Microdata documentation, protection, and dissemination

TARGET AUDIENCE

The training program is intended for:

- Staff from statistical institutes and training centers in charge of training programs related to microdata management and processing.
- Mid-level survey statisticians and data processing specialists in national statistical agencies or line ministries, in charge of the management, processing, analysis, or dissemination of their organization's microdata.

QUALIFICATIONS

Participants are expected to have experience in the collection, processing, or analysis of microdata from censuses, surveys, or other sources and a solid background in statistics. Proficiency in the use of statistical software like Stata, SPSS, or R, and/or with data capture applications like CSPro or Survey Solutions, will be an advantage but is not a requirement.

COURSE DESCRIPTION

The course covers

(i) an introduction to tools and methods for microdata anonymization.

(ii) the use of international metadata standards and best practice for the documentation and cataloguing of microdata, and

Much of the training will consist of hands-on sessions. All software applications used in the training are open-source software and freely downloadable.

The first part of the course will introduce concepts, methods, and tools used for the anonymization of microdata. Participants will become familiar with statistical disclosure scenarios, with the measures of disclosure risk, and with some of the main methods available to reduce the risk of respondent reidentification. They will also learn about the limitations and alternatives to traditional anonymization methods. Hands-on sessions will allow participants to acquire practical experience on microdata anonymization, using the sdcMicro software application (an R package provided with a graphical user interface; no expertise in R is required from participants).

In the second part of the training, participants will learn about the DDI Codebook and Dublin Core metadata standards used for the documentation of microdata. They will be trained on the use of a specialized Metadata Editor software application and will learn about the benefits of metadata standards for data users and data producers. During hands-on sessions, they will document a survey dataset and publish standard-compliant metadata in an on-line data catalog.

Both parts of the training will make use of a synthetic data set provided by the training facilitators. This dataset can be openly distributed and subsequently used by the participants in their own training programs.

The program will also include a session by the Italian Statistics Office (Istat), on the practical implementation of statistical disclosure control and limitations in the context of a national statistical agency.

COURSE OBJECTIVES

Upon completion of this course, participants will:

- Understand the core concepts and methods of microdata anonymization, and their limitations.
- Acquire practical experience in the anonymization of microdata using the software package sdcMicro.
- Understand the importance of rich and structured metadata for data usability and discoverability.
- Be familiar with the purpose and use of the DDI Codebook and Dublin Core metadata standards.
- Acquire practical experience in the documentation, cataloguing, and dissemination of microdata.

DURATION

4.5-day course including lectures and hands-on sessions.

IT REQUIREMENTS

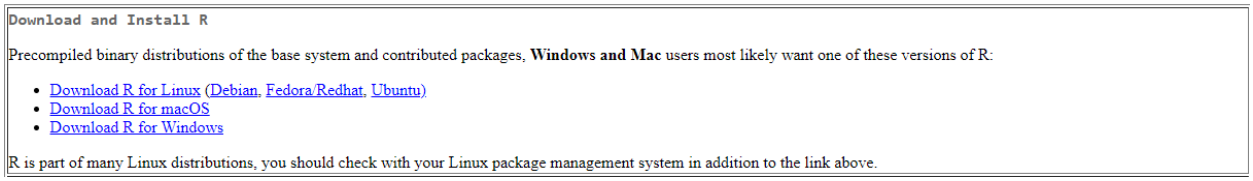
Each participant must bring her/his own laptop, with the following specifications:

- Operating System: Windows 11 or iOS (Mac)
- Minimum of 8Gb RAM, preferably 16 Gb
- At least 3 Gb of free hard-disk space
- Administrator access (participants must be allowed to install software on their laptop)
- Access to internet through a wireless network (WLAN)

Installation instructions

Downloading and installing R

The hands-on sessions on microdata anonymization will make use of R. They do not require participants to have prior expertise or experience using R. Participants are requested to install the latest version of R (version 4.2.3 as of March 2023) on their laptop prior to the training. The software (free, open source) is available from https://cran.r-project.org and is available for Linux, Windows, and Mac OS X.



Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux (Debian, Fedora/Redhat, Ubuntu)
- Download R for macOS
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Participants familiar with R should also install *devtools* available from https://devtools.r-lib.org/

The sdcMicro application (R library for the anonymization of microdata) will be installed during the first hands-on session on anonymization.

Microdata anonymization, documentation, and dissemination

SYLLABUS

Course overview

This course is intended to provide practical knowledge in standards and tools for the anonymization, documentation, and dissemination of survey microdata.

Learning objectives

By the end of the course, participants will have gained a good understanding of and hands-on experience with the **Statistical Disclosure Control (SDC) process** for microdata and with **metadata preparation, data cataloguing, and on-line survey data dissemination**. This includes:

- On microdata anonymization:
  - understanding why confidentiality matters,
  - knowing what role confidentiality plays in the data dissemination process,
  - understanding the commonly used disclosure risk measures and their use
  - understanding commonly used anonymization methods, their differences, applicability and choice of parameters and limitations,
  - assessing information loss resulting from anonymization,
  - implementing all the above in the software sdcMicro using a case study dataset.
- On microdata documentation and dissemination:
  - understanding the importance of rich and structured metadata for data usability and discoverability,
  - familiarization with the purpose and use of the DDI Codebook and Dublin Core metadata standards,
  - acquiring hands-on experience on the use of a Metadata Editor,
  - cataloguing a survey in an on-line microdata catalog.

Agenda / Outline of the lectures

The course consists of lectures and hands-on sessions over 4.5 days. The agenda is provided below. Two days will be devoted to the training on microdata anonymization. One day and a half will be devoted to data documentation and dissemination.

Materials available to participants

Participants will be provided with:

- **Software**: the training requires the use of R, sdcMicro, and of a Metadata Editor (all free software). R and sdcMicro will have to be installed on the participants' laptop. The Metadata Editor will run on a server and will be pre-installed by the training organizers.
- **Data**: a synthetic dataset with related metadata will be provided. The same dataset will be used for both the training on anonymization and the training on data documentation and dissemination.
- **Training and reference materials**:

- Statistical Disclosure Control for Microdata: A Practice Guide (Online document; a PDF copy will also be provided)
- Statistical Disclosure Control for Microdata: Theory (Powerpoint presentation; in English and French)
- sdcApp manual (Online document; in English only; a PDF copy will also be provided)
- Metadata Editor user guide (Online manual; in English only)
- Metadata standards and microdata dissemination good practice (Powerpoint presentation, in English and French)

All materials listed above are openly available and participants will be allowed to make use of them in their own training programs.

Dataset and software

Dataset for hands-on sessions

- For hands-on sessions on data documentation: a sample dataset of two data files will be provided to the participants and used for the hands-on sessions on data documentation and cataloguing. The dataset comprises one data file with variables at the household level, and one data file with variables at the individual level. The data have respectively 7,975 and 30,986 observations. They will be provided in Stata format. The variables and values in the data are labelled (in English and in French). The total size of the (uncompressed) files is 8 Mb. The data are extracted from a fully synthetic population dataset created by the World Bank. The data are for an imaginary country.
- For hands-on sessions on data anonymization: the same sample extracted from the synthetic population dataset will be used for the hands-on sessions on microdata anonymization.

Software applications

- The anonymization software and the Metadata Editor rely on the R software. Participants must therefore have R installed on their laptop. R is an open-source application. Instructions for installation are available in the "IT Requirements" section. Participants will preferably install R prior to the training. Those who cannot do that will be provided with support at the training.
- sdcMicro: sdcMicro is an R package. The package will be installed during the first hands-on session on data anonymization.
- Metadata Editor: The Metadata Editor is a web-based application. It will be installed on a server prior to the training. All participants will be provided with access to the application. A user account will be created for each participant prior to the training.
- NADA 5: NADA is a web-based application. It will be installed on a server prior to the training. All participants will be provided with access to the application.

Instructors' names, bios, and contacts

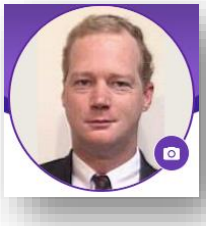
Mehmood Asghar is a senior data Mehmood Asghar is a Senior Data Scientist with the World Bank's Development Data Group. His work has focused on metadata standards, natural language processing, and developing open-source tools for data documentation and dissemination. He is the technical lead and the developer of the NADA software (web-based data cataloging tool) which National Statistics Offices are using in over 80 countries and international organizations.

Email: masghar@worldbank.org

Thijs Benschop is a Statistician with the World Bank's Development Data Group and works on the World Bank Microdata Library and the World Development Indicators (WDI) database. His expertise is in data confidentiality and anonymization, and his focus is on climate related indicators in the WDI.

Prior to the World Bank, Thijs worked as a research associate at CPB Netherlands Bureau for Economic Policy Analysis and FAO. A Dutch national, Thijs holds a B.Sc. in Econometrics and Operations Research and a B.A. in European Studies, both from Maastricht University, The Netherlands, as well as a M.Sc. in Statistics and a Ph.D. in Statistics from Humboldt University Berlin, Germany.

Email: tbenschop@worldbank.org



Olivier Dupriez is a Lead Statistician with the World Bank's Development Data Group. He is an advisor to the Data Group on issues related to data management and analytics. Recently Olivier has been leading a work program focusing on exploiting metadata standards, machine learning, and natural language processing models to improve data and knowledge discoverability. He is also a contributor to the development of software applications for data documentation, anonymization, and dissemination.

Olivier joined the World Bank in 2004 to establish the International Household Survey Network (IHSN). He worked with multiple international partners on the development and implementation of tools, guidelines, and good practice on microdata curation, dissemination, and protection.

Prior to joining the World Bank, Olivier worked for four years as a Poverty Statistician with the Asian Development Bank, for five years as a resident expert in population census for the United Nations and European organizations in Cape Verde, Chad, and North Macedonia, and for four years as a consultant (survey statistician) for the World Bank's Africa region in Djibouti and in Washington, DC. Olivier is experienced in designing and conducting surveys and censuses, in data analysis using Stata and R, and in data management and dissemination.

Email: odupriez@worldbank.org

C4D2 TRAINING

ROME | CENTER FOR DEVELOPMENT DATA

Microdata anonymization, documentation, and dissemination

July 10 – July 14, 2023
Perugia, Italy

PROGRAM[1]

Facilitated by

Mehmood Asghar, Senior Data Engineer, World Bank
Thijs Benschop, Data Scientist, World Bank
Olivier Dupriez, Lead Statistician, World Bank

[1] The week-long training features two kinds of activities:
- *Traditional Teaching*, lectures by the main instructors.
- *Hands-On Sessions*

Day 1 (Monday, July 10, 2023)

09:00 – 09:15	Opening remarks Tbd
09:15 – 09:30	Course overview Olivier Dupriez
09:30 – 10:15	Lecture 1[1] – What is Statistical Disclosure Control (SDC) and why do we need SDC? In English: Thijs Benschop In French: Olivier Dupriez
10:15 – 10:30	Tea/Coffee Break
10:30 – 12:45	Lecture 2 – Measuring disclosure risk In English: Thijs Benschop In French: Olivier Dupriez
12:45 – 14:00	Lunch
14:00 – 15:30	Hands-on session 1 - Software installation and exploration (Installation of the R package sdcMicro and exploration of the Graphical User Interface) - Importing a dataset - Measuring risk in the case study dataset (household data) Thijs Benschop, Olivier Dupriez, Mehmood Asghar
15:30 – 15:45	Tea/Coffee Break
15:45 – 17:15	Lecture 3 Anonymization methods - Perturbative and non-perturbative anonymization methods In English: Thijs Benschop In French: Olivier Dupriez

Day 2 (Tuesday, July 11, 2023)

09:00 – 10:45	Hands-on session 2 - Anonymizing a dataset: applying methods Thijs Benschop, Olivier Dupriez, Mehmood Asghar
10:45 – 11:00	Tea/Coffee Break
11:00 – 12:45	Lecture 4 - Balancing risk and utility - The SDC process In English: Thijs Benschop In French: Olivier Dupriez
12:45 – 14:00	Lunch

14:00 – 15:30	Hands-on session 3 - Assessing information loss (household data) - Transitioning to R from the Graphical User Interface (GUI) - Reporting Thijs Benschop, Olivier Dupriez, Mehmood Asghar
15:30 – 15:45	Tea/Coffee Break
15:45 – 17:15	Hands-on session 4 - Merging household and individual data - Anonymizing individual data Thijs Benschop, Olivier Dupriez, Mehmood Asghar

[1] Some of the sessions are delivered both in English and French in parallel. These sessions are indicated by two presenters ("In English:", "In French:")

Day 3 (Wednesday, July 12, 2023)

09:00 – 10:45	Hands-on session 5 - Anonymizing individual data Thijs Benschop, Olivier Dupriez, Mehmood Asghar
10:45 – 11:00	Tea/Coffee Break
11:00 – 12:45	Brainstorming session - Next steps - How can trainers be supported? Thijs Benschop, Olivier Dupriez, Mehmood Asghar, and organizers
12:45 – 14:00	Lunch
14:00 – 15:30	Lecture 5 – Introduction to metadata standards: what, why, how? – Data discoverability, accessibility, usability In English: Thijs Benschop In French: Olivier Dupriez
15:30 – 15:45	Tea/Coffee Break
15:45 – 17:15	Hands-on session 5 - Importing and exploring the dataset and related materials - Introducing the template manager Mehmood Asghar, Thijs Benschop, Olivier Dupriez

Day 4 (Thursday, July 13, 2023)

09:00 – 10:45	Hands-on session 6 - Metadata editor – Documenting a survey dataset Mehmood Asghar, Thijs Benschop, Olivier Dupriez
10:45 – 11:00	Tea/Coffee Break
11:00 – 12:45	Hands-on session 7 - Metadata Editor – Documenting a survey dataset (cont'd) Mehmood Asghar, Thijs Benschop, Olivier Dupriez
12:45 – 14:00	Lunch
14:00 – 15:30	Hands-on session 6 - Metadata editor – Documenting a survey dataset (cont'd) - Publishing the data and metadata in a data catalog Mehmood Asghar, Thijs Benschop, Olivier Dupriez
15:30 – 15:45	Tea/Coffee Break
18:00 - …	Social program – Discovering Italy

Day 5 (Friday, July 14, 2023)

09:00 – 10:45	Hands-on session 8 - Documenting and publishing other types of data: documenting a publication and an image Mehmood Asghar, Thijs Benschop, Olivier Dupriez
10:45 – 11:00	Tea/Coffee Break
11:00 – 12:30	Discussion / debrief - Building and maintaining a data library - More features of the tools / future developments - How to train / support trainers? - Closing Mehmood Asghar, Thijs Benschop, Olivier Dupriez
12:30 – 14:00	Lunch