata file. There are two types of SDC-masking methods. ***Perturbing-masking methods*** edit and modify the data before publication by introducing an element of error purposely for confidentiality reasons. ***Non-perturbing-masking methods*** reduce the amount of information released by suppressing or aggregating data.

Below is an overview of the most widely-used SDC techniques and the type of data to which each can be applied: continuous, categorical or both. A variable is considered continuous if it is numerical, and arithmetical operations can be performed with it – for instance, variables such as 'income', 'age' and 'household size'. If standard arithmetical operations cannot be performed and the variable is defined only over a finite set, then it is considered categorical – for instance, variables either with ordinal scales, such as 'highest level of educational attainment', or with nominal scales, such as 'marital status', where the order between values is irrelevant.

Non-perturbing-masking techniques

Data reduction or non-perturbing-masking techniques modify the microdata files in order to eliminate variables or records that can be associated uniquely with an individual. Alternatively, categories can be created in such a way as to increase the number of possible respondents in a category. For example, an NSO may decide to use a rule of thumb that requires there to be a minimum number of responses in a cell. There are six commonly-used non-perturbing-masking techniques:

1. **Sampling**, instead of disseminating the entire microdata file, is advisable when releasing population census data. When the released sample is sufficiently small (e.g. five per cent of the population) and all direct identifying variables are removed, this can reduce sufficiently the risk of disclosure when the released data contains only categorical data. However, when the census microdata file also contains continuous variables that could be matched more easily with an externally-available data file, it can become necessary to apply additional SDC techniques.

2. **Global re-coding** involves aggregation of the observed survey values into pre-defined classes in such a way that individual responses are not visible. This approach can be applied

to continuous or discrete variables and to geographical codes. For example, age can be collapsed into age intervals and occupation and industry codes into broad categories, and geographical detail can be removed below the level at which the sample design is representative. Global re-coding is suitable for both continuous and categorical data.

3. **Top-coding and-bottom coding** are techniques applicable when, for numerical or ordinal variables, the highest and lowest values are very rare and could reveal the identity of respondents. Top coding involves creation of 'catch-all' categories such as 'age greater than x' or 'income greater than y'. Bottom coding involves creating catch-all categories for small values. This technique is suitable for continuous variables as well as categorical variables defined on ordinal scales, i.e. those that can be ranked in a meaningful manner.

4. **Local suppression** is a basic technique used when two variables taken together could lead to identifying a unique person. In other words, when combining these variables would result in a re-identification key for a particular record. For example, a record with the information 'age = 85' and 'school-attendance status = currently enrolled in primary school' is likely be unique, since there are not many 85-year-old people enrolled in primary school.[12] However, suppressing either 'school-attendance status' or 'age' for this particular record most likely would eliminate the problem. Local suppression is most useful when applied to categorical data.

5. **Removing variables** from a microdata file for dissemination is necessary if particular information is regarded as too sensitive to be released, for example ethnicity or religion.

6. **Removing records** is sometimes necessary to protect the anonymity of respondents with a unique set of variables. When a record is removed entirely from a microdata file, it is necessary to compute and include adjusted weighting factors. Use of this approach should be minimised as removal of a significant number of records will distort the data.

12 see, for example, http://news.bbc.co.uk/2/hi/africa/4244520.stm

Perturbing SDC techniques

Data perturburbing techniques involve modifying the data so their matching becomes difficult and less certain. If re-identification is attempted, the values thus modified create uncertainty about whether the match is a true one. A brief summary of seven perturbing techniques follows:

1. **Additive noise** is a technique that involves generating random values that can be added to those reported by the respondent. This can be done in a number of ways depending on whether applied to single or multiple variables or the noise is added so the means, variances and co-variances are preserved. In addition, linear-programming techniques can minimise differences between the altered values and the true ones.

2. **Data swapping** describes methods that transform a microdata file by exchanging values of confidential variables among individual records. Records are exchanged so that low-order frequency counts or marginals are maintained. This approach can be used for continuous and categorical data.

3. **Rank-swapping** is an approach whereby variables needing to be protected are sorted in ascending order, and groupings are constructed. Random pairs are selected from each group and their values swapped with values from other pairs within a pre-defined range. Creation of different group sizes leads to different data views.

4. **Micro-aggregation** involves replacing an observed value in the sample with the average of a small group of units (small aggregate or micro-aggregate), including the one under investigation. Units in the same group are represented in the released file by the same value. The groups contain a minimum pre-defined number k of units. The k minimum accepted value is 3. For a given k, the issue consists in determining the partition of the whole set of units in groups of at least k units (k-partition), minimising the information loss usually being expressed as a loss of variability. Therefore, the groups are constructed according to a criterion of maximum similarity between units. The micro-aggregation mechanism achieves data protection by ensuring there are at least k units with the same value in

the data file. This technique is sometimes applied to continuous variables.

5. **Rounding** techniques can be applied in a number of ways: to ensure that totals and certain summation properties are preserved; or, alternatively, randomly to ensure the cell counts in aggregate tables do not reveal counts of one or two observations.

6. **Re-sampling** involves taking a number of different independent samples of the values of the variables being masked. These are sorted using the same ranking criterion. The masked variables are created by taking, as first value, the average of the first values of the samples; and, as second value, the average of the second values, and so on…

7. **Post-randomisation** is a randomised version of data swapping. This technique induces uncertainty in the values of some variables by exchanging them according to a probabilistic mechanism. As with data swapping, data protection is achieved because users cannot ascertain with certainty if a released value is true. Consequently, attempts to match the record to external identifiers can lead easily to a mismatch or attribute misclassification. This method is used mainly for categorical variables, but is applicable likewise to continuous numerical variables.

Synthetic microdata files

A synthetic microdata file is generated randomly subject to the constraint that certain statistics or internal relationships of the original file are preserved. Providing entirely simulated microdata files with no disclosure risk, in place of the corresponding original microdata files, is an appealing proposition. Indeed, several techniques have been developed to generate such files. However, vis-à-vis data masking approaches, the different techniques of generating a synthetic microdata file all have one thing in common: they are extremely cumbersome and complex to implement. Typically, surveys collect data on hundreds of variables whose distribution and relationship are not modelled easily using standard parametric tools. Further research to improve and implement such techniques continues.

7.5 Managing the SDC trade-off: disclosure risk v information loss

Applying statistical disclosure control techniques to a microdata file results in information loss. An NSO must strike a proper balance when trading information loss for reduced disclosure risk. In the same way that disclosure risk can be determined, the NSO might be interested in assessing information loss associated with applying different SDC approaches. Information loss from categorical data can be assessed using techniques that include direct comparison, comparison of contingency tables and entropy-based measurement. For continuous data, comparisons of mean squares, absolute means, and mean variation can provide a measure of information loss.[13] However, as potential uses of microdata files are vast, it is simply impossible to undertake an exhaustive assessment of information loss. In practice, it more useful to identify which subset of users will be most affected by SDC application measures that minimise disclosure. Typically, these will comprise researchers skilled in conducting advanced statistical analysis with microdata files. Since generally this group is small and their research can result, one way or another, in substantive and important contributions of public benefit, it underlines further the need to disseminate not only PUFs but also the less anonymised licensed files. An NSO should be able to learn a great deal about information loss from SDC by examining the licensed files application forms that outline why the PUF version of a given microdata file is deemed unsuitable for particular research projects.

7.6 Documenting the SDC process

The SDC methods used should balance loss of information against the likelihood of individual information being disclosed. Users should be aware if a disseminated dataset has been assessed for disclosure risk and whether methods of protection have been applied. Dataset users should be provided with an indication of the nature and extent of any modification due to the application of disclosure-control methods. Any SDC technique(s) used may be specified, but the level of detail made available should not allow a user to apply reverse-engineering techniques to reconstruct the original microdata files.

13 See the IHSN website (www.ihsn.org) for references to information-loss assessment techniques.

Box 15 **How the US Census Bureau Reports SDC Measures Applied to the Census 2000 Public-use Microdata Sample Files**

"Confidentiality will be protected by the use of the following processes: data swapping, top-coding of selected variables, geographic population thresholds, age perturbation for large households, and reduced detail on some categorical variables.

Data swapping is a method of disclosure limitation designed to protect confidentiality in tables of frequency data (the number or percent of the population with certain characteristics). Data swapping is done by editing the source data or exchanging records for a sample of cases. Swapping is applied to individual records and, therefore, also protects microdata.

Top-coding is a method of disclosure limitation in which all cases in or above a certain percentage of the distribution are placed into a single category.

*Geographic population threshold*s prohibit the disclosure of data for individuals or households for geographic units with population counts below a specified level (see descriptions of Public Use Microdata Areas (PUMAs) and super-PUMAs in Section III).

Age perturbation, that is, modifying the age of household members, will be required for large households (households containing ten people or more) due to concerns about confidentiality.

*Detail for categorical variable*s will be collapsed if the categories do not meet a specified national minimum population threshold."

Source: http://www.census.gov/population/www/cen2000/pums/index.html accessed on August 10, 2010.

8. Should microdata be sold or provided free of charge?

Data producers may consider selling microdata products to recover their production costs, which are not always budgeted for. As statistics offices have become more autonomous, they are often faced with the obligation to identify and implement income-generating activities. Microdata are one of their high-value products.

While availability of microdata files enhances a survey's value, there will be costs incurred in the documentation and anonymisation process. If production of microdata for dissemination is not part of the survey budget, NSO might have little choice but to try and recover the cost of it.

A complete microdata dissemination service includes other costs such as a review process for release of files, licensing of data files, operating a data enclave, supporting users, and maintaining infrastructure. Generally this requires dedicated staff and infrastructure. If these are not in the agency's budget or the responsibility cannot be assigned to an existing part of the organisation, it may be difficult to provide a sustainable service of appropriate quality.

Selling data is one way of making the beneficiaries of the data service bear some of these costs.

8.1 Some countries' experience

There is a long history of cost recovery by NSOs around the world. It became common in the 1980s, generally in response to budgets cuts and pressure from central agencies to shift the burden of cost of statistics from taxpayer to user. While a complete review of NSOs' cost-recovery activities is beyond the scope of this document, here are a few observations:

- Statistics New Zealand (SNZ) undertook to recover 25 per cent of its total budget from the sale of products and services. This happened at a time when New Zealand underwent a rigorous administrate review. While this undertaking by SNZ staved off programme cuts, it was unable to reach its target, and eventually it was judged inappropriate so to do. Instead of attempting to recover costs, NSZ has undertaken a cost-reduction strategy. Visiting SNZ's website reveals that about 90 per cent of their information is freely available. Charges are levied for custom tabulations and detailed tables. There are fees for access to microdata files but these are not specified on the website. [14]

- Statistics Canada implemented a comprehensive cost-recovery programme in the 1980s in response to budgetary and political pressure. It undertook this challenge positively despite severe opposition from users. The programme met with uneven success and many parts of it have been eliminated or greatly reduced. When charges for microdata products were increased substantially, Canadian researchers turned to using data from the United States whose data were readily available through the Inter-university Consortium for Political and Social Research (ICPSR); the Data Liberation Initiative was born when it was seen that this was not serving Canadian objectives. The following quote provided the impetus for the DLI:

  > "...the genuine exercise of democracy increasingly requires that citizens get access to complex information and have the skills required to understand it" While he realises there are pressures on Statistics Canada to reduce costs and increase income, he feels the outcome has been the restriction of `...access to information only to groups that have the solid ability to pay". Bernard feels that this may "...hamper the participation in public debates of groups whose contribution is not backed up by much money' as well as `those who have no prospect of turning a profit or reaping some tangible and relatively immediate benefit from using it'. This, he states, is `...likely to lead, in the long run, to suboptimal development and less than full-blown democracy'." [31]

This led to creation of a network of 74 educational institutions with access to all public data from Statistics Canada. These data include approximately 300 public-use microdata files and thousands of other files, databases and geographical files. The subscription fee covers support costs as well as development of an enhanced technical infrastructure serving subscribers as well as the agency. It is not intended to cover cost of the data.

14 See http://www.stats.govt.nz/about-us/making-more-information-free/default.htm accessed on October 18, 2007.

The DLI access portal is not available to government departments, many of which have agreements with Statistics Canada to share the costs of certain surveys. Generally, their access to microdata data is covered by such agreements and thus they may not need this service.

8.2 Free or for a fee?

There is no definitive answer to this question. There are many reasons to minimise the charges for accessing microdata files; that is not to say there should be no charge at all.

Free

The main argument for no fees is that, in many cases, NSOs are in the process of strengthening their

data dissemination practices for microdata files. They fear too high a price may be a barrier to attracting users.

- Selling reduces considerably the number of potential users and hence the real value of the data.
- In developing countries, it may be an obstacle to users with the most interest in the data: students, local research centres, universities, etc.

The other argument against fees is the cost of collecting them. Also, it matters whether the fees revert to the NSO or a central agency. This affects staff incentives to recover fees: to work property, the system must be efficient.

- Selling imposes an obligation of quality and service.
- Selling generates little income: much of the demand is from academia, which has limited resources and the option of conducting their research elsewhere with other data.

While experience in other, mainly developed, countries shows some degree of cost recovery is possible, it is unlikely that all the additional costs of microdata dissemination can be recovered. Also, it can be shown that aggressive cost recovery discourages the use of microdata files and, in the long run, reduces a survey's potential value.

The most desirable approach for an NSO is to have all costs covered in the survey budget. This maximises accessibility. Such costs can be borne by sponsors as a way of maximising surveys' benefits. This is especially important in countries where researchers have limited finances and producers few resources for analysis.

For a fee

It is more than likely that charging for data files generates some income. But other factors must be kept in mind. The following points must also be considered in deciding whether or not to charge:

- Does the NSO have the legal right to charge a fee for their products?
- Which costs does the NSO wish to recover?
- Are these costs identifiable: will users understand and accept them?
- Can users afford to pay? Can a user-consortium be formed to recover the costs up-front? This entails identifying which costs should be recovered and dividing them among user organisations.
- Can fees be collected efficiently?
- Complex files freely available on a website may result in access by people lacking the ability to use microdata. This can lead to increased demand for support.

Another important consideration when developing a pricing strategy is that it should be in harmony with the pricing policy/philosophy for other products such as paper publications and Internet access. Most NSO websites are freely available to users as there are few, if any, incremental costs involved. The same, however, is not true for paper products. If publication pricing is based on covering the incremental costs of producing and shipping additional copies, the same principle could be applied to microdata files and the additional cost of supporting or servicing additional users.

9. When in the dissemination cycle should microdata files be released?

Production and release of microdata files must form part of the dissemination cycle. There is a strong argument to say that information aimed at broad audiences should be released first so the NSO can meet its immediate objectives and provide public feedback. Such information includes descriptive survey reports and analysis by the data producer. For official data producers, it is important to establish these official results and disseminate/publish them at the outset of the dissemination cycle.

Production of a microdata file demands additional time, uses specialised resources that must scheduled, and requires some sort of vetting. The NSO might need to satisfy certain internal or external research/analytical objectives and thus may choose to delay creation of a microdata file for some months following official release of survey results. Regardless of the timetable, researchers like to be informed of planned release dates so they can schedule their own work. Delays must be reasonable. Those of several years make results much less relevant.

"Even when micro-data are disseminated as promptly as possible, there are situations where it would be beneficial to release a portion of the micro-data or aggregated data prior to the time when the full set of micro-data can be made available. Such requirements [should be] included in the planning stages or raised as soon as the need for them is apparent." [14]

Box 16 **Policy Statement on the Timing of Data Release – US National Health Statistics Center**

"Science and the public good are best served by an open exchange of findings and views. Toward that end, NCHS policy is to disseminate micro-data *as soon as possible* following data collection, subject only to limits imposed by resources, technology, and data quality. NCHS will not impede the prompt dissemination of micro-data in order to preserve publication rights of its staff, collaborators, or the staff of other organizations.

1. Public-use data files will be released as soon as they have been prepared and the necessary reviews and approvals have been obtained, including review by the NCHS Disclosure Review Board.

   Depending on interest and expertise, NCHS may enlist the assistance of collaborators (including funding agencies) in the process of preparing data for public release, including editing and re-coding and on final file structure. To the extent that it is necessary to accomplish these steps, NCHS may provide a collaborator with data that are not yet released to the public. Any such release of files that are not yet ready for public release to a collaborator must be permissible under NCHS confidentiality policy and be consistent with NCHS legislative authority, informed consent, and submissions for human subjects reviews. Such releases are normally carried out under an agreement specifying how appropriate confidentiality protections are to be provided by the collaborator.

2. NCHS does not "embargo" data that are otherwise ready for public release and does not provide collaborators with preferential early access to data files or tabulations that are otherwise ready for public release, nor provide preferential release of tabulations based on data files that have not been publicly released.

   In cases where non-public data are made available to one user, it should be considered releasable to others who might request the same data, subject only to confidentiality provisions. Files or tabulations that might not be approved for public release due to the risk of disclosing confidential data may be made available in the NCHS Data Center (or, where consistent with NCHS confidentiality policy, through special use agreements) to ensure the widest possible access to the data. In a limited number of cases, such as for Departmental publications with lengthy publication lead times, tabulations may be made available in advance of a general data release.

Exceptions to this general policy will be rare and can be justified on a case-by-case basis. Requests should be made prior to the onset of data collection. They should be submitted to the NCHS Confidentiality Officer and will require the approval of the Director, NCHS."

Source: http://www.cdc.gov/nchs/about/policy/data_release.htm accessed on June 7, 2010.

10. What are technical infrastructure requirements for disseminating microdata files?

"Access to [microdata], and their optimum exploitation, requires appropriately designed technological infrastructure, broad international agreement on interoperability, and effective data quality controls. (...) [The] long-term sustainability of the infrastructure required for data access is particularly important. Research institutions and government organisations should take formal responsibility for ensuring that (...) data are effectively preserved, managed and made accessible in order that they can be put to efficient and appropriate use over the long term. (...) Specific attention should (also) be devoted to supporting the use of techniques and instruments to guarantee the integrity and security of research data. With regard to guaranteeing the integrity of a data set, every effort should be made to ensure the completeness of data and absence of errors. With regard to security, the data, along with relevant meta-data and descriptions, should be protected against intentional or unintentional loss, destruction, modification and unauthorised access in conformity with explicit security protocols." [17]

A proper technological infrastructure must be put in place for the various components of microdata archiving, i.e. for data documentation, cataloguing and dissemination, anonymization and preservation.

Microdata documentation

International metadata standards have been developed for the documentation of microdata and related resources. The Data Documentation Initiative (DDI) and the Dublin Core standards, described in Chapter 2, provide a practical solution. Documenting datasets in compliance with these standards is made easy by the availability of specialised metadata editors such as the IHSN Microdata Management Toolkit (Box 17) and the Nesstar Publisher software from the Norwegian Social Science Data Services.

Microdata cataloguing and dissemination

Interested users need to be properly informed about the existence and characteristics of the datasets made available. Many potential users will have very little if any information about the available datasets. Good metadata must be made available, preferably in the form of a searchable on-line catalogue.

Box 17 **IHSN Microdata Management Toolkit**

The Microdata Management Toolkit developed by the Norwegian Social Science Data Services (NSD) and the World Bank Data Group for the International Household Survey Network aims to promote the adoption of international standards and best practices for microdata documentation, dissemination and preservation.

The Toolkit comprises two modules. The **Metadata Editor** is used to document data in accordance with international metadata standards (DDI and Dublin Core). The **Explorer** is a free reader for files generated by the Metadata Editor. It allows users to view the metadata and to export the data into various common formats (Stata, SPSS, etc). The Metadata Editor and Explorer are based on Nesstar technology and developed by the Norwegian Social Science Data Services (NSD). The **CD-ROM Builder** is used to generate user-friendly outputs (CD-ROM, website) for dissemination and archiving.

See http://www.ihsn.org/toolkit

The objective of a microdata catalogue is to provide easy access to data and documentation in a format most convenient for users. A survey catalogue will provide tools for:

- Finding the data file most appropriate to the user's needs. This may be relatively trivial when the number of microdata files is small. But, as the number of files increases, a tool that can search data files at the variable level becomes essential.
- Evaluating information that has been identified to ensure compatibility with the researcher's needs, e.g. the universe, concepts and definitions employed in the survey. This role is supported by the metadata used to document the file.
- Accessing the data. This involves an extraction and/or delivery system of some sort. Commonly, such files can be delivered via a website/portal and an FTP server; or such tools can be used within the NSO to make a CD/DVD to deliver the data.
- Using the data. There is no such thing as a single tool for researchers to undertake their analytical

work. Rather, they prefer data available in a variety of formats as this allows them to use tools of their choice. Typically, these include formats for SPSS, STATA, SAS and ASCII formats.

The characteristics of a good microdata cataloguing system are:

- ***From the user point of view***, a good catalogue:
  - Complies with international metadata standard.

    International XML metadata standards such as the DDI and the Dublin Core standards considerably facilitate the production and maintenance of such catalogues.

  - Is web-based to facilitate discovery.
  - Provides rich metadata, including at the variable level.

    Survey catalogues become particularly relevant and powerful when the survey metadata provides not only a detailed description of the survey itself (with information on title, primary investigator, sampling, date of data collection, topics, geographic coverage, etc), but also a detailed description of each variable (with information on variable name and label, categories, literal question, interviewer's instructions, definitions).

    A variable-level catalogue can be relatively easily established using the DDI metadata standard, and the IHSN tools (in particular the IHSN Microdata Management Toolkit and the free NAtional Data Archive (NADA) application, available at www.ihsn.org).

  - Is searchable within all fields of the study. Within the DDI framework, this means the catalogue should be searchable within both the study (title, year, country, organization) and variable (variable name, variable label, variable value label) description fields. The catalogue should provide user-friendly full text search functionalities.
  - Provides clear information on the policy and procedure for accessing the data.
  - Provides a list and direct access to reference materials (questionnaires, manuals, reports).
  - Includes a "search by topic" compliant with a standard taxonomy of topics.

    To facilitate the exchange of information among catalogs, the data archive community has developed thesaurus to describe the topics covered by the datasets listed in their respective catalogs. A thesaurus is a set of terms or concepts used to describe objects like datasets, variables, books etc. The terms in a thesaurus are normally organized as a tree or hierarchy with broader terms being parents to narrower terms. Usually, a thesaurus will also include parallel terms and synonyms allowing users to find what they are looking for, even when they are not using the preferred terms.

    Many archives use a thesaurus when adding keywords at the study level or concepts at the variable level. The use of a thesaurus will encourage consistency by making sure that the same terms are selected when describing identical objects. Moreover, if users have access to the thesaurus when searching for data, there is a greater chance that they will use terms and concepts returning the most relevant list of hits.

    An example of the use of a thesaurus is the catalogue maintained by the Council of European Social Science Data Archives (CESSDA), an umbrella organization for social science data archives across Europe.

  - Is capable of displaying the results of searches quickly, even in large catalogs. This implies an efficient indexing system.
  - Provide a means to compare catalogue items. This is useful in comparing variables in standardized surveys or surveys for which multiple versions of the same study have been uploaded.

- Provides easily visible information on access policies for each study. For example, is the micro-data available and if so, provide clear instructions on how to obtain it.
- Provide good on screen help for users.
- Provide a means to link catalogue items to external web site resources as well as to allow the attaching of additional information, such as bibliographic references to publications that have used the study.

- ***From the catalogue administrator's point of view***, a good cataloguing system:
  - Provides a secure environment for storing and sharing data and metadata.
  - Provide the tools to manage the micro-data access process. This ranges from automated approval for micro-data with no access restrictions up to systems for managing and processing applications for which vetting is required before access is granted.
  - Provides a solution for sharing public use files and licensed files.
  - Provides a secure means for the sharing of micro-data and documentation, thus increasing end-user access.
  - Collect information on users of the catalog, the data they download and, where required, the purpose for which they are using the data. Such records are useful for the sponsors of the studies as they provide a means to gauge the use of the micro-data. Such records are also useful to end-users as they allow for users to be informed when new versions of the data are published or when changes are made to studies which they have downloaded.

Microdata anonymisation

Data anonymisation requires staff knowledgeable about statistics and software packages such as Stata or SPSS. Some specialized software is available to measure or reduce disclosure risk. None of these software applications provide an integrated and satisfactory solution for complex hierarchical data files. Practically, anonymization remains very much an ad-hoc process. Work is going on (among others by the IHSN) to develop tools and guidelines which may eventually contribute to make microdata anonymization more user friendly.

As already mentioned, survey-file anonymisation involves two steps: detection of potential instances of disclosure risk and some form of data reduction or perturbation in order to reduce the risk. This latter step requires input from someone with subject-matter knowledge who can recommend data reductions that will be least damaging to the researchers who will be using the files.

Microdata (and metadata) preservation

Digital data and metadata are vulnerable to software obsolescence, hardware and media obsolescence, physical threats, and human errors. Long-term preservation of data and metadata therefore requires proper procedures and infrastructure. Principles and good practices for preserving data are described in detail in an International Household Survey Network working paper produced by the Interuniversity Consortium for Political and Social Research [8].

11. What are the institutional and financial requirements for disseminating microdata files?

Disseminating microdata files may be a new activity for many NSOs and one that might lead to new and important uses of their data. In their *Principles and Guidelines for Access to Research Data from Public Funding*, the Organisation for Economic Co-operation and Development (OECD) defines a set of principles to which data providers should adhere:

> ***"Openness***
> Openness means access on equal terms for the international research community at the lowest possible cost, preferably at no more than the marginal cost of dissemination. (...)
>
> ***Flexibility***
> Flexibility requires taking into account the rapid and often unpredictable changes in information technologies, (...), legal systems and cultures of each (...) country. (...)
>
> ***Transparency***
> Information on research data and data producing organisations, documentation on the data and specifications of conditions attached to the use of these data should be internationally available in a transparent way, ideally through the Internet. (...)
>
> ***Legal conformity***
> Data access arrangements should respect the legal rights and legitimate interests of all stakeholders (...).
>
> ***Protection of intellectual property***
> Data access arrangements should consider the applicability of copyright or of other intellectual property laws that may be relevant to publicly funded research databases. (...)
>
> ***Formal responsibility***
> Access arrangements should promote explicit, formal institutional practices, such as the development of rules and regulations, regarding the responsibilities of the various parties involved in data related activities. These practices should pertain to authorship, producer credits, ownership, dissemination, usage restrictions, financial arrangements, ethical rules, licensing terms, liability, and sustainable archiving. (...)
>
> ***Professionalism***
> Institutional arrangements for the management of research data should be based on the relevant professional standards and values embodied in the codes of conduct of the scientific communities involved. (...)
>
> ***Interoperability***
> Technological and semantic interoperability is a key consideration in enabling and promoting international and interdisciplinary access to and use of research data. Access arrangements, should pay due attention to the relevant international data documentation standards (...).
>
> ***Quality***
> The value and utility of research data depends, to a large extent, on the quality of the data itself. Data managers, and data collection organisations, should pay particular attention to ensuring compliance with explicit quality standards. (...)
>
> ***Security***
> Specific attention should be devoted to supporting the use of techniques and instruments to guarantee the integrity and security of (...) data. (...)
>
> ***Efficiency***
> One of the central goals of promoting data access and sharing is to improve the overall efficiency of publicly funded (data collection) to avoid the expensive and unnecessary duplication of data collection efforts. (...)
>
> ***Accountability***
> The performance of data access arrangements should be subject to periodic evaluation by user groups, responsible institutions and (...) funding agencies. (...) [17]

It is likely that complying with these principles will require new procedures and outlooks. In their Guidelines [7], OECD defines the major issues inherent in providing data access as follows (these same issues apply to statistical microdata generated for statistical purposes by official data producers):

- "Institutional and managerial issues: while increased accessibility is important to all science communities, the diversity of the scientific enterprise suggests that a variety of institutional models and tailored data management approaches are most effective in meeting the needs of researchers.

- Financial and budgetary issues: scientific data infrastructure requires continued and dedicated budgetary planning and appropriate financial support. The use of research data will not be maximised if access, management, and preservation costs are an add-on or after-thought in research projects. It is important to note, however, that the cost of storing and managing data has decreased dramatically in recent years, and lack of knowledge about such changes can, in itself, be a barrier to advancement.

- Legal and policy issues: national laws and international agreements, particularly in areas such as intellectual property rights and the protection of privacy, directly affect data access and sharing practices, and must be fully taken into account in the design of data access arrangements.

- Cultural and behavioural issues: appropriate educational and reward structures are a necessary component for promoting data access and sharing practices. These considerations apply to those who fund, produce, manage, and use (…) data."

(…) Responsibility for the various aspects of data access and management should be established in relevant documents, such as descriptions of the formal tasks of institutions, grant applications, research contracts, publication agreements, and licenses." [17]

An alternative: trusted data repositories

For some data producers, establishing and maintaining such a data archive and dissemination service might be an unrealistic objective – for budgetary, legal or other reasons. An alternative option is to entrust an existing data archive. An example of a data archive is the UK Data Archive[15] located at the University of Essex. It manages and disseminates data from statistical agencies, research organisations and researchers themselves. Another example is the Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan, which performs a similar function in the United States.[16]

These data archives not only provide competent licence management but are also leaders in data curation and innovation. A recent example of this is the UK Data Archive's new suite of web pages providing guidance on data management and sharing. These aim to provide data creators, data managers and data curators with best practice strategies and methods for creating, preparing and storing shareable datasets. [17]

15 The United Kingdom Data Archive; see http://www.data archive.ac.uk/

16 See http://www.esds.ac.uk/aandp/create/research.asp or http://www.nsf.gov/pubs/2005/nsb0540/

17 See http://www.data archive.ac.uk/sharing/

12. How to promote use of microdata files?

Is the existence of properly-constituted public or licensed microdata files sufficient to develop a user base for such products? Unfortunately, the answer is, not necessarily. Users may have to be persuaded to become involved, and trained so to do.

National sample surveys and public-use datasets generated out of population censuses are however datasets that are of interest for a broad community of researchers and analysts. If properly documented and advertised, such datasets are most likely to be used extensively. The Demographic and Health Survey (DHS) programme provides a strong evidence of the large demand for such datasets.[18] DHS datasets are easily and freely accessible, and have been downloaded by a considerable number of users. This resulted in a diverse and rich collection of studies and publications.

In countries with a history of producing of such files, the richness of policy discussion based on their use is also well known. However, elsewhere the use of microdata files may not be well-understood. Those just beginning to disseminate microdata files might have to employ various means to promote their use and to educate users about their value – and also their limitations.

A culture of data sharing and collaboration should result in a wealth of new knowledge. NSOs and their partners are urged to promote the use of microdata files by way of, for example, seminars and training events, nationally and internationally. There are many such opportunities.

Microdata files are important for research and education. Their use in developing policies and programmes is important for both national governments and international agencies. These should be viewed as natural allies of the NSOs in supporting and promoting proper use of microdata files. In addition, universities can a play a key role in training new users.

Ensuring the right people are aware of the product and its benefits – and advocacy –influencing the attitude of a person or group to a particular issue – are necessary to the success of NSO microdata dissemination. They involve identifying and approaching the organisations and individuals therein likely to be potential users. In many cases, NSOs will be aware already of this need as they will have been approached by such users previously. Web links, pamphlets and inviting users to seminars are examples of steps that can be taken.

As mentioned above, training of potential users is also needed. The Canadian Data Liberation Initiative for example uses training sessions to promote and support the use of microdata files and other data products in research and teaching.[19] While many researchers in Canada were familiar with microdata files, those responsible for supporting them were not. This has been ameliorated using electronic mailing or e-mail lists via which researchers, NSO personnel, data intermediaries and other interested subscribers are able to ask questions and share expertise. The archive of answers provides an invaluable resource and alleviates the pressure on the NSO to answer all the questions it receives.

Part of the road to success in disseminating files to researchers is having NSO staff in the research and archive community network. This helps both parties understand each other's needs and concerns better and opens up two-way communication.

18 http://www.measuredhs.com

19 Readers are encouraged to visit the DLI training repository at https://ospace.scholarsportal.info/handle/1873/69 accessed on October 18, 2007.

Appendix 1

Application for access to a licensed dataset for a specific research purpose

The following provides a model template, which needs to be adapted to specific contexts.

Information you provide on this form will not be shared with others, unless a breach of the legal agreement is confirmed, in which case the NSO may inform partner statistical agencies in other countries.

This form is to be mailed or faxed with a covering letter on the sponsoring agency's letterhead, to:

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

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

_____

_____

Terms

In this agreement,

1. 'Principal researcher' refers to the researcher who will serve as the main point of contact for all communications involving this agreement. The Principal researcher assumes responsibility for compliance with all terms of this Data Access Agreement. The principal researcher must be an individual with authority to represent the receiving organisation in agreements of this sort.

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

3. 'Receiving organisation' refers to the organisation/university/establishment that employs the Primary Researcher.

Section A. Primary Researcher

- First name _____
- Last name _____
- Title _____
- Organisation _____
- Position in organisation _____
- Postal address _____
- Telephone (with country code) _____
- Fax no (with country code) _____
- E-mail _____

Section B. Other Researchers

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 Organisation

Organisation's name _____

Type of organisation (tick one)

- ☐ Line ministry/public administration
- ☐ University
- ☐ Research centre
- ☐ Private company
- ☐ International organisation
- ☐ Non-governmental agency (national)
- ☐ Non-governmental agency (international)
- ☐ Other (specify) _____

Organisation's website (URL) _____
Postal address _____

Section D. Description of Intended Use of the Data

Please provide a description of your research project (questions, objectives, methods, expected outputs, partners). If information is insufficient, your request may be rejected or additional information requested. This information may be provided in an appendix to this request.

List of expected output(s) and dissemination policy

_____
_____
_____

Section E. Identification of Data Files and Variables Needed

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

This request is submitted to access (tick one):

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

Section F. Data Access Agreement

The Primary Researcher and the other researchers agree to comply with the following:

1. Access to the restricted data will be limited to the Primary Researcher and other researchers identified in this Agreement.
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 anyone other than those mentioned in this Data Access Agreement, unless formally authorised by the NSO.
3. The data will be processed only for the stated statistical and research purposes. They will be used solely for reporting aggregated information and not for investigating specific individuals or organisations. Data will not be used in any way for administrative, proprietary or law-enforcement purposes.
4. The Primary Researcher undertakes that no attempt will be made to identify any individual person, family, business, enterprise or organisation. If such a unique disclosure occurs inadvertently, no use will be made of the identity of any person or establishment discovered and full details will be reported to the NSO. The identification will not be revealed to any person not included in the Data Access Agreement.
5. The Primary Researcher will implement security measures to prevent unauthorised access to licensed microdata acquired from the NSO. The microdata must be destroyed upon the completion of this research, unless NSO obtains a satisfactory guarantee that the data can be secured, and provides written authorisation to the receiving organisation to retain them. Destruction of the microdata will be confirmed in writing to the NSO by the Primary Researcher.
6. Any books, articles, conference papers, theses, dissertations, reports or other publications that employ data obtained from the NSO will cite the data source in accordance with the citation requirement provided with the dataset.
7. An electronic copy of all reports and publications based on the requested data will be sent to the NSO.
8. The NSO and the relevant funding agencies bear no responsibility for use of the data or for interpretation or inferences based upon it.
9. This agreement comes into force on the date approval is given for access to the restricted dataset and remains in force until the end-date of the project or earlier if the project is completed ahead of time.
10. If there are any changes to the project specification, security arrangements, personnel or organisation detailed in this application form, it is the Primary Researcher's responsibility to seek NSO's agreement to such changes. Where there is a change to the employer organisation of the Primary Researcher this will involve a fresh application and termination of the original project.
11. Breaches of the agreement will be taken seriously and the NSO will instigate action against those responsible for the lapse, if either wilful or accidental. Failure to comply with NSO's directions will be deemed a major breach of the agreement and may involve recourse to legal proceedings. The NSO will maintain and share with partner data archives a register of those individuals and organisations responsible for breaching the terms of the Data Access Agreement, and will impose sanctions on release of future data to these parties.

Signatories

The Principal researcher or an authorized representative of the receiving organization has read and agree with the Data Access Agreement as presented in section F above:

Name _____

Signature _____
Date _____

Request reviewed by … on [date]
Decision by the committee:

- ☐ Approved
- ☐ Deny [reason] _____
- ☐ More information needed: _____

Appendix 2

Model of a data enclave access policy

This formulation is for guidance only. It must be adapted by each country

Objectives

The National Data Enclave (NDE) was established by the National Data Archive to allow researchers with certain qualifications, and under strict supervision, to access confidential statistical microdata files. NDE provides a mechanism whereby researchers can access detailed data files securely, without jeopardising respondents' confidentiality.

Location

The NDE is located at [provide physical location, plus tel, fax, email and website]

NDE Operations

Researchers can access the data on-site and be provided with computer equipment, software, office space and NDE staff supervision.

Data

- NDE staff constructs the necessary data files before the guest researcher arrives, and ensure no restricted data leave the facility.
- Researchers proposing multiple analyses employing multiple datasets have access to only one dataset at a time. Under no circumstance are researchers permitted an opportunity to merge datasets on their own.
- NDE allows researchers to supply their own anonymous data to link to NDE datasets and create merged datasets for storage at NDE. The researcher-supplied data may consist of proprietary data collected and 'owned'' by the researcher, or other publicly-available data legally obtained by the researcher. Researchers MUST provide NDE staff with complete documentation of any data proposed for merger with NDE data. Researchers expecting to use merged files are responsible for interacting with NDE staff to ensure their data can be merged with NDE data. NDE accepts user data files in SAS, SPSS or Stata format.
- NDE periodically creates and maintains back-up copies of all computer files. Back-up files are stored securely and accessible by NDE staff only, although they may be made available to researchers needing to return for additional analyses. These back-up files contain user-supplied data as well as merged files, and will be destroyed at the written request of the user.

Computer Equipment

- NDE has [N] user work-stations and a black-and-white laser printer in a secure room. NDE computers are not linked to the Internet and are configured so that removable media such as CD-ROM or DVD writers, floppy disks or USB ports are inaccessible to users.
- NDE work-stations consist of [Pentium X XXX MHz] computers running [Windows NT / other?].

Software

- CSPro, EPI-Info, SAS, SPS and Stata are installed in the work-stations in addition to MS-Office applications. Additional programming/analytical languages can be supplied as needed. For more information on the software versions available at NDE, please contact us.
- Researchers must have sufficient expertise to conduct their own analyses with one of the software applications provided. NDE does not provide technical support for this.

Office Space

- Researchers must work under the supervision of NDE staff and only during normal working hours (Monday-Friday, 8:30 a.m. - 5:00 p.m.).
- Admittance to NDE is limited to researchers whose names are included in the research proposal. They are required to show photo-identification before admittance.
- A maximum of three collaborating researchers can sit at a computer station.
- Scheduling time at NDE is on a first-come, first-served basis.

NDE Staff Supervision (For Disclosure Review)

- External researchers are not allowed to bring documents, manuals, books, etc, that may enable them to identify and disclose confidential information accessed at NDE. Neither are they allowed to bring cell phones, pagers or other devices that would enable them to communicate outside NDE.
- Researchers may not save output, files or programs to transportable electronic media. NDE staff can copy output or programs to transportable media if requested.
- Researchers may take the results of their analyses off-site only after a disclosure review by NDE staff. Disclosure reviews consists of looking for tabular cells less than five, tables with geographical variables in any dimension, models with geographical variables (or variables tantamount to geographical variables) as outcome variables, or case listings.
- All logs must be printed or electronically archived and are kept by NDE, which will retain only programs and procedures run by external researchers. The logs will not include results of their own research.
- All computer output generated by statistical programs and all hand-written notes based on this are subject to disclosure review

by NDE staff before removal from NDE. Output is restricted to summary tables. In no case may any table contain cells with fewer than five observations. If found, these small cells are suppressed, generally by obliterating them. To ensure that small cells cannot be calculated from other cells in the same row or column, staff make illegible the totals for the rows and columns corresponding to the small cell. Once the disclosure review is completed, researchers receive a photocopy of the final tabulations. NDE staff use best practice in determining whether tabular data are identifiable and are conservative in their decisions. NDE decisions are final and not subject to negotiation by researchers.

Admission Costs

Researchers using NDE are charged for space and equipment rental and for staff time necessary for supervision, disclosure limitation review, maintenance of computer facilities (including hardware and software) and creation and maintenance of data files required by the researcher. Cost of accessing NDE is given below:

Affiliation of the primary data investigator	Set-up charge and file creation (fixed cost)	Use of facilities (per day, per computing station)
National Users		
Staff from NDE member agency	Free	Free
Other public agency	[Cost/Currency]	[Cost/Currency]
University/research centre	[Cost/Currency]	[Cost/Currency]
NGO	[Cost/Currency]	[Cost/Currency]
International Users		
Research in partnership with NDE	[Cost/Currency]	[Cost/Currency]
International organisation	[Cost/Currency]	[Cost/Currency]
University/research centre	[Cost/Currency]	[Cost/Currency]
NGO	[Cost/Currency]	[Cost/Currency]

An additional amount may be charged as needed for special handling – such as merging additional data, creating custom file formats or the acquisition and installation of specific non-standard software. The amount will be determined by discussion between the researcher and NDE staff. Payment is expected in advance of the using NDE.

Payments should be made to: [Provide instructions on mode of payment]

Submission of Research Proposals

Researchers must submit proposals using the form below. Prospective researchers are encouraged to check with NDE staff before writing their proposals to ensure the data of interest to them is available.
Researchers should develop their proposals in a way that helps NDE staff create the analytical files required for the project. Proposals should be explicit about the variables needed and any case selection required. Only data items needed for the proposed analyses will be included in the analytical data file and proposals should say why the requested data are needed. Overly large and complex projects or those poorly defined require extensive communication between NDE staff and the proposers. This can make the process move slowly. Work to prepare data files can be accomplished most expeditiously if large, complex projects are sub-divided into manageable parts and requested data are clearly defined.

Researchers wishing to link NDE data with external data should provide the latter to NDE staff before coming to NDE.

Upon receipt, the research proposal is evaluated by a review committee convened for that purpose.

The following criteria apply to a proposal review:

- Scientific and technical feasibility of the project.
- Availability of NDE resources.
- Risk of disclosure of restricted information.

Researchers should note that approval of their application does not constitute endorsement by NDE of the substantive, methodological, theoretical or policy relevance or merit of the proposed research. NDE approval is only a judgement that the research described in the application is not illegal use of the requested data file, and that it is highly probable the project can be completed successfully at NDE.

Appendix 3

Application for access to data in the National Data Enclave (NDE)

The following provides a model template, which needs to be adapted to specific contexts.

Information you provide on this form will not be shared with others unless a breach of the legal agreement is confirmed, in which case the National Data Enclave may inform partner statistical agencies in other countries.

This form is to be mailed or faxed, with a covering letter bearing the sponsoring agency's letterhead, to:

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

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

_____

_____

Terms

In this agreement,

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

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

3 'Receiving organisation' refers to the organisation/university/establishment that employs the Primary Data Investigator.

4 'Representative of the receiving organisation' refers to an individual with authority to represent the receiving organisation in agreements of this sort.

Section A. Primary Data Investigator

- First name _____
- Last name _____
- Title _____
- Prof/Dr/Mr/Mrs/Ms _____
- Organisation _____
- Position in organisation _____
- 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 _____

A current resume or *curriculum vitae* for each person who will participate in the research must be provided with this request. Resumes or CVs must specify nationality.

Section C. Receiving Organisation

Organisation name _____

Type of organisation (tick one)

- ☐ Line ministry/public administration
- ☐ University
- ☐ Research centre
- ☐ Private company
- ☐ International organisation
- ☐ Non-governmental agency (national)
- ☐ Non-governmental agency (international)
- ☐ Other (specify) _____

Organisation's website (URL) _____
Postal address _____

Section D. Representative of the Receiving Organisation

- First name _____
- Last name _____
- Title _____
- Prof/Dr/Mr/Mrs/Ms _____
- Organisation _____
- Position in organisation _____
- 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 (questions, objectives, methods, expected outputs, partners). Explain why publicly-available datasets are not sufficient for your purposes. If information

is insufficient, your request may be rejected or additional information required. This information may be provided in an attached appendix to this request.

List of expected output(s) and dissemination policy:
Will you need to merge the dataset with other data? ☐ YES ☐ NO
If YES specify all other datasets needing to be merged.

_____
_____
_____
_____
_____

Section F. Identification of Data Files and Variables Needed

The NDE provides detailed metadata on its website, including a description of data files and variables for each dataset. Researchers are requested to indicate which subset of variables or cases they are interested in, to allow the NDE to prepare the data files.

This request if submitted to access:

- ☐ The whole dataset (all files, all cases)
- ☐ A subset of variables and/or cases as described below (note that variables such as the sample weighting coefficients and records' identifiers will always be included in sub-sets):

Section G. Software Requirements

The following software will be used by the researchers:
☐ CSPro ☐ SAS ☐ SPSS ☐ Stata

Other software (specify): _____

Notes:

- *NDE regularly upgrades its software. Contact us if you need more information on the version of each application available.*
- *Researchers who need software not provided as a standard by NDE will have to provide NDE with a valid licence of the application, which will be installed by NDE staff for the duration of the research work (the licence will remain the property of the researcher). Please contact NDE prior to finalising this request to confirm technical feasibility.*

Section H. Data Access Agreement

If approved, following agreement will be signed:

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

1. Access to the confidential data will be limited to the Primary Data Investigator and Other Investigators listed in the application form, and who will sign the Affidavit of Confidentiality.
2. The data will only be processed for the stated statistical purpose. They will be used solely for reporting aggregated information and not for investigation of specific individuals or organisations. Data will not be used in any way for any administrative, proprietary or law-enforcement purposes.
3. The Primary Data Investigator undertakes that no attempt will be made to identify any individual person, family, business, enterprise or organisation. 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 will be reported to the NDE. The identification will not be revealed to any other person not included in the Data Access Agreement.
4. Any books, articles, conference papers, theses, dissertations, reports or other publications that employ data obtained from the NDE will cite the source of data in accordance with the citation requirement provided with the dataset.
5. An electronic copy of all reports and publications based on the requested data will be sent to the NDE.
6. The original collector of the data, the NDE, and the relevant funding agencies bear no responsibility for use of the data or for interpretation or inferences based upon such uses.
7. Breaches of the agreement will be treated seriously and the NDE will take action against those responsible for the lapse if either willful or accidental. Failure to comply with the directions of the NDE will be deemed a major breach of the agreement and may involve recourse to legal proceedings. The NDE will maintain and share with partner data archives a register of those individuals and organisations responsible for breaching the terms of the Data Access Agreement and will impose sanctions on release of future data to these parties.
8. The NDE reserves the right to terminate any project at any time it deems an investigator's actions will compromise confidentiality or ethical standards of behaviour in a research environment.
9. No printouts, electronic files, documents, written notes or media will be removed from the NDE until scanned for disclosure risk by NDE staff.
10. The Primary Data Investigator and other investigators may be barred from any future use of the NDE upon review and determination by the Director of the NDE that this is necessary to protect the integrity and confidentiality of the NDE.

Signatories

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

The Principal Data Investigator

Name _____
Signature _____ Date _____

Representative of the Receiving Organisation

Name _____
Signature _____ Date _____

NDE expects that all researchers will adhere to established standards and principles for carrying out statistical research. and analyses. Researchers must conduct only those analyses which have received approval. Failure to comply will result in cancellation of the research activity and potential disbarment from future research activities in the NDE.

References

[1] Altman, M. and King, G. 2006. "A Proposed Standard for the Scholarly Citation of Quantitative Data". http://gking.harvard.edu/files/cite.pdf

[2] Boyko, E. and Watkins, W. 2003. "Safe Data, Safe Places: Not Either/Or Solutions". In Eurostat. 19th CEIES seminar - Innovative solutions in providing access to microdata - Lisbon, 26 and 27 September 2002, pp 109-118. http://www.pedz.uni-mannheim.de/daten/edz-ma/eus/03/KS-PB-03-001-EN.PDF

[3] CENEX-SDC. 2007. "Handbook on Statistical Disclosure Control". http://neon.vb.cbs.nl/cenex/CENEX-SDC_Handbook.pdf

[4] Dupriez, O. and Greenwell, G. 2007. "Quick Reference Guide for Data Archivists". IHSN paper. http://www.ihsn.org/home/download.php?file=DDI_IHSN_Checklist_OD_06152007.pdf

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

[6] Hamilton, E. and Humphrey, C. 2000. "Measuring the Impact of DLI: Use of the NPHS Public Use Microdata File in Academic Outcomes".

[7] Hamilton, E, and Humphrey, C. 2002. "C.DLI and the NPHS: A Study in Compatibility". Fall 2002. http://www.statcan.gc.ca/dli-ild/doc/update52-miseajour52-eng.pdf

[8] Inter-university Consortium for Political and Social Research (ICPSR). 2009. "Principles and Good Practice for Preserving Data", International Household Survey Network, IHSN Working Paper No 003. December 2009. http://www.ihsn.org/home/index.php?q=focus/principles-and-good-practice-preserving-data

[9] ISO-IEC. 1999. "ISO-IEC 11179-1 – Information technology – Specification and standardization of data elements – Part 1: Framework for the specification and standardization of data elements". http://metadata stds.org/11179-1/ISO-IEC_11179-1_1999_IS_E.pdf

[10] King, Gary. 1995. "Replication, Replication. *PS: Political Science and Politics*, with comments from nineteen authors". http://gking.harvard.edu/files/replication.pdf

[11] King, Gary. 1995. "A Revised Proposal, Proposal". Vol. XXVIII, No. 3,September, 1995, pp 443-499. http://gking.harvard.edu/files/abs/replication-abs.shtml

[12] Lambert, D. 1993. "Measures of Disclosure Risk and Harm", Journal of Official Statistics, Vol 9, 407-426.

[13] Madsen, P. 2003. "The Ethics of Confidentiality: The Tension Between Confidentiality and the Integrity of Data Analysis in Social Science Research", mimeo, Carnegie Mellon University, USA, June 2003.

[14] National Center for Health Statistics (NCHS). 2002 . "Policy on Micro-data Dissemination". http://www.cdc.gov/nchs/data/NCHS%20Micro-Data%20Release%20Policy%204-02A.pdf

[15] National Center for Health Statistics (NCHS). 2004. "NCHS Staff Manual on Confidentiality". http://www.cdc.gov/nchs/data/misc/staffmanual2004.pdf

[16] National Center for Health Statistics (NCHS) – Research Data Center. 2008. "Guidelines for Proposal Submission". http://www.cdc.gov/nchs/data/r&d/guidelines_10_14_08c.pdf

[17] Organisation for Economic Cooperation and Development (OECD). 2007. "OECD Principles and Guidelines for Access to Research Data from Public Funding". http://www.oecd.org/dataoecd/9/61/38500813.pdf

[18] Statistics Canada. How to Cite Statistics Canada Products. http://www.statcan.gc.ca/pub/12-591-x/12-591-x2006001-eng.htm (accessed June 22, 2010)

[19] Tambay, J. L., Goldmann, G., and White, P. 2001. "Providing Greater Access to Survey Data for Analyses at Statistics Canada", Proceedings of the Annual Meeting of the American Statistical Association.

[20] UK Data Archive, University of Essex. 2002. "Good Practices in Data Documentation. Revised Version". http://www.esds.ac.uk/news/goodPractice.pdf

[21] UK Data Archive, University of Essex. 2009. "Managing and Sharing Data. A Best Practice Guide to Researchers", second edition. http://www.data archive.ac.uk/news/publications/managingsharing.pdf

[22] UK Statistics Authority. 2009. Code of Practice for Official Statistics. Edition 1.0. January 2009. http://www.statisticsauthority.gov.uk/assessment/code-of-practice/code-of-practice-for-official-statistics.pdf

[23] United Nations Economic Commission for Europe (UNECE). 2000. "Terminology on Statistical Metadata". Conference of European Statisticians Statistical Standards and Studies, No. 53, Geneva. http://www.unece.org/stats/publications/53metadaterminology.pdf

[24] United Nations Economic Commission for Europe (UNECE), Conference of European Statisticians. 2007. "Managing Statistical Confidentiality and Microdata Access: Principles and Guidelines of Good Practice". http://www.unece.org/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

[25] United Nations Economic Commission for Europe (UNECE). 2009. "Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes". http://www.unece.org/stats/publications/Confidentiality_aspects_data_integration.pdf

[26] United Nations Economic Commission for Europe (UNECE) and Statistics Sweden. 2003. "Statistical Confidentiality and Access to Microdata. Proceedings of the Seminar Session of the 2003 Conference of European Statisticians". http://www.unece.org/stats/publications/statistical.confidentiality.pdf

[27] United Nations Statistical Office. 1994. "Principle 6, Principles Governing International Statistical Activities". http://unstats.un.org/unsd/methods/statorg/Principles_stat_activities/principles_stat_activities.pdf

[28] United States Bureau of the Census, Software and Standards Management Branch, Systems Support Division. 2008. "Survey Design and Statistical Methodology Metadata". Washington DC, Section 3.4.4.

[29] US Federal Committee on Statistical Methodology. 2005. "Statistical Policy Working Paper 22 (Revised 2005) – Report on Statistical Disclosure Limitation Methodology". http://www.fcsm.gov/working-papers/spwp22.html

[30] Watkins, W, and Boyko, E. 1996. "Data Liberation and Academic Freedom", Government Information in Canada/Information gouvernementale au Canada 3, no.2. http://www.usask.ca/library/gic/v3n2/watkins2/watkins2.html

[31] Watkins, W. "The Data Liberation Initiative: A New Cooperative Model". Unpublished paper written for the Canadian Social Science Federation. http://library2.usask.ca/gic/v1n2/watkins/watkins.html

Websites

African Association of Statistical Data Archivists (AASDA)
http://www.aasda.net/home_dev/index.php

Australian Bureau of Statistics (ABS)
http://www.abs.gov.au/

American Statistical Society, Privacy, Confidentiality, and Data Security
http://www.amstat.org/comm/cmtepc/index.cfm?fuseaction=main

Central Statistics Office, Ireland (CSO)
http://www.cso.ie/

Council of European Social Science Data Archives (CESSDA)
http://www.cessda.org/

Data.Gov (UK)
http://*data.gov.uk*

Data.Gov (USA)
http://www.data.gov/

Data Liberation Training Depository (housed at the Ontario Universities Scholars Portal Economic and Social Data Service)
https://ospace.scholarsportal.info/handle/1873/69

Data Documentation Initiative Alliance (DDI)
http://www.ddialliance.org

Department of Census and Statistics, Sri Lanka
http://statistics.sltidc.lk

Dublin Core Metadata Initiative (DCMI)
http://dublincore.org/

Economic and Social Data Service (ESDS)
http://www.esds.ac.uk/aandp/create/research.asp

European Social Survey website
http://www.europeansocialsurvey.org/index.php?option=com_content&task=view&id=78&Itemid=190

Statistical Office of the European Commission (Eurostat)
http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/

Institute for Social and Economic Research, The British Household Panel Survey
http://www.iser.essex.ac.uk/ulsc/bhps/.

International Household Survey Network (IHSN)
http://www.ihsn.org

Inter-university Consortium for Political and Social Research (ICPSR)
http://www.icpsr.umich.edu

Luxembourg Income Study (LIS)
http://www.lisproject.org/

Measure DHS
http://www.measuredhs.com

Michigan Census Research Data Center (MCRDC)
www.isr.umich.edu/src/mcrdc/

National Center for Health Statistics (NCHS), Research Data Center (USA)
http://www.cdc.gov/nchs

National Opinion Research Center (NORC) at the University of Chicago
www.norc.org/DataEnclave

National Science Foundation (USA)
http://www.nsf.gov/index.jsp

Research Data Centres (RDC) Program of Statistics Canada
www.statcan.gc.ca/rdc-cdr/index-eng.htm

Statistics Canada, Data Liberation Initiative
http://www.statcan.ca/english/Dli/dli.htm

The Latin American and Caribbean Demographic Centre, CEPAL, United Nations Statistical Office
http://www.cepal.org.ar/software/icepa8c.html

United Nations Statistical Office
http://unstats.un.org/unsd/methods/statorg/Principles_stat_activities/principles_stat_activities.asp

UK Data Archive (UKDA)
http://www.data archive.ac.uk/sharing/metadata.asp
http://securedata.ukda.ac.uk/about/about.asp

UK Statistics Authority
http://www.statisticsauthority.gov.uk/assessment/code-of-practice/index.html

US Census Bureau
http://www.census.gov/population/www/cen2000/pums/index.html
www.census.gov/srd/sdc/

Wikipedia
http://en.wikipedia.org

About the IHSN

In February 2004, representatives from developing countries and development agencies participated in the Second Roundtable on Development Results held in Marrakech, Morocco. They reflected on how donors can better coordinate support to strengthen the statistical systems and monitoring and evaluation capacity that countries need to manage their development process. One of the outcomes of the Roundtable was the adoption of a global plan for statistics, the Marrakech Action Plan for Statistics (MAPS).

Among the MAPS key recommendations was the creation of an International Household Survey Network. In doing so, the international community acknowledged the critical role played by sample surveys in supporting the planning, implementation and monitoring of development policies and programs. Furthermore, it provided national and international agencies with a platform to better coordinate and manage socioeconomic data collection and analysis, and to mobilize support for more efficient and effective approaches to conducting surveys in developing countries.

The IHSN Working Paper series is intended to encourage the exchange of ideas and discussion on topics related to the design and implementation of household surveys, and to the analysis, dissemination and use of survey data. People who whish to submit material for publication in the IHSN Working Paper series are encouraged to contact the IHSN secretariat via info@ihsn.org.

www.ihsn.org
E-mail: info@ihsn.org