CRP.1
ENGLISH ONLY
22 September 2017

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Skopje, 20-22 September)

REPORT OF THE MEETING

Prepared by the UNECE secretariat

PARTICIPATION

1. The Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality was held in Skopje, from 20-22 September 2017. It was attended by participants from:, Australia, Austria, Bosnia and Herzegovina, Bulgaria, Canada, Croatia, Finland, France, Germany, Hungary, Italy, Japan, Lithuania, Mexico, Montenegro, Netherlands, New Zealand, Norway, Romania, Russian Federation, Serbia, Slovenia, Sweden, the former Yugoslav Republic of Macedonia, Turkey, United Kingdom of Great Britain and Northern Ireland, United States of America as well as by representatives from the European Central Bank, Eurostat, and UNECE. Participants from numerous universities and institutes attended the work session at the invitation of the UNECE secretariat.

ORGANIZATION OF THE MEETING

2. The agenda of the work session consisted of the following substantive topics, the outcomes of which are documented in the annex:

(i) Utility and disclosure risk in anonymised data
(ii) Microdata and output protection
(iii) Confidentiality of big data and special types of data
(iv) Access to microdata
(v) Methods and tools for tabular data protection
(vi) Census 2021 - Confidentiality issues
(vii) Public use files / open data - Future availability of information

3. Mr. Apostol Simovski, Acting Director General of State Statistical Office, the former Yugoslav Republic of Macedonia, opened the workshop and welcomed the participants. Peter-Paul de Wolf (Netherlands) and Eric Schulte Nordholt (Netherlands) were elected as co-chairs of the Work Session.

4. The provisional agenda was adopted.

5. The following persons acted as Session Organizers:

Topic (i): Josep Domingo-Ferrer (Universitat Rovira i Virgili);
Topic (ii): Krish Muralidhar, (University of Oklahoma);
Topic (iii): Peter-Paul de Wolf (Netherlands) and Josep Domingo-Ferrer (Universitat Rovira I Virgili);
Topic (iv): Aleksandra Bujnowska (Eurostat) and Annu Cabrera (Finland),

Topic (v): Sarah Giessing (Germany);
Topic (vi): Eric Schulte Nordholt (Netherlands); and
Topic (viii): Eric Schulte Nordholt (Netherlands).

RECOMMENDATIONS FOR FUTURE WORK

6. The participants reviewed the recommendations for future work on the basis of a proposal put forward by an ad hoc working group composed of Maël Buron (France), Josep Domingo-Ferrer (Universitat Rovira i Virgili), Prof. Mark Elliot (University of Manchester), Mr. Tobias Enderle (Germany), Mr. Antony Gomez (New Zealand), and Dr. Krish Muralidhar (University of Oklahoma).

7. The participants considered it useful to continue the exchange of experiences in the field of statistical data confidentiality, and recommended that a future work session on statistical data confidentiality be convened in 2019. The following topics were proposed:

- Crisis Management – what happens when there is a breach of confidentiality?
- Confidentiality issues with administrative data
- Census 2021
- Increasing risk appetite – what are the implications?
- Case studies of remote access to microdata
- Legal requirements - how to comply with them
- Software tools for statistical data confidentiality
- Case studies of on-demand tabulation
- Software tools for secure statistical computing
- Merging several anonymised data sources (and other big data issues)
- Data user perspective
- Confidentialising other types of data – e.g. streaming, unstructured, location, and mobility
- First practical steps to implement statistical data confidentiality
- Formal privacy
- Future challenges for statistical data confidentiality

FURTHER INFORMATION

8. The conclusions reached during the discussion concerning the substantive items on the agenda are contained in the Annex. All background documents, presentations and the final report for the meeting are available on the website of the UNECE Statistical Division: **https://statswiki.unece.org/display/SDC2017**

9. On behalf of the participants, Ms. Thérèse Lalor (UNECE) expressed her great appreciation to the State Statistical Office of the former Yugoslav Republic of Macedonia for hosting this meeting and providing excellent facilities for the work.

ADOPTION OF THE REPORT

10. The participants adopted the present report before the Work Session adjourned.

Annex: Summary of discussions on substantive topics

A. Topic (i): Utility and disclosure risk in anonymised data

12. This topic was organized by Josep Domingo-Ferrer (Universitat Rovira i Virgili). It included the following presentations:

- University of Oklahoma - Mahalanobis distance-based record linkage revisited
- Chuo University - Investigating New Methods for Creating Anonymized Microdata Based on Japanese Census Data
- Universitat Rovira i Virgili - A Methodology to Compare Anonymization Methods Regarding Their Risk-Utility Trade-Off
- University of Manchester - A Study of the Impact of Synthetic Data Generation Techniques on Data Utility using the 1991 UK Samples of Anonymised Records.
- The former Yugoslav Republic of Macedonia - Measures for information loss in protected data

13. The following points were raised in the discussions:

- The parameterization of PRAM needs to be carefully considered.
- There is a tension between official statistics and researchers regarding the transparency of masking techniques. It is useful to reverse engineer the masking methods to find the best linking method.
- Data about people and data about enterprises are different and require different approaches. Some National Statistical Organizations (NSOs) do not provide access to enterprise microdata, because it is difficult to anonymize the data.
- A point raised in the discussion was whether different methods are better suited for different types of data. There is no short cut as whether a method works better depends on how correlated the data is, how much variation there is, and the number of attributes in the data set.
- The participants discussed the status of synthetic data in NSOs. It is complicated and labour-intensive to create synthetic data. There is a risk when using synthetic data that conclusions that are not real are drawn. Synthetic data can be useful for preliminary analysis, but full analysis should be done on real data.

B. Topic (ii): Microdata and output protection

14. This topic was organized by Krish Muralidhar, (University of Oklahoma). It included the following presentations:

- Austria - sdcApp - a shiny new GUI for sdcMicro
- University of Manchester - Evolutionary Methods on Synthetic Data
- University of Edinburgh - Recognising real people in synthetic microdata: risk mitigation and impact on utility
- University of Manchester - A unified approach to the assessment of both identification and attribution risks

15. The following points were raised in the discussions:

- It is useful to have software that evaluates many alternative masking mechanisms using a user-friendly interface and assesses the trade-off between utility and risk.
- There are a number of techniques that have existed for many years. It is important to think how to modify and enhance those techniques to suit current needs.

C. Topic (iii): Confidentiality of big data and special types of data

16. This topic was organized by Peter-Paul de Wolf (Netherlands) and Josep Domingo-Ferrer (Universitat Rovira i Virgili). It included the following presentations:

- The former Yugoslav Republic of Macedonia - Data confidentiality and statistical registers in the Macedonian statistical system
- Secure Data Access Center - CASD-TeraLab Secure Access to Big Data
- Netherlands - Location related risk and utility

17. The following points were raised in the discussions:

- When choosing a solution for statistical data control it is important to think about how the solution might evolve in the future.
- It was clarified that the method developed by CASD is for use in distributed computing environments.
- When looking at location related geographic data, it is possible to protect the map and not the microdata file.
- The protection of other data types was discussed.
    - Streaming data: The role of streaming data is not yet clear.
    - Network data: There are simple examples of this in NSOs, e.g. linked employee/employer data.

D. Topic (iv): Access to microdata

18. This topic was organized by Aleksandra Bujnowska (Eurostat) and Annu Cabrera (Finland). It included the following presentations:

- The United Kingdom - The Five Safes Framework
- Mexico - Best International Practices Translated into Local Context
- Canada - Access to Statistics Canada's Microdata
- Japan - On-site Service and Safe Output Checking in Japan
- France - Digital act in France: Impact on NSI
- Hitotsubashi University/Japan - A Proposal of a Simple and Secure Statistical Processing System using Secret Sharing
- University of the West of England - Lessons learned in training 'safe users' of confidential data

19. The following points were raised in the discussions:

- When creating a microdata access system, it is important to consider the perspectives of both the NSO and research environments. This will lead to more efficient systems.
- Many organisations give access to microdata if the project is for public good. There is no standard definition of public good. It differs between organisations, depending on how the organisation currently wants to define it.
- It is impossible to track everything that a researcher does when using microdata, organisations must decide to trust them or control them.
- A common approach to training related to confidentiality has been to focus on rules and the consequences of (not) following them. Training that focuses on attitudes rather than behaviour may gain more buy-in from users.
- It is not just the NSO that has a role in training users and researchers in confidentiality. It would be useful if universities taught data handling and data ethics.

E. Topic (v): Methods and tools for tabular data protection

20. This topic was organized by Sarah Giessing (Germany). It included the following presentations:

- Eurostat - Statistical Confidentiality in European Business Statistics
- France - Center of Excellence on Statistical Disclosure Control
- New Zealand - Establishing an Automated Confidentiality Service in Stats NZ
- Finland - Harmonization of the protection of social statistics at Statistics Finland
- Universitat Politècnica de Catalunya - On using an improved Benders method for cell suppression
- Australia - Constrained optimisation for tabular suppression in the Australian Bureau of Statistics
- Canada - Disclosure control that accounts for survey realities: assessing the risk using G-Confid

21. The following points were raised in the discussions:

- While regulations could be set for statistical disclosure control, there is a preference to have guidelines and recommendations. There have been some excellent examples of the harmonisation of methods across Europe (for example, on the protection for censuses).
- Many countries are using active confidentiality. The use of waivers is compatible with this. To implement waivers, some countries send requests for waivers with the questionnaire, others target enterprises that require a lot of suppressions, and others get waivers from the top contributors.
- Data about people can remain sensitive over time. However, data about business can become less sensitive over time. Perhaps there is some scope to set an expiry data on confidentialised business data.
- There are many tools available from the open source community. Organisations have different comfort levels and success with using open source tools as part of the statistical production process.

F. Topic (vi): Census 2021 - Confidentiality issues

22. This topic was organized by Eric Schulte Nordholt (Netherlands). It included the following presentations:

- Germany - Testing CTA as Additivity Module for Perturbed Census 2021 EU Hypercube Data
- Norway - The European Census Hub Hypercubes 2011- Norwegian SDC Experiences
- The United States of America - The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau
- Slovenia – SDC considerations of publishing data in the application STAGE
- The United Kingdom - Creation of synthetic microdata in 2021 Census Transformation Programme (proof of concept)
- Nippon Telegraph and Telephone - Secure statistical computation system on encrypted data: An empirical study of secure regression analysis for official statistics
- The United Kingdom - Progress towards a table builder with in-built disclosure control for 2021 Census

23. The following points were raised in the discussions:

- Differential privacy is a measure of disclosure based on the assumption that the ratio of the probabilities of observing a response to a query in the presence or absence of an individual should be within a small bound expressed as the exponent of epsilon (the security

parameter). Using this approach should be considered carefully as some problems have been found with its utility.

- Small values must be masked in some countries. Several approaches are used (removal, introducing uncertainty, etc). Organisations may choose to approach small values differently based on user consultations and legal issues.
- If a member of the public thinks they can identify themselves in the data (correctly or not), this causes debate in the media and can harm the office.
- The release of perturbed tabular data can cause additional effort in the research data centres to ensure that there is no inconsistency in the outputs. Some countries avoid this by using a sample for their census microdata.

G. Topic (viii): Closing Panel Discussion

24. This topic was organized by Eric Schulte Nordholt (Netherlands), and concerned Public use files / open data and future availability of information. It included the following presentations:

- University of Essex - Public use files – are they obsolete
- Finance Think - Open data – High impact or struggle to access?
- University of the West of England - Open data: who needs it?

25. The following points were raised in the discussions:

- Is it necessary to register and provide personal information to download the data? The registered person might not be the end-user of the data.
- It is useful for organisations to understand who their users are. This information can help to understand the return on investment. Feedback can be gathered using other mechanisms, such as a feedback form on the website.
- The production of public use files is expensive; therefore it could be worth investigating whether they are used enough to justify their production.
- As it is difficult to predict what users will want in the future, it is better to provide incremental improvements in tools. Otherwise, by the time organizations have developed a tool, the users may have already moved on.
- Open data would be the ideal option for users, but organizations need to balance user needs, data providers and privacy laws.
- Case studies about what has and has not worked are valuable to statistical organisations investigating how they might release open data.
- When there is a breach, the focus is not on the user but on the statistical organisation. Often, we only hear the negative experiences. There are many positive stories of the data being used. These should be shared more widely. It is important to educate the public on disclosure issues.