Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development

14-Mar-2014

English - Or. English

STATISTICS DIRECTORATE
COMMITTEE ON STATISTICS AND STATISTICAL POLICY

Declassified

EXPERT GROUP FOR INTERNATIONAL COLLABORATION ON MICRODATA ACCESS

FINAL REPORT

Meeting of the Committee on Statistics and Statistical Policy - 11th Session
7-8 April 2014, OECD Conference Centre, Paris

This document contains the full final report of the CSSP Expert Group for International Collaboration on Microdata Access. The Executive Summary of this report has been issued under STD/CSSP(2014)9.

This document is for BACKGROUND INFORMATION under Agenda item 9b.

For further information please contact Paul J. Jackson, Office for National Statistics, United Kingdom (email: paul.j.jackson@digital.ons.gov.uk) or Mariarosa Lunati, OECD Statistics Directorate (email: mariarosa.lunati@oecd.org).

JT03354274

Complete document available on OLIS in its original format
This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

STD/CSSP/RD(2014)2
For Official Use

English - Or. English

NOTE FROM THE SECRETARIAT

The Expert Group for International Collaboration on Microdata Access of the OECD Committee on Statistics and Statistical Policy (CSSP) was created in June 2011. The membership includes 25 countries: Australia, Belgium, Canada, Chile, Czech Republic, Denmark, Estonia, France, Greece, Germany, Hungary, Israel, Italy, Japan, Korea, Mexico, Netherlands, New Zealand, Norway, Slovenia, Sweden, Switzerland, Turkey, United Kingdom, United States; and Eurostat.

The Mandate of the Expert Group is to propose concrete solutions to encourage and facilitate cross-border access to official microdata.

The Expert Group has taken into account a number of other international and national initiatives relevant to the aim of maximising the safe use of microdata, including UNECE High-Level Group for the Modernisation of Statistical Production and Services, the International Household Survey Network (IHSN), the European Commission projects Data without Boundaries (DwB) and Decentralised and Remote Access to Confidential Data (DARA), the Science as an Open Enterprise report of the Royal Society, the 2013 Seminar on Microdata of the Conference of European Statisticians, the Data Liberation Initiative of Statistics Canada, the Statistical Data Integration initiative of Australian Bureau of Statistics, the UK Economic and Social Research Council (ESRC) data analysis initiative, the Paris Microdata Group. Some of the Group's members are/were active in these other initiatives, and the benefits have flowed in both directions.

This document contains the final report[1] prepared by the Expert Group. It presents the conclusions of the work of the Group and its recommendations.

ACTIONS REQUIRED

Delegates are asked to:

EXPRESS THEIR VIEW on the proposed paradigm to improve international access to official microdata, which assigns an important role to the concept of trust.

PROVIDE comments on the main recommendations included in the report.

CONSIDER whether the Mandate of the Expert Group for International Collaboration on Microdata Access should be extended to investigate the implementation, and feasibility, of recommendations.

[1] The Executive Summary of this report has been issued under STD/CSSP(2014)9.

PREFACE

Providing access to microdata for statistical and research purposes is now the norm for most statistical organisations. However, most microdata access services stop at the first national boundary. International collaboration with microdata remains an exception rather than the rule.

The challenge in the 21st Century is to change practices in access to microdata so that the access services can cross borders and support trans-national analysis and policy making. This is necessary to reflect the increasingly international (global) reach and impact of comparative analysis and shared policy making. This report, prepared by the OECD Expert Group for International Collaboration on Microdata Access, does not need to elaborate on the need for international collaboration in statistical development, production, dissemination and analysis as this is now fully established by other forums and initiatives. The purpose of this report is to make recommendations which, if adopted, will allow statistics organisations to change 20th century national microdata access processes into services that will support 21st century international collaboration in microdata.

None of the recommendations are in themselves radical. Each can be adopted with relatively small changes in current practice. No new legislation, nor substantial new infrastructure, nor new technology is called for. Instead, the report seeks for the smarter deployment of what already exists in most OECD countries.

Firstly, the Group recommends adopting an objective approach to selecting the best modality for international collaboration in microdata.

Three simple and familiar models of international collaboration in microdata are described in terms of *trust*, *method* of access, and *legal* relationship. These terms are selected because the barriers to providing access to confidential microdata across borders are usually found in one of them.

Secondly, the Group recommends using a maturity model for assessing current practices in international collaboration in microdata, and for delivering strategic change.

Maturity models provide a practical way for senior leaders in statistics offices to self-assess, set targets, and strategically plan for change. Maturity modeling has proved its worth in other complex areas of business change and is described in this report in the context of international collaboration in microdata for the first time. Statistics offices with an "embedded" level of maturity will be able to participate, and innovate, in international collaboration in microdata.

Thirdly the Group has concluded that change to improve international collaboration in microdata is best achieved if:

a. Microdata assets and practices use a common terminology derived from a recognised glossary.
b. Trust in international delivery partners and data users is objectively established according to a circle of trust model.
c. The costs, benefits, and risk management of microdata access are made more transparent.
d. International collaboration with microdata is made part of core statistical business planning.

The detailed recommendations of the Expert Group are presented in the report to assist statistics offices that wish to make positive and planned movements in their international microdata collaboration maturity.

Paul J. Jackson
Office for National Statistics, United Kingdom
Chair of the OECD Expert Group for International Collaboration on Microdata Access

ACKNOWLEDGMENTS

The Executive Summary of this report presents the conclusions of the work of the OECD Expert Group for International Collaboration on Microdata Access. Members of the Expert Group include representatives from National Statistical Offices and Eurostat: Jenine Borowik and Merry Branson (Australian Bureau of Statistics), Youri Baeyens and Mona Kombadjian (Belgium), Heather Dryburgh (Statistics Canada), Ximena Clark (Banco Central de Chile) and Juan Radrigán (INE, Chile), Ondřej Vozár (Czech Republic), Ivan Thaulow (Statistics Denmark), Ioannis Nicolaidis (Statistics Greece), Tuulikki Sillajõe (Statistics Estonia), Michel Isnard (INSEE, France), Maurice Brandt (Destatis, Germany), Zoltán Vereczkei (Central Bureau of Statistics of Hungary), Brian Negin (Central Bureau of Statistics of Israel), Luisa Franconi (ISTAT, Italy), Hideaki Nakamura (Statistics Japan), Gihyeong Ryoo (Statistics Korea), Natalia Volkow (INEGI, Mexico), Leo Engberts (Statistics Netherlands), Hamish James (Statistics New Zealand), Johan-Kristian Tønder (Statistics Norway), Tomaz Smrekar (Statistics Slovenia), Claus-Göran Hjelm and Eva Nilsson (Statistics Sweden), Anne Balzli-Prysi (Statistics Switzerland), Ilker Guven and Nalan Eviren (Statistics Turkey), Paul J. Jackson (United Kingdom Office for National Statistics), Brian Harris-Kojetin (Office of Budget and Management of the United States); and Aleksandra Bujnowska (Statistical Office of the European Union, Eurostat).

Mariarosa Lunati from the OECD Statistics Directorate assisted in the preparation of this report. Young-tae Son from Statistics Korea compiled the Glossary for International Microdata Access during his secondment in the Statistics Directorate. Merry Branson (Australian Bureau of Statistics) and Hamish James (Statistics New Zealand) provided detailed comments. François Fonteneau (FAO and OECD Paris 21) and Steve Vale (UNECE) contributed with input and advice.

Expert Group members contributed to individual chapters of this report, either as lead authors or contributors: Zoltán Vereczkei (Chapter 2); Maurice Brandt (Chapter 4); Brian Negin, Paul Jackson and Aleksandra Bujnowska (Chapter 5); Natalia Volkow (Chapter 6); Claus-Göran Hjelm and Eva Nilsson (Chapter 7); Aleksandra Bujnowska (Chapter 8); Luisa Franconi (Chapter 9); Brian Negin (Chapter 10); Paul Jackson (Chapter 11); Heather Dryburgh, Tuulikki Sillajõe and Tomaz Smrekar (Chapter 12); Leo Engberts (Chapter 13); Luisa Franconi and Daniela Ichim (Chapter 14); Ivan Thaulow, Johan-Kristian Tønder and Claus-Göran Hjelm (Chapter 15); Tuulikki Sillajõe (Chapter 16); and Melissa Gare and C. Chien (Chapter 17). The views expressed in the Chapters are those of the authors and do not necessarily reflect those of their respective organisations.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Key points

In 2011, the OECD Expert Group for International Collaboration on Microdata Access was formed to examine the challenges for cross-border collaboration with microdata. The challenges are many and in its meetings the Group came to some general conclusions:

- Change to improve cross-border collaboration with microdata will be incremental.

The challenge should be addressed in small, deliberate, and achievable steps.

- There are risks but they can be managed. The risks in international collaboration in microdata should be managed as other corporate risks are managed, deploying solutions that best fit the task.

It is possible to manage exposure to risks that are located beyond a country's national borders provided a suitable mode of access is selected and trusted delivery partners are identified.

- The challenges are different in every statistical system. The variations in current practices and culture between countries means that change will be internally-driven and bespoke to a self-assessment of local and national circumstances. The Group's recommendations are related to maturity indicators, in order that a path to enhanced cross-border collaboration can be mapped in a way that is directly relevant to each national statistical office.

A maturity model can be used to self-assess and to set targets for change to improve cross-border collaboration with microdata.

- Building trust in partners and users of microdata is essential. A trusted partner across a border can be a route to international collaboration with users of microdata in other countries.

Trust, objectively gained and evidence based, is a necessary condition for managing exposure to risks across borders.

These general conclusions give the Executive Summary its structure: a discussion on the treatment and the value of trust in collaboration with microdata; a summary of the use of different methods of international collaboration with microdata; a review of using maturity modelling to improve practices in microdata collaboration; and detailed recommendations which, if adopted, will enable a statistical office to improve its maturity relative to microdata collaboration.

- The first set of recommendations concerns "**Speaking the same language**". The Group believes a shared glossary of terms is a prerequisite for effective collaboration. The use of metadata standards enables flexibility and interoperability.

- The second set of recommendations addresses "**Trust**". Collaboration with microdata is enabled when partner organisations understand and respect each other's frameworks of risk management. Where that understanding and respect is achieved the partners are in a circle of trust and collaboration with microdata can dramatically improve. A shared understanding of user accreditation, penal and administrative sanctions, licensing, and the relationship with users is

needed to be inside the circle of trust.

- The third set of recommendations encompasses "**Information as an economic resource**". Different methods of access have different cost/benefit ratios. The Group makes a general recommendation that the costs (and benefits) of collaboration with microdata should be visible and justifiable.

- A final set of recommendations involves "**Making microdata collaboration as our business**". The Group makes recommendations relating to the Generic Statistical Business Process Model, and provides references examples where international scientific re-use of statistical microdata has been made business as usual.

Adopting the recommendations will move a statistical organisation's maturity of collaboration in microdata to the point where the practices are embedded and collaboration across borders become business as usual with a net economic benefit, without a disproportionate increase in the risk management burden.

Background

1. Between 2005 and 2007, the OECD conducted exploratory work to investigate the feasibility of making official microdata more accessible to policy makers and analysts. The OECD Conference "Assessing the Feasibility of Micro-Data Access" (Luxembourg, 26-27 October 2006) and subsequently the report "Study on the Feasibility of Micro-Data Access for the OECD" (STD/CSTAT(2007)3/ANN) were the two key visible outputs of this initiative. The final report investigated a wide variety of options including Public Use Files, Licensed Files, Remote Access Facilities, and also the scope for NSOs to construct new, disclosive indicators based on the underlying microdata. The conclusions of this initial effort revealed that for many countries it was too early to be able to provide the OECD with some of the access modes envisaged but there was strong encouragement for countries to continue to develop their dissemination capacities.

2. The work of the Expert Group, described in this document, has built on this earlier effort and has expanded the scope of the investigation to also address the more general issue of facilitating microdata access for international collaborations. Also, past discussions were partly centred on assessing the relative suitability and costs of different microdata access modes to identify best solutions (including the IT infrastructure) taking into account the legislative framework. The Expert Group has instead privileged an approach based on the coexistence of alternative solutions depending on users' needs and resource availability in National Statistical Offices (NSOs). This paradigm, illustrated in this Executive Summary, has the advantage of accommodating different countries' legislative and technical settings in the area of microdata access.

Trust for improving international collaboration with microdata access

3. Where there is to be collaboration in microdata, there is risk. Where there is risk, there must be trust in the international partners and end users who help manage the risk. But what is meant by trust? How is it to be recognised? How is it to be achieved? What value does trust have in risk management, in practice? How does an objective and evidence-based approach to trust help with the challenges of cross-border collaboration with microdata?

4. For the purposes of this report, 'trust' is a shared understanding of relevant standards and behaviours with respect to the risks and responsibilities in managing risk in microdata access. The higher the risk in the data, the stronger that shared understanding must be.

5. Conceptually, partners can be placed in concentric rings, arranged according to the degree of trust they are held in. Placing the data owner (i.e. the statistical office) at the centre, its most trusted partners are in the central ring. These are the international partners and end users that the data owner has good reason to believe can, and will, manage the data owner's risks in confidential microdata - even without its direct management control over those data, and even outside its direct legal jurisdiction.

6. In outer ring(s) are the partners and users that the data owner has reason to believe can, and will, respect the undertakings they must make to use its scientific use files and its research data access facilities.

7. Beyond the outer ring(s) are those who are not, or cannot be made subject to any meaningful management control of risks in microdata. For these, and many other types of users including those not interested in microdata for research purposes, there are Public Use Files or other types of files made available to all such as Open Datasets, indicators, and bespoke aggregates.

8. The concentric rings create the 'circles of trust'. Every statistical office will have circles of different constructs, and of different size, number, and membership, because each will have different risk attitudes, partnerships, user communities, modes of access, and data availability.

9. The biggest gains in cross-border collaboration in microdata access are when two or more statistical offices in different countries agree, objectively, that they are partners in each other's central circle of trust. Exchange of confidential microdata between them may be enabled. Each can then act as a national point of access for the microdata of the other partner. As this network builds, cross-border collaboration with microdata could expand. The Expert Group has learned from the European project "Data without Boundaries" that trials between partner countries are underway using this model. Interestingly, it is the inclusion of expert Social Science Data Archives as trusted partners that is enabling these trials. A partnership is also the basis of future exchange of confidential data in the European Statistical System – indeed, the word "partnership" is used to define the European Statistical System in the legal framework for European statistics (Statistical Law: Regulation (EC) No 223/2009 of the European Parliament and of the Council, of 11 March 2009).

10. For networks of trusted partners across borders to grow there needs to be a shared approach to building the circles of trust. The Group recommends a step by step approach:

First step: Decide what constitutes 'trust' in microdata collaboration and build a circle model

Statistical offices should as a first step design the parameters of each circle using those indicators of standards and behaviours thought important to collaboration in microdata access. Many OECD countries will have organisations within their borders that already have each other in their inner circles of trust, in particular where there is a decentralised statistical system and/or long history of partnership with the research community. These relationships may not be expressed according to a

model template, nor even explicitly, but nevertheless they can be examined in order to derive useful indicators for the circle of trust model. To that end, an example of effective cross-country collaboration is provided in Chapter 7.

It is probable that some statistical offices are using RMADS (Risk Management and Accreditation Document Set) as a way of managing information security risks in shared information. RMADS would be a very good starting point for building a broader trust model.

Second step: Match current and potential partners against these dimensions of trust.

A second step would be to consider partners (or potential partners) in microdata collaboration against these indicators, using what is known about them. It is important to accept that this is according to an *objective* assessment of what is *known* about the other organisations. Inevitably, this involves gathering information about the other organisation, which should be done rigorously and with as little burden as possible. Designing the model in step 1 first will minimise the burden of evidence-gathering.

Third step: Use the trust model to build new or stronger relationships

The important third step is to share the indicators of the circles of trust with prospective partners in collaboration with microdata. The information needed to place the requesting organisation or end user in the proper circle will often be easy to provide once it is clear what is sought and for what purpose. Once potential partners know what it takes to be inside a circle of trust, they can take measures to meet those indicators and "win the trust" of the statistical organisation holding the microdata to be accessed.

Fourth step: Explain your decisions about access using the trust model

This deals with the awkward situation where, for instance, a reputable international organisation or end user asks for access to microdata and is refused - and wants to know why. Where there is an absence of a detailed knowledge of the standards and behaviours of the requesting organisation, there is no objective basis for trust and no concrete reason to believe that risks can be properly managed. Explaining this enables the requesting organisation to supply the missing information.

Fifth step: Use the trust model to provide a route to positive collaborative relationships

Too often just a single reason is given for refusing access, and no route is offered to address the matter. Worse, it is sometimes only when the given reason is resolved that the next obstacle is explained, and this can be deeply frustrating and very time consuming for the organisation that would like access to the data.

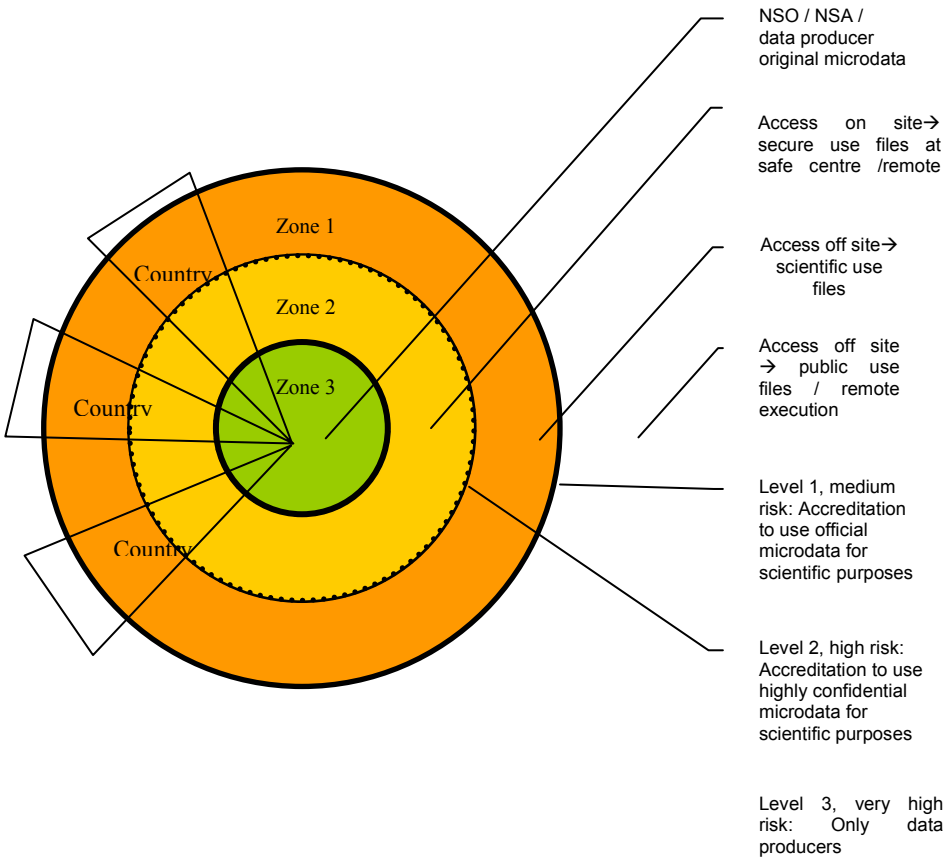
The trust-based modelling of international collaboration for microdata allows for all the relevant reasons to be given, promptly, for denying access; but more importantly it ensures the pathway to addressing those reasons can be provided at the same time. Instead of an unexplained "no", the applicant receives useful information that can be used to resolve the matters preventing access taking place. The objective is a more positive, goal-oriented conversation, with a potentially successful destination in sight.

Sixth step: Addressing 'Trust' can provide solutions to other barriers

Once a partner organisation in another country is fully trusted, solutions to international access to other parties present themselves. In effect, the trusted party provides a subsidiary function for the confidential microdata. The adoption of a systematic and evidence-based method to determine how

to solve new challenges in access to microdata in this way is itself an indicator of high levels of maturity in the microdata access.

Trust in itself cannot overcome express legislative barriers to transmission of confidential data to users in other countries. However, it may be possible to find lawful routes through to that user by deploying trusted partners in that country. In particular, where an inability of the data owner to prosecute a loss of confidentiality in the second country is the blocking issue, a trusted partner in the second country who is prepared and able to use its own legal framework may offer the solution. Thus, providing the data to the trusted NSO partner in the second country, who in turn provides a data access service to the user under their shared national law, might complete the necessary legislative cover.



Modalities for international collaboration in microdata access

11. Regional and international shared policy making needs the support of evidence drawn from comparative analysis and/or the combined data of the national parties to the collaboration. Since the 2007 UNECE report on *Managing Statistical Confidentiality & Microdata Access. Principles and Guidelines of Good Practices* it has become widely accepted that official microdata is a legitimate, important and necessary data source for analytical work provided guarantees of confidentiality can be met. However, it is currently the case that most international collaboration with microdata is achieved through the use of published, or publishable, aggregate statistics only. The advantages of using microdata are not being fully

realised in the world of cross-border microdata collaboration. However, the familiar modes of microdata access can be used in cross-border collaboration.

12. Practices in microdata access follow, with variations across countries, three basic modes. Each mode is built up in three components: the risk design of the data; the design of the access service; and the legal status of the relationship. Data owners typically use one or more of three basic modes for their access services:

1. Creation of *low-risk datasets (e.g. Public Use Microdata files - PUFs)* that can be distributed safely to non-accredited users who agree to licence and/or copyright conditions.

2. Creation of *medium-risk datasets (e.g. Scientific Use Files - SUFs*) provided through *remote access or remote execution systems* to *accredited users* who are subject to *administrative sanctions*.

3. Use of *high-risk datasets* (e.g. *Secure/Master Use Files*) directly on their own site or in laboratory conditions to *wholly trusted users* bound by *contract law and/or legislation*, as well as *administrative sanctions,* to protect the confidentiality of the data.

13. Each of these modes can be adapted to enable international collaboration with microdata. The Report does not recommend any one of these modalities above the others as the best approach to international collaboration in microdata. An objective approach to selecting the most suitable mode of access is more valuable. The development of a portfolio of solutions should be an objective outcome of a risk-based analysis and the resources available. A 'mature' statistical office will have a range of options available to it. That said, the analysis conducted by the Expert Group suggests that in the short to medium term it might be expected further development of Public Use Files, supported by a number of relevant innovations (e.g. new excellent pre-tabulation disclosure control techniques, the Creative Commons and Open Government licences providing useful models for PUF licensing, improved capacity in web-based data delivery enabling statistical offices to maintain much larger online catalogues of datasets). It is telling though that it is this mode, where trust is least important as an enabler, which is seeing the most growth. If we are to see similar developments in international exchange in *confidential* microdata, it is necessary to address the issue of trust in access provider partners and end users.

14. **The Group recommends that "Trust" should be a key determinant of the modality to be selected**, which together with the legal setting will determine the final choice of type of access. Thus:

- Where there is **no evidence base for trust** (due to users being unknown or otherwise beyond the management control of the provider of the microdata), non-disclosive PUFs or remote execution will be the preferred type of access, due to the risk management being in the statistical disclosure control design of the dataset, alone. It follows that the legal relationship will be limited to copyright or licensed conditions of use, only.
- Where there is **some, but limited, evidence for trust** derived from, for example, a case-by-case user accreditation procedure, then remote access may be the preferred type of access. Remote access enables risk management control over the data in the service as well as management control over the user of the service. Users will be placed in an agreement with some legal force, but emphasising significant administrative sanctions.
- Where there is **strong and evidence-based relationship of trust** managed at a permanent and institutional level, collaboration through a direct supply of, or direct access to, confidential microdata may be preferred. Users will be under an institutional agreement often backed up by specific legislation or international contract law, including both administrative and penal sanctions. A highly trusted user relationship could of course allow for the deployment of remote access and PUFs solutions as these are lower risk.

15. These descriptions can be visualised in 'circles of trust'.

TRUST	*Legal relationship*	*Type of dataset*	*Type of access*
Inner circle of trust	Internationally binding contractual terms and/or specific legislative provisions.	High risk Secure Use Files, containing unique identifiers if necessary.	Direct supply to the site of the user, or access under data laboratory conditions.
Outer circle of trust	User accreditation and agreements containing directly applicable administrative sanctions for breaches	Medium risk Licensed Files	Remote access.
No evidence	Copyright and licence conditions only	Low risk anonymised Public Use Files (Open (micro)Data)	Remote execution, distribution of datasets on CD, electronic transfer, or internet download.

16. National preferences and policies may cause some OECD members to prefer one mode of collaboration over another, but each mode has its own balance of cost/benefit and of risk/utility. The overarching goal of this report is to provide assistance to collaborators in microdata so that the most suitable mode of exchange can be quickly and easily adopted.

17. It is important to be able to adopt the most suitable mode of collaboration in data. Trying to meet demanding international analytical needs with published statistics can pressurise statistics offices to include more risk in its published output than it would like. The usual practice of international collaboration through published statistical outputs can itself be put under unfair pressure when the demand for ever finer breakdowns of data categories, geography, or time are sought by users. The statistical producer may have to work ever closer to the margins of safety in statistical disclosure control settings in order to keep using the publishing dissemination channel, all the time knowing that the threats to confidentiality when data are in the public domain are increasing. It may be safer to meet those demanding analytical needs with an appropriate mode of microdata access. When any mode of data collaboration is put under undue pressure, a failure of service, or analytical error, or excessive costs, or even a breach of confidentiality, can occur. Properly managed microdata collaboration may meet user needs with less risk than attempting to meet those needs with highly detailed published statistical output.

18. A statistical office that achieves a high level of maturity in cross-border collaboration in microdata, by providing a mix of microdata access products, is in a position to adopt the mode that presents the lowest cost, lowest risk, highest significance balance possible in the circumstances.

19. The owner of the data could benefit from involving other organisations to assist them in collaboration with microdata where the other organisation has particular expertise. In some countries social science data archives have highly-developed expertise in same key areas of risk management such as metadata, researcher accreditation, the preparation of PUFs, and more recently in remote access services too. Social science data archives, in countries where they exist, can play a significant role, and many will already have high levels of maturity in some of the recommendations in this report. In particular, many data archives have considerable expertise in resource discovery, researcher accreditation, and metadata standards. Where a statistical office decides to adopt national partners in microdata access, then many of the recommendations in this report could be carried out by national social science data archives or other service providers, and producers of official data should discuss these recommendations with them.

A maturity model for international collaboration in microdata access

20. The Expert Group derived and deployed a maturity model to give focus to its recommendations. A maturity model is a matrix of indicators structured by levels of practices, attitudes or behaviours in each of a number of dimensions of a topic. Maturity models are used to initiate, set targets, and to monitor strategic change. An organisation can use a maturity model to assess itself, and compare itself with others using the same benchmarks. It can be used to set targets for maturity and to design its policies and processes to meet those targets. The linearity of the model is meant to simplify its presentation, but it does not imply that one organisation has necessarily to go through one level to achieve the next one. Its purpose is to highlight desirable characteristic in the development of transborder access to microdata, so organization can know in what direction they have to stir their actions to improve it.

21. The Group is not aware of a suitable pre-existing maturity model for collaboration in microdata. When thinking about its recommendations, a maturity model began to emerge. The Group felt its recommendations should be linked to indicators of low, heroic, and embedded levels of maturity in those areas thought to be barriers to improving international collaboration in microdata access. Thus each theme of this report contributes to the construction of a maturity model for international collaboration in microdata access. The recommendations are directly linked to moving maturity from low, through heroic, to an embedded status.

22. Typically, low levels of maturity indicate that an organisation is aware of the matter at hand, but has done little as yet to prompt practices, processes, policies, and behaviours that would remove barriers to implementation. Strategies, policies, and practices are not coherent across the organisation, and management control of any outcomes is difficult, or even impossible. Individual senior officers will be risk-averse in this context. Change will not be deliberate or directed to any particular outcome.

23. Heroic levels of maturity are found in organisations where good practice is found in some parts of the business, typically those where a passionate and committed individual has invested considerable time and effort into “blazing a trail”. The term ‘Heroic’ is used for this reason - organisations in this level of maturity are reliant on the actions and expertise of one or two passionate individuals. They are valuable people, but also represent a single point of failure. If they leave the organisation, the processes and knowledge that support the activity are lost.

24. Embedded maturity indicates that an organisation has adopted all the attributes necessary to make international collaboration in microdata access its business as usual. Such an organisation will still have its heroes, as there will always be innovations and improvements to explore and in time adopt, but the main business of microdata collaboration is owned and delivered at all levels of the organisation, and is a feature of all relevant business areas within it. The knowledge and information necessary to run this part of the business is owned and supported by auditable corporate procedures.

25. Embedded maturity enables an organisation to be confident in its ability to manage its business as usual efficiently and effectively. It also allows for creativity and innovation in response to new demands and new challenges. Most importantly, it allows an organisation to use the most suitable modality for a new collaboration in microdata. In this way, the risks inherent in overburdening one particular modality are minimised.

26. In the following section, relevant indicators of maturity are presented for the different areas addressed by the Expert Group, together with related recommendations that will enable those levels to be achieved.

Recommendations

27. The members of the Expert Group each led on a challenge to cross-border collaboration in microdata access. The following is a summary of their work.

Part I. Speaking the same language

28. The first challenge for the Expert Group was to use a shared vocabulary in its own discussions. In its first meeting it became clear that terms such as "anonymised" had different meanings and uses between the members of the group. It was not difficult to conclude that this must be a wider issue.

29. A common glossary is essential if there is to be trust between organisations. Misunderstandings over terminology may result in the selection of the wrong mode of access or other incorrect risk management treatment for the microdata.

30. The Group adopted a collection of relevant terms from as wide a collection of official sources as possible, and then selected its preferred terms and definitions. These are the terms used in this report and they are highlighted in Annex I.1.

31. At a more technical level a similar issue arises. Statistical metadata are essential to ensure the best use of the data, to avoid analytical errors, and to enable resource discovery. Practices in metadata vary widely, and the Group makes recommendations relating to the availability, accessibility, and format of metadata to enhance cross-border collaboration in microdata access.

	Maturity indicators		
	Low level of maturity	**Heroic level of maturity**	**Embedded maturity**
Common glossary	Teams of staff use a common vocabulary locally in their working conversation, which is sometimes shared informally with others in international forums and with secondary users of the data. There is no written and maintained glossary.	Some business areas have formally adopted a working glossary of terms. Local glossaries use internationally recognised standard terms and definitions. Local glossaries are maintained in business area information systems, only. Local glossaries may be available on request to secondary users of the data.	A corporate glossary of terms is agreed and changes to it are made in a formal and agreed manner. The corporate glossary is benchmarked against internationally recommended definitions. All staff use the terms in the corporate glossary consistently, in both working conversations and in formal documentation. The glossary is available and accessible to all users of microdata.
	Recommendations		
	To move from low to heroic:	***To move from heroic to embedded:***	
	All statistical business areas to adopt a glossary of terms for internal use.	Adopt a Glossary based on internationally-harmonised definitions. Ensure all documentation uses terms according to the glossary.	

		Make the glossary available and accessible to all staff and all users of microdata.	
Metadata	**Maturity indicators**		
	Low level of maturity	**Heroic level of maturity**	**Embedded maturity**
	Metadata is limited to the immediate local requirements of statistical production.	Where there are known secondary users of micro data an export of the producer's metadata is provided. Metadata follows relevant domain standards and guides; the Metadata Common Vocabulary is used.	Metadata is maintained according to a recognised international standard such as SDMX, or DDI. Metadata files are machine readable and accessible to all users, including a version in the English language Metadata files are available through international resource discovery portals for pre-examination before data access requests.
	Recommendations		
	To move from low to heroic:		***To move from heroic to embedded:***
	Develop statistical metadata for all available microdata sets. Introduce automatic process to produce and disseminate statistical metadata. Adopt the Metadata Common Vocabulary, independently on the standard used to compile metadata.		Structure microdata information according to international metadata standards. Produce statistical metadata both in the national language and English. Publish metadata, or make available through a resource discovery service.

Part II. Circles of Trust

32. Compliance with legislation is the first duty of the holder of a public office. Directors General for official statistics can have different risk appetites, but they cannot knowingly or recklessly breach the laws they are appointed to uphold.

33. Where national legislation expressly prohibits any secondary use of confidential data, there is no potential for cross-border collaboration with those data in that form. Legislation of that sort is rare. It may be silent on the subject or it may provide specific provision for legitimate scientific/statistical research uses. Where there are provisions, they are different in every jurisdiction, and often different within jurisdictions for different sources of data. This Group cannot make specific recommendations when the legal framework is so complex and specific to national settings.

34. To the best of the knowledge of the Expert Group, there is no legal mechanism whereby a criminal breach of statistical confidentiality can be prosecuted directly across an international boundary. This is one feature of the legal frameworks for statistics that does seem to be shared across the world. This means that the statistical office of Country A is not able directly to bring a prosecution of an offence against their confidential data where that offence takes place in Country B. Quite simply, offences in

Country B must be prosecuted by the agencies of Country B. Breach of statistical confidentiality is very unlikely ever to qualify for an extradition. If confidential data from Country A is ever to be used across the border in Country B, the *availability* of sanctions of *equivalent effect* must be known, as must the preparedness of the authorities in Country B to apply them.

35. The Group also concluded that the importance of *penal* (criminal) sanctions as opposed to administrative sanctions is sometimes over emphasised. There is considerable difficulty in making a criminal sanction a meaningful deterrent in practice. Before a successful prosecution there must be an assessment of public interest in bringing the prosecution; there is the burden of proof that is beyond reasonable doubt; the NSO would usually have the status of a witness, and would not be able to gather evidence but would merely provide it as directed by the prosecutor; news coverage may be harmful; and there is the possibility that the breach is the result of an improper decision of the NSO itself. Even in the event of a successful prosecution, the sanction may be a relatively small fine or suspended sentence. Taking this into account, an administrative sanction may well be a more effective deterrent in practice. In contrast to penal sanctions, administrative sanctions can be applied by the data provider; they can be relevant to the cross-border user directly; the burden of proof is the balance of probabilities; and suspension from future access to data and loss of employment status may be a more meaningful deterrent than a court fine. In conclusion, an effective administrative sanction across a border may well be equivalent in its deterrent (risk management) effect as the penal sanction in the home jurisdiction.

36. Using the circle of trust model allows data owners to map the trust status of cross-border users against the available legal framework for managing risks in cross-border data access.

- A cross-border user with a low (or not known) trust status cannot receive or even view confidential microdata. No legislation known to the Expert Group allows a data owner to provide access to confidential data to persons who are unknown or who cannot be subjected to management controls. Those outside the circles of trust must use Public Use Files or a remote job submission service. The laws of copyright and licensing may be the only element of law that applies.
- A cross-border user who is a member of a circle of trust may represent a risk that can be managed when he/she is subject to the controls imposed by contract law, legislation, and/or equivalent administrative sanction in the location of their access. It can be seen that there are two obligations on the data owner: i) the user has to be correctly designated according to the parameters of trust, and ii) the contract law, legislation, and/or administrative sanction applicable to the user must be known to be *available and known to be equivalent* to the sanctions framework of the data owner.
- These two obligations are clearly linked. It is an important parameter of trust that the cross-border user is known to be subject to an available and equivalent framework of sanctions.

37. The Group also considered the importance of harmonising some of the basic documentation for the accreditation of users of microdata. To that end, recommendations for a standardised application process are provided.

<table>
<tr><td rowspan="6">Circles of Trust</td><td colspan="3">Maturity indicators</td></tr>
<tr><td>Low level of maturity</td><td>Heroic level of maturity</td><td>Embedded maturity</td></tr>
<tr><td>Decisions to provide access to microdata to a partner organisation or a data user are ad-hoc and subjective.

Standards required for successful accreditation are not available to potential partners and users of data.</td><td>Some business areas of an organisation maintain a list of precedent-setting decisions with respect to certain partners and data users who have completed an accreditation procedure.

The parameters of accreditation and the method of approval are available to potential data users on request.</td><td>Partner organisations and microdata users have a trust status that is objectively based on a corporately-agreed set of parameters.

Decisions of microdata access mode are made with direct reference to the corporately-agreed trust status of the applicant.

Potential data users are informed of the parameters for trusted status and the path to achieving this status is transparent.</td></tr>
<tr><td colspan="3">Recommendations</td></tr>
<tr><td colspan="2">To move from low to heroic:</td><td>To move from heroic to embedded:</td></tr>
<tr><td colspan="2">Decide what constitutes ‘trust’ in a microdata collaboration.

Set indicators of trust.

Match current and potential partners against the established features of trust.</td><td>Use indicators of the circles of trust to build up evidence-based relationships with cross-border users.

Use the trust model to provide a route to positive collaborative relationships.</td></tr>
<tr><td rowspan="6">Sanctions for confidentiality breaches</td><td colspan="3">Maturity indicators</td></tr>
<tr><td>Low level of maturity</td><td>Heroic level of maturity</td><td>Embedded maturity</td></tr>
<tr><td>The relevance and availability of administrative and penal sanctions in other jurisdictions is unknown.

No recognition is awarded to the sanctions in other jurisdictions.</td><td>Assessments are made of the relevance of sanctions in peer-group organisations.

Some collaboration with microdata is enabled through a case-by-case assessments and an exceptions procedure approval process.</td><td>Partner organisations and data users with equivalence in administrative and penal sanctions are listed.

A map of equivalent sanctions is used to make cross-border collaboration decisions routine.</td></tr>
<tr><td colspan="3">Recommendations</td></tr>
<tr><td colspan="2">To move from low to heroic:</td><td>To move from heroic to embedded:</td></tr>
<tr><td colspan="2">Establish clear terms of access and sanctions for breach of contract between the parties.</td><td>Develop “Mutual Statistical Assistance Agreements” with partner statistical offices in other countries</td></tr>
<tr><td rowspan="3">Standardised application process</td><td colspan="3">Maturity indicators</td></tr>
<tr><td>Low level of maturity</td><td>Heroic level of maturity</td><td>Embedded maturity</td></tr>
<tr><td>The documentation for applying</td><td>Some business areas are</td><td>A standard approach to</td></tr>
</table>

	for access to microdata is created on an ad-hoc basis. No information about procedure or timeliness of decision-making is available. No glossary and no metadata is available to assist applicants in the application process,	using a standardised application form and process designed with local data users in mind Applications from other countries can be accepted. Glossaries and metadata for applications are available on request. Some indication of timing and procedure service levels is provided on request.	applications and the application process is used which is consistent with practice in other countries There is a documented process for handling requests from data users in other countries. A service level agreement of timeliness and good practice is visible to applicants. Glossaries and metadata including for the application process are available through resource discovery services and the web. Reasons for approval or rejection are transparent and clearly related to the conditions expressed in the application procedure.
	Recommendations		
	To move from low to heroic: Structure the process of access to microdata into standardised steps, applicable to any mode of access. The proposed standard sub-processes are: publication of microdata; (submission) of application; provision of service; outputs; and evaluation.		***To move from heroic to embedded:*** Adopt an internationally harmonised application form and process for microdata access. Issue forms and information in the national language and a common language (English)

Part III. Microdata as an economic resource

38. The micro-datasets of a statistical office are among its most valuable assets. They are recognised as core input to the statistical output production, and as valuable assets in an NSO's information security management and disaster recovery planning. Many statistical offices will have precise estimates of the costs of collection and secure maintenance of these assets.

39. As public authorities exercising functions of national importance and serving a public good, it is incumbent upon NSOs to consider the wider economic case for exploitation of its microdata assets, both in controlled-access and in Public Use domains. The argument needs to be made for a funding line in an NSO's budget for exploitation of microdata assets. A statistical office that is mature in microdata access will be able to quantify its input costs of providing access to microdata, but importantly it will also be able to measure in some relevant way the uses to which its data are put.

40. Statistical offices should consider carefully precisely how and where this investment should be made. Depending on national circumstances, the investment might be focussed upon the production of safe Public Use Files, or in the establishment of a remote access facility, or in the construction of data laboratories on the sites of key users, or in a national social science infrastructure based on data archives. Each will have a different return on investment.

41. Yet, despite the economic benefits, risks, and costs involved in microdata access, the Expert Group found little evidence of choices of modes of access being determined on a benefit/costs ratio consideration. This is to some extent unsurprising, as it would imply a maturity of the function of providing microdata access that virtually all NSOs still need to gain.

42. The Expert Group considers in particular that there is more work to be done in estimating the benefits of providing third party access to NSO microdata. It is difficult to do, but some economic modelling is possible and where it has been done, the case is compelling.[2] It almost goes without saying that a breach of confidentiality would damage the benefit/cost ratio significantly, as well as have other wider negative effects. The Expert Group would like to explore this issue more deeply in its future work. An economic model could be produced that will enable ratios to be calculated on a comparable basis across OECD members.

43. The Group also discussed of the value of increasing the production of non-confidential microdata files to be made easily available on the web, e.g. PUFs, as a way to foster general international microdata access through a simplified channel. PUFs represent one of the channels of microdata access that an NSO can implement, the perspective in this report being the availability of a range of solutions that allow for access in an international setting. The production of PUFs should aim at maintaining the high standards of official statistics in terms of significance of the released files; this can be achieved by designing products that are able to preserve crucial analytical characteristics of the data accompanied by metadata and methodological documentation and by ensuring transparency.

44. The Group also considered the issue of commercial use of microdata. The privilege of access to confidential microdata is rightly restricted to statistical research purposes only. The question is whether a commercial organisation can satisfy that condition. Many statistical offices rule out commercial organisations, whether or not there is a commercial value to the product of the research, and whether or not the purpose is statistical research only.

45. There is a legitimate concern to ensure the public good is served when the privilege of access to confidential microdata is granted to a user. It is less clear to the Group whether the work of a privately funded organisation is necessarily incompatible with that principle.

46. This issue is becoming more pertinent as public administrations rely ever more upon external sources of analytical expertise. Think-tanks, charities, lobby groups, public-private partnerships, commercially sponsored research in academia, collaborations between government departments and commercial consultancies are increasingly involved in research - and these initiatives put pressure on those organisations that hold to a "non-commercial only" condition for access.

47. Opinions within the Group varied on this issue, but common ground was found when the status of the *product* of the access to confidential microdata was considered. The Group recommends that all users of confidential microdata should be bound, as a condition of access, to make their research product publicly available. The gold standard is that the research results derived from privileged access to confidential microdata should be Open, in that they are available in a machine-readable format, free of charge, and free to be used and reused without restriction other than proper attribution. A lower but still acceptable standard would be that results are publicly available through payment of a subscription fee, set at a cost-recovery rate only, or through a journal or other accessible route. The Group agreed that the privilege of access to confidential microdata obtained from private individuals or businesses should not be

[2] For example, the UK's Economic and Social Research Council conducted an economic impact analysis of the UK Data Service, which is a core information infrastructure service providing researchers with access to Public Use Files and Secure Use Files from UK's statistical offices and other sources. http://www.esrc.ac.uk/_images/ESDS_Economic_Impact_Evaluation_tcm8-22229.pdf.

granted where the results are withheld for exclusive private gain - whether that gain is through personal professional advancement, or through exclusive exploitation of its commercial value. This is in contrast to the use of Open (micro)Data and Public Use Files, where the commercial exploitation of these non-disclosive statistical products is permitted or even encouraged.

48. Once it is accepted that the *results* of statistical research and analysis of microdata should be Open, then the non-exclusive economic benefits of the product can be realised without threat to statistical principles. The Open products of research use of microdata should have the economic benefit as Open official statistics, including direct commercial economic benefit and indirect benefit through better public administration.

<table>
<tr><td rowspan="6">Portfolio of access solutions -PUFs</td><td colspan="3">Maturity indicators</td></tr>
<tr><td>Low level of maturity</td><td>Heroic level of maturity</td><td>Embedded maturity</td></tr>
<tr><td>Complete absence of PUFs or other modes of microdata access easily available for cross border users in the dissemination plan of an NSO.</td><td>Sporadic development and release of PUFs without any specific treatment at the design stage for quality aspects.

NOTE: A distinct case is that of countries or organisations that make a serious appraisement of different designs of PUFs but, taking into account the problem of statistical confidentiality in small populations, conclude that PUFs are not a realistic solution for the time being.</td><td>PUFs are considered as particularly important microdata dissemination products and developed coherently with other products of the portfolio of modes of access to microdata.

Statistical Disclosure Control (SDC) methods that satisfy the overarching principle of preserving important statistical properties of the data as well as benchmarking statistics are implemented

Users are made aware through resource discovery services and the web of the content and the properties of the released microdata file with respect to the original one and of all the possibilities in the portfolio of modes of access offered by the NSO.</td></tr>
<tr><td colspan="3">Recommendations</td></tr>
<tr><td colspan="2">To move from low to heroic:</td><td>To move from heroic to embedded:</td></tr>
<tr><td colspan="2">Adopt, at least in a case by case manner, the SDC paradigm (i.e. disclosure risk assessment, implementation of disclosure control methods, evaluation of information loss and significance).

Conduct and present to users quality checks carried out on the released microdata.

Make license terms and license agreements visible on the NSO’s website.

Use license agreements on a website only when the end user is required to provide positive proof of agreement (e.g. such as by clicking on a statement attesting to that agreement).</td><td>Implement the systemic use of the SDC paradigm guided by data utility where care is taken in adopting methods that preserve crucial analytical characteristics of the data.

Define a clear governance for microdata access

Propose a portfolio of modes of access available to national and international users.

Implement a user-centric dissemination system with fully implemented resource discovery features coupled with training programs to foster the development of culture and knowledge around microdata.

Take the leadership in the coordination of PUFs and open microdata releases inside national statistical systems, and raise awareness on the interaction between open microdata available from</td></tr>
</table>

		other sources, big data and the released microdata. Be proactive in increasing dialogue with new actors of Open microData dissemination. At a regional and ultimately at a global level, seek harmonisation in the design standards and licence/copyright controls over PUFs, enabling international exchange.	
Open data and the challenges of transparency	**Maturity indicators**		
	Low level of maturity	**Heroic level of maturity**	**Embedded maturity**
	No analysis of the potential value of Open (micro)Data.	Internal challenges seek to remove excessive management controls over safe data. Some Open (micro)Data products are created in response to legal or policy challenges placed upon the organisation.	Internal and external challenges to management control over micro-data are accepted and responded to. A portfolio of potential Open micro)Data products is maintained and systematically worked through. All microdata products that are known to be safe in the public domain are made available through a resource discovery portal.
	Recommendations		
	To move from low to heroic:		***To move from heroic to embedded:***
	A template design for what characterises Open (micro)Data should be designed. Seek advice from the national data protection supervisor.		For each statistical source an Open (micro)Data product is considered, and produced where possible.

Part IV. Making microdata collaboration our business

49. The final group of recommendations relate to making cross-border microdata collaboration part of our business as usual.

50. The Group analysed the General Statistics Business Process Model (GSBPM) descriptions and have made recommendations to the responsible UNECE body to revise some of the descriptions in the model. The general purposes of the creation and use of the GSBPM – the harmonisation of methods and practices to enable the more efficient adoption of tools and technology serving a more coherent market – apply strongly to the advancement of cross-border collaboration in microdata and the broader recommendations of this report.

51. As the production of statistics moves increasingly to more use of administrative sources, it is important that this business change is reflected in the provision of access to this new input data. Many statistical offices are limiting research access to those microdata derived from survey sources, only. The Group considers it important to move the information base for microdata access files at the same pace as for statistical production when an office increases its use of administrative data. This is not straightforward, as the design of the data, the consent of providers, and the legal framework may all be more complex.

Experts in administrative data, in particular the Nordic countries, have given their advice to the Group and the recommendations are included below and in the report.

52. A "heroic" microdata access business will not have transparent measures of its costs, nor its benefits. They will be hidden within budgets for the parent activity, such as the survey from which the data are derived. Costs are unlikely to be recovered, whether directly from the user or indirectly from a funding council or other grant. However, a mature microdata access business will be able to itemise its costs to enable transparent cost recovery either from within the budget of the parent organisation or from the beneficiaries of the service. An example of costs in the process flow of a remote access mode of service is included in the report.

53. Part of making "microdata collaboration our business" is to develop mature relationships with those who are most able to reduce statistical offices' costs and risks when providing cross-border access. In particular, a mature organisation will be prepared and able to collaborate with others who, in a foreign country, have greater capacity and/or capability in some or all of the process flow of microdata access. In some OECD countries, the national Social Science Data Archive may have greater capacity and/or capability in many of the GSBPM descriptions with respect to microdata. Using maturity model indicators:

- a low maturity organisation will be unaware of the existence, or non-existence, of the expertise in other organisations;

- a heroic maturity organisation will know something about relevant potential national partners in access to microdata. Where potential partners exist, the organisation will have some relationship with them, perhaps extending for example to cooperation with metadata, and inclusion of the organisation's data sources and services in the partner's resource discovery catalogue;

- an organisation with embedded maturity will collaborate with national partners on those descriptions of the GSBPM relevant to microdata access that are better delivered by that partner. A mature organisation that discovers no relevant national partner will work with candidate partners to ensure there is no single point of failure at the national level for access to microdata.

	Maturity indicators		
	Low level of maturity	**Heroic level of maturity**	**Embedded maturity**
GSBPM	NSO will be providing access to microdata by sending approved researchers Scientific Use Files. This activity, while potentially desirable to some researchers, is only viable in this early stage when demand is small and NSOs can accept the risk of a few files that are no longer held on the secure NSO servers.	The NSO develops the legal framework for expanded access to microdata for research purposes and commits to at least one of the other access types (remote execution, remote access, secure centres, or fully masked PUFs).	NSO provides multiple modes of access, including options that will allow for cross-border access to the NSO's microdata holdings.
	Recommendations		
	To move from low to heroic:		***To move from heroic to embedded:***
	Use the (revised) GSBPM descriptions to ensure access to microdata is planned from the beginning stages of the survey life cycle, and all costing is included in the planning		Work toward an embedded access model where multiple access modes are available to meet the needs of different users, and where international access is made

	stages. Allow researchers access to microdata by one or more of the access types as a first step.	possible through at least one of the available access types	
Releasing the value of administrative sources	**Maturity indicators**		
	Low level of maturity	**Heroic level of maturity**	**Embedded maturity**
	NSO only responsible for its own data.	NSO integrates administrative data in the statistical production process but do not provide external access to administrative data.	NSO is entitled and equipped to provide researchers with access to the administrative data held.
	Recommendations		
	To move from low to heroic:	***To move from heroic to embedded:***	
	Integrate administrative data as input for production of statistics. Design or modify the national Statistics Act so that it provides the NSO the permission to give access to administrative microdata for research purposes.	Implement legislation on data protection, in a way that reassures the public on the benefits of using administrative data in statistics and research. Provide access to comprehensive documentation on administrative microdata to researchers.	
Recovering the costs	**Maturity indicators**		
	Low level of maturity	**Heroic level of maturity**	**Embedded maturity**
	No formal arrangements to apportion or recovery microdata access costs when access occurs internationally.	Bi-lateral or one-off limited agreements made to assign costs and revenue to parties to international access to microdata.	An agreed model for sharing costs between international partners in microdata access.
	Recommendations		
	To move from low to heroic:	***To move from heroic to embedded:***	
	Develop a cost model for current access services enabling attribution of costs to the relevant elements of the data access service(s).	Make a fully costed business plan to integrate microdata access services in the wider business planning of the statistics office. Share costed business plans with the research community, in particular research funding councils, and jointly prepare a sustainable financial plan for future access services.	

PART I. SPEAKING THE SAME LANGUAGE

CHAPTER 1. GLOSSARY OF TERMS

1. First challenge in international collaboration is using the same words to mean the same thing. Talking at cross-purposes is not just inconvenient. It can affect our confidence in each other, and at worst it can result in error. Since the first meeting of the Expert Group in June 2012 it appeared evident that there was no common understanding of terms, not even of basic concepts and terms that are usually utilised in the context of microdata. An illustrative example is the term "remote access", which experts from different countries and statistical agencies used, at the meeting, with diverse meanings. Interestingly, although an international definition of the term "microdata" exists, some experts interpreted it as a narrower or, on the opposite, broader concept. In light of this, and noticing that no such standardised glossary of terminology on microdata currently exists at the international level, the Group considered the creation of a glossary of terms a necessary precondition for its own work. It became apparent that others would find the glossary just as useful.

2. The Group recommends to publish the *Glossary*, and to open it to use and reuse for all. The *Glossary* is presented in Annex I.A1.

Presentation of the selected terms

3. The *Glossary* compiled by the Expert Group covers all the terms that the members identified as part of a key vocabulary on microdata. Specifically, the criteria to retain "terms" were the following:

- Terms used in official methodological and terminology documents of National Statistical Offices (NSOs) and international organisations;
- Terms used in the meeting of the OECD Expert Group;
- Additional basic terms that seem to pertain to the area of microdata.

4. The terms are presented according to the following criteria:

- By alphabetical order.
- With definition, source(s) and hyperlink(s).
- Some terms with no definition are included, if they appear to be part of a commonly used vocabulary in the field of microdata.

5. The text contains cross-references when two different terms refer to the same concept.

6. Only the main methodological terms related to Statistical Disclosure Control (SDC) are included.

REFERENCES

For the preparation of the *Glossary*, only documents from NSOs and international organisations' sources were consulted. In particular, the terms included in this draft collection were drawn from the following main references:

A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf

EC Regulation No 223/2009 of the European Parliament and of the Council, http://eurex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:01:EN:HTML

EC Regulation No 831/2002 on access to confidential data for scientific purposes, http://eur-lex.europa.eu/LexUriServ/site/en/oj/2002/l_133/l_13320020518en00070009.pdf

EU regulation 557/2013 on access to confidential data for scientific purposes and repealing regulation (EC) No 831/2002, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:164:0016:0019:EN:PDF

International Household Survey Network - IHSN (2010), Dissemination of Microdata Files - Principles, Procedures and Practices, Dupriez O. and E. Boyko, IHSN Working Paper No. 005, http://www.surveynetwork.org/home/sites/default/files/resources/IHSN-WP005.pdf

OECD (2002) Frascati Manual: Proposed Standard Practice for Surveys on Research and Experimental Development, 6th edition

OECD (2007), Glossary of Statistical Terms, http://stats.oecd.org/glossary/index.htm

OECD (2007), OECD Principles and Guidelines for Access to Research Data from Public Funding, http://www.oecd.org/science/scienceandtechnologypolicy/38500813.pdf

UNECE (2009), Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes, http://www.unece.org/fileadmin/DAMstats/publications/Confidentiality_aspects_data_integration.pdf

UNECE (2007), Managing Statistical Confidentiality & Microdata Access – Principles and Guidelines of Good Practice, http://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

UNECE (2000), Conference of European Statisticians Statistical Standards and Studies – No. 53, "Terminology on Statistical Metadata", http://www.unece.org/stats/publications/53metadaterminology.pdf

UNECE and UNSC (1995), Guidelines for the Modelling of Statistical Data and Metadata, Conference of European Statisticians Methodological Material, http://www.unece.org/fileadmin/DAM/stats/ publications/metadatamodeling.pdf

To be noted, the original sources of the definitions of the terms drawn from the *OECD Glossary of Statistical Terms* include:

Eurostat (1996), "Manual on disclosure control methods", Office for Official Publications of the European Communities, Luxembourg, http://ec.europa.eu/eurostat/ramon/statmanuals/files/manual_on_disclosure_control_methods_1996.pdf

IMF (2000), Code of Good Practices on Transparency in Monetary and Financial Policies, Part 1- Introduction; approved by the IMF Executive Board on July 24, http://www.imf.org/ external/np/mae/mft/sup/part1.htm#appendix_III

IMF, "Guide to the Data Dissemination Standards, Module 1: The Special Data Dissemination Standard", Washington, May 1996, http://dsbb.imf.org/sddsindex.htm

ISO/IEC FDIS 11179-1, "Information technology - Metadata registries - Part 1: Framework", March, http://metadata-standards.org/11179/

OECD, IMF, ILO - Interstate Statistical Committee of the Commonwealth of Independent States (2002), Measuring the Non-Observed Economy: A Handbook, Annex 2, Glossary, http://www.oecd.org/std/nationalaccounts/1963116.pdf

Statistics Canada (1998), Statistics Canada Quality Guidelines, 3rd edition, October, http://www.statcan.ca/english/freepub/12-539-XIE/12-539-XIE.pdf

Statistics Canada, "Statistics Canada Quality Guidelines", Fifth Edition, October 2009, http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.pdf

Statistical Data and Metadata Exchange (SDMX) – BIS, ECB, Eurostat, IBRD, IMF, OECD and UNSD – Metadata Common Vocabulary, http://www.sdmx.org/

Statistics New Zealand, "Classifications and Standards", http://www.stats.govt.nz/surveys_and_methods/methods/classifications-and-standards.aspx

UNECE (1995), Guidelines for the Modelling of Statistical Data and Metadata, Conference of European Statisticians, Methodological material, United Nations, Geneva, http://www.unece.org/ stats/publications/metadatamodeling.pdf

UNESCO, OECD, Eurostat (2001), Data Collection on Education Systems: Definitions, Explanations and Instructions, http://stats.oecd.org/glossary/detail.asp?ID=10

United States Bureau of the Census, Software and Standards Management Branch, Systems Support Division, Survey Design and Statistical Methodology Metadata, Washington D.C., August 1998, Section 3.3.17, http://www.census.gov/srd/www/metadata/metada18.pdf

CHAPTER 2. METADATA FOR INTERNATIONAL MICRODATA ACCESS

Introduction

7. Producers of official statistics have been sharing and exchanging statistical information among each other and with users for decades. The exchange of aggregated statistical information in fixed format (e.g. pre-defined tables) or in *ad-hoc* manner proved to be adequate to meet different stakeholder needs. In recent years, however, the need for more detailed statistical information has been constantly increasing. There are today several initiatives and ongoing developments at the international level to foster the exchange of detailed statistical information not in the form of aggregated information (tables) but as detailed unit-level databases (microdata). National Statistical Organisations (NSOs), the international statistical systems, Data Archives and the researcher community are all concerned by the question of microdata access.

8. In order to be shared, statistical information needs to be accompanied by descriptions that provide further details about the data: these are the "metadata". This also applies to microdata; indeed, statistical organisations with experience and history of microdata sharing know very well that microdata sets should be accompanied by appropriate metadata descriptions to facilitate their use.

9. This need is also recognized by the international statistical systems, in particular the *United Nations Principles of Official Statistics*. The European Statistics Code of Practice, Principle 15 "Accessibility and Clarity" states that "statistics and the corresponding metadata are presented [and archived] in a form that facilitates proper interpretation and meaningful comparisons".

10. Statistical metadata are "data about statistical data, and comprise data and other documentation that describe objects in a formalised way" (SDMX Metadata Common Vocabulary). In this sense, statistical metadata usually describe information relating directly to the given dataset and also other relevant information related to the collection, processing and the dissemination of statistical information. Metadata are crucial for data exchange, especially in the international context where the chosen dissemination formats and language are key factors to understand the datasets to the necessary detail.

11. Even though the need to share good quality statistical metadata together with microdata sets is generally acknowledged, there are still many barriers in this area that make microdata exchange at the international level difficult. The Expert Group for International Collaboration on Microdata Access considered solutions to overcome these problems and move toward a system where all stakeholders in microdata sharing could speak the same language.

Statistical metadata for microdata – major stakeholders and their needs

12. In order to identify the main barriers that make international microdata exchange difficult, the first step is to define the key stakeholders and their needs in this context: the producers and disseminators of official microdata (NSOs and the international statistical systems) and the users of statistical information (e.g. Data Archives and researchers).

13. Regarding Data Archives, it has to be mentioned that their cooperation with NSOs varies greatly among countries (Data without Boundaries, 2013). In some countries, Data Archives are actively involved in the production and/or dissemination of microdata sets, while in others they are not at all involved. Data Archives can be regarded both as producers and disseminators and as users of microdata.

Producers/disseminators of microdata

14. Traditionally, producers of microdata are focusing on statistical metadata from two perspectives. First, to have good quality metadata available to drive the internal processes to produce official statistics, including microdata. Second, to have good quality metadata ready for the dissemination of statistical information, using various dissemination channels (in case of microdata, safe centres, remote access, remote execution and public use files (PUFs)). The Chapter 9 on PUFs also identifies various needs for the statistical metadata provided for these datasets.

15. The management of metadata ensures that good quality metadata are maintained throughout the whole statistical business process. In that sense, metadata management is a real over-arching activity.

16. Aggregated official statistics are typically compiled by tabulations based on detailed unit-level data; thus, the quality of the microdata used for tabulations is of key importance during the whole statistical business process. From this perspective, different levels of attention is given to different variables, thus it results in quality differences for the variables of the final microdata set, especially when more emphasis is put on dissemination of aggregated information. Therefore, due to lack of resources or other constrains, quality assurance is not necessarily enforced for variables not used for tabulations. This is usually an outstanding issue when the whole microdata set is used for statistical and/or scientific purposes?

17. As microdata are increasingly considered as a key output by themselves and not as a "byproduct" of the statistical processes, the quality of the whole microdata set (especially, the variables included in the microdata, ready for dissemination) should be in focus during the business processes. Paying attention to the needs and requirements regarding the microdata sets from the very beginning of the statistical processes (from planning phase) requires a new way of thinking from the producers; compared to the approach of focusing solely on aggregated statistical information.

18. Producers of official statistics are interested not only in the production of microdata but also in maximizing the usability of the microdata sets disseminated; therefore more and more emphasis is put on the metadata descriptions provided for these datasets.

19. Disseminating microdata using various channels serves both national and international needs. In the context of international microdata sharing, the issues of harmonisation, standardization and comparability across countries and domains are important requirements. For this purpose, different metadata standards have been developed over the years and are currently being used or introduced to the dissemination processes of producers of official statistics.

Users of microdata

20. Official statistics are traditionally intended for wide range of users who can be classified into different groups. From the perspective of international microdata exchange, users from the research community are a relevant group, as access to microdata is foreseen traditionally for scientific purposes.

21. Researchers are typically interested in finding information on what microdata sets are available for scientific purposes and on their access conditions. Good documentation on where and how to find this statistical information facilitates substantially microdata access.

22. Apart from the availability and access conditions, the content and quality of the microdata sets is also of high importance. For that reason, detailed description (metadata) on the content of the dataset and the quality of the information stored in the microdata (separate quality report and/or statistical metadata on quality) is necessary for them.

23. More and more research projects involve using different microdata sets for one research purpose and use data linkage techniques to create a joined microdata set. This is needed as different microdata sets cannot address given issues of research and linking this information to other information (specifically, additional microdata) is needed. This also requires detailed, good quality statistical metadata on all microdata sets.

24. Different researchers have different needs; therefore it is quite implicit that different researchers are interested in different aspects of the microdata sets, such as different variables or different level of detail. They would also like to assess the usability of the same microdata sets for different research purposes. Due to efficiency and resource reasons, it cannot be expected from the producers of microdata to provide information for all the different and special needs of researchers.

25. Researchers are also interested to access detailed statistical metadata on microdata sets in a clear and understandable form. This also fosters the need to use standardized solutions for the published metadata on microdata sets.

26. Regarding the different dissemination channels of microdata, it is also important to have precise information (name, time coverage, etc.) for variables in the microdata files. This can be extremely important for remote execution, for example, when researchers have to prepare syntax files and write programmes that will then be executed by the microdata holders. Due to high level of automation, machine-readability (next to human-readability) of disseminated statistical metadata has become a fundamental requirement.

Recommended principles to provide statistical metadata on microdata for international data exchange

27. In light of the above, it is proposed to consider the following seven principles. Their implementation by NSOs should facilitate microdata exchange at international level.

P1. Availability. *The first and most basic recommendation is to have statistical metadata available for all microdata sets that are ready to be disseminated using various dissemination channels.*

Microdata without statistical metadata cannot be interpreted by the users thus the microdata sets cannot be used and the potential in disseminated microdata sets cannot be exploited.

P2. Accessibility. *Users should have access to statistical metadata on microdata.*

Metadata on microdata should preferably be shown together with the list of available microdata.

P3. Bilingualism. *It is recommended to provide the metadata descriptions not exclusively in the national language but also in English.*

When statistical metadata is in the national language only, the use of the microdata is likely to be limited to speakers of the given language (more often, national users).

P4. Fitness for use. *As microdata could serve different needs, statistical metadata should be as complete as possible*.

For this purpose, stakeholder needs should be constantly monitored and the content of the metadata descriptions be regularly assessed against user needs. The United Nations Economic Commission for Europe [1] collected statistical metadata content that is vital for the end users

of statistical metadata and data. These requirements should be used as guidance or minimum checklist for the content of statistical metadata.
In order to judge the usability of microdata, the following two *"major dimensions of the quality of statistical data*" should be covered by the statistical metadata [2]:

- the **content** (meaning) of the statistical data, making it possible for the user to judge the relevance of the data with regard to his/her questions or problem;

- the **accuracy** (precision, reliability) of the statistical data, that is, how well the measurement/estimations actually measure/estimate what was intended (by the designers) to be measured/estimated.

P5. Quality aspects. *Complete information on the quality of the microdata set should be provided.*

Due to their importance, quality aspects are listed under this separate principle, even though, technically, they can be considered as a dimension of the principle "fitness for use".
Quality dimensions affect the usability of the microdata for a given research purpose. Information on quality is usually provided as integral part of the statistical metadata. Sometime separate quality reports/descriptions on the microdata set is provided. In order to have good quality microdata sets available for international access, it is highly recommended to consider microdata sets as important outputs of the statistical business process from the very beginning of the planning phase.

P6. Harmonisation. *In order to compare information from different sources, it is important that metadata are structured according to a common format.*

This leads to requirements on common structure and terminology. This requirement is also underlined by the recent report of the EC project Data without Boundaries [3]. This principle is usually fulfilled by providing statistical metadata on microdata according to one of the available metadata standards. The Metadata Common Vocabulary (MCV) and metadata formats (SDMX - Statistical Data and Metadata Exchange, DDI - Data Documentation Initiative, or other) especially target this issue by providing common terminology and structure for the meta descriptions.

P7. Machine-readable. *Statistical metadata should be provided also in a machine-readable format (through interfaces and web services).*

As a wide range of IT tools are available, statistical metadata should be structured in a way that it can automatically be downloaded into different repositories from websites. Machine-readable is also a requirement for the coordination of access through single gateways for metadata (and statistical data). To further coordinate such access, SDMX-based joint hubs are currently being developed.

28. The recommended principles are applicable at the national and international level, as the statistical metadata required by users at the international level is usually the same as that needed by users in a single country. Though, the purpose of this collection of principles to draw attention to issues that currently make the exchange of microdata difficult or impossible at the international level.

29. These seven principles also contribute to the more efficient and transparent operation of statistical processes. Ultimately, following these principles mean that the producer of official statistics is more openly contributing to the "circle of trust" concept (Chapter 4).

Three levels of maturity

30. The seven recommended principles are intended to reflect on the different levels of the data collaboration maturity level model.

31. At the initiating phase, it is more important to provide statistical metadata on statistics and make them available to the users. At this stage, the level of detail may vary by statistical domains or datasets but it is the intention and motivation of the producer of official statistics to provide as much statistical metadata to users as possible. Principles 1 and 2 are the most relevant in this phase of maturity.

32. At the level of pioneering, more automation is involved in the production and dissemination of statistical metadata and more emphasis is put on providing all statistical metadata not only in the national language but also in English. Principles 3 to 5 become important in this stage.

33. At the embedded level, all statistical metadata is provided by automation, English statistical metadata is created as integral part of the statistical business processes and overall standard solutions are sought and applied in order to increase efficiency in the management of metadata. All the seven Principles are followed, but more emphasis is put on the further development and harmonisation, thus on Principles 6 and 7.

Practical considerations for metadata standards used for international microdata exchange

34. Using metadata standards for transborder access to microdata would allow addressing most of the issues covered by the Principles presented above. Currently, different national and international metadata standards are used across organisations and countries; also, these standards are being constantly updated. It is not obvious for an organization producing microdata to decide which standard and/or version of it should be used. At present, most countries use their own metadata standards in their internal systems and apply one or more international metadata standards for dissemination purposes to foster exchange of statistical information.

35. The lack of international harmonization is also highlighted by a recent report on metadata standards by an EC project (Data without Boundaries, 2013). The report states that the use of metadata standards does not mean that data producers need to adopt these standards in their own working environment but there should be an option to map their own internal metadata standards into the agreed format to provide the necessary information in a comparable form to the users. The project carried out a detailed survey on the use of metadata standards in European countries; the survey findings stress the variety of metadata standards used across European countries.

SDMX and DDI

36. In the context of exchange of tabular/aggregate data and microdata, the initiatives of SDMX and DDI are usually mentioned. With recent developments of both initiatives (SDMX 2 and DDI 3), the issue of complementary and parallel use of both standards is now widely discussed.

37. SDMX and DDI are used by countries at different levels. Canada, for example, uses the DDI-Codebook for public use microdata, coding at the end of the statistical processes, while Turkey uses the same initiative in the production databases to drive their processes. Other countries are still experimenting with the application of the different standards and their application; others have no experiences on the implementation level. Altogether, it is not obvious how and to what extent these standards should be applied in the business processes.

38. The debate on SDMX versus DDI is currently going on in the framework of several projects and Working Groups, for instance, the SDMX Expert Group meetings, the Statistical Metadata group (METIS), the Statistical Data and Metadata Exchange, DDI Alliance and the Open Data Foundation. Other recent developments in the area include the Generic Statistical Information Model (GSIM), and ESSnet projects CORE (Common Reference Environment), SDMX I and II.

39. In a paper about DDI-SDMX integration and implementation, at the UNECE Work Session on Statistical Metadata on 6th-8th May 2013, the cooperative operation of the two initiatives was endorsed. Also, in the European Statistical System, the coexistence of SDMX and DDI is addressed by the ESS.VIP programme, especially in a cross-cutting project on Information Models and Standards (CRC.IMS). As this initiative is also coordinated with the UNECE High-Level Group for the Modernisation of Statistical Production and Services, Modernisation Committee on Standards, it is desirable that a broader consensus on the issue would eventually emerge.

40. In particular, the two-year-long CRC.IMS project now includes a phase specially dedicated to making proposals for an integrated standard to handle microdata and aggregated data. This programme aims at analysing and evaluating various standards to deliver a proposal for an integrated micro-macro framework based on SDMX and DDI.

41. The development of SDMX and DDI is also in line with the strategic vision of High-Level Group on Modernisation of Statistical Production and Services and both standards are considered very important tools for the industrialization of statistics.

42. The bodies responsible for the development of SDMX and DDI have attempted to create standards which do not duplicate efforts; DDI and SDMX are indeed more complementary than competing standards and they are both intentionally aligned with each other (Gregory and Heus, 2007). This harmonisation is motivated by the awareness that users need to deal with several different standards. SDMX was created with the awareness of DDI versions 1.XX and 2.XX (they were both available before SDMX was created) and that DDI version 3.XX benefited from having SDMX as a published specification.

43. It is generally understood that if an NSO attempts to support the entire GSBPM using only one standard (either SDMX or DDI) then the results will be sub-optimal. Neither SDMX nor DDI fully covers the entire production for national statistical organisations. As SDMX focuses on the management of regular data collection and dissemination, it is usually considered that SDMX corresponds to the phases 5 to 7 of GSBPM (process, analyse, disseminate), while DDI seems more appropriate for the first phases that cover microdata and not tabular or aggregated information (Boško and Hudec, 2012).

44. In view of the above considerations, it is suggested that countries, based on their needs, adopt one of the metadata standards for data description to address the issues raised under the seven Principles; and that they share their experience with the international community. The Expert Group also supports to adopt the Metadata Common Vocabulary (MCV), independently on whether SDMX, DDI or other standard is used. This would facilitate easier mapping of metadata between systems.

REFERENCES

Boško, P. and Hudec, M. (2012), Study of SDMX and DDI in Data Collection. Part of Work Package 3 of ESSnet on SDMX II: Support of SDMX Application for Microdata Handling, within Statistical Business Process at the NSI Level

Data without Boundaries (2013), "Metadata Standards – Usage and Needs in NSIs and Data Archives", July, 2013

Gregory, A. and Heus, P. (2007), DDI and SDMX: Complementary, not Competing Standards. Open Data Foundation. In: DDI/SDMX Workshop, Wiesbaden, Germany, June, 2008

United Nations Statistical Commission and Economic Commission for Europe (1995), *Guidelines for the Modelling of Statistical Data and Metadata*. United Nations, New York and Geneva

United Nations Economic Commission for Europe (2009), *Statistical Metadata in a Corporate Context: A Guide for Managers*. United Nations, Geneva

United Nations Statistical Commission and Economic Commission for Europe – Conference of European Statisticians (2013), DDI-SDMX Integration and Implementation. Geneva, Switzerland, 6th-8th May, 2013

CHAPTER 3. REVIEW OF INTERNATIONAL INITIATIVES ON MICRODATA

45. One of the first actions performed by the OECD Expert Group for International Collaboration on Microdata Access was to survey the activities of other international initiatives and groups in the field of microdata with a view of identifying areas of complementarity and minimising possible duplications of work. Whenever relevant, contacts and cooperation were established with other groups, at the regional or international level, that explore(d) the feasibility of cross-border access to official microdata. This Chapter presents synthetic information on the international initiatives surveyed by the Expert Group. Most of the projects are conducted within the European Statistical System (ESS)[3] and involve EU and EFTA member states. A few other initiatives are developed at the broader international scale.

Table 3.1. Projects/Groups whose work is directly related to the Mandate of the Expert Group

1. DATA WITHOUT BOUNDARIES

Structure

The project Data without Boundaries (DwB) was set to enhance transnational access to official microdata within Europe by promoting co-ordination of existing infrastructures, i.e. data producers (statistical offices and other members of the European Statistical System) and data archives.

Launched in May 2011 for a duration of four years, DwB is funded by the European Commission under the 7th Framework Programme for RTD.

The approach of DwB is to build on the existing infrastructures and establish and/or further develop the essential ingredients for cross-border access, notably cooperation and trust between institutions, common view and agreements on standards between the European Statistical System (led by Eurostat), other stakeholders (such as the Central Banks), CESSDA and researchers who are the final users.

Membership

28 partners belonging to the European Statistical System (10 National statistical Institutes or statistical departments), to the CESSDA (11 Data Archives) and to the Research Community (6 universities and 1 SME involved in methodological research).

Objectives

DwB aims to create an integrated model where the best solutions for accessing microdata become available irrespective of national boundaries, and are flexible enough to fit national regulations and constraints.

Specific objectives include: i) develop and promote a widely-recognised standard for researcher accreditation; ii) collect information on and describe the legal frameworks for researchers' access to official data; iii) provide a comprehensive overview of available official data in Europe; iv) build cooperation between national statistical institutes and data archives for developing compatible metadata standards; v) provide remote access to confidential data for foreign researchers across Europe; vi) develop methods for achieving the best ratio of utility against disclosure risk for datasets.

2. ESSNET DARA

Structure

The ESSnet project funded by the European Commission on Decentralised and Remote Access to Confidential Data in the ESS (DARA) deals with the implementation of remote access from safe centres in NSOs in the EU Member States to confidential microdata sets in Eurostat, with no need of

Objectives

DARA aims to implement remote access to confidential data held by Eurostat.

Key objectives include:

i) Rules and standards for accredited safe centres

[3] The European Statistical System (ESS) is a partnership between Eurostat, and the National Statistical Institutes and other national authorities, responsible in each Member State for the development, production and dissemination of European statistics (Regulation No 223/2009 on European statistics). ESS covers the 28 EU Member States and the four EFTA member states: Iceland, Liechtenstein, Norway and Switzerland.

consulting them at the safe centre of Eurostat in Luxembourg. It started in October 2011 for a duration of two years, as a result of the recommendations of, the ESSnet Decentralised Access to EU Microdata Sets. The feasibility study completed by this previous project had concluded on the possibility of remote access to confidential data held in Eurostat and had suggested to conduct a pilot.

DARA focused on the analysis of technical aspects and safety requirements for implementing remote access. It conducted a pilot to test the viability of the solutions proposed.

Membership

Participants represented five countries and six statistical institutes: ONS, CBS, ISTAT, HCSO, Destatis and SSO (Germany).

ii) A Manual for the scientific community on how to access EU statistics from a safe centre, including a detailed request form.

iii) A Manual for NSOs describing the workflow when data access is requested.

iv) A Manual of all the required specifications for security, interoperability, workflows, including the check-list of security constraints (from client to host).

3. ESSNET SDC HARMONISATION

Structure

The ESSnet project funded by the European Commission on Common Tools and Harmonised Methodology for Statistical Disclosure Control in the ESS lasted from December 2010 to April 2012 and addressed the issue of the release of harmonised microdata in multiple countries

The project mainly conducted case studies where different SDC methods were assessed, taking into account the risk of disclosure and users' needs, in particular with respect to data comparability.

Membership

Participants represented five countries: CBS, Istat, Destatis, Statistics Sweden, Statistics Norway

Objectives

The project produced:

- A proposed framework for the release of harmonised microdata from multiple countries that are useful for the final users.

- Recommendations for future directions of the development of SDC tools

4. WORKSHOP ON DATA ACCESS (previously Nuremberg Group)

Structure

The Workshop on Data Access (WDA) brings together researchers who are active all over the world in promoting innovations regarding data access and the management of research facilities, in particular Research Data Centres (RDC). Researchers are invited to share knowledge and cooperate to build an international network for the diffusion of innovative modes of data access.

WDA is organised by a group of RDCs but is not a formal group.

Membership

There are no official representatives of WDA. Participants to WDA meetings included researchers from North America, Australia, China, Japan, Mexico, New Zealand, South Africa and several countries in Europe.

Objectives

WDA aims to promote access to confidential government data for research purposes.

Its objective is to build a network of experts who could help to provide advice and identify and disseminate best practices on access to confidential microdata.

Table 3.2. Other projects/groups

Project/Group	*Objectives*
High-Level Group for the Modernisation of Statistical Production and Services (HLG) The Group was set up by the Bureau of the Conference of European Statisticians in 2010 to oversee and coordinate international work relating to statistical modernisation.	To develop a vision and strategy for the future of official statistics. One element of the strategy is to improve access to "semi-finished" statistical products, including microdata sets.

The UNECE provides the secretariat to this group; the membership includes chief statisticians from national and international organisations around the world.	
Paris 21 - Partnership in Statistics for Development in the 21st Century Consortium of users and producers of statistics from both developing and developed countries and multi-lateral organisation.	To encourage and assist low income and lower middle income countries to develop an overall vision of the development of their national statistical system. Paris21 creates and promotes tools to improve data management and dissemination, including: microdata management toolkit, metadata application, microdata anonymisation toolkit; and the establishment of a central survey catalogue to better inform microdata users of the availability of survey and census data from all data sources.
GSIM - Generic Statistical Information Model GSIM is developed by the High Level Group for the Modernisation of Statistical Production and Services.	To provide a reference framework of information objects, which enables generic descriptions of the definition, management, and use of data and metadata throughout the statistical production process. GSIM forms a key part of the strategic vision of HLG.
OECD Working Party on Information Security and Privacy (WPISP) Members are the 34 OECD countries.	To develop policy options to sustain trust in the Internet Economy, working in areas such as: Critical information infrastructure, Digital identity management and e-authentication, Malware, Radio-frequency identification (RFID), Sensor networks, OECD Privacy Guidelines, Protecting children online, Privacy law enforcement co-operation.
CESSDA - Council of European Social Science Data Archives Members: 20 EU Member States: Austria, Czech republic, Denmark, Estonia, Finland, France, Greece, Germany, Hungary, Ireland, Italy, Luxembourg, Netherlands, Norway, Romania, Slovenia, Spain, Sweden, Switzerland, UK.	To promote the acquisition, archiving and distribution of data throughout Europe. To promote projects and procedures for enhancing exchange of data and technologies among data organizations.
EUDAT – Collaborative Data Infrastructure Members: Representatives from research communities from 15 specific research disciplines across all major fields of science. It comprises 25 European partners from 13 countries. Duration: October 2011 –36 months	To build a sustainable pan-European infrastructure for scientific data. To contribute to the production of a Collaborative Data Infrastructure (CDI) which will allow researchers to share data within and between communities and enable them to carry out their research effectively. To provide a pan-European solution to the challenge of data proliferation in Europe's scientific and research communities.
DASISH - Data Service Infrastructure for the Social Science and Humanities Members: 19 participants from 11 EU countries Duration: January 2012 – 36 months	To provide solutions to a number of common issues relevant for projects in social science and humanities, notably as concerns data quality, data archiving, data access and legal and ethics issues.
OECD Expert Group on Data and Research Infrastructure for Social Sciences Group established in April 2010 by the OECD Global Science Forum. Final report submitted in April 2012. Members: 31 participants from OECD and non-OECD countries.	To review developments in international data availability, consider their suitability for comparative research, detail the challenges to be addressed and make recommendations to respond to these new opportunities.

ANNEX I.A1. GLOSSARY FOR INTERNATIONAL MICRODATA ACCESS

Academic staff

Academic Staff (International Standard Classification of Education (ISCED) 5-6) includes personnel whose primary assignment is instruction, research, or public service. This includes staff personnel who hold an academic rank with titles such as professor, associate professor, assistant professor, instructor, lecturer, or the equivalent of any of these academic ranks. The category includes personnel with other titles, (e.g. dean, director, associate dean, assistant dean, chair or head of department), if their principal activity is instruction or research. It does not include student teachers or teacher aides

Source: 2001 Data Collection on Education Systems: Definitions, Explanations and Instructions, UNESCO, OECD, Eurostat, page 45, http://stats.oecd.org/glossary/detail.asp?ID=10

Access arrangements

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.

Access arrangements, whether at the governmental or institutional levels, should be developed in consultation with representatives of all directly affected parties. In collaborative research programmes or projects, and especially in international scientific co-operation or in research projects based on public/private partnerships where there are differences in regulatory frameworks, the parties involved should negotiate research data sharing arrangements as early as possible in the life of the research project, ideally at the initial proposal stage. This will help ensure that adequate and timely consideration will be given to issues such as the allocation of resources for sharing and sustainable preservation of research data, differences in national intellectual property laws, limitations due to national security, and the protection of privacy and confidentiality.

Access arrangements also should be responsive to factors such as the characteristics of the data, their potential value for research purposes, the level of data processing (raw versus partially processed versus final), whether they are homogeneous data from a facility instrument or sensor versus heterogeneous field data collected by single researchers, data on human subjects or physical parameters, and whether the data are generated directly by a government entity or as a result of government funding. These variations in the origin or type of data should be taken into consideration when establishing data access arrangements. Further, consideration should be given to the following:

• Many of the problems related to access, dissemination and sharing of data result from the lack of explicit institutional agreements on the terms of access and use. With data management becoming ever more complex in certain areas of research, traditional informal arrangements between researchers may no longer be adequate and may need to be complemented by formally agreed practices and procedures.

• 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.

• 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

research data are effectively preserved, managed and made accessible in order that they can be put to efficient and appropriate use over the long term.

Source: OECD (2007), *OECD Principles and Guidelines for Access to Research Data from Public Funding*, Paris, p. 17. http://www.oecd.org/science/scienceandtechnologypolicy/38500813.pdf

Accessibility of statistical information

It refers to the ease with which it can be obtained from *a statistical agency*. This includes the ease with which the existence of information can be ascertained, as well as the suitability of the form or medium through which the information can be accessed. The cost of the information may also be an aspect of accessibility for some users.

Source: Statistics Canada, "Statistics Canada Quality Guidelines", Fifth Edition, October 2009, http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.pdf

Note: The text in Italics was modified by the Expert Group. Accessibility is one of the criteria of statistical quality. It is defined as "conditions and modalities by which users can obtain, use and interpret data"

Source: Regulation (EC) No 223/2009 on European statistics, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:0173:en:PDF

Accessibility of documentation

It refers to the availability of documentation of various aspects of the data (sources and methods documents) and the content of such documentation.

Source: Statistical Data and Metadata Exchange (SDMX) – BIS, ECB, Eurostat, IBRD, IMF, OECD and UNSD – Metadata Common Vocabulary, http://www.sdmx.org/

Accountability

The performance of data access arrangements should be subject to periodic evaluation by user groups, responsible institutions and research funding agencies. Although each party is likely to use somewhat different evaluation criteria, the sum total of the results should provide a comprehensive picture of the value of data and of data access regimes. Such evaluations should help to increase the support for open access among the scientific community and society at large. The following should be considered in establishing evaluation criteria:

- Overall public investments in the production and management of research data.
- Management performance of data collection and archival agencies.
- Extent of re-use of existing data sets.
- Knowledge generated from the re-use of existing data.

The use of targeted foresight exercises to determine the nature and scope of data preservation activities and the types of data most likely to be needed in the future. Even if gaining clear insight into the cost, benefit and performance of data access arrangements will not be an easy task, those in charge of data access arrangements should put effort into showing the benefits of open data access to justify and help ensure sustained support from all levels of government.

Source: OECD (2007), *OECD Principles and Guidelines for Access to Research Data from Public Funding*, Paris, 2007, p. 21, http://www.oecd.org/science/scienceandtechnologypolicy/38500813.pdf

Administrative data

Administrative data is the set of units and data derived from an *administrative source*.

Source: OECD, IMF, ILO, Interstate Statistical Committee of the Commonwealth of Independent States, "Measuring the Non-Observed Economy: A Handbook", Annex 2, Glossary, Paris, 2002, http://www.oecd.org/dataoecd/9/20/1963116.pdf

Administrative data/Administrative records

The data collected by sources external to statistical offices.

Source: UNECE, Conference of European Statisticians Statistical Standards and Studies, No. 53, "Terminology on Statistical Metadata", Geneva, 2000, http://www.unece.org/stats/publications/53metadaterminology.pdf

Administrative source

Administrative source is the organisational unit responsible for implementing an administrative regulation (or group of regulations), for which the corresponding register of units and the transactions are viewed as a source of statistical data.

Source: OECD, IMF, ILO, Interstate Statistical Committee of the Commonwealth of Independent States, "Measuring the Non-Observed Economy: A Handbook", Second Draft, Annex 2, Glossary, Paris, 2002, http://www.oecd.org/dataoecd/9/20/1963116.pdf

Anonymisation[4]

Anonymisation is the set of methods applied to microdata in order to minimise the risk of identification of the statistical units concerned.

Source: Commission Regulation (EC) No 831/2002 on access to confidential data for scientific purposes.

Anonymised microdata

Individual statistical records which have been modified in order to minimize, in accordance with current best practice, the risk of identification of the statistical units to which they relate.

Source: Commission Regulation (EC) No 831/2002 on access to confidential data for scientific purposes, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2002:133:0007:0009:EN:PDF

[4] The term "anonymisation" is sometimes used to indicate the process of removing direct identifiers from the confidential data. The group however having the objective of sharing the same language and being coherent in the development of the final report decided to use the term "de-identification" to define the removal of direct identifiers and the term "anonymisation" to define the process of application of statistical methods in order to minimise the risk of disclosure.

Anonymised record

A record modified in order to minimize, in accordance with current best practice, the risk of identification of the statistical unit.

Source: Commission Regulation (EC) No 831/2002 on access to confidential data for scientific purposes, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2002:133:0007:0009:EN:PDF

Approximate disclosure or Partial disclosure

Approximate disclosure happens if a user is able to determine an estimate of a respondent value that is close to the real value. If the estimator is exactly the real value the disclosure is exact.

Source: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf

Attribute

An inherent characteristic of an object.

Source: UNECE, Conference of European Statisticians Statistical Standards and Studies, No. 53, "Terminology on Statistical Metadata", Geneva, 2000, http://www.unece.org/stats/publications/53metadaterminology.pdf

Attribution

Attribution is the association or disassociation of a particular attribute with a particular population unit.

Source: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf

Circle of trust

The concept of "circle of trust" is based on the agreement that each member of the circle is accepted according to the same rules and conditions that are approved by all members. In the context of statistical activities, trust involves confidentiality rules and security requirements but also competence and legal aspects.

Source: OECD Expert Group for International Collaboration on Microdata Access.

Composite microdata

Unit record data resulting from data integration

Source: UNECE, Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes, Geneva, 2009, http://www.unece.org/fileadmin/DAM/stats/publications/Confidentiality_aspects_data_integration.pdf

Confidential cells

The cells of a table which are non-publishable due to the risk of statistical disclosure.

Source: Eurostat, "Manual on disclosure control methods", Office for Official Publications of the European Communities, Luxembourg, 1996, p. 8-9, http://ec.europa.eu/eurostat/ramon/statmanuals/files/manual_on_disclosure_control_methods_1996.pdf

Confidential data

Confidential data means data which allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information. To determine whether a statistical unit is identifiable, account shall be taken of all relevant means that might reasonably be used by a third party to identify the statistical unit.

Source: Regulation (EC) No 223/2009 of the European Parliament and of the Council on European statistics, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:01:EN:HTML

Confidentiality (see Statistical confidentiality)

An obligation to the provider of information to maintain the secrecy of that information.

Source: UNECE, Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes, Geneva, 2009, http://www.unece.org/fileadmin/DAM/stats/publications/Confidentiality_aspects_data_integration_n.pdf

Confidentiality agreement/statement

A non-disclosure agreement (NDA), also known as a confidentiality agreement (CA), confidential disclosure agreement (CDA), proprietary information agreement (PIA), or secrecy agreement, is a legal contract between at least two parties that outlines confidential material, knowledge, or information that the parties wish to share with one another for certain purposes, but wish to restrict access to or by third parties. It is a contract through which the parties agree not to disclose information covered by the agreement. An NDA creates a confidential relationship between the parties to protect any type of confidential and proprietary information or trade secrets. As such, an NDA protects non-public business information. NDAs are commonly signed when two companies, individuals, or other entities (such as partnerships, societies, etc.) are considering doing business and need to understand the processes used in each other's business for the purpose of evaluating the potential business relationship. NDAs can be "mutual", meaning both parties are restricted in their use of the materials provided, or they can restrict the use of material by a single party. It is also possible for an employee to sign an NDA or NDA-like agreement with an employer. In fact, some employment agreements will include a clause restricting employees' use and dissemination of company-owned "confidential information".

Source: Wikipedia, http://en.wikipedia.org/wiki/Non-disclosure_agreement

Data

The physical representation of information in a manner suitable for communication, interpretation, or processing by human beings or by automatic means.

Source: UNECE, Conference of European Statisticians Statistical Standards and Studies, No. 53, "Terminology on Statistical Metadata", Geneva, 2000, http://www.unece.org/stats/publications/53metadaterminology.pdf

Database

A data file or set of data with relationships expressed among data. Data stored in the database are independent of any particular application.

Source: UNECE, Conference of European Statisticians Statistical Standards and Studies, No. 53, "Terminology on Statistical Metadata", Geneva, 2000, http://www.unece.org/stats/publications/53metadaterminology.pdf

Data confidentiality

A property of data, usually resulting from legislative measures, which prevents it from unauthorized disclosure.

Source: UNECE, Conference of European Statisticians Statistical Standards and Studies, No. 53, "Terminology on Statistical Metadata", Geneva, 2000, http://www.unece.org/stats/publications/53metadaterminology.pdf

Data cubes

Data cubes are the main vehicle for releasing all statistical information. Statistical confidentiality protection is applied in a routine fashion. Moreover, data cubes can be easily linked and compared on a meso level. Conversely, a lack of coherence is easily discovered. Adding data cubes to database ensures that statistical information is produced and published to serve the public at large. Data cubes are primarily made and used to serve the public at large. Even if they are produced and paid for by a third party, as a matter of policy the resulting data cubes are available for all.

Source: UNECE, Managing Statistical Confidentiality & Microdata Access - Principles and Guidelines of Good Practice, 2007, p. 32; Annex 1.3. Case Study: Data Cubes – Netherlands, http://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

Data dissemination

Dissemination is the release of data obtained from a statistical activity to users through various media.

Source: Statistics Canada, "Statistics Canada Quality Guidelines", Fifth Edition, October 2009, http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.pdf

Data enclave (On-site facility, Safe centre)

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.

Source: Olivier Dupriez and Ernie Boyko. 2010, "Dissemination of Microdata Files; Principles, Procedures and Practices" IHSN Working Paper No. 005, August 2010, p. 7, http://www.surveynetwork.org/home/sites/default/files/resources/IHSN-WP005.pdf

Note: The text in Italics was added by the Expert Group.

Data file

An organised collection of related records of data.

Source: UNECE, Conference of European Statisticians Statistical Standards and Studies, No. 53, "Terminology on Statistical Metadata", Geneva, 2000, http://www.unece.org/stats/publications/53metadaterminology.pdf

Data integration

The process of combining data from two or more sources to produce new outputs.

Source: UNECE, Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes, Geneva, 2009, http://www.unece.org/fileadmin/DAM/stats/publications/Confidentiality_aspects_data_integration.pdf

Data intruder

A data user who attempts to disclose information about *a statistical unit or* a population unit through identification or attribution.

Source: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf

Note: The text in Italics was added by the OECD Expert Group.

Data laboratories (Data enclave, on-site facilities)

This involves working on-site at the National Statistical Office, or one of its Branches, to obtain access to microdata. Access could be direct or indirect through staff of the National Statistical Offices.

Source: UNECE, "Managing Statistical Confidentiality & Microdata Access - Principles and Guidelines of Good Practice", 2007, p. 107, http://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

Data matching

The linkage of microdata from different sources based on common features present in those sources.

Source: UNECE, Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes, Geneva, 2009, http://www.unece.org/fileadmin/DAM/stats/publications/Confidentiality_aspects_data_integration.pdf

Data processing (Personal data processing)

The operation performed on data in order to derive new information according to a given set of rules.

Source: UNECE, Conference of European Statisticians Statistical Standards and Studies, No. 53, "Terminology on Statistical Metadata", Geneva, 2000, http://www.unece.org/stats/publications/53metadaterminology.pdf

Note: The text in Italics was added by the OECD Expert Group.

Data protection

An activity aimed at covering or shielding data from physical damage or unauthorized access *or disclosure.*

Source: UNECE, Conference of European Statisticians Statistical Standards and Studies, No. 53, "Terminology on Statistical Metadata", Geneva, 2000, http://www.unece.org/stats/publications/53metadaterminology.pdf

Note: The text in Italics was added by the OECD Expert Group.

Data provider

An organisation which *provides* data or metadata. *(..)* This term includes providers of data files from statistical or non-statistical sources, but not individual respondents to statistical surveys.

Source: UNECE, Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes, Geneva, 2009, http://www.unece.org/fileadmin/DAM/stats/publications/Confidentiality_aspects_data_integration.pdf

Note: The text in Italics was modified by the OECD Expert Group.

Data security

The measures taken to prevent unauthorised access or use of data.

Source: UNECE, Conference of European Statisticians Statistical Standards and Studies, No. 53, "Terminology on Statistical Metadata", Geneva, 2000, http://www.unece.org/stats/publications/53metadaterminology.pdf

Data set

Any organised collection of data.

Source: UNECE, Conference of European Statisticians Statistical Standards and Studies, No. 53, "Terminology on Statistical Metadata", Geneva, 2000, http://www.unece.org/stats/publications/53metadaterminology.pdf

Data type

A category used to classify the collection of letters, digits, and/or symbols to depict values of a data element based upon the operations that may be performed on the data element.

Source: UNECE, Conference of European Statisticians Statistical Standards and Studies, No. 53, "Terminology on Statistical Metadata", Geneva, 2000, http://www.unece.org/stats/publications/53metadaterminology.pdf

De-identification

De-identification is the process of removing direct identifiers from the original confidential data.

Source: OECD Expert Group for International Collaboration on Microdata Access.

De-identified microdata

Microdata file without direct identifiers.

Source: OECD Expert Group for International Collaboration on Microdata Access.

De-identified record

A record from which direct identifiers have been removed.

Source: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf.

Direct identifier (*Formal identifier*)

Direct identifier is a name or address or any other publicly accessible identification number.

Source: Regulation (EC) No 223/2009 of the European Parliament and of the Council on European statistics, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:01:EN:HTML

Direct identification

Direct identification means the identification of a statistical unit from its formal identifiers, i.e. from its name or address, or from a publicly accessible identification number.

Source: Regulation (EC) No 223/2009 of the European Parliament and of the Council on European statistics, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:01:EN:HTML

Disclosure (Statistical disclosure)

Disclosure relates to the inappropriate attribution of information to a data subject, whether an individual or an organisation. Disclosure has two components: identification and attribution.

Sources: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf

Disclosure analysis

Disclosure analysis is the process of protecting the confidentiality of data. It involves *estimating disclosure risk and proposing methods to reduce such risk by* limiting the amount of detailed information disseminated and/or masking data via noise addition, data swapping, generation of simulated or synthetic data, etc.

Source: United States Bureau of the Census, Software and Standards Management Branch, Systems Support Division, "Survey Design and Statistical Methodology Metadata", Washington D.C., August 1998, Section 3.3.17, page 28, http://www.census.gov/srd/www/metadata/metada18.pdf

Note: The text in Italics was added by the OECD Expert Group.

Disclosure control methods

Disclosure control methods are statistical procedures that aim at reducing the risk of disclosure. There are two main approaches to control the disclosure of confidential data. The first is to reduce the information content of the data provided to the external user. For the release of tabular data this type of technique is called restriction based disclosure control method and for the release of microdata the expression disclosure control by data reduction is used. The second is to change the data before the dissemination in such a way that the disclosure risk for the confidential data is decreased, but the information content is retained as much as possible. These are called perturbation based disclosure control methods.

Sources: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf.

Note: The text in Italics was added by the OECD Expert Group.

Disclosure risk

A disclosure risk occurs if an unacceptably narrow estimation of a respondent's confidential information is possible or if exact disclosure is possible with a high level of confidence.

Sources: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf.

Disclosure scenarios

Depending on the intention of the intruder, his or her type of a priori knowledge and the microdata available, three different types of disclosure or disclosure scenarios are possible for microdata: disclosure by matching, disclosure by response knowledge and disclosure by spontaneous recognition.

Source: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf.

Exact disclosure

Exact disclosure occurs if a user is able to determine the exact attribute for an individual entity from released information.

Sources: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf

Formal identifier (Direct identifier)

Any variable or set of variables which is instrumentally unique for every population unit, for example a population registration number. If the formal identifier is known to the intruder, identification of a target individual is directly possible for him or her, without the necessity to have additional knowledge before studying the microdata. Some combination of variables such as name and address are pragmatic formal identifiers, where non-unique instances are empirically possible, but with negligible probability.

Source: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf

GSBPM

The GSBPM (General Statistical Business Process Model*)* is intended to apply to all activities undertaken by producers of official statistics, at both the national and international levels, which result in data outputs. It is designed to be independent of the data source, so it can be used for the description and quality assessment of processes based on surveys, censuses, administrative records, and other non-statistical or mixed sources.

Sources: Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS) Generic Statistical Business Process Model Version 4.0 – April 2009,
http://www1.unece.org/stat/platform/download/attachments/8683538/GSBPM+Final.pdf?version=1&modificationDate=1241066597110

Indirect identifier (or quasi identifier or key variable)

These are variables which identify the respondent with some degree of ambiguity. A combination of indirect identifiers may lead to unambiguous identification. Example of indirect identifiers: age, country of birth, marital status etc.

Source: Regulation (EC) No 223/2009 of the European Parliament and of the Council on European statistics, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:01:EN:HTML.

Indirect identification

Indirect identification means the identification of a statistical unit by any other means than by way of direct identification.

Source: Regulation (EC) No 223/2009 of the European Parliament and of the Council on European statistics, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:01:EN:HTML.

Internal access

Internal access refers to giving full transparency to any necessary pre-release access within government, as deemed appropriate by the government.

Context: Under the SDDS *(Special Data Dissemination Standard),* this entails the listing of persons or officials holding designated positions within the government, but outside the agency producing the data, who have pre-release access to the data and the reporting of the schedule according to which they receive access.

Source: Statistical Data and Metadata Exchange (SDMX) - BIS, ECB, Eurostat, IBRD, IMF, OECD and UNSD - Metadata Common Vocabulary, www.sdmx.org

Intruder

A data user who attempts to link a respondent to a microdata record or make attributions about particular population units from aggregate data.

Sources: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf

Key variable

See Indirect identifier.

Licensed Files (or Research Files or Scientific Use Files or Microdata files for research)

Licensed Files– are distinct from PUFs *(Public Use Files)*: their dissemination is restricted to users who have received authorization 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 minimized when used in isolation, they may contain potentially identifiable data if, *e.g. matched with other information or* linked with other data files. Direct identifiers such as respondents' names must be removed from a licensed dataset.

Source: Olivier Dupriez and Ernie Boyko. 2010, "Dissemination of Microdata Files; Principles, Procedures and Practices" IHSN Working Paper No. 005, August 2010, p. 7, http://www.surveynetwork.org/home/sites/default/files/resources/IHSN-WP005.pdf

Note: The text in Italics was added by the OECD Expert Group

Licensing agreement

A permit, issued under certain conditions, for researchers to use confidential data for specific purposes and for specific periods of time. This agreement consists of contractual and ethical obligations, as well as penalties for improper disclosure or use of identifiable information. These penalties can vary from withdrawal of the license and denial of access to additional data sets to the forfeiting of a deposit paid prior to the release of a microdata file. A licensing agreement is almost always combined with the signing of a contract. This contract includes a number of requirements: specification of the intended use of the data; instruction not to release the microdata file to another recipient; prior review and approval by the releasing agency for all user outputs to be published or disseminated *if applicable* ; terms and location of access and enforceable penalties.

Sources: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf

Note: The text in Italics was added by the OECD Expert Group.

Macrodata

See Tabular data

Master use file

See Secure use file

Metadata

Metadata provides information on data - and about processes of producing and using data. Metadata are data which are needed for proper production and use of the data.

Source: UNECE and UNSC, Guidelines for the Modelling of Statistical Data and Metadata, Conference of European Statisticians Methodological Material, 1995, http://www.unece.org/fileadmin/DAM/stats/publications/metadatamodeling.pdf

Metadata Common Vocabulary (MCV)

It contains concepts and related definitions that are normally used for building and understanding metadata systems and SDMX data exchange arrangements of international organisations and national data producing agencies. The MCV covers a selected range of metadata concepts.

Source: http://sdmx.org/wp-content/uploads/2009/01/04_sdmx_cog_annex_4_mcv_2009.pdf

Metainformation system

A metainformation system uses and produces metadata, informing about data, and it fulfills its tasks by means of functions like “metadata collection”, “metadata processing”, “metadata storage”, and “metadata dissemination”. A metainformation system may be active or passive.

- An active metainformation system is physically integrated with the information system containing the data that the metadata in the metainformation system informs about.

- A passive metainformation system contains only references to data, not the data themselves.

Source: UNECE and UNSC, Guidelines for the Modelling of Statistical Data and Metadata, Conference of European Statisticians Methodological Material, 1995, http://www.unece.org/fileadmin/DAM/stats/publications/metadatamodeling.pdf

Microdata (or statistical microdata)

Microdata is the file consisting of the set of records where each record represents individual statistical unit.

The term microdata can refer to data about an individual person, household, business or other entity. It may be data directly collected by the NSO or obtained from other sources, such as administrative sources.

Source: UNECE, Managing Statistical Confidentiality & Microdata Access - Principles and Guidelines of Good Practice, 2007, p. 1, http://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

From a data producer's perspective, microdata is the form from which all other data outputs are derived and is the primary form that data is stored in.

Source: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf

National Statistical Office (NSO) or National Statistical Institute (NSI)

The National Statistical Office is the leading statistical agency within a national statistical system.

Source: Measuring the Non-Observed Economy: A Handbook, OECD, IMF, ILO, Interstate Statistical Committee of the Commonwealth of Independent States, 2002, Annex 2, Glossary, http://www.oecd.org/dataoecd/9/20/1963116.pdf

In the *European* context the National Statistical Institute is the authority designated by each Member State as the body having the responsibility for coordinating all activities at national level for the development, production and dissemination of European statistics. The NSI shall act as the contact point for the Commission (Eurostat) on statistical matters. (Regulation (EC) No 223/2009 on European statistics).

National Statistical Authority (NSA)

Although the term is used in the singular, it is meant to incorporate all statistical agencies, or statistical departments within administrations, who produce official statistics.

Source: OECD Expert Group for International Collaboration on Microdata Access.

National statistical system (NSS)

The national statistical system (NSS) is the ensemble of statistical organisations and units within a country that jointly collect, process and disseminate official statistics on behalf of national government.

Source: Measuring the Non-Observed Economy: A Handbook, OECD, IMF, ILO, Interstate Statistical Committee of the Commonwealth of Independent States, 2002, Annex 2, Glossary, http://www.oecd.org/dataoecd/9/20/1963116.pdf

Official statistics

Official statistics are statistics/*data* disseminated by the national statistical system, excepting those that are explicitly stated not to be official.

Source: Measuring the Non-Observed Economy: A Handbook, OECD, IMF, ILO, Interstate Statistical Committee of the Commonwealth of Independent States, 2002, Annex 2, Glossary, http://www.oecd.org/dataoecd/9/20/1963116.pdf.

Note: The text in Italics was added by the OECD Expert Group.

On-line access

See Remote access and Remote execution

On-site facility *(Data enclave)*

A facility that has been established on the premises of several NSIs. It is a place where external researchers can be permitted access to potentially disclosive data under contractual agreements which cover the maintenance of confidentiality, and which place strict controls on the uses to which the data can be put. The on-site facility can be seen as a 'safe setting' in which confidential data can be analysed. The on-site facility itself would consist of a secure hermetic working and data storage environment in which the confidentiality of the data for research can be ensured. Both the physical and the IT aspects of security would be considered here. The on-site facility also includes administrative and support facilities to external users, and ensures that the agreed conditions for access to the data were complied with.

Source: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf.

Open data

Open Data are data (datasets) that are:

- accessible to anyone and everyone, ideally via the internet,
- in a digital machine readable format that allows interoperation with other data,
- available at reproduction cost or less, and
- free from restrictions on use and re-use.

Source: OECD Expert Group for International Collaboration on Microdata Access.

Personal data processing

Personal data processing mean any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organisation, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction.

Source: Regulation (EC) No 45/2001 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data

Perturbation based disclosure control methods

Techniques for the release of data that change the data before the dissemination in such a way that the disclosure risk for the confidential data is decreased but the information content is retained as far as possible. Perturbation based methods falsify the data before publication by introducing an element of error purposely for confidentiality reasons. For example, an error can be inserted in the cell values after a table is created, which means that the error is introduced to the output of the data and will therefore be referred to as output perturbation. The error can also be inserted in the original data on the microdata level, which is the input of the tables one wants to create; the method will then be referred to as data perturbation - input perturbation being the better but uncommonly used expression. Possible perturbation methods are:

· rounding;

· perturbation, for example, by the addition of random noise or by the Post Randomisation Method;

· disclosure control methods for microdata applied to tabular data.

Source: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf.

Perturbation-based methods

Perturbation-based methods falsify the data before publication by introducing an element of error purposely for confidentiality reasons. This error can be inserted in the cell values after the table is created, which means the error is introduced to the output of the data and will therefore be referred to as output perturbation, or the error can be inserted in the original data on the microdata level, which is the input of the tables one wants to create; the method with then be referred to as data perturbation - input perturbation being the better but uncommonly used expression. Possible methods are: rounding; perturbation, for example, by the addition of random noise or by the Post Randomisation Method; disclosure control methods for microdata applied to tabular data.

Source: Eurostat, 1996, "Manual on disclosure control methods", Office for Official Publications of the European Communities, Luxembourg, p. 15, http://ec.europa.eu/eurostat/ramon/statmanuals/files/manual_on_disclosure_control_methods_1996.pdf

Privacy

Someone's right to keep their personal matters and relationships secret, involving an obligation of the holder of information to the subject of the information to do so.

Source: UNECE, Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes, Geneva, 2009, http://www.unece.org/fileadmin/DAM/stats/publications/Confidentiality_aspects_data_integration.pdf

Context: Privacy is a concept that applies to data subjects, while confidentiality applies to data. There is a definite relationship between confidentiality and privacy. Breach of confidentiality can result in disclosure of data which harms the individual. This is an attack on privacy because it is an intrusion into a person's self-determination on the way his or her personal data are used.

Process/Procedure

A series of actions or operations that make gradual changes leading towards a particular result.

Source: UNECE, Conference of European Statisticians Statistical Standards and Studies – No. 53, "Terminology on Statistical Metadata", Geneva, 2000, http://www.unece.org/stats/publications/53metadaterminology.pdf

Public disclosure

Public disclosure refers to the act of making information or data readily accessible and available to all interested individuals and institutions. Some examples of the different forms that public disclosure may take include: verbal or written statements released to a public forum, to the news media, or to the general public; publication in an official bulletin, gazette, report, or stand-alone document; and information posted on a website.

Source: "Code of Good Practices on Transparency in Monetary and Financial Policies", Part 1-Introduction; approved by the IMF Executive Board on July 24, 2000, http://www.imf.org/external/np/mae/mft/sup/part1.htm#appendix_III

Public Use Files (PUF)

A Public Use File (PUF) is a machine readable file containing microdata that:

- allows users to make sensible inferences on the phenomenon for which the data were collected, and
- has been subject to statistical disclosure control methods that render the data non-confidential according to the legal and methodological standards applicable to the NSO that has produced it.

Due to its non-confidential nature, it is provided:

- either in the form of Open data,
- or with some restrictions to some aspects of use of the file or provision of access to it, including registration and express agreement to terms of use.

Source: OECD Expert Group for International Collaboration on Microdata Access.

Raw microdata files

Raw microdata files contain all replies by each respondent obtained immediately after data entry.

Source: Olivier Dupriez and Ernie Boyko. 2010, "Dissemination of Microdata Files; Principles, Procedures and Practices" IHSN Working Paper No. 005, August 2010, p. 5, http://www.surveynetwork.org/home/sites/default/files/resources/IHSN-WP005.pdf

Remote access

Remote access systems combine the flexibility for researchers to do all their analysis in a research centre while removing the constraints of travelling to the NSIs. Modern developments in the internet make it possible to set up a safe controlled connection, a VPN (virtual private network). A VPN is a technique to set up a secure connection between the server at the NSIs and a computer of the researcher. It uses firewalls and encryption techniques. Also additional procedures to control the login procedure like software tokens or biometrics can be used to secure the connection.

The main idea of a remote facility is that it should resemble the 'traditional' on-site research centres as much as possible, concerning confidentiality aspects. The following aspects have to be taken into account:

- Only authorized users should be able to make use of this facility;
- Microdata should remain at the NSI;
- Desired output of analyses should be checked on confidentiality;
- Legal measures have to be taken when allowing access.

Sources: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf

Remote execution

Submitting scripts on-line for execution on disclosive microdata stored within an institute's protected network. If the results are regarded as safe data, they are sent to the submitter of the script. Otherwise, the submitter is informed that the request cannot be acquiesced. Remote execution may either work through submitting scripts for a particular statistical package such as SAS, SPSS or STATA which runs on the remote server or via a tailor made client system which sits on the user's desk top.

Source: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf

Research community

Although this mainly refers to people working in research institutions such as universities, it also includes researchers working in government agencies, NGOs, international agencies and the private sector. Some countries may want to define the research community more narrowly and only include those working in research institutions.

Source: UNECE, "Managing Statistical Confidentiality & Microdata Access - Principles and Guidelines of Good Practice", 2007, p. 106, http://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

Research files

See Licensed files.

Research purposes

Ad-hoc activities to investigate or explain economic or social phenomena, which result in statistical outputs. These activities may be undertaken by a statistical organization (in which case the results may not necessarily be published), or by external researchers (following the Conference of European Statisticians "Principles and guidelines on managing statistical confidentiality and microdata access").

Source: UNECE, Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes, Geneva, 2009, http://www.unece.org/fileadmin/DAM/stats/publications/Confidentiality_aspects_data_integration.pdf

Researchers

Professionals engaged in the conception or creation of new knowledge, products, processes, methods and systems, and in the management of the projects concerned.

Source: Frascati Manual: Proposed Standard Practice for Surveys on Research and Experimental Development, 6th edition, OECD, 2002.

Risk avoidance

This approach tries to eliminate all risks. In the case of microdata confidentiality, it requires the confidentiality of the data to be absolute, not only in its own right, but *also* in association with other available data.

Source: UNECE, "Managing Statistical Confidentiality & Microdata Access - Principles and Guidelines of Good Practice", 2007, p. 107, http://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

Note: The text in Italics was added by the OECD Expert Group

Risk management

Within the constraints provided by legislation, it involves identification of the risks and managing them in accordance with their significance (impact) and their likelihood. More effort is put into managing the high impact, strong likelihood risks. Microdata confidentiality may not be absolute when considered in association with other data. Confidentiality could be considered in association with other means of reducing the risk.

Source: UNECE, Managing Statistical Confidentiality & Microdata Access - Principles and Guidelines of Good Practice, 2007, p. 107, http://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

Safe centre

See also data enclave, data laboratory

Scientific use file

See also Licensed file.

Scientific-use files means confidential data for scientific purposes to which methods of statistical disclosure control have been applied to reduce to an appropriate level and in accordance with current best practice the risk of identification of the statistical unit.

Source: Regulation (EU) No 557/2013 on access to confidential data for scientific purposes, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:164:0016:0019:EN:PDF

Secure use file (or Master use file)

Confidential data for scientific purposes to which no further methods of statistical disclosure control have been applied.

Source: Regulation (EU) No 557/2013 on access to confidential data for scientific purposes, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:164:0016:0019:EN:PDF

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. Data sets and the equipment on which they are stored should be protected as well from environmental hazards such as heat, dust, electrical surges, magnetism, and electrostatic discharges.

Source: OECD (2007), *OECD Principles and Guidelines for Access to Research Data from Public Funding,* Paris, p. 20, http://www.oecd.org/science/scienceandtechnologypolicy/38500813.pdf

Statistical activity

The collection, storage, transformation and distribution of statistical information.

Source: UNECE, Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes, Geneva, 2009, http://www.unece.org/fileadmin/DAM/stats/publications/Confidentiality_aspects_data_integration.pdf

Statistical catalogue

A code list of statistical indicators, metadata, questionnaires, tables and other defined elements of a statistical information system.

Source: UNECE, Conference of European Statisticians Statistical Standards and Studies – No. 53, "Terminology on Statistical Metadata", Geneva, 2000, http://www.unece.org/stats/publications/53metadaterminology.pdf

Statistical confidentiality

See also Confidential data.

The protection of data that relate to single statistical units and are obtained directly for statistical purposes or indirectly from administrative or other sources against any breach of the right to confidentiality. It implies the prevention of unlawful disclosure.

Source: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf

Statistical data

Statistical data refers to data from a survey or administrative source used to produce statistics.

Source: Measuring the Non-Observed Economy: A Handbook, OECD, IMF, ILO, Interstate Statistical Committee of the Commonwealth of Independent States, 2002, Annex 2, Glossary, http://www.oecd.org/dataoecd/9/20/1963116.pdf

Statistical Data and Metadata Exchange (SDMX)

Statistical Data and Metadata Exchange (SDMX) is an initiative sponsored by BIS, ECB, Eurostat, IMF, OECD, UN and World Bank to address standardization of the exchange of statistical information.

Source: Statistical Data and Metadata Exchange (SDMX) – BIS, ECB, Eurostat, IBRD, IMF and OECD – Metadata Common Vocabulary, Release 1, December 2003, www.sdmx.org

Statistical Data Protection (SDP)

Statistical Data Protection is a more general concept which takes into account all steps of production. SDP is multidisciplinary and draws on computer science (data security), statistics and operations research.

Source: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf

Statistical disclosure

See also Disclosure.

Statistical disclosure is said to take place, if the dissemination of a statistics enables the external user of the data to obtain a better estimate for a confidential piece of information than would be possible without it.

Source: Eurostat, "Manual on disclosure control methods", Office for Official Publications of the European Communities, Luxembourg, 1996, p. 7, http://ec.europa.eu/eurostat/ramon/statmanuals/files/manual_on_disclosure_control_methods_1996.pdf

Statistical Disclosure Control (SDC)

See also Disclosure Control Methods

Statistical Disclosure Control techniques can be defined as the set of methods to reduce the risk of disclosing information on individuals, businesses or other organisations. Such methods are only related to the dissemination step and are usually based on restricting the amount of or modifying the data released.

Source: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf

Statistical macrodata

An observation data gained by a purposeful aggregation of statistical microdata conforming to statistical methodology.

Context: Macrodata is data derived from microdata by statistics on groups or aggregates, such as counts, means, or frequencies. (Survey Design and Statistical Methodology Metadata, Software and Standards Management Branch, Systems Support Division, United States Bureau of the Census, Washington D.C., August 1998, Section 3.4.4, page 39).

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

Statistical metadata (see also Metadata)

Metadata provide information on data and about processes of producing and using data. Metadata describe statistical data and - to some extent - processes and tools involved in the production and usage of statistical data.

Source: UNECE, "Guidelines for the Modelling of Statistical Data and Metadata", Conference of European Statisticians, Methodological material, United Nations, Geneva, 1995, http://www.unece.org/stats/publications/metadatamodeling.pdf

Statistical metadata is a data about statistical data, and comprise data and other documentation that describe objects in a formalised way.

Source: SDMX Metadata Common Vocabulary, http://sdmx.org/wp-content/uploads/2009/01/04_sdmx_cog_annex_4_mcv_2009.pdf

Statistical microdata (see also microdata)

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

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

Source: UNECE, "Terminology on Statistical Metadata", Conference of European Statisticians Statistical Standards and Studies, No. 53, Geneva, 2000, http://www.unece.org/stats/publications/53metadaterminology.pdf

Statistical purposes

It is particularly important to make a distinction between statistical, *scientific* and administrative uses. In the case of statistical use, individual data are used as an input to derive statistics that refer to a group of persons or legal entities. It may also incorporate support for other activities within a NSO (*e.g.* sample selection off a business register). Administrative uses concern decisions about a particular person or legal entity which may bring benefit or harm to the individual. The statistics referred to above include statistical aggregates, statistical distributions, parameters for models and other forms of statistical analysis that may refer to groups of individuals or organizations without identifying them. Microdata used for research is consistent with statistical purposes if it is being used to produce of statistics.

Source: UNECE, "Managing Statistical Confidentiality & Microdata Access - Principles and Guidelines of Good Practice", 2007, p. 106, http://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

Note: The text in Italics was added by the OECD Expert Group.

Statistical standard

A statistical standard provides a comprehensive set of guidelines for surveys and administrative sources collecting information on a particular topic. Components of a standard include:

- definition(s)
- statistical units
- classification(s)
- coding process(es)
- questionnaire module(s)
- output categories

Context: The use of statistical standards permits the repeated collection of statistics on a consistent basis. They also enable the integration of data over time and across different data sources, allowing the use of data beyond the immediate purpose for which it was produced. Standards also reduce the resource requirements associated with many aspects of survey development and maintenance.

Source: Statistics New Zealand, "Classifications and Standards", http://www.stats.govt.nz/domino/external/web/prod_serv.nsf/092edeb76ed5aa6bcc256afe0081d84e/35b11e7066c13db1cc256ca5006f44e4?OpenDocument

Statistical unit

Statistical unit means the basic observation unit, namely a natural person, a household, an economic operator and other undertakings, referred to by the data

Source: Regulation (EC) No 223/2009 on European statistics

Statistical unit for macrodata

A statistical unit which is a carrier or a supplier of statistical macrodata in statistical system.

Source: UNECE, Conference of European Statisticians Statistical Standards and Studies – No. 53, "Terminology on Statistical Metadata", Geneva, 2000, http://www.unece.org/stats/publications/53metadaterminology.pdf

Statistical unit for microdata

A statistical unit which is a carrier or a supplier of statistical microdata in statistical system.

Source: UNECE, Conference of European Statisticians Statistical Standards and Studies – No. 53, "Terminology on Statistical Metadata", Geneva, 2000, http://www.unece.org/stats/publications/53metadaterminology.pdf

Supplementary data

In SDMX, "Supplementary Data" refers to a description of data not routinely disseminated that are made available to users upon request. It may include customized tabulations that can be provided (perhaps for a fee) to meet specific requests. Also include information on procedures for obtaining these supplementary data.

Source: Statistics Canada, "Statistics Canada Quality Guidelines", 3rd edition, October 1998, page 59, http://www.statcan.ca/english/freepub/12-539-XIE/12-539-XIE.pdf

Sustainability

Due consideration should be given to the sustainability of access to publicly funded research data as a key element of the research infrastructure. This means taking administrative responsibility for the measures to guarantee permanent access to data that have been determined to require long-term retention. This can be a difficult task, given that most research projects, and the public funding provided, have a limited duration, whereas ensuring access to the data produced is a long-term undertaking. Research funding agencies and research institutions, therefore, should consider the long-term preservation of data at the outset of each new project, and in particular, determine the most appropriate archival facilities for the data.

Source: OECD (2007), *OECD Principles and Guidelines for Access to Research Data from Public Funding*, Paris, p. 22, http://www.oecd.org/science/scienceandtechnologypolicy/38500813.pdf

Tabular data

Aggregate information on entities presented in tables.

Source: A network of Excellence in the European Statistical System in the field of Statistical Disclosure Control – ESSNet SDC (2010), Handbook on Statistical Disclosure Control, Version 1.2, http://neon.vb.cbs.nl/casc/.%5Csdc_handbook.pdf

Teaching File

A Teaching File shares the same characteristics of the Public Use File in terms of provisions to users but it is different in terms of content. Due to the basic statistical disclosure control procedures used for its creation is not meant to make proper statistical analysis or to make inferences based on its microdata but it can be used solely to teach specific statistical methods or to make data manipulation.

Source: OECD Expert Group for International Collaboration on Microdata Access.

Test Data

Test data are microdata files that share the same logical structure of complex confidential microdata sets that statistical agencies make available to researchers through remote execution or remote access. The scores and values of the variables in the test data are either generated through stochastic processes or through procedures that destroy any relationship with real statistical units. These data aim at preserving consistency inside the record as well as complex structures and relationships between variables in order to allow users to develop correct code for the analysis of the microdata.

Source: OECD Expert Group for International Collaboration on Microdata Access.

Timeliness

Speed of dissemination of the data, *i.e.* the lapse of time between the end of a reference period (or a reference date) and dissemination of the data.

Context: In SDMX, "Timeliness and Punctuality" is a single entity. Timeliness refers to the speed of dissemination of the data. It reflects many factors, including some that are related to institutional arrangements, such as the preparation of accompanying commentary and printing. Punctuality refers to the possible time lag existing between the actual delivery date of data and the target date when it should have been delivered, for instance, with reference to dates announced in some official release calendar or previously agreed among partners.

Source: International Monetary Fund (IMF), "Guide to the Data Dissemination Standards, Module 1: The Special Data Dissemination Standard", Washington, May 1996, http://www.imf.org/external/bopage/pdf/mar2000.pdf

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. Lack of visibility of existing research data resources and future data collection poses serious obstacles to access. Factors to consider in ensuring transparency include:

- Information on data-producing organisations and their holdings, documentation on available data sets *and their production process* and conditions of use should be easy to find on the Internet.

- Research organisations and government research agencies should actively disseminate information on research data policies to individual researchers, academic associations, universities and other stakeholders in the publicly funded research process.

- Whenever relevant, all members of the various research communities should assist in establishing agreements on standards for cataloguing data. The application of existing standards should be considered, whenever appropriate, in order to avoid placing additional burdens on research resources and work loads of researchers and their institutions.

- Information on data management and access conditions should be communicated among data archives and data producing institutions, so that best practices can be shared.

Source: OECD (2007), *OECD Principles and Guidelines for Access to Research Data from Public Funding,* Paris, p. 15, http://www.oecd.org/science/scienceandtechnologypolicy/38500813.pdf

Note: The text in Italics was added by the OECD Expert Group.

PART II. TRUST

CHAPTER 4. ESTABLISHING TRUSTED PARTNERS IN DELIVERING MICRODATA SERVICES

by Maurice Brandt

Introduction

46. Access to official microdata is basically available in many countries for residents. On the contrary, trans-border access to microdata, *i.e.* access to the official microdata that are held by NSOs and statistical agencies of foreign countries, is still an emerging activity. There are several reasons for that, including legal issues, appropriate technical infrastructure and missing standards.

47. The objective of this chapter is to propose a "concept of trust" for international microdata access which relates to the different possible levels of data confidentiality and security requirements necessary to access those data. According to the level of data confidentiality there are different grades of risk assessments: while original data (non-anonymised data without direct identifiers) have a very high risk of disclosure, public use files have almost no risk at all.

48. The producers of official statistics in most countries would only provide data to a third party or another country if they are obliged by law, or if it is explicitly allowed by law. In this context, there can be a big difference between delivering microdata (physically transferring microdata to the custody of another organisation based in a different country) and giving access to microdata. Even when data producers have to deliver data physically to a super-ordinate institution, they would like to keep control of who is accessing the microdata. At the same time, the data producers are willing to give access to their data to international researchers on the basis of their Statistics Act. Depending on the national law, it is already possible for some countries to do so. For other countries it is only allowed under special conditions and for some it is completely prohibited.

49. There seems to be a solution that makes it possible to guarantee access to the data and keep the data still in a secure environment. Under safe settings, the possibility of remote access to confidential microdata can be considered as a secure and convenient way of providing access for scientific purposes. Virtual PCs and cloud computing can support the technology of remote access even when there are still details to solve, *e.g*. the assessment of IT securities for the different existing systems. However, the mode of remote access is not implemented and accepted by all countries, and just a very few have a Statistics Act that allows to provide access in this way.

Possibilities for sharing microdata services

50. Even without delivering any data physically, it is possible to give access to the microdata whereas the microdata itself will remain in the secure environment of the data producing authority. There are different modes of access possible which require certain preconditions for the accessing institutions or persons. Usually the access to official microdata is provided for scientific purposes to independent research entities.

51. For confidential data this can be done via a safe centre inside the access facility or via remote access solutions. Scientific use files (anonymised files that are still confidential) are often not considered as sufficient from a researcher's point of view, because they are not detailed enough to run sophisticated analyses or good comparisons between countries. The future development of user demands tends to

detailed microdata that can be accessed preferably from a researcher's own workstation inside his own institution. The remote access solutions are taking this user demands into account. The other modes of access have their justification as well, depending on the purpose of use. Sometimes it is not even necessary to use highly confidential data, or the service and support in a safe centre is just more convenient for the researcher than using remote access on his own.

52. For security or organisational reasons it is also possible to combine the different ways of access, like a safe centre and remote access. Usually the microdata are physically in a safe centre, but in combination with remote access it is also possible that the data are accessed from a safe centre whereas the secure server where the data are stored is located somewhere else. This might be a good approach if the data producing authority does not send out any data and the access is not allowed to be granted at the researchers' institution. In this case, a researcher can visit a certified safe centre that is connected remotely to the data at another countries server. Based on the different modes of access combined with a different level of data confidentiality, several zones of trust can be implemented.

The concept of "circle of trust"

53. The concept of "circle of trust" is based on the agreement that each member is accepted according to the same rules and conditions that are approved by all members. In the context of statistics, this mainly refers to confidentiality rules and security requirements but also to competence and legal aspects. This makes it possible to create a group/membership of trust. According to the level of confidentiality and risk of disclosure, one may envisage different zones of trust in the circle, whereas the inner circle is the most sensitive one with the highest disclosive risk.

54. A valid question is why a concept of trust is needed, if everything in terms of international microdata access is regulated by the national Statistical Act. If the Statistical Act prohibits granting access to foreign third parties, there is no trust needed; if it allows giving access, there might be no trust needed as well. However, there is also the situation where access for another country or a third party is not mentioned in the national law, implying that it is not forbidden from a legal point of view. There are two interpretations of the law in this case. Some countries treat everything that is not explicitly forbidden as allowed and some countries treat everything that is not explicitly allowed as forbidden. Therefore, even when giving access to third parties is not forbidden, some of the National Statistical Authorities still do not want to give access as long as they are not obliged by law. From this perspective, the "circle of trust concept" is needed for three reasons.

55. The first one is that a "circle of trust" can bridge the gap for those countries where the law does not mention anything at all. For these countries it would be helpful to have a basic concept of requirements so that they are reassured that their data are accessed in an organised and secure manner, comparable with the security requirements in their own country.

56. The second reason is that a circle of trust can reassure those countries that basically allow access to foreign third parties. A law can be broken, but it is possible to prevent a breach of confidentiality with an organisational or technical solution and a reasonable amount of effort. Their data will be accessed at least under the same security standards as in their own country.

57. The third reason is that those countries that do not allow microdata access to foreign third parties are encouraged to reconsider their position; if the data are accessed in a very secure way at least comparable with their own standards, there is actually no practical reason to refuse microdata access in the long run. There is always the opportunity to deliver anonymised datasets at the beginning.

58. The concept of trust does not imply that data producers simply have to trust that their data will not be misused or revealed. The trust is generated according to strict rules and standards that each participating party has to fulfil. These rules could include:

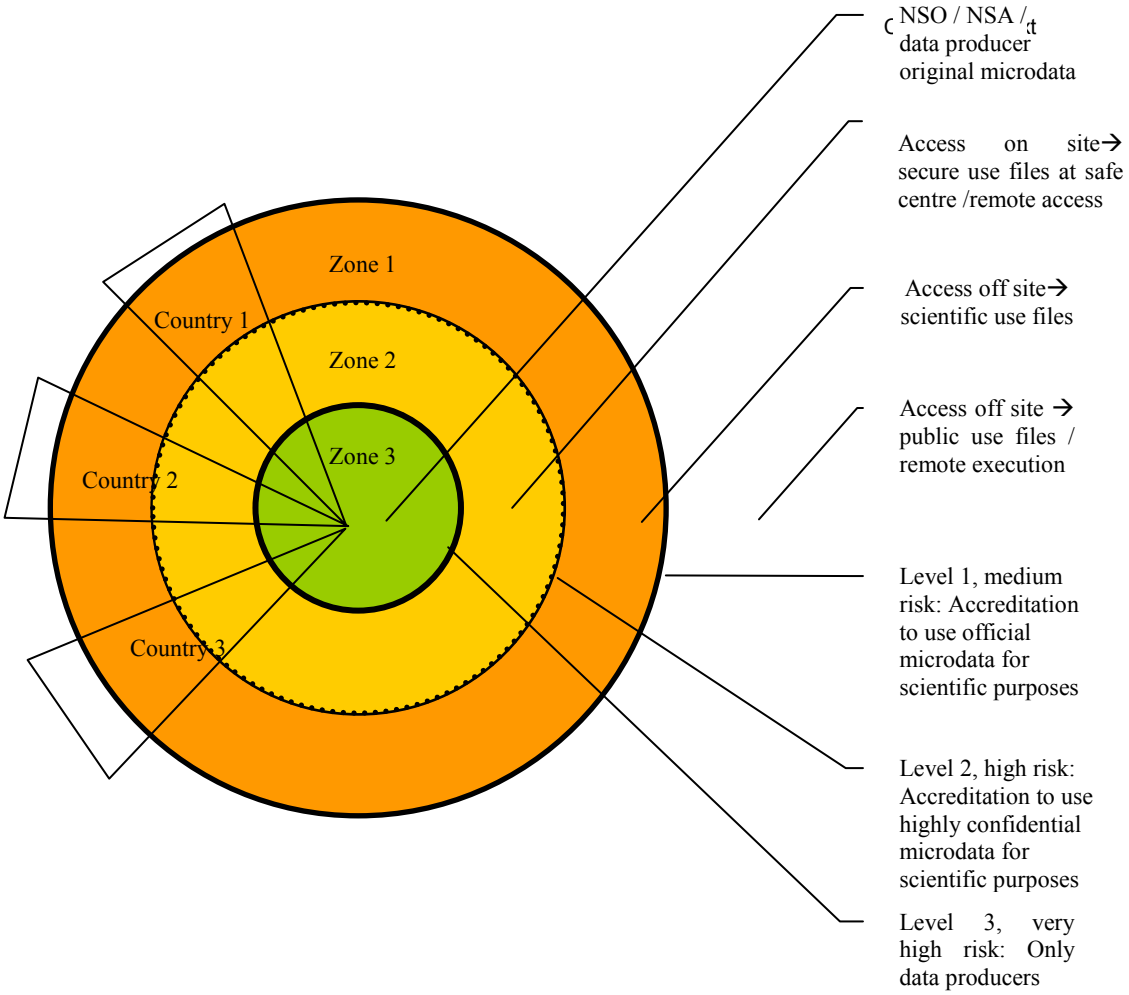
- Glossary of terms and definitions
- Share best practices
- Collection and documentation of rules and protocols for transparency
- Cooperation agreement
- Harmonised contracts for microdata access
- Guidelines for the treatment of microdata requests
- Catalogue of rules to check what institution is approved to access microdata
- Security concept and accreditation for safe centres
- List of security and user demands for a remote access system
- Anonymisation concept for scientific use files
- Rules and protocols for the transmission of microdata
- Guidelines for statistical disclosure methods and output checking
- Training courses
- Common understanding of responsibilities

59. Preconditions can be set up with regard to the institution requesting access and to the specificities of the technology for providing access. The minimum requirement for most data producers is that data that are taken from another body are treated at least under the same or higher level of confidentiality as in their own premises. Figure 1 shows the different zones of trust, based on the degree of data confidentiality with their corresponding risk of disclosure, and the kind of access adequate to each zone. The degree of confidentiality goes from highly-confidential microdata with a very high disclosive risk in the inner circle to less confidential with a very low or no risk of disclosure in the outer circle. Outside the circle only non-confidential with almost no risk at all are provided (such as Public Use Files) to virtually any user.

60. The zones themselves represent the microdata with their different levels of risk, their corresponding modes of access and the institutions that are accredited to access the microdata. The lines between the circles represent the different levels of confidentiality with their associated risk and therefore the requirements that an institution has to fulfil to access the data in the corresponding zone. To enter the circle at all to Zone 1, a minimum of criteria has to be fulfilled and level 1 needs to be passed. This can be done by the application of a set of measures to assure that the disclosure risk will be minimised. Some countries may not differentiate between Zone 1 and 2 (dashed line) and there is only a binary decision. Whether an institution is accredited as scientific or not and if they are considered as scientific, they can access all kind of data. The triangles represent the countries whereas a country can host the data producer and the data requesting institution as well. The apex of the triangle in the green Zone 3 is only for the data producing authority and can usually not be accessed by a scientific institution.

61. As a first trust building mechanism, an agreement on the rules and standards for joining the "circle of trust" could be helpful and also a common understanding on which parties are expected to be in the inner or outer circle. A starting point would be to define a group of countries, *e.g.* only the OECD member countries, as authorised to enter the circle.

Figure 4.1. The circle of trust and its sub-zones



62. It would be advantageous, if a research institution that is accredited by one country for Zone 2, would be automatically accredited in another country for Zone 2. This means the institution can use the secure use files of all countries in the ring of Zone 2. Before that, an investigation is necessary whether all countries have the same understanding of the confidentiality levels of the different zones. Also, a harmonised accreditation system need to be applied that has been agreed by all countries.

63. Overall, there are still several open questions to be answered, and such an ambitious system of international microdata access should be future-proof in terms of technology, so that it can be also used in five or ten years without too much additional investment.

64. In case of new technologies, such as cloud computing, data can be stored virtually and temporarily in a central place to prepare an international dataset starting from data of single countries. This implies processing of the data and might be seen as very sensitive, because the microdata are somehow transmitted and can be seen and edited at another place, even if it is a virtual space. An evaluation of the

legal aspects involved in these new technologies, including virtual working environments, needs to be performed at the national level in order to clarify the legal situation.

Concluding remarks

65. In terms of trust, it is very important to consider what already exists and to build on that. This can be an already existing infrastructure or a best practice system in a single country. The authorities producing the data also depend on the trust of the respondents, which means that a high quality production system for official statistics can only be guaranteed as long as this trust can be maintained. Trust is of course not a one way street that goes only in one direction. It needs to be proved over many years and one single bad experience can destroy all the investments and efforts built in the past.

66. The responsibility for that lays with the data producers, because respondents do not care who is responsible for the breach of confidentiality at the end. Their only possibility of protection is to refuse to participate in a survey or, in case they are obliged by law, provide incorrect answers. This scenario would decrease the quality of official statistics and would affect both data producers and users. Therefore, it is important that the producers have control of the admissibility criteria and of the decision on the parties admitted to accessing their data.

67. The proposed approach for addressing the complexity of microdata access, especially at the international level, is not a system that can be established rapidly; instead, it entails progressive developments in gaining mutual trust, experience and competences in the area of microdata management.

CHAPTER 5. SANCTIONS FOR BREACH OF CONFIDENTIALITY

by Brain Negin, Paul Jackson and Aleksandra Bujnowska

Introduction

68. Systems for official statistics established in compliance with the *UN Fundamental Principles of Official Statistics* operate within public laws and regulations, and guarantee the confidentiality of the data they collect (Principles 7 and 6 respectively). The principles must be maintained when confidential data are transferred from one statistical system to another.

69. Transnational access to confidential data is inhibited when a statistical agency considers that the confidentiality guarantee cannot be maintained at the destination of the data. An important feature of the confidentiality guarantee is the ability to sanction any breach of confidentiality. To that end, Principle 10 is engaged, *i.e.* bilateral and multilateral cooperation is essential.

70. Considering that, this document addresses the frequently-expressed need for a legal solution to permit exchanges of confidential data between the statistics offices of different countries. That legal solution might be expressed in three parts: i) *Power, i.e.* a lawful authority to supply confidential data to another party is necessary; ii) *Permission, i.e.* the power must be exercised where there is permission to do so in regulatory and policy frameworks and where no prohibition on such disclosures exists; and iii) *Penalty, i.e.* sanctions for the protection of confidential data must be applicable. This document addresses the third component, only.

Box 5.1. Sanctions, offenders and exchange

The meaning of "sanction" is important. Breaches of confidentiality can be placed on a sliding scale, from the failure of staff or researchers to observe a rule of conduct, through to the criminal misuse of the privacy of information. "Administrative sanctions" are relevant to the lower end of the scale, and may include suspension from further access to data, or suspension from employment. "Contractual sanctions" may lie in the middle of the scale, where the breach of an agreement can be addressed by civil actions. "Criminal sanctions" lie at the high end of the scale, where the offence is against the state and convictions in a court carry fines or even imprisonment.

The concept of "offender" is important too. It is easy to conceptualise breaches by natural persons outside the NSO holding the data. However, the NSO itself might be the offender - either one of its staff, or institutionally. This can affect the nature of identifying a breach and the course of action to take.

The concept of “exchange” of data is relevant. Confidential data is exchanged when there is even a fleeting fixation of those data, or the confidential information derived from those data, in a second location. Thus the bulk transfer of data to another party is clearly in the scope of this document, but so is remote access to data where the user can see the confidential data or derived confidential information on their screen through a linked terminal.

71. From a legal perspective, the best way for data to cross international borders is through publication or Open Licence. The benefits and challenges of creating and disseminating Open (micro)Data are presented in Chapter 11. The analysis presented in the current document is relevant only if Open Data cannot meet user needs and/or the risks in producing a suitable Open product are too great.

72. This chapter proposes that a statistical agency should not be inhibited to exchange data where it is assured that the agency at the destination of the data is both willing and able to sanction any breach of

confidentiality to a standard that meets or exceeds its own public laws and regulations. A model relationship is proposed. Recognising that these assurances may not be available for all OECD countries, the chapter also analyses whether the use of remote execution can obviate the needs for such assurances.

Legislative and regulatory framework for applying sanctions

73. Systems for official statistics are typically established by a core legislative framework, a key component of which is statutory protection for the privacy of respondents to statistical inquiries. The framework is made up of sanctions expressed specifically in statistical laws, and laws with a regulatory effect such as Data Protection legislation. The precise nature varies from system to system, but typically the following necessary features of the legislative and regulatory framework for applying sanctions are found:

- The data and their protection are defined.
- The legal person(s) responsible for protecting confidentiality is established.
- The offence of a breach of this protection is defined.
- The statutory sanctions in the event of such a breach are described.

74. Operationally, all public administrations establish regulations, rules, guidance, and public commitments to assist them in applying sanctions where a breach occurs. They establish:

- Whether certain data are within the scope of the definition "confidential".
- Whether the breach has a statutory or administrative sanction.
- The process for reporting statutory offences to a prosecuting authority.
- The process for addressing breaches of rules and procedures relating to maintaining confidentiality.

75. Confidence in using legislative and regulatory sanctions to maintain confidentiality is a pre-condition for giving access to confidential data. The experience of this Expert Group and other initiatives to explore transnational exchange of confidential data (for example the European Commission funded project “Data Without Boundaries” and discussion in the European Statistical System Committee) is that this confidence is lost when the potential breach of its confidentiality is outside the territory of the original statistical system.

76. The lack of confidence arises from one of, or a combination of, five substantial issues:

- Absence of an express legislative and/or regulatory provision for managing confidentiality outside the territory of the original statistical system.
- Differences in definitions, modalities, and access regimes in data exchange between statistical systems.
- Differences in sanctions and regulatory protections, and whether the original statistical system can employ them to penalise a breach of confidentiality outside their territory.
- Difficulties inherent in the extraterritorial enforcement of any penal law, including issues of cooperation between investigating authorities.
- Difficulties in identifying who and where the offender is, and whether sanctions can be applied to them. In particular, where the breach has an institutional cause rather than the deliberate action of a natural person.

77. Clearly, the sources of concern are directly related to the necessary legislative features.

78. The exchange of confidential data between countries has evidently not been a priority for the development of the statutory and regulatory frameworks for national statistical systems, because express provision is rarely made for this scenario; yet, on the other hand, it is rarely expressly prohibited either. Outside of the European statistical system there has been little or no cause to ensure harmonisation of the key features of the legislative and regulatory frameworks of national statistical organisations. There are substantial variations in detail in the way the key features of statutory and regulatory sanctioning of breach of confidentiality is carried out. For most statistical systems, these issues are barriers to international exchange of confidential data.

Transnational access to microdata: Analyses of two options

79. Due to the complexity of the legal framework for multi-participant, multi-location, multi-purpose transnational exchange of confidential data, practitioners cannot expect a single simple solution. There is no prospect of a new transnational statistical sanction law with direct and equal effect in every country. Organisations with a global or regional remit for statistics (UNECE, OECD, ESS) and which may have laws and administrations that cross boundaries are prohibited through legal principles such as subsidiarity, and by policy, from direct statistical actions in the territory of states. Nor is there any prospect of the systematic harmonisation, or integration, of national laws to remove differences.

80. Instead, this chapter explores two options for transnational access.

- ***Option 1.*** It rests on exploiting the relatively homogenous national legal frameworks for Official Statistics. This option relies on bilateral agreements between institutions that include undertakings to enforce criminal and regulatory sanctions for breach of statistical confidence occurring within their own borders irrespective of the origin of the data. In effect, the parties to the agreement agree to place their data in each other's' care, and to use, protect, and sanction the use of the other party's data *as if they were their own data*.
- ***Option 2.*** This option relies on remote execution mode, which minimises reliance on cross-border enforcement of legal norms.

81. Heads of NSOs make confidentiality undertakings and guarantees to the subjects of their statistical inquiries, which they feel they cannot maintain when their data are outside their national jurisdiction. They cannot and should not authorise a transnational exchange of confidential data unless these undertakings and guarantees can be maintained. The two options discussed in this chapter are therefore relevant if NSOs meet that simple challenge.

Option 1: "Mutual Statistical Assistance Agreements"

82. A mutual legal assistance treaty (MLAT) is an agreement between two countries for the purpose of gathering and exchanging information, which may include evidence, in an effort to enforce the law. MLATs provide a useful concept for addressing the inhibitions in international exchange of confidential statistical data. This type of agreement is often used in tax matters, in particular as part of international double taxation agreements wherein the parties agree to deliver information for tax purposes.

83. It is proposed to consider the creation of Mutual Statistical Assistance Agreements (MSAAs) that could perform the following functions:

- Identify the NSOs in question
- Describe the data to be exchanged and the purposes for which those data may be used

- Describe the confidentiality undertakings and guarantees, and the corresponding legislative and regulatory sanctions these data in each NSO are subject to (including relevant non-statistical sanctions such as data protection regulations)
- Express the equivalence of the undertakings and guarantees, and legislative and regulatory sanctions, in each territory
- Clarify that the confidential data of signatory Country A is protected by the relevant laws of Country B, and vice versa. It is of the first importance is that the recipient NSOs can bring the data provided by the donor NSOs into the scope of its undertakings, guarantees, and sanctions.
- Describe how use of data is to be authorised when those data are in the possession of the other party
- Express the commitment of each NSO to identify and act upon breaches of the confidentiality of data originating from the other party *as if they were their own data*
- Describe how breaches of confidentiality in each Country are sanctioned once identified, including identifying the relevant prosecuting authority
- Acknowledge that the recipient NSO may themselves be the cause of the breach of confidentiality, and an offender, and subject to the criminal and administrative sanctions of their state.
- Describe how each NSO will inform the other about breaches, and progress with sanctioning them
- Reference other important documents such as information risk assessments and information security standards

84. Importantly, the concept of a MSAA has two main preconditions for its adoption: i) that the data can be exchanged lawfully in the first place; and ii) that the parties to it trust each other to uphold the agreement and act according to its terms.

85. Once agreed, the MSAA would govern the use of multiple exchanges and uses of data that fall within the terms of its agreement.

86. Clearly, in preparing such an agreement the parties may discover that one or more of its articles cannot be achieved. For example, it may be discovered that the legislation of the recipient NSO cannot protect through criminal sanction any confidential data it did not itself collect. The parties could agree to proceed knowing this limitation by giving extra weight to the non-criminal sanctions that are available, or could decide that no exchange of confidential data will be possible until that issue is addressed. In this way, preparing a MSAA is beneficial even if exchanges of data are not immediately enabled. Their preparation would focus the parties on the issues that need to be addressed.

87. Perhaps, the most important element of the agreement is the ability to adopt the confidential data of another party outside your national border and include it in your legislative and regulatory framework for sanctions. There are examples where this is known to be the case. The UK's Statistics and Registration Service Act 2007, for example, places all the data held by the Office for National Statistics under the same non-disclosure rule and the same criminal sanction for wrongful disclosure. The legislation does not discriminate on the basis of the origin of those data held by ONS, neither by instrument (survey, register, administrative record, census, etc) nor by supplier (ONS's own instrument, another department, local authority, private sector supplier, etc), nor by territory (devolved administration of the United Kingdom, another country). If another party is willing and able to supply data to ONS, ONS is obliged by its legislation to protect it.

The special case of research access to data

88. A 2012 survey of countries subject to the *UN Fundamental Principles for Official Statistics* reports 109 countries have practices for granting access to confidential data for researchers, and of these 64 have such access provided for and regulated in their legislation.[5] In most cases the legislation is supplemented with contractual obligations and administrative sanctions agreed with the researcher. In detail the legislative and contractual arrangements are different from one country to other, and also differ from one authority to another within a country. However, it is highly likely their intention and effect is intended to be similar.

89. Most statistics agencies have provisions in their statutory framework for research access to a version of the data they have obtained, and statutory protections for those data when used for research purposes. If a purpose of the international exchange of data is to enable researchers in the recipient country to use those data, this should be explicitly addressed in the MSAA.

The special case of confidential data held by Eurostat and supplied for research purposes

90. European legislation with direct effect in all EU member states provides producers of European statistics with the power and the permission to provide confidential data to Eurostat. It also provides Eurostat with the power and permission to further supply those confidential data to recognised researchers working in accredited facilities on approved projects. However, due to the way European law operates, European legislation cannot contain provisions to penalise a breach of confidentiality during research use of data. Member states are obliged to bring forward national legislation and administrative procedures to sanction breaches of confidentiality of European statistics data, including data re-supplied for scientific research use. Annex II.A1 provides detailed information about provisions on sanctions.

91. A recent survey of the European Statistical System has discovered that only 12 member states say that their legislation can sanction the breach of confidentiality of European statistics data being used in their state that originates from other countries, through Eurostat. Only 14 member states say that they have administrative sanctions in place for addressing misuse of European statistics data used in this way.

Concluding remarks on Mutual Statistical Assistance Agreements

92. It is interesting that even in the most coherent, most comprehensive, and longest-established multinational legal framework for data exchange, *i.e.* the EU legal framework - one that is expressly designed in its statistical and regulatory laws to enable the exchange of confidential data - less than half of the NSOs of the partnership can bring a legislative sanction for a breach of confidentiality in their territory when the data originate in another partner country, by their own admission.

93. This may argue in favour of the greater use of MSAAs. Annex II.A2 presents further arguments to support the establishment of MSAAs.

94. However, the limitations of the EU multinational legal framework for data exchange also argue strongly in favour of enabling the derivation of the necessary information from the confidential data without those confidential data themselves being exchanged at all. Remote execution provides a solution of this type.

[5] http://unstats.un.org/unsd/statcom/doc13/BG-FP.pdf, paragraphs 109, 110.

Option 2: Remote execution

95. Option 1 relies heavily on legalities and therefore has limitations. An alternative solution is remote execution, which virtually eliminates the reliance on legalities and puts the burden on technological means for protecting confidential data.

96. With remote execution, a researcher never sees confidential data and therefore cannot breach statistical confidentiality through permitted use of the remote execution system. The responsibility of ensuring confidentiality falls on the NSO allowing access to its resources, which must vet research output to ensure that it does not violate the confidentiality laws of its own country. This solution, while posing challenges to researchers and NSO staff, can make confidential data available for use in any reasonable circumstance.

97. Remote execution relies on the submission of scripts on-line for execution on confidential data stored within an NSO's protected network. The researcher never sees the confidential data themselves. Output is vetted to ensure that it is safe before it is returned to the researcher.

98. Since the researcher is never exposed to confidential data, and since output is vetted to ensure that no confidential data is returned to the researcher, the researcher should never be in a situation in which it is possible for him or her to breach statistical confidentiality. Thus penal sanctions for breach of confidentiality should never be an issue. Subsequently, one need not look into the sufficiency of legal systems and penal sanctions in the country from which a researcher submits a script for remote execution. Nonetheless, terms of access and sanctions for breach of contract must be established between the parties. The term "parties" is left vague here, since a researcher submitting a script for remote execution does not necessarily have to do so via an NSO. Thus a contract could be created between the NSO performing the remote execution and the researcher submitting the script. While such a direct contract between researcher and NSO is possible, for cross-border access, it would be preferable that the submission of the script and the receipt of the vetted output be carried out via a trusted NSO. This is to affect supervision over the researcher to ensure that he/she does not attempt to “hack” the system or introduce hostile code into the host computer and to ensure that expenses for running the script and vetting the output are covered.

99. The drawbacks of this proposal are technological and, to some extent, methodological. Researchers must be exposed to information about the data sets that is sufficient to enable them to write their scripts. Writing the scripts also may be challenging to ensure that they actually do what is desired. As a result, the extent to which research can be carried out could be hindered.

Box 5.2. Canada: Real Time Remote Access

Statistics Canada has a remote execution system in place, called "Real Time Remote Access" ("RTRA"). Information is managed in the RTRA System through a combination of some corporate systems and some specially designed processes. Following the 4 stages of the process flow:

a. Input: RTRA data holdings are anonymized and very slightly masked, stored at Statistics Canada

b. Process for access: All information around the researcher approvals are captured in Statistics Canada's Client Relationship Management System, a corporate tool for managing client information. Access permission is verified against this information in the authorization step. Code/syntax is submitted through an air gap using Statistics Canada's Electronic File Transfer system for secure transmission to the pre-scan process.

c. Process for research: Code/syntax is sent through a custom built pre-scan process to ensure no harmful practices are being used, then run against secure confidential data in Statistics Canada using SAS (Stata also in future). Researchers use Statistics Canada's metadata repository (Integrated Metadata Base) on the website for information about the data. Dummy files to test syntax are available on request. No temporary files are retained.

d. Output: Confidentiality vetting occurs automatically by a Disclosure control post scan developed by our methodologists to handle disclosure and residual risk. The processing also produces quality indicators, since the researcher cannot see data and therefore cannot assess quality without the indicators. Once the non-confidential outputs are returned to the researcher, no data outputs are retained.

International researchers are not required to undergo any additional steps for access beyond those required of all researchers.

Source: STD/CSTAT/MICRO(2012)7, The Process Flow for Microdata Exchange: Two Canadian Examples.

Concluding remarks on remote execution

100. Remote execution provides an effective tool for cross-border researcher via technology that exists and continues to be developed. Its reliance on technology instead of on penal sanctions to protect confidentiality seems to better respond to the challenges that emerged from the review of national statistical laws criminalising the breach of statistical confidentiality.

101. On the basis of the analysis conducted, the Expert Group concluded that: i) NSOs and other statistical agencies should not overly rely on penal sanctions as a means of protecting confidential statistical data, in particular in trans-national exchanges of microdata; ii) Access to microdata should be carried out in a manner that effectively protects the data from misuse and also prevents breach of confidentiality based on normal legal use of a computer system; iii) Administrative sanctions are recommended for researchers or their institutions for breach of confidentiality.

102. In addition, the proposal provides a solution for all OECD countries, and does not distinguish between personal data of natural persons and personal data of legal persons.

CHAPTER 6. STANDARIDISED APPLICATION PROCESS FOR MICRODATA ACCESS

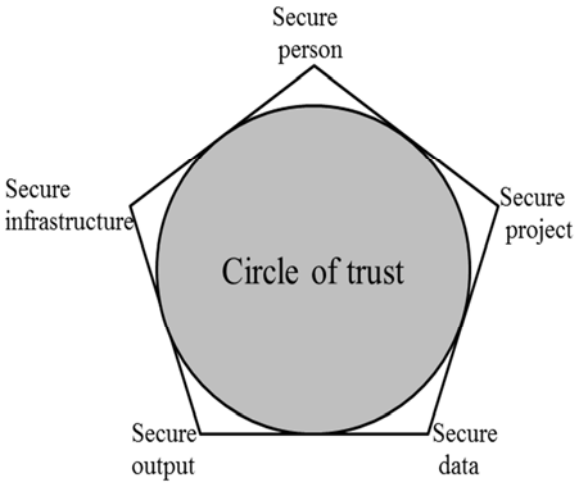
by Natalia Volkow

Introduction

103. As part of their data dissemination policy, many national statistics offices (NSOs) and other statistical agencies provide access to microdata for research purposes; transborder access is, in some cases, also allowed. NSOs and other statistical agencies recognise the need of researchers to work with microdata, but they need to safeguard data confidentiality.

104. In the Generic Statistical Business Process Model (GSBPM), access to microdata can be considered within the phase "Build", in particular the sub-phase "3.2 Build or enhance process components".[6] The IT infrastructure for microdata access will vary depending of the type of access that each NSO and other statistical agencies provide. Also, each type of access (*e.g.* on-site, remote access, remote execution, PUFs) implies different requirements for the user.

105. This document proposes a model for structuring the process of providing access to microdata, based on the approach to risk management presented by Desai and Ritchie in their work on "Effective Researcher Management" (2009). In the context of microdata access, risk management is aimed at creating a "circle of trust" between data providers and users. Five elements are relevant: *infrastructure, person, project, data and output*; each of them need to be "secure" in order to create the secure space where researchers can have access to microdata. The five elements are complementary, and constitute different dimensions of the same problem (Ritchie, 2009).[7]



[6] http://www1.unece.org/stat/platform/download/attachments/8683538/GSBPM+Final.pdf

[7] http://www.academia.edu/1336375/UK_Official_Microdata_Release_Practices

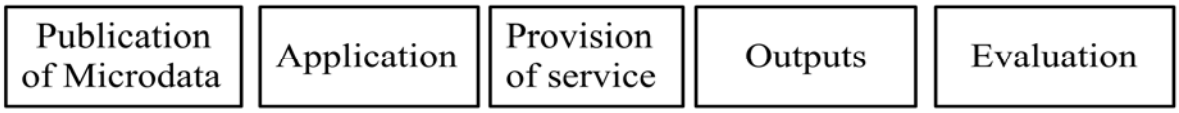
106. In particular, "secure" should be interpreted as follows:

- *Secure infrastructure* - Secure IT environment through which access is provided. This can also include physical secure premises within an NSO or other statistical agencies or outside.
- *Secure person* – An approved researcher is someone that meets the requirements outlined (e.g. he has filled in an application form, provided personal data and a CV, and be accredited by a research institution recognised as such) and whose results will be publicly available. Users need to.
- *Secure project* - In their application for data access, users must describe the study they intend to undertake; the objective of the research project has to be valid and serve the public good.
- *Secure data* – Data have no direct identifier, and or have been treated to limit disclosure risk. Data are provided just upon need, need that is justified by the aim of the research project.
- *Secure output* – Once users finish their processing of the data, the concerned NSOs or other statistical agencies check the output to validate disclosure risk and apply statistical disclosure control methods, if necessary.

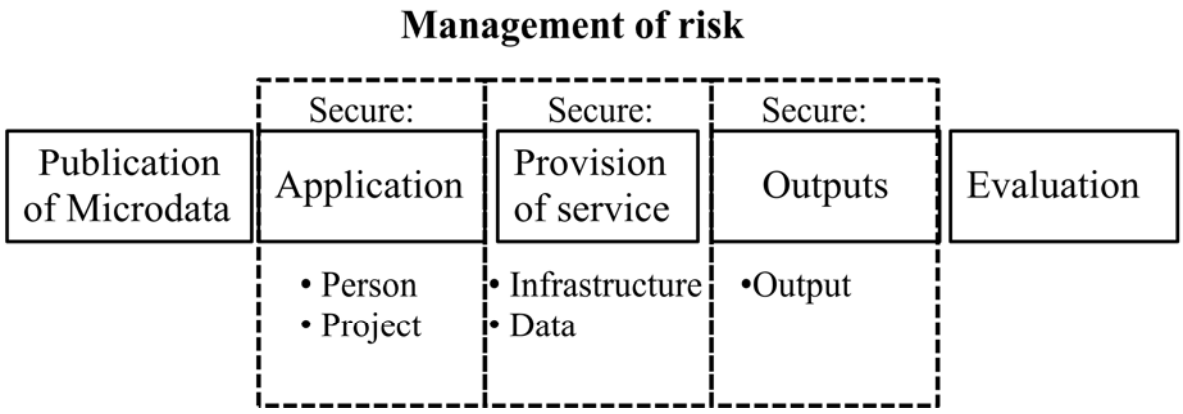
Structuring the process of access to microdata

107. While the modes of access to microdata can differ, there are always common elements in the process that leads to providing microdata access. The standardisation of the core elements of the process can simplify and reduce time and costs for microdata access; also, this could benefit substantially transborder access. The aim of standardisation is to create common steps, which researchers would know they will have to fulfil for any microdata requests, in any countries.

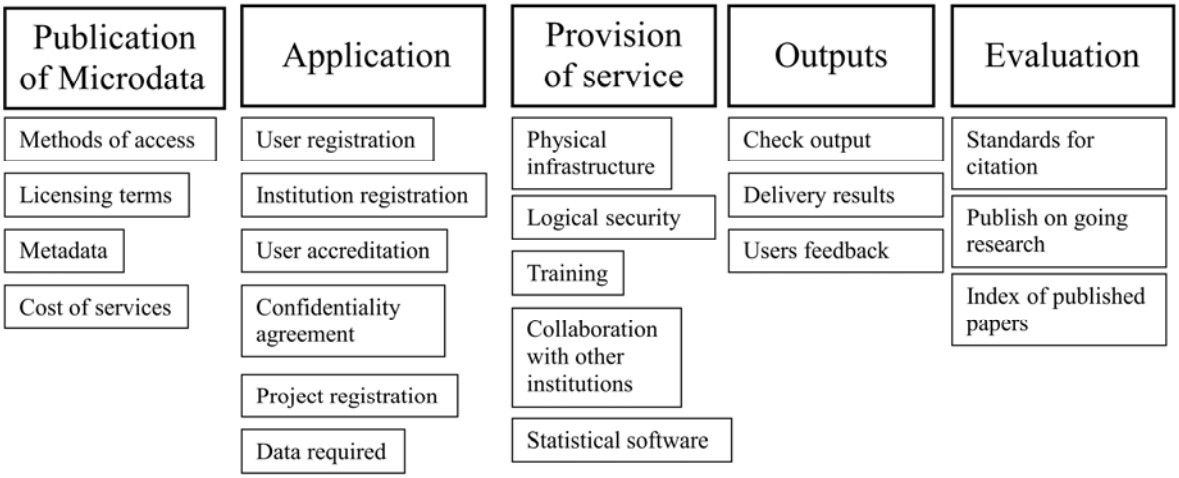
108. The review of the process followed by NSOs or other statistical agencies to provide access to microdata helps to identify the core elements that are typically present and could be standardised. The process generally compromises five sub-processes that can be located within the GSBPM in the phases: dissemination, archive and evaluate.



109. In fact, the five sub-processed are normally carried out in a way that ensures risk management in the provision of microdata access. In that respect, the general processes put in place by NSOs or other statistical agencies are very similar, even if specific aspects are applied or particular steps are called differently by different organisations.



110. Also, a review of the processes implemented in NSOs or other statistical agencies show that there are specific activities within each of the five sub-processes:



Publication of Microdata

111. Once a NSO or other statistical agencies has taken the decision to allow access to microdata sets, it will need to publish the list of microdata sets available as well as the terms and conditions for obtaining access, for each microdata set. In this way, users will be able to discover what information is available for their research and the requirements they have to fulfil. Accessibility of information is important, so that users can easily find on the website of a NSO or other statistical agencies, the microdata sets available. Ideally, the website should display a single entry point that leads to the website section dedicated to the provision of access to microdata. The section should include all microdata sets available, from census, surveys and administrative registers, through all forms of access, and tools that could help users identify what is useful for their research like DDI catalog, and the dates of release of the microdata sets.

Methods of access

112. The NSO or other statistical agencies should clearly indicate the different alternatives of access available for each microdata set and the characteristics of each dataset, especially if the quality is altered by SDC methods.

Licensing

113. For each method, the NSO or other statistical agencies should clearly establish the terms and conditions users have to fulfil in order to have access to them. For the files that can be downloaded, it should be specified if there is a licensing agreement or not and what are the user' responsibilities.

Metadata

114. The metadata of microdata datasets have to be published, because the description of the datasets is necessary for users in order to be able to assess the potential of using them for their research. DDI is an international standard widely used for the documentation of microdata sets, and many NSO or other statistical agencies use it. In the absence of standards for documenting the metadata, NSOs or other statistical agencies should publish all methodological information available (*e.g.* questionnaire, file descriptor, a methodological document and glossary).

Cost of services

115. Most NSOs or other statistical agencies charge cost of recovery for the service; in other cases is free. It should be clearly specified the level of resources a researcher will need to carry out her/his research project based on microdata.

Application

User registration

116. In most cases users have to fill a registration to have access to microdata. The core fields usually required for personal information on the user are:

- Name
- Surname
- Country of birth
- Country of residence
- ID
- Document of institutional affiliation
- Curriculum Vitae

Institution registration

117. The institution that has been recognised as research entity (according to a transparent definition of research entity) provides accreditation of the user that applies for microdata access as a person that can be trusted. In many cases the NSO or other statistical agencies require the signature of a legal agreement. The core fields requested in the application regarding the data of the institution are:

- Legal name
- Short name acronym
- Legal address
- Web url
- Country

Project registration

118. NSOs or other statistical agencies need to check research proposal as adequate. The objective of the research project has to have a legitimate purpose and be of public interest. The data are provided solely for the specific research project for which is requested. In the application form, the core fields the user has typically to fill are:

- Objective of the research project
- Data set required
- Justify need to use microdata
- Research methodology
- Expected results
- Expected duration of project
- Forms of dissemination

Signature of confidentiality declaration

119. Users are required to sign an agreement in which he or she is responsible of safeguarding confidentiality and agrees to conform to the conditions and terms of the service of access to microdata.

User accreditation

120. The procedures for the accreditation can vary depending on the legal framework and institutional arrangements of the academic and research sector in each country. Generally there are two types of accreditation provided by:

- Research institution registered as such by a NSO or other statistical agencies. Each NSO or other statistical office has to define the rules and procedures that have to be undertaken by a research institution to be considered registered and able to provide accreditation to its students and researchers. Once a research institution is registered, it will need to define a mechanism by which the NSO or other statistical agencies can validate the institutional affiliation of the user and his or her permanence in the research institution that is providing their accreditation.
- The National Research Council, if such an institution exists, together with the NSO or other statistical agencies will have to define the rules by which it can provide the accreditation and the mechanism to operate it.

Provision of service

121. The provision of service encompasses: physical infrastructure (the premises where access is provided), hardware and software. The service of microdata access can involve on-site access within the NSOs or other statistical agencies buildings, or remote access in other institutions premises, which can imply the collaboration with other institutions and the signature of a legal contract covering specific hardware and software for the identification of users. In some cases, training is offered to users previous to the first session in the microdata laboratory.

Outputs

Check outputs

122. Once the user has carried out his microdata processing, the NSO or other statistical agencies check that results do not have confidential disclosure risks; in some cases, SDC methods are applied before the output is delivered to the user.

Delivery of results

123. Checked results are given to the user by email, ftp or with an external memory device.

Users' feedback

124. Some NSOs or other statistical agencies require users to provide feedback regarding the quality of the microdata they used for their research so the areas that generate statistical information can use it as input in their continuous process of improvement.

Evaluation

Citation and index of published papers

125. The NSO or other statistical agencies define how the citation of research carried out with microdata has to be done in order to facilitate the monitoring of microdata use. Providing access to microdata for research purpose brings benefits to the society. One way to value the benefits is by scientometrics, for instance through indicators of the scientific papers published based on research carried out with microdata. Standardized structure for citations facilitated the integration of indicators to value the impact of the service.

Publish ongoing research

126. In order to show how data are used, some NSOs or other statistical agencies publish in their websites the list of ongoing research project relaying on microdata provided by the access service.

Conclusions and proposed recommendations

127. Describing the process for providing access to microdata can help standardise the steps of the application that a researcher has to submit to request data. Knowing in advance what to expect from this type of service and the requirements a researcher will have to comply facilitates the process of access. The idea is to make the application process as similar as possible across NSOs or other statistical agencies, even if for certain cases compliance with special requirements will be necessary. Overall, requirements are built upon the need to create the conditions for the "circle of trust", where researchers are allowed to work with the microdata they need.

128. Access to microdata, including transborder access, could be facilitated by:

- One single entry point in every NSO or other statistical agencies website (with a heading "Microdata").
- Users are informed on and understand the general steps of the process they have to undertake.
- Users can reuse the information entered in the application form when they need to apply to similar microdata sets in different countries for a same research project.

129. The single entry point in a NSO or other statistical agencies website should contain all the relevant information. The first page would need to present the microdata sets in a way that provide users with a clear idea of datasets available and the requirements for requesting access to them.

130. It is important that NSOs or other statistical agencies present information in a similar way, in particular as concerns three sub processes involved in the process of providing access: publication of datasets available, application procedure and provision of service.

A standardised application form

131. Finally, it is proposed that a typical application form for microdata access contains the following core standardized fields:

User personal data

Name
Surname
Country of birth
Country of residence
Copy of ID - "passport or equivalent"
Document of institutional affiliation (receiving a copy as pdf or in the automated interface facility of integrating a pdf)
Curriculum Vitae

Data of the institution of affiliation

Legal name
Short name acronym
Legal address
Web url
Country

Project data

Objective of the research project
Data set required
Justify need to use microdata
Research methodology
Expected results
Expected duration of project
Forms of dissemination

132. Having a standard form presents considerable advantages. Once the user fills the standardised fields of an application form for getting microdata access from a NSO or other statistical agencies, the information can be reused again when submitting a request for the same microdata to a NSO or other statistical agencies in a different country for the same research project. In this way, a user would only need to input the information of the standardised fields once, and he/she would then complete the specific field of data that each NSO or other statistical agencies require on top of the common core standardised fields.

133. Also, knowledge of the main steps that a NSO or other statistical agencies follow for the provision of microdata access would help a researcher to better prepare his/her requests and manage the research process, despite the specific differences that might exist among the microdata service of each NSO or other statistical agencies.

CHAPTER 7. CASE STUDY: A CIRCLE OF TRUST IN NORDIC COUNTRIES

by Claus-Göran Hjelm and Eva Nilsson

History of the Nordic cooperation

134. The Nordic cooperation in the statistical field between Denmark, Norway and Sweden goes back to 1964; since the 1990s around 40 contact groups have worked on different subject matters, technologies and common issues. As a result of the contact groups and the fact that there are no or small language barriers among Scandinavians, contacts at the personal and organisational level between the different National Statistical Institutes (NSIs) have been successful. The similarities in legislation among the Nordic countries have also contributed to facilitate cooperation.

Legislation

135. Data confidentiality is guided by two major aspects, which are necessary in order to meet access requests from researchers:

a) General rules (guidelines, screening procedures, contracts, regulations and laws, etc.)

b) Technical and practical measures.

136. The legislation concerning confidentiality and protection of individual's integrity establishes the criteria to provide access to microdata: it sets the limits for release of data, for example for research purposes, and provides the legal foundation for administrative and technical safeguards. Specific legislation of importance in the Nordic countries includes the respective Statistics Act and the Data Protection Act. In addition, the current EU legislation on statistical confidentiality is also relevant.

EU legislation

137. The Regulation (EC) No 223/2009 of 11 March 2009 on European Statistics contains rules that are important for the use of information collected for European statistics. According to the Regulation, data used by the national authorities and the Community authority for the production of European statistics shall be considered confidential when they allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information.

138. To determine whether a statistical unit is identifiable, account shall be taken of all the means that might reasonably be used by a third party to identify the statistical unit. Confidential data obtained exclusively for the production of European statistics shall be used exclusively for statistical purposes unless the respondents have unambiguously given their consent to the use for any other purposes. However, it is possible to allow access for scientific purposes to confidential data obtained for European statistics. The modalities, rules and conditions of such access are laid down in the Commission Regulation (EU) No 557/2013.

139. The data processing and release is also regulated by the Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (the Data Protection Directive). The object of the Directive is to strengthen data protection, *e.g*. the legal protection of individuals with regard to automatic processing of personal information relating to them. The Directive has been implemented in all the Nordic countries.

140. The Directive applies to computerized personal data and personal data held in structured manual files. It applies to anything at all done to personal data processing. The term "processing" covers all types of processing of personal data, including registration, storing, disclosure, merging, changes, deletion, etc. According to the Directive data must be:

- Processed fairly and lawfully.
- Collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes. However, further processing of data for historical, statistical or scientific purposes is not considered as incompatible.
- Adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed.
- Accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that data which are inaccurate or incomplete, with regard to the purposes for which they were collected or for which they are further processed, are erased or rectified.
- Kept in a form, which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed. Personal data can be stored for longer periods for historical, statistical or scientific use.

National legislation

141. The protection measures applied to confidential data obtained for statistical purposes are based on several legal acts and directives. However, it should be noted that access to statistical microdata for research or other purposes is part of NSIs duty service and not an obligation established by legislation on the use of statistical information. In the Nordic countries, the Statistics Acts regulate the use of statistical information.

142. Data collected for statistical purposes, in accordance with any prescribed obligation to provide information, or which is given voluntarily, should in principle only be used for the production of statistics. There are exceptions that enable access to data for research purposes and public planning. However, a condition for the use for research is that there is no incompatibility between the purpose of such processing and the purpose for which the data were collected. The processing of data, which includes release of data, must also be in accordance with the regulation concerning protection of individual's integrity.

143. Besides the Statistics Acts there are specific Personal Data Acts that apply to the production of statistics and the release of microdata. The Acts are based on the Data Protection Directive and contain rules about the fundamental requirements concerning the processing of personal data. These demands include, inter alia, that personal data may only be processed for specific, explicitly stated and justified purposes.

144. Very stringent rules apply to the processing of sensitive personal data. They may be processed for research and statistics purposes, provided the processing is necessary and the public interest in the project manifestly exceeds the risks of violation of personal integrity. Furthermore, in Denmark, Norway and

Sweden processing of sensitive data for research purposes needs approval. A scientific project involving processing of sensitive personal data is subject, in these countries, to notification to and approval by the Data Inspection Agency before such processing can commence. This applies to all surveys, whether conducted by a public administration, individuals or enterprises. In Sweden, the approval of the National Data Inspection Agency is not necessary if a research committee has approved the processing. If the Data Inspection Agency approves the processing, personal data may be provided to be used in research projects unless otherwise provided by the rules on confidentiality. This means that the NSI may take other issues into consideration even if the Data Inspection Agency (or research committee) has approved the processing of data.

145. Data obtained for statistical purpose are declared as confidential, when they allow statistical units to be identified, directly or indirectly and thereby disclosing individual data. Also anonymous data can be considered confidential. Statistical data are confidential irrespective of their source. Moreover, data taken from public administrative sources are confidential while in the possession of the NSI. The confidentiality rules are the same irrespective of whether data concerns individuals or enterprises.

146. Under the main rules, access may be granted in forms which do not allow direct or indirect identification of people or other data subjects such as enterprises.

147. Confidential data may be released to a third party for the purpose of statistical surveys and scientific research. According to the legislation in Denmark, Norway and Sweden, statistical data may even be released with identification data for these purposes. One condition in all the three countries is that access to confidential data for statistical or research purposes must not cause any damage or be detrimental to the data subjects. In practice, this means that the NSIs only provide access to anonymous data or de-identified data.

148. The countries also have special public business registers that contain some common primary information about enterprises. These registers are (except in Denmark) administered by the NSIs and can also be used for other purposes than statistics or research.

149. When data have been collected through a voluntary survey, respondents must give consent to the release of the data.

150. It is the NSO that decides whether data may be released for research purposes.

151. In Norway, access for other purposes than statistical must be approved by the Data Inspection Agency. The Agency has given general permission to Statistics Norway to provide access to microdata for research purposes. The Data Inspection Agency may nevertheless make exceptions to such obligation of confidentiality for certain types of information if they find it in conflict with the Data Protection Act.

152. The obligation of confidentiality will also apply to the recipient of the data, according to the law or by imposition of a duty of non-disclosure. The NSO may also impose a restriction limiting the researchers' right to re-communicate or use the information. Breach of confidentiality restrictions is punishable by simple detention or imprisonment.

153. In Sweden, it is not possible to impose restrictions when data are released to another authority. Therefore, it is important for Sweden to take into consideration if the data will be treated as confidential, according to the Secrecy Act, also by the authority receiving data. If not, anyone who so desires could have access to the data because of the authority's obligation under Chapter 2 of the Freedom of the Press Act to provide personal data that are not confidential. For that reason, there are rules providing that confidentiality accompanies data to another authority in special situations; for instance, if an authority, for research purposes, receives information from another authority where the data are confidential, the confidentiality

will apply also within the receiving authority. There are no such rules yet concerning the release of data for statistical purposes or public planning.

Providing access to microdata

154. In addition to laws and regulations on data confidentiality, Denmark, Norway and Sweden follow some kind of screening procedure requiring written confirmation that the researcher has signed a general confidentiality statement. Legal contracts are made that include various limitations to the access to microdata by specifying the people, research projects, variables and periods during which data can be used. As mentioned above, Sweden does not impose restrictions when data are released to another authority.

155. The Nordic NSOs mainly provide access to microdata to public authorities and people or organisations performing scientific research (universities and research institutions). Sweden also provides access to microdata to other authorities and municipalities producing statistics. Governmental or municipal institution in Norway can have access to anonymised microdata for planning purposes.

156. Generally, the uses of microdata for commercial purposes are ruled out.

Requesting access to microdata: two separate situations

a. Requesting access to microdata from Nordic NSIs for the purpose of producing statistics

157. The National Statistical Offices in Denmark, Norway and Sweden exchange identifiable personal data (date of birth and name) to facilitate the identification of commuters across borders for the joint production of regional workforce flows across the national borders. In Sweden, these statistics provide a supplement to the Swedish national register-based labour market statistics for Swedish residents who are gainfully employed in another country.

158. Prior to the commencement of this joint initiative, an exchange was tested based on the prevailing national legislation in each country. It was assessed that the scope existed for the sharing of data for the statistical purposes according to each country's statistical legislation. Each annual exchange of data, however, becomes the object of a separate disclosure assessment process.

b. Requesting access to microdata from Nordic NSIs for the purpose of research

159. To enable the release of data from the Nordic countries to researchers outside their NSIs legal support is required. The crucial questions are: what information is relevant for the research project, what the information will be used for and by whom. Therefore, information must accompany the application to be able to decide whether the data should be released or otherwise that there are laws or regulations dictating such procedures. In the case where data are released for Nordic research (meaning register-based research on Nordic data), they are made available by way of each National Statistical Office; the purpose of this is to maintain protection of the data. Identifiable data are never in question for release.

160. In the bilateral cooperation between Denmark and Sweden, a Trans-regional register is being produced for the purpose of analysis where migration flows are also provided on a personal level. A serial number is assigned to the microdata which are anonymised prior to possible release for research purposes. The conditions and possibilities for the release of data to supplement existing research registers, or to set up new registers, will be assessed according to a separate procedure.

CHAPTER 8. CASE STUDY: MICRODATA ACCESS IN THE EUROPEAN STATISTICAL SYSTEM

by Aleksandra Bujnowska

Introduction

161. Regulation (EC) No 223/2009 of the European Parliament and of the Council on European Statistics ("Regulation on European statistics") establishes a legal framework for the development, production and dissemination of European statistics which are defined in the statistical programme. Eurostat and national statistical institutes (NSIs) and other national authorities responsible in each Member State for the development, production and dissemination of European statistics constitute a European Statistical System (ESS).

162. The Regulation on European statistics allows "transmission of confidential data from an ESS authority that collected the data to another ESS authority provided this transmission is necessary for the efficient development, production and dissemination of European statistics or for increasing the quality of European statistics".

163. The Regulation on European statistics specifies also the conditions for access to confidential data for scientific purposes. Researchers may be granted access to confidential data which only allow for indirect identification of the statistical units. The approval of the NSI or other national authority which provided the data is required for each submitted research proposal.

164. This document presents how access to European microdata is provided by Eurostat ensuring full collaboration of the data providers (national statistical institutes and authorities in the ESS).

Access to microdata at the European Union level

165. According to Regulation on European statistics Eurostat may grant access to confidential data to researchers carrying out statistical analysis for scientific purposes. The Regulation allows researchers' access under three conditions:

1. The confidential data for scientific purposes can not contain any direct identifiers;

2. The approval of the national statistical authorities that transmitted the data to Eurostat is required for each research project requiring access to microdata;

3. The modalities, rules and conditions for access have to be established by the separate Commission Regulation;

166. Eurostat in collaboration with national statistical authorities and representatives of research community has been working on the new Commission Regulation since 2009 when the statistical law was adopted.

167. The final version of the new regulation on access to confidential data for scientific purposes was adopted by the representatives of the EU national statistical institutes (European Statistical System

Committee) in February 2013. The new Commission Regulation (EU) No 557/2013 on access to confidential data entered into force in July 2013[8].

168. The main principles of the new Regulation 557/2013 are:

- access to EU data may be granted to "research entities"; recognition of research entities is based on the assessment of the following criteria: purpose of the entity, publication of the results of research, appropriate organisation structure, safekeeping of confidential data (appropriate security measures in place);
- new modes of access (including remote access) are enabled;
- external partners (NSIs in the first place) may be involved in the provision of access to researchers (access facilities);
- contract is replaced by the licence (confidentiality undertaking) covering all future access requests;

169. In order to be granted access to EU confidential data the researcher's organisation must be first recognized by Eurostat as a research entity and sign the licence (confidentiality undertaking). Once Eurostat has recognised the researcher's organisation, the application for access (research proposal) can be submitted[9].

170. The research proposal must include information on the person requesting access, his or her research entity, the data requested and the mode of access. The criteria require that the research proposal state the legitimate purpose of the research, i.e. scientific purpose, and that the results of the research are to be made public. The planned outputs (articles, presentations, books, etc.) have to be specified in the research proposal. The need to use microdata for the research project should be justified.

171. The approval of the NSIs or other national authorities which provided the data is required for each research proposal.

Types of EU microdata available for scientific purposes

172. Confidential data for scientific purposes are available in two forms:

- "scientific-use files" partially confidentialised data delivered to researchers on electronic devices (CD-Rom, DVD, etc.);
- "secure-use files" available in Eurostat's "safe centre" in Luxembourg (non-confidentialized data);

173. Scientific use files are especially prepared to make the identification of survey respondents more difficult. Statistical disclosure control (SDC) methods are applied to this data to reduce to an appropriate level and in accordance with best practices the risk of identification of the statistical unit. SDC methods are not applied to secure use files. These files can be made available to researchers only in the secure environment ensuring that the results of the statistical analysis are not released prior to output checking.

---

8 Commission Regulation (EU) No 557/2013 of 17 June 2013 implementing Regulation (EC) No 223/2009 of the European Parliament and of the Council on European Statistics as regards access to confidential data for scientific purposes and repealing Commission Regulation (EC) No 831/2002. OJ L 164, 18.6.2013, p. 16.

9 The list of recognized research entities is available on Eurostat website under the following link: http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/introduction

174. This arrangement allows for different modes of access:

- Remote execution
- Remote access
- Decentralized access
- On-site access

175. In addition Regulation (EU) No 557/2013 permits other parties to be involved in the provision of access to confidential data. These parties (access facilities) have to fulfil the specific criteria and be accredited by Eurostat.

Micro dataset available for scientific purposes

176. All confidential data that national statistical authorities transmit to Eurostat for the purpose of compiling EU statistics can be made available for scientific purposes provided that appropriate protection methods are drawn up and applied.

177. Methods of protection are decided in collaboration with national statistical authorities, taking into account the mode of access, the probability of re-identification, utility, harmonisation and the impact of unlawful disclosure.

178. In general the agreed SDC methods are applied on all countries' data but specific adaptations are also possible. Data providers may also decide not to participate in the release of particular data for scientific purposes at all.

179. The actual application of the SDC methods is performed by Eurostat (in-house or subcontracted) or by the statistical authority providing the data.

180. Scientific use files are available for the following data collections:

- European Community Household Panel (ECHP);
- Labour Force Survey (LFS);
- European Union Statistics on Income and Living Conditions (EU-SILC);
- Adult Education Survey (AES);
- Community Innovation Survey (CIS);
- Structure of Earnings Survey (SES);
- European Health Interview Survey (EHIS);
- European Road Freight Transport Survey (ERFT);
- Community Statistics on Information Society (CSIS) – as from December 2013;
- Continuous Vocational Training Survey (CVTS) – as from December 2013;

- Household Budget Survey (HBS) –as from 2014.

181. Secure use files are available in the Eurostat safe centre (on-site) in Luxembourg for:

- Community Innovation Survey (CIS);
- Structure of Earnings Survey (SES).

182. Regulation (EU) No 557/2013 introduces as well the concept of accredited access facilities, which can provide access to EU microdata on behalf of Eurostat at the national level. In order to be accredited the access facilities have to comply with specific legal, administrative, technical and security requirements.

183. As agreed with the Member States the development of access facilities shall be implemented in a stepwise manner starting with national statistical institutes (NSI, institutes coordinating production of statistics at the national level). Based on this experience, access facilities may be extended to other statistical authorities and then other facilities.

On-going projects and initiatives on improving microdata access

184. The **ESSnet on "Decentralized and remote access to confidential data in the ESS" (ESSnet DARA)** gathered together representatives of various national statistical authorities to implement the concept of a network of accredited safe centres located in NSIs where access to EU confidential data will be provided. Whereas data will physically remain in the secure environment in Eurostat, the researchers will access this data remotely through secure channels linking safe centres with Eurostat. The following partners collaborate in the ESSnet project:

- Federal Statistical Office (co-ordinator), Germany, DESTATIS
- National Institute of Statistics and Economic Studies, France, INSEE / Groupe des Écoles Nationales d'Économie et de Statistique, GENES
- Hungarian Central Statistical Office, Hungary, HSCO
- Office for National Statistics, United Kingdom, ONS
- Instituto Nacional de Estatistica, Portugal, INE
- State Statistical Institute Berlin-Brandenburg, Germany, AfS Berlin

185. The solution developed by French national statistical authority CASD (Le Centre d'Acces Securise Distant aux Donnees – Centre for secure remote access to the data) was tested by the project team. CASD provides "full" remote access solution to confidential data in France.

186. ESSnet team recommends that in the short term Eurostat should remain the central point of the system but more distributed model can be considered in the longer term with NSIs checking the results of the scientific analysis carried out in their safe centre, regardless of the source of the data and on behalf of the countries whose data are used.

187. The 7th Framework Programme project "**Data without Boundaries (DWB)"** works on the broader infrastructure, which includes not only NSIs but also data archives and research bodies. It covers also broader range of datasets, i.e. not only European statistics but also statistics available at national level only.

188. The purpose of the DWB project is to enhance researchers' access to national and European official microdata. It aims at bringing together NSIs, data archives and researchers, to agree on standards (for eligibility of researchers and research projects, for metadata, for statistical disclosure control methods etc.) and common views, and to build mutual trust. The project enables researchers to test the access infrastructure and to gain access to other countries' data. One of the interesting ideas brought forward by the DWB project is to establish a Service Centre for Official Statistics, a single entry point providing information on the microdata (with appropriate metadata) available at the national and EU level. The information on the data provided by other agencies (like IPUMS) could also be available in the Service Centre.

Exchange of confidential data between members of the European Statistical System

189. The implementation of the Regulation on European statistics (March 2009) and the reinforcement of the European Statistical System triggered further discussion on the ultimate model for production of EU statistics.

190. In accordance with the so called "vision" document [10] published soon after the entry into force of the new Regulation on European statistics, all the ESS members shall be allowed to have access to the data collected by other countries under European legislation if this contributes to the improvement of the data quality and efficiency of the system. It concerns especially data on the cross-borders flows of persons (migration, tourism data) and goods.

191. An on-going initiative in this direction is SIMSTAT (Single Market Statistics, statistics on trade between EU countries). The idea of SIMSTAT is to simplify the reporting requirements on intra-EU imports and to make the exchange of micro-data on EU intra-exports compulsory for the ESS members concerned so that the export data can be used to compute the import data of other countries.

[10] Communication from the Commission to the European Parliament and the Council on the production method of EU statistics: a vision for the next decade (COM/2009/0404).

ANNEX II.A1. SANCTIONS FOR BREACH OF CONFIDENTIALITY OF THE EU DATA

Place for sanctions of violations of statistical confidentiality in the legal framework

192. Regulation (EC) No. 223/2009 on European statistics obliges Member States to establish measures to prevent and sanction violations of statistical confidentiality **at national level**. It is for each Member State to decide the actual contents of these provisions. The current wording of Article 26 of the Regulation 223/2009 allows an introduction at national level of sanctions applicable to individuals and entities.

Article 26 Violation of statistical confidentiality

Member States and the Commission shall take appropriate measures to prevent and sanction any violations of statistical confidentiality.

193. Very similar provisions, *i.e.* clearly indicating that penalties/penal sanctions should be foreseen in national law are suggested in the proposal for a new basic regulation on the protection of personal data, the *General Data Protection Regulation* (GDPR). According to Article 78 of the proposal, Member States shall lay down the rules on penalties, applicable to infringements of the provisions of the regulation and take all measures necessary to ensure that they are implemented.

194. As for the administrative sanctions proposed in the GDPR (Article 79), they require the setting up of a national supervisory authority in each Member State. The actual implementation of these provisions on administrative sanctions would be ensured at national level. Once the regulation on protection of personal data is adopted, it will apply also to processing of personal data for statistical purposes (and for scientific purposes).

Provisions on sanctions in the EU countries

195. Eurostat grants access to confidential data for scientific purposes transmitted by national statistical institutes (NSOs) or other national authorities. The modalities, rules and conditions for access at the EU level are spelled out in the Regulation (EC) No. 831/2002 on access to confidential data for scientific purposes. This Regulation has now been revised and the new one will enter into force in July 2013.

196. Eurostat and delegates from the EU Member States conducted recently a questionnaire survey of the existing national provisions on sanctions for statistical confidentiality. The survey revealed that penal sanctions for violations of statistical confidentiality with respect to their own confidential data within their own territory are in place in 23 out of 24 Member States that filled in the questionnaire. At the same time, only half of the Member States confirm that their penal provisions can be applied in cases where access within their territory is granted by the Commission (Eurostat).

197. As far as the administrative sanctions are concerned, they exist in 14 Member States. In five countries these provisions can be used to sanction violation of confidential EU data provided on the basis of EU legal framework by the Commission (Eurostat).

198. In a majority of EU countries there exist also nationally established provisions in place for disciplinary sanctions. The actual penalties (imprisonment, fines) vary from country to country but their scope is rather similar.

Application of the sanctions

199. The recent discussion at the EU level about sanctions for breaches of confidentiality of the EU data to which access was granted by the Commission (Eurostat) leaded to the following conclusions:

- Penal sanctions exist but they are also the most difficult to apply; the process of application of penal sanctions may be lengthy; the criminal law could possibly be used to sanction most severe violations;
- The sole event of a breach of confidentiality should suffice to prosecute the violation of statistical confidentiality by one legal system, even if the access to the data was provided by the institution not belonging to that legal system;
- Administrative sanctions will be strengthened with the entry into force of the GDPR; the system for sanctions foreseen in this Regulation will be applicable also to statistical data (based on natural persons, not on the entities); the administrative sanctions should cover breaches of confidentiality of the EU data;
- The identification of the applicable law in case of cross-border access to confidential data shall be done in accordance with applicable measures (private international law), on a case by case basis, taking into account place of violation, the nationality of the researcher, source datasets etc.;
- The measures to prevent violation of statistical confidentiality shall be implemented in the first place: recognition of researchers/research entities, appropriate and secure access modes, appropriate anonymisation of the datasets; contractual arrangements;
- The obligations to protect confidential data shall be made clear to the researchers granted access to confidential data; researchers should be aware of the confidentiality obligations and of the applicable sanctions in case of deliberate or negligent misuse of the data;
- Contractual arrangements and provisions of the civil law should be used first (before penal and/or administrative sanctions) to sanction violations of statistical confidentiality; they allow for immediate and direct sanctioning of the violation;

Measures that can be taken by the Commission

200. The Commission can take action in the event of a breach of confidentiality in the following way:

- by withdrawing from the offending researcher, and if necessary from his/her research entity, the possibility to access microdata;
- by inviting the research entity to take disciplinary action against the researcher;
- by claiming civil-law compensatory damages from the research entity; the confidentiality undertaking in this regard includes a reference to the applicable law and competent court;
- by filing a complaint or by reporting the breach to the police on the basis of national legislation; the Commission could possibly participate in national proceedings as plaintiff.

201. Depending on the situation the sanctions may be applied on researchers or their research entities.

ANNEX II.A2. LEGISLATION IN OECD COUNTRIES AND MUTUAL STATISTICAL AGREEMENTS

202. The institutions of the European Union and its member states have an expectation that personal data should move freely within the EU. The Data Protection Directive of 1995 (95/46/EC) summarises the issues of transnational data sharing in its preamble:

"...the free movement of goods, persons, services and capital...require not only that personal data should be able to flow freely from one Member State to another, but also that the fundamental rights of individuals should be safeguarded. (Recital 3)

...the increase in scientific and technical cooperation...in the Community necessitate and facilitate cross-border flows of personal data. (Recital 6)

...in order to remove the obstacles to flows of personal data, the level of protection of the rights and freedoms of individuals with regard to the processing of such data must be equivalent in all Member States... (Recital 8)

...given the equivalent protection resulting from the approximation of national laws, the Member States will no longer be able to inhibit the free movement between them of personal data on grounds relating to protection of the rights and freedoms of individuals, and in particular the right to privacy; whereas Member States will be left a margin for maneuver, [and] will therefore be able to specify in their national law the general conditions governing the lawfulness of data processing; [and] within the limits of this margin for maneuver and in accordance with Community law, disparities could arise in the implementation of the Directive, and this could have an effect on the movement of data within a Member State as well as within the Community. (Recital 9)"

203. Recitals 3, 6, and 8 established the expectation of the Directive, and set out how Member States were expected to implement their national laws. However, Recital 9 predicted the current situation very well. The intention of the Directive is to prevent member states citing different national laws for privacy protection as their reasons for not allowing data to cross EU borders. At the same time, it also acknowledges that the room for maneuver left to Member States as they implement the Directive may negatively affect the movement of data in the EU. This has come to pass, and is a motivation for current Commission proposals to replace the Directive with a *General Data Protection Regulation* (GDPR) with direct effect in Member States and no 'room for maneuver' that may inhibit transnational data flow. The Commission's proposal for a GDPR includes the following in its preamble:

"Differences in the level of protection of the rights and freedoms of individuals, notably to the right to the protection of personal data, with regard to the processing of personal data afforded in the Member States may prevent the free flow of personal data throughout the Union." (Recital 7)

204. The aim of the data protection regime for the European Union is the uninhibited free flow of personal data between Member States and other states with equivalent protections for personal data wherever they are processed.

205. Under the current Directive, the territorial extent of its protections are the EU Member States directly, but also third countries formally recognised as having equivalent levels of protection.[11] These will

[11] http://ec.europa.eu/justice/data-protection/document/international-transfers/adequacy/index_en.htm

continue to enjoy such recognition under the GDPR. They are: Andorra; Argentina; Australia; Canada; Switzerland; Faeroe Islands; State of Israel; Isle of Man; Jersey; Eastern Republic of Uruguay. OECD countries that currently do not enjoy of the recognition of equivalent levels of protection include: Chile, Japan, Korea, Mexico, New Zealand, and Turkey. Transfer of personal data to the United States is restricted to transfer of air passenger name record (PNR) data and organisations included in the "safe harbor" exception.[12]

206. There is one further way of extending the reach of the principles of data protection in the Directive. Data can be freely exchanged where there are binding contractual provisions that replicate the protections in the Directive itself. A MSAA can contribute to this provision.

207. Underlying the Directive and the GDRP are two principles. The first is that EU Member States and recognised third countries have similar legal and supervisory regimes regarding the protection of personal data. The second, following from the first, is that personal data transferred from one country to another will be effectively protected in the country to which it has been transferred. There is a certain element of trust in this framework, but it is based on clearly defined criteria, which can be demonstrated by proposed Articles 41 and 42[13] of the GDRP, regarding the conditions of recognising third countries and international organisations for the purpose of transfer to them of personal data of EU data subjects.

208. Article 41 of the GDPR sets out the following conditions:

- "(a) the rule of law, relevant legislation in force, including concerning public security, defense, national security and criminal law, the professional rules and security measures which are complied with in that country or by that international organization, as well as effective and enforceable rights including effective administrative and judicial redress for data subjects, in particular for those data subjects residing in the Union whose personal data are being transferred;
- (b) the existence and effective functioning of one or more independent supervisory authorities in the third country or international organization in question responsible for enforcing compliance with the data protection rules including sufficient sanctioning powers, for assisting and advising the data subjects in exercising their rights and for co-operation with the supervisory authorities of the Union and of Member States;
- (c) the international commitments the third country or international organization in question has entered into."

209. Article 42 of the GDPR states that "where the Commission has taken no decision pursuant to Article 41, a controller or processor may not transfer personal data to a third country, territory or an international organization unless the controller or processor has adduced appropriate safeguards with respect to the protection of personal data in a legally binding instrument." In addition, those appropriate safeguards shall, at least:

- guarantee the observance of the principles of personal data processing as established in Article 5;
- safeguard data subject rights as established in Chapter III and provide for effective redress mechanisms;

12 For the US "safe harbor" list, see: https://safeharbor.export.gov/list.aspx

13 The original proposal for the GDPR was prepared by the European Commission, and is dated 25.1.2012. It was referred to the Committee on Civil Liberties, Justice and Home Affairs of the European Parliament, which published its first draft report, with proposed amendments to the original proposal, on 17.12.2012. Subsequently, further amendments have been tabled. At the time of preparation of this report, it is not possible to ascertain the final wording that will be presented for approval to the European Parliament. Reference to language in the GDPR proposal, for the purpose of this report, will be to the 17.12.2012 draft report referred to above.

- ensure the observance of the principles of privacy by design and by default as established in Article 23;
- guarantee the existence of a data protection officer pursuant to Section 4 of Chapter IV.

210. Regarding enforcement within the EU, Chapter VII (Remedies, liabilities and sanctions) sets out: the rights of data subjects to judicial remedies; the obligations of Member States to lay down rules on penalties; and the power of supervisory authorities to impose administrative sanctions.

211. Based on all of the above, the GDPR seems to offer an effective framework in which OECD countries **that are also members of the EU** can transfer personal data from one NSO to another for statistical research, subject to the conditions set out in the GDPR. In addition, such transfers must only be carried out based on binding contractual obligations between the NSOs, establishing the conditions of use of such data and the undertakings of the NSO that receives the data to enforce the protection of the data according to all of its relevant laws, including laws regarding statistical confidentiality.

212. The above also applies to the transfer of personal data from an NSO in an EU Member State to and NSO in a non-EU member state that has been recognised under the current Directive or the GDPR.

213. There are limitations of the proposal for the GDPR that mitigate its effectiveness regarding all OECD countries.

Personal data of natural persons versus data of legal persons

214. The GDPR controls the transfer or exchange of personal data, *i.e.* meaning any identified or identifiable information relating to **natural persons**. Excluded from its application is information relating to legal persons, such as corporate entities. Therefore, the supervisory administration mandated under the GDPR will not be available for the protection of enterprise/corporate data, including that being used for statistical research. It will be a matter of each NSO to decide if another NSO can guarantee the security of such data transferred to it and to make appropriate contractual arrangements. If such arrangements are made, they should only be between NSOs of EU Member States or between NSOs of EU Member States and third countries recognised under the GDPR.

Transfer of data versus remote access to data

215. The GDPR controls the transfer of personal data from one data controller to another for the purpose of processing that data. Underlying this framework is the assumption that data will be transferred in the sense that it will reside on the receiving controller's computer system for processing. Once the data is transferred and resides on the receiving controller's computers, it is axiomatic that the laws of the country of that controller will apply to the protection of that data, including its confidentiality.

216. The GDPR, and in general laws of statistical confidentiality, do not specifically relate to the legal status of confidential data accessed remotely between NSOs in different countries. By remote access, mean that the confidential data reside on a computer in one country (Country A), but a researcher is able to view that data from a computer terminal in another country (Country B) and give instructions from that computer terminal (in Country B) that will cause the computer in Country A to perform research operations on the data. The output of those operations will be vetted by Country A before it is provided to the researcher in Country B, to ensure that it contains no confidential data.

217. It is not clear, under the GDPR (or under the Directive today) the status and responsibility of each of the parties involved in such remote access researcher. The NSO in Country A would certainly be considered the data controller. What would be the status of the NSO in Country B and the researcher of the NSO in Country B? Do they "control" the data? Are they "processing" the data?

218. Apart from the GDPR, how should one interpret the national legislation controlling statistical confidentiality as regards such data accessed from Country B? In the example above using Article 39 of the UK Act, should ONS serve as a site of accessing confidential data that resides on a computer in Germany, would such data be considered to be "held" by ONS in the "exercise of its functions"?

219. As of today, we know of no legislation or legal precedent that answers this question. At present, each NSO would have to answer that question individually. One can argue, though, that in principle, the national laws of the NSO in country B should apply to confidential data accessed in the above manner. In addition, one can argue that accessing confidential data via remote access should be considered tantamount to transferring that data to the remote access site, if even temporarily.

220. A similar question in the context of copyright on the Internet arose in 1995. Among many questions asked, one was if viewing copyright protected work of a website via one's computer creates a copy of that content on the viewer's computer, making that copy subject to copyright protection. The legal question was focused on the demand in some countries, such as the United States, that a work will only be protected under copyright law if it is "fixed" in a tangible medium of expression, from which they can be perceived, reproduced, or otherwise communicated (17 USC Section 102(a)). In other words, when a person accesses a website via a personal computer, is the content being viewed on the personal computer sufficiently "fixed" to make it subject to the laws of copyright?

221. In 1995, this matter was discussed in the Report of the US Working Group on Intellectual Property Rights, "Intellectual Property and the National Information Infrastructure"[14] and the following conclusion was stated on page 28:

> "A simultaneous fixation (or any other fixation) meets the requirements if its embodiment in a copy or phonorecord is "sufficiently permanent or stable to permit it to be perceived, reproduced, or otherwise communicated for a period of more than transitory duration." Works are not sufficiently fixed if they are "purely evanescent or transient" in nature, "such as those projected briefly on a screen, shown electronically on a television or cathode ray tube, or captured momentarily in the 'memory' of a computer." Electronic network transmissions from one computer to another, such as e-mail, may only reside on each computer in RAM (random access memory), but that has been found to be sufficient fixation."

222. Data being accessed remotely for the purpose of statistical research should reasonably be considered as "fixed" in the random access memory of the accessing computer, since it must be sufficiently stable to be perceived for a period of more than a transitory duration in order for a researcher to make use of it. That data should therefore be considered as being transferred, if only temporarily, to the NSO where the remote access computer is located. The consequence of this argument is twofold. First, all such transfers will be subject to the GDPR, even if the processing of that data is not done on the computer to which the data has been transferred. Second, the data will be seen to be in the control and possession of the NSO where the remote access computer is located, and its use will be subject to the confidentiality and privacy laws of the country in which that NSO is located. Thus breach confidentiality by a researcher remotely accessing the data will implicate both the confidentiality laws of the country from which he/she is accessing the data and the country that is hosting and processing the data.

223. One could further argue that the researcher accessing the data remotely should be considered the "processor" for the purpose of the GDPR, even if the actual processing is being carried out remotely. Such a researcher has access to the data and gives instructions how to process it. The fact that the actual processing is done in a different country does not derogate from the researcher's role and responsibility. At

[14] http://www.uspto.gov/web/offices/com/doc/ipnii/ipnii.pdf

the same time, the NSO hosting the data, on whose computers the processing actually takes place, could be seen as controller and joint processor of the data.

Transfer of confidential data from non-EU member states

224. While the GDPR sets the rules for the transfer of EU data to non-EU member states, it does not (and cannot) establish when a non-EU member state is permitted, under the laws of that non-EU member state, to transfer data to an EU member state. In addition, the GDPR does not (and cannot) control the transfer of data between non-EU member states. But even if a non-EU member state's laws do not prohibit the transfer of confidential data to another country, that non-EU country might find that there is a lack of parity between the legal systems (including penal sanctions for breach of confidentiality) that would make such transfer undesirable.

225. For example, would an agency in the United States agree to transfer confidential statistical data to ONS for research purposes? The "Confidential Information Protection and Statistical Efficiency Act of 2002" (CIPSEA)[15] controls use of executive agency data for statistical research.

226. CIPSEA allows, inter alia, for research by external researchers ("agents"). The term "agent" refer, inter alia, to an individual, defined in Section 502(2)((A)(i) of CIPSEA "who is an employee of a private organization or a researcher affiliated with an institution of higher learning (including a person granted special sworn status by the Bureau of the Census under section 23(c) of title 13, United States Code), and with whom a contract or other agreement is executed, on a temporary basis, by an executive agency to perform exclusively statistical activity under the control and supervision of an officer or employee of that agency."

227. Under CIPSEA, could a researcher at ONS be made an agent who would be "under the control and supervision of an officer or employee" of a US executive agency? This is a question that only the relevant executive agency would be able to answer. There could be situations in which the answer would be positive -- for example, if the United States Census Bureau were to set up a data research centre on the premises of the ONS, to be administered, controlled and supervised by an officer or employee of the Census Bureau.

228. However, the US Census Bureau could decide not to set up such a data research centre due to the lack of parity between the United States and the United Kingdom regarding sanctions for breach of statistical confidentiality.

229. Section 513 of CIPSEA defines the criminal act of breach of statistical confidentiality as follows:

> "Whoever, being an officer, employee, or agent of an agency acquiring information for exclusively statistical purposes, having taken and subscribed the oath of office, or having sworn to observe the limitations imposed by section 512, comes into the possession of such information by reason of his or her being an officer, employee, or agent and, knowing that the disclosure of the specific information is prohibited under the provisions of this title, wilfully discloses the information in any manner to a person or agency not entitled to receive it, shall be guilty of a class E felony and imprisoned for not more than 5 years, or fined not more than $250,000, or both."

230. Article 39(9) of the UK Statistics and Registration Service Act 2007 states:

---

15 http://www.eia.gov/oss/cipsea.pdf. See also OMB "Implementation and Guidance for Title V of the E-Government Act, Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA), http://www.gpo.gov/fdsys/pkg/FR-2007-06-15/pdf/E7-11542.pdf

> "A person who contravenes subsection (1) is guilty of an offence and liable --
>
> (a) on conviction on indictment, to imprisonment for a term not exceeding two years, or to a fine, or both.
>
> (b) on summary conviction, to imprisonment for a term not exceeding twelve months, or to a fine not exceeding the statutory maximum, or both."

231. In the United States, breach of statistical confidentiality under CIPSEA is a felony offence, with a maximum sentence of five year. In the United Kingdom, it is a misdemeanour offence, punishable by a maximum sentence of two years. A US executive agency could reasonably decide not to transfer data to the United Kingdom, where that data would not enjoy the same protection that Congress mandated for it.

232. The above example also holds for the transfer of data (including remote access) between non-EU Member States.

Concluding remarks

233. OECD countries that are also EU Member States may transfer *personal data of natural persons* to each other (including by remote access) for research purposes, subject to the GDPR and agreements that set out the terms of use and obligations of each side. The laws of the country to which the data has been transferred shall apply to all aspects of use of that data, including penal sanctions for breach of confidentiality. It is our opinion that data accessed remotely shall be treated as if it has been transferred during the period it is being accessed. Thus all the laws of the country from which the data is being accessed, including for breach of confidentiality, shall apply to that access.

234. The above is also applicable to the transfer of personal data of natural persons from OECD countries that are also EU Member States to OECD countries that are not EU Member States but that have been recognised under the Directive as "third countries" entitled to receive such data.

235. The transfer of confidential *information on legal entities* for research purposes between OECD countries can take place when the parties have ascertained a parity of legal systems that enable effective enforcement upon transfer, subject to agreements that set out the terms of use and obligations of each side. As above, the laws of the country to which the data has been transferred, or from which the data is being accessed remotely, shall apply to all aspects of use of that data, including penal sanctions for breach of confidentiality. Data accessed remotely shall be treated as if it has been transferred during the period it is being accessed. Thus all the laws of the country from which the data is being accessed, including for breach of confidentiality, shall apply to that access.

PART III. INFORMATION AS AN ECONOMIC RESOURCE

CHAPTER 9. INTERNATIONAL ACCESS THROUGH PUBLIC USE FILES

by Luisa Franconi

Introduction

236. The release of datasets containing individual records stemming from social surveys was the first channel of microdata access used by National Statistical Offices (NSOs). This mode of access has developed through the years till the current situation where many NSOs, especially in Europe, have developed different types of files for different types of users. In recent years "Licensed files" for researchers (once called Microdata Files for Research purposes – MFRs) have been the most common way to access microdata covering the whole of Europe (see *Chapter 8*).

237. As one of the desired features expressed by users is the possibility to acquire data directly from the Web avoiding cumbersome procedures, there is a massive request for Public use microdata files, simply called Public Use Files (PUFs). A PUF is characterised by two principal factors: the content of the file that, although non-confidential, allows users to make sensible inferences on the phenomenon for which the data were collected, and the simple manner in which it is possible to reach it (see definition of PUF in the *Glossary*). Indeed, because of these characteristics, this type of file is ideal for a general purpose use in an international setting.

238. Other types of microdata are also freely available for users but they are not meant to make proper statistical analysis or to make inferences based on the data: these are Teaching Files and Test Data. Teaching Files are data sets containing very few records and few variables, in many cases they are oversimplified versions of survey microdata. Teaching Files share the same characteristics of the PUF in terms of its provision to users but it is different in terms of content. Due to the basic statistical disclosure control procedures used for its creation the Teaching File is not meant to make proper statistical analysis or to make inferences based on its microdata but it can be used solely to make data manipulation or to practice statistical methods. Examples of teaching data are presented in Section 2.6. Test data are microdata files that share the same logical structure of complex confidential microdata sets that statistical agencies make available to researchers through remote execution or remote access. The scores and values of the variables in the test data are either generated through stochastic processes or through procedures that destroy any relationship with real statistical units. These data aim at preserving consistency inside the record as well as complex structures and relationships between variables in order to allow users to develop correct code for the analysis of the microdata. They are becoming more and more common as they are very useful in dealing with microdata through remote execution channel; for example, IAB (the German Institute for Employment Research) has developed test data of the LIAB (Linked Employer Employee Dataset). In this case the aim was the preservation of complex relationships inside the data such as those between employers and employees, or employment histories, longitudinal structures and consistency between different spells and covariates.

Public use files: practices in selected OECD countries

239. This section aims at presenting methods adopted by statistical agencies in various OECD countries to limit disclosure when releasing Public Use Files; it is largely based on an Istat internal report (Ichim *et al*., 2012) and refers mainly to information collected between November 2011 and February 2012.

240. The examples analysed in this section concern the practices in the United States, Canada, France, Italy and Spain. Examples of teaching files from Germany and the United Kingdom are also presented.

United States

241. In the USA there is a long history of releasing public use files, the so called Public Use Microdata Samples (PUMS). US statistical agencies - US Bureau of Census, Bureau of Labor Statistics, Bureau of Transportation, etc. - release public use microdata files both on individuals and households as well as dwellings. In general, the geographical level released is the state; the US Census Bureau uses territorial units, PUMA, containing at least 100,000 inhabitants or super-PUMA (400,000 inhabitants), depending on the level of details of other possibly identifying variables.

242. Four different methods are used altogether to limit disclosure:

a) *Global recoding*, especially used for geographical information: the level of detail mostly used is the state level (the smallest state is Wyoming, about 500,000 inhabitants);

b) *Top-coding*, used mainly for ages and skewed continuous variables such as income;

c) *Sub-sampling*, always applied to all PUMS;

d) *Data swapping*, always applied, although with some variations, to all PUMS.

243. Other methods are routinely used as well such as rounding, noise addition and, lately, the use of synthetic data (see Zayatz, 2005).

244. To increase quality in the released microdata Rubin (1993) suggested that agencies release partially synthetic data, which comprise the original units surveyed with some collected values replaced with multiple imputations. The imputations are drawn from distributions designed to preserve important relationships in the confidential data. As reported by Drechsler and Reiter (2011) recently some statistical agencies have introduced disclosure limitation methods that maintain in the released public use files some statistical properties of the original microdata. The U.S. Federal Reserve Board in the Survey of Consumer Finances replaces monetary values at high disclosure risk with multiple imputations, therefore releasing a mixture of these imputed values and the unreplaced, collected values (Kennickell, 1997); the U.S. Census Bureau has released a partially synthetic, public use file for the Survey of Income and Program Participation that includes imputed values of Social Security benefits information and other highly sensitive variables (Abowd et al., 2006); the Census Bureau protects the identities of people in group quarters (e.g., prisons, shelters) in the American Community Survey by replacing quasi-identifiers for records at high disclosure risk with imputations (Hawala, 2008).

245. Table 1 (from Ichim *et al*. 2012) presents surveys carried out by statistical agencies in the United States, the indication whether a PUMS is released, methods applied and, where possible, link to further material available on the web.

Table 9.1. Public Use Microdata Samples (PUMSs) in the United States: Type of survey and main characteristics (Ichim et al. 2012)

Survey	Survey sponsor	PUMS	Note
The American Community Survey	Census Bureau	yes	Various editions, link: http://www.census.gov/acs/www/data_documentation/pums_data/ Geography: PUMA and super-PUMA (minimum 100.000 inhabitants). Data-swapping and top-coding. Sub-sampling (1% of the population instead of 2.5% used in the survey).
American Housing Survey	Department of Housing and Urban Development	no	
Current Population Survey and supplements	Bureau of Labor Statistics	yes	http://www.census.gov/cps/ Top-coding of Usual Hourly Earnings
Housing Vacancy Survey (module of CPS)		no	
National Survey of Fishing, Hunting, and Wildlife-Associated Recreation	U.S. Fish and Wildlife Service, U.S. Department of the Interior	yes	CD-ROM http://www.census.gov/prod/www/abs/fishing.html
New York City Housing and Vacancy Survey	New York City Department of Housing Preservation and Development	yes	http://www.census.gov/hhes/www/housing/nychvs/2008/data.html Dwellings. Available years: 1999, 2002, 2005, 2008. Top-coding on variables age and income.
Property Owners and Managers Survey	Department of Housing and Urban Development	yes	Year: 1995 Sub-sampling: Initial sample size: 16,300, Microdata file: 2504 records.
Residential Finance Survey	Department of Housing and Urban Development	yes	http://www.huduser.org/portal/datasets/rfs/censr-27.pdf Dwellings. Year: 2001 Methods used: Sub-sampling: Initial sample size: 69,000 residential addresses, Eligible: 48,000 properties, Data collected: 40,000 properties; PUMS: 22700 records. Local suppression Top-coding Page 310"The U.S. Census Bureau has modified or suppressed some data in this data release to protect confidentiality…… Data Swapping is a method of disclosure limitation designed to protect confidentiality in tables…"
Survey of Income and Program Participation	Bureau of Labor Statistics	yes	http://www.census.gov/sipp/pub_use.html Years: 2008, 2004, 2001, 1996, …
Survey of Program Dynamics		yes	http://www.census.gov/spd/ Available years: 1999, 2000, 2001 and 2002. http://www.census.gov/spd/pubs/spd98.pdf --- page 17 Geography: State level "income from every source is topcoded so that no individual amounts above $100,000 are revealed". "Other economic variables are topcoded at the 97 percentile level, meaning the top 3 percent of values are not disclosed." "We top-code age by bottom coding year of birth. For the 1998 SPD file no age will be older than 88."
American Time Use Survey	Bureau of Labor Statistics	yes	Various years available at http://www.bls.gov/tus/data.htm No geographical detail More than 50 variables are aggregated (recoded). rank swapping: "… and some responses were edited to protect the confidentiality of ATUS respondents.". variable suppression "Not all ATUS variables are on the files." Top coding of age and income
Consumer Expenditure Survey	Bureau of Labor Statstics	yes	Various editions; http://www.bls.gov/cex/#products

			CD-ROM; fees apply. Geography: State level.
National Health and Nutrition Examination Survey (NHANES) National Health Care Surveys (NHCS) National Vital Statistics System (NVSS) National Survey of Family Growth (NSFG) National Health Interview Survey (NHIS) National Immunization Survey (NIS)	Centre for Disease Control and Prevention	yes	Various surveys and years are available. Very little geographical details (the maximum level present in a file is: North, South, East and West). Policy http://www.cdc.gov/nchs/data_access/ftp_data.htm "… Users of NCHS public-use data files must comply with data use restrictions to ensure that the information will be used solely for statistical analysis or reporting purposes." "…Any effort to determine the identity of any reported case is prohibited by this law." "By using these data you signify your agreement to comply with the above-stated statutorily based requirements." ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NAMCS/doc00.pdf "It should be noted that, in an effort to preserve confidentiality, all geographic names, personal names, commercial names, and exact dates of injury have been stripped from the verbatim text." Top-coding of the variable age. Recoding (1=Northeast, 2=Midwest, 3=South, 4=West) ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2010/srvydesc.pdf "In addition, statistical noise at both the variable level and record level may have been added to allow for the protection of respondent confidentiality, and, at the same time, allow for release of files with as many variables as possible." Recoding (born in one of the 50 United States or the District of Columbia; born in a U.S. territory; or not born in the U.S. or a U.S. territory.) "Beginning in 2009 and continuing in 2010, for confidentiality reasons a decision was made to include only selected 4-digit external cause of injury codes in the public use file". "However, a review of NHIS data suggested that the level of detail contained in the codes could compromise respondent confidentiality. Consequently, beginning in 1997, the Census codes were restricted to in-house NHIS data files, and DHIS staff created several 2-digit industry and occupation recodes that could be included on the public use data files." "Age questions … are "top coded" to 85+ years to insure confidentiality among the oldest respondents" "Since 1997, the heights for men were top-coded to 76 inches and women's heights top-coded to 70 inches for confidentiality reasons." http://www.cdc.gov/nchs/data_access/data_linkage/mortality/nhis_linkage_public_use.htm perturbation: "The public-use versions of the NHIS Linked Mortality Files were subjected to data perturbation techniques to reduce the risk of respondent re-identification while increasing the accessibility of NCHS micro data files. Synthetic data was substituted for the actual date and underlying cause of death data for selected decedent records. Information regarding vital status was not perturbed."

Canada

246. Following the DLI (Data Liberation Initiative) a subscription-based service has been set up that offers institutional access to the collection of available Statistics Canada public use microdata files (PUMF). For a yearly fee, designated contacts at subscribing Canadian institutions can have unlimited access to all microdata and documentation available in the PUMF collection.

247. As far as the methods used and the production process are concerned, according to Boyko and Watkins (2003): "…The survey master files are processed in such a way as to reduce the probability of disseminating information about unique respondents. The survey managers are responsible for carrying out the disclosure control procedures by reducing the geographic detail, by suppressing variables and by collapsing response categories. The resulting public file is reviewed by the Microdata Release Committee consisting of a diverse range of senior managers. The survey managers are also expected to produce a codebook which documents the survey. In most cases, command files for popular statistical packages are also produced. Users are expected to sign an end user agreement that outlines their responsibilities and obligations. The main users of these files have been researchers in universities."

248. In some cases, *e.g.* for the Population Census, also example files not fitted for inferences and with limited number of records (2000) are available.[16]

France

249. The French National Statistical Institute, INSEE, releases freely on its web site public use files both on individuals and enterprises.[17] The data are available in Beyond20/20 or dBase format, metadata are in French. Table 2 shows the surveys for which data are available and the SDC methods used.

Table 9.2. Public use files available from INSEE: Name of the survey and their characteristics

Survey	Information on SDC methods used
Population census	20% sample or 25% sample.
Enterprise demography	15 variables on geographical location, NACE code, size class, legal form
Description des emplois privés et publics et des salaires en 2010	Sampling 1/12.
Labour Force Survey	Geography: region. 5-digit NACE code. Top-coding of age. Recoding of variable income, Several variables are derived from original ones.

Italy

250. The Italian National Statistics Institute (Istat) releases public use files free of charge from its web site[18] from surveys both on individuals/households and enterprises; metadata are in Italian. As Istat is already releasing Microdata File for Research purposes (MFR), the release of PUF needs to be consistent and coherent with what already released (see for example Trottini *et al*. 2006). This means that the information registered in the PUF, available to everyone, should be contained in the corresponding MFR, available to researchers only. The solution adopted is based on a subsampling from the corresponding MFR following two steps (see Casciano *et al*, 2011 and Foschi *et al*., 2012). In the first step both

16 http://www.statcan.gc.ca/kits-trousses/microdata-microdonnees/census-recensement2001/edu06g1_0001-eng.htm

17 http://www.insee.fr/fr/bases-de-donnees/fichiers-detail.asp

18 http://www.istat.it/en/products/microdata-files#file_microstat

disclosure risk and some data utility requirements are taken into account when determining the optimal sample allocation. The second step consists in drawing a random balanced sample, thus aiming at the approximate preservation of some weighted totals. The PUF and the MFR share the same structure. Such hierarchical structure of the two data sets greatly simplifies assessment of the disclosure risk and information loss associated with the anonymisation procedure and preserves the hierarchical detail as well as the internal consistency of the records. As a random subsample is drawn, the randomness feature of the microdata file is maintained as well. Moreover, the hierarchical structure between the two files allows for a reduction in the cost of preparation therefore increasing efficiency.

Spain

251. The Spanish National Statistics Institute releases free of charge public use microdata files from its webpage.[19] The statistical units contained in the files are mainly individuals but few files present also enterprises. In total 29 surveys are covered; metadata are in Spanish.

252. According to the notice present on the web, data "have been filtered appropriately to achieve anonymous information so as to ensure confidentiality"; no other information on methods used is available. The full text of the information on the webpage is available in the footnote[20].

Teaching datasets

253. Sometimes datasets are directly accessible from the web but, either because of the methods applied to them, or because of the small number of variables available, they do not allow to make any sensible inference from them. For this reason they are not classified as PUFs. We mention here a couple of examples.

Germany

254. In Germany, DESTATIS releases Campus file freely; the documentation is available in German.[21] Since 2004 sixteen files have been produced stemming from nine surveys.

255. The statistical disclosure limitation methods mostly used are:

a) *Global recoding*: geography in the Campus file is available only at NUTS1 (macro regions) or it is a binary variable (East-West). Information on economic activity, occupation, education and so on are highly aggregated.

b) *Top-coding* applied to age and income;

c) *Micro-aggregation* applied to continuous variables;

d) *Sub-sampling* always used in all files.

---

19 http://www.ine.es/en/prodyser/microdatos_en.htm

20 "This page includes a list of statistics that can be used to obtain microdata files. Microdata files comprise individual data for one given statistic, which have been filtered appropriately to achieve anonymous information so as to ensure confidentiality. These ASCII files have a field-based structure and garner the values for each variable for each individual survey recording. When using microdata files, any published information including data obtained thereby must quote the INE as the primary data source. Furthermore, the level of accuracy or reliability of the information derived drawn up by the authors is exclusively their responsibility".

21 http://www.forschungsdatenzentrum.de/en/campus-file.asp

256. Due to the heavy anonymisation of the data, DESTATIS recommend not to use this file for any statistical inference[22]. The German Institute for Employment Research, IAB, which applies the same methods, publishes on its website a similar note:[23]. Table 3 shows the surveys for which Campus files are available.

Table 9.3. Campus files available from DESTATIS: Type of survey and main characteristics

Survey	Available years	# records in the file	Sampling fraction
Microcensus	1976, 1998, 2002, 1996-1999	25.000 individuals	3.5%, except for 1976 (10%)
Continuing Vocational Training Survey	2000, 2006	2300	80%
Statistics of students winter semester	2000	200.000	10%
Statistics of examinations winter semester	2000	10.000	10%
Sample data of compulsory health insurants	2002	55.000	**0.7%, only 13 variables available**
Statistics of public assistance	1998	35.000	5%
Structure of earnings survey 2006	2001, 2006	60.000	5%, East-West
Cost structure survey at small and medium-sized businesses	1999	**500**	
Wage and income tax statistics 2001	1998, 2001	270 000	1%
Census of Agriculture	1999	4200 enterprises	

United Kingdom

257. The ESDS (Economic and Social Data Service) distributes from its website, on provision of an email address, few microdata files for teaching purposes ("The teaching dataset is a subset which has been

22 "CAMPUS Files have been created by the research data centres of the statistical offices of the Federation and the Länder especially for academic teaching at institutions of higher education. CAMPUS Files contain absolutely anonymised microdata which can be used by students to acquire methodological knowledge and to examine questions of social science or economics.

Each CAMPUS File is available as a download from the internet or can be obtained for a small charge on CD or DVD.

The CAMPUS Files have been specially developed for teaching purposes. Due to the strong reduction in information, they are usually not suitable for detailed analyses. As regards scientific analyses as part of diploma or doctoral dissertations, datasets are available in a less strictly anonymised form as Scientific Use Files (for domestic researchers) or for on-site use."

23 **"**The campus files have been specifically designed for the purpose of academic teaching and serve only for vivid and practical training of survey, data management or various other data analysis techniques applied to different social research problems**. Due to the utilized comprehensive anonymisation techniques the data must not be used for any statistical inference** concerning the contents or for any kind of publication (including seminar papers or bachelor theses). With this data it is not possible to make generalised statements about individual characteristics or relationships between different characteristics." http://fdz.iab.de/359/view.aspx

subjected to certain simplifications and additions for the purpose of learning and teaching").[24] They can be used only if the following statistical software is available: SPSS or STATA.

258. The SDC methods used to create such teaching dataset include:

a) *Global recoding*: geography is at National or macro-regional level;
b) *Top-coding*: mainly for variable age;
c) *Sub-sampling*: always used;
d) *Complete suppression of variables*: always used;
e) *Suppression of records*.

259. Table 4 gives an overview of the four surveys available at the time of the review, together with some information on their content.

Table 9.4. United Kingdom: Teaching files and their characteristics

Survey	Available years	Notes
Health Survey for England	2000, 2001	15 variables
British Crime Survey	2007-2008	35 variables, 11,676 record (sub-sampling 20%), Geography: UK
Quarterly Labour Force Survey	1 quarter: January-March 2011	13 variables 30% random sample (no aged 0-15); 25,162 respondents Age bands in 5 year intervals – 12 categories plus top coding Marital status – 3 categories (recoded) Geography: 13 regions http://www.esds.ac.uk/findingData/snDescription.asp?sn=7140
ONS Opinions Survey, Well-Being Module	April 2011	24 variables (13 well being questions and socio economic variables), 1124 records Marital status – 3 categories (recoded) Ethnicity white-other (recoded) Age (recoded), 6 categories, plus top coding Geography: 11 regions http://www.esds.ac.uk/findingData/snDescription.asp?sn=7146

3. Some guiding principles for the release of PUFs

Increase democracy and statistical literacy

260. The development of a public use file that shares quality and complexity of the corresponding original microdata file is based on the principles of democracy of access and right to research as well as the certainty that only by allowing students to be trained on complex official data statistical literacy will increase.

261. As Teaching Files are becoming more and more common as a quick and easy solution to the creation of files to be offered on the Web, Chapter 16 details the necessity for the development of proper PUF as opposed to teaching files. Here we only stress the features that are essential to be taught for

[24] http://www.ccsr.ac.uk/esds/data/

general statistical literacy: how to make reasoning and extract knowledge out of data. The value added of a PUF is the consequence of its quality standards defined as the ability to simulate real applications.

Improve impartiality, transparency and public trust

262. Dissemination and communication of “value added" analytic products require making specific choices about what data or patterns to show, what findings to interpret, what implications to discuss, but also what to bypass, ignore, or downplay. By increasing impartiality and neutrality of official statistics, microdata-based products definitely contribute to increase the transparency of the NSIs and, much more important, their credibility and public trust. Indeed, only access to microdata allows users to “clone” the NSIs estimates, perform analyses and comparisons, thus contributing also to a continuous innovation and customisation of the statistical system to the society information needs. The high standards of quality, as well as strict ethical and professional principles followed by the NSIs in the production of the statistical information should encourage more NSIs to open their microdata banks, in full compliance with confidentiality laws. Transparency, impartiality and neutrality can be increased by adopting microdata as a product on its own right and by distributing microdata in a more open manner like it is the case of PUFs.

Foster cross border access to microdata

263. Due to the basic terms under which PUFs are made available to users they can be considered the natural candidates to foster international access to microdata. The simple requirements needed to acquire such files make it perfectly feasible to allow easy cross border access. The adoption of PUFs guarantees that at least one access channel in the portfolio developed in the NSI can reach international users.

Quality PUFs to achieve the guiding principles

264. The guiding principles analysed so far can be fulfilled by developing PUFs that satisfy quality features: use SDC methodologies that preserve analytical characteristics of the data and inform users about the features that are preserved and those that are not maintained w.r.t. the original data.

265. The increase of transparency and public trust through the release of microdata can be achieved by adopting sound statistical methodologies for the statistical disclosure control (SDC) process which are clearly embedded in a data utility view. This means the adoption of the SDC paradigm (Hundepool *et al*. 2012)- disclosure risk assessment, implementation of disclosure control methods, evaluation of information loss – as well as the implementation of methods whose impact on the secondary analysis for the specific phenomenon under study is limited. So a careful implementation of methods that keep into account the specificity of the data is needed.

266. Transparency is achieved also by reporting the quality and the statistical properties of released microdata w.r.t. the original ones. All SDC methods most often used to produce public use files have some drawbacks as reported by Drechsler and Reiter (2011):

- Global recoding of geographic variables: aggregation of geography to high levels disables small area estimation and hides spatial variation;
- Top-coding: this method eliminates learning about tails of distributions which are often most interesting and degrades analyses reliant on entire distributions (Kennickell and Lane, 2006);
- Swapping: this method when used at high rates destroys correlations among swapped and not swapped variables (Winkler, 2007);

- Noise addition: adding random noise introduces measurement error that distorts distributions and attenuates correlations (Fuller, 1993).

- Stochastic perturbation: Elliott and Purdam (2007) show empirically that the quality of statistical analyses can be degraded even when using swapping or stochastic perturbation at modest intensity levels. These problems would only get worse with high intensity applications."

267. Sub- sampling techniques, always used to produce PUF, reduce the risk of disclosure by adding some uncertainty on the number of population units sharing the same score on the identifying variables (Hundepool *et al*. 2012). However, these methods if not carefully tailored to the specific need of SDC do not guarantee, by default, neither a controlled reduction of the disclosure risk nor the preservation of some data utility indicators.

268. The documentation of PUFs should clearly report on the consequences of the adopted SDC methods by describing possible different features of the estimates stemming from the original microdata and those coming out the released file or providing information of possible differences. The metadata released on a microdata file should contain information on which statistics are maintained, which inferences could be affected by these methods and at what extent.

269. As far as international access is concerned, in general released files are designed as single products to be made available to the "internal market" and issues of comparability with other countries and harmonisation at supranational level are not tackled. As reported in Section 2 different countries use different SDC methods with different levels of parameterisation. All this is perfectly acceptable as different countries have different laws, different cultural attitudes, and different availability of external information to be used for possible re-identification. Due to the variety of types of data and types of information to be released as well as release policy, it is not possible to think of a single set of SDC methods to be applied routinely and irrespectively of the type data. However, given the effort provided for the production of such files, a further step could be made to allow for better international comparisons. For example the results of a recent Essnet project on *Common tools and harmonised methodology for SDC in the ESS* [25](see Franconi and Ichim, 2012) showed that it is possible to adopt an overarching framework where different methods can be used by different countries without jeopardising utility for final users. This aim can be reached by agreeing on working under a common principle: for the development of microdata to be released NSIs agree to use SDC methods that maintain unchanged, w.r.t. the original data, a set of predefined benchmarking statistics/indicators (Ichim and Franconi, 2010). To this end if NSIs share the same working framework then the cooperation through the development of thematic network of expertise could bridge the knowledge gap between different NSIs and agencies and help in reaching the global goal of comparability at transnational level.

270. To achieve the principle of public trust, NSI should consider the opportunity of coordinating both internally, by releasing different products to different users coherently , and externally, inside the national statistical system or with other institutions delivering data to the public.. In fact, when releasing different files (teaching files, PUFs, MFRs) a clear strategy in the adoption of multiple releases and in the design of different products from the same survey is needed in order to avoid disclosure by differencing (Ichim and Franconi, 2010). Moreover, as open microdata files from various different government agencies, data archives, local authorities and smart cities projects[26] are becoming more and more visible on the web coordination inside national statistical systems and among data producers is recommended. Also awareness and continuous monitoring for new big data initiatives from possible relevant sources is crucial.

[25] http://neon.vb.cbs.nl/casc/..%5Ccasc%5CESSNet2index.htm

[26] http://setis.ec.europa.eu/implementation/technology-roadmap/european-initiative-on-smart-cities

271. Lastly, transparency and trust can be fulfilled also by developing efficient ways in which the microdata is provided to users and changing the relationship with them. These themes are developed in *Chapter 14* of this report where a new dissemination system and its development through the approach of industrialisation is proposed. However, as detailed in *Chapter 14*, the presence of powerful tools to find and analyse vast data stores is not sufficient: tools are not the final answer. Adequate competences should be developed to choose suitable methods, apply them correctly and understand and interpret results. Statistical literacy should be constantly promoted and supported by NSOs; indeed, it is expected that cooperation initiatives will pave the way for changes in the relationship between producers and users of official statistics. Users will not just be data analysts but will be more and more called for an active contribution toward the improvement of official statistics.

Recommendations

272. This Chapter identifies different levels of maturity as the production and release of PUFs is concerned and analyses how to improve international access to microdata through systemic production of high quality PUFs.

273. The complete absence of PUFs in the dissemination plan of an NSO or its sporadic presence without any specific treatment at the design stage for quality and utility aspects is an indicator of low maturity in international microdata access.

274. The first step to be made in order to move up to a medium level of maturity is the adoption, at least in a case by case manner, of the SDC paradigm - disclosure risk assessment, implementation of disclosure control methods, evaluation of information loss – and the initial use and subsequent presentation to users of quality checks carried out on the released microdata.

275. NSOs at a high level of maturity have implemented the systemic use of the SDC paradigm guided by data utility where care is taken in adopting methods that preserve crucial characteristics of the data (such as the randomness of the sample) and the possibility to reproduce, as much as possible, published tables/main statistics. These NSOs consider PUFs an important dissemination product for international access to microdata, and for the sake of comparability they widely implement SDC methods that satisfy the overarching principle of preserving important statistical properties of the data/benchmarking statistics.

276. It is recommended to adopt transparency along the production process. This means, *firstly*, report to users the effects that implemented SDC methods have on main and more relevant analyses (for the underlying phenomenon) and clearly state how inferences could be affected. *Secondly*, to invest in the development/adoption of methods allowing complete transparency without increasing disclosure risk. In order to reach transparency and comparability at transnational level the establishment of structured thematic centers of expertise to share knowledge, increase collaboration and network NSIs staff and interested organisations are highly recommended. The implementation of user-centric dissemination system with fully implemented resource discovery features coupled with training programs will foster the development of culture and knowledge around microdata.

277. Finally, NSOs should take the leadership in the coordination of PUFs and open microdata releases inside their respective national statistical systems, to raise awareness on the interaction between big data and the released microdata, be proactive in increasing dialogue with new actors of open microdata dissemination such as data archive and govern microdata dissemination inside their national statistical system.

REFERENCES

Abowd, J.M., Stinson, M. and Benedetto, G. (2006). Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. Available at: http://www.census.gov/sipp/FinalReporttoSocialSecurityAdministration.pdf

Boyko, E. and Watkins, W. (2003). Safe data, safe places: not either/or solutions. In 19th CEIES seminar *Innovative solutions in providing acces to microdata* Lisbon, 26 and 27 September 2002, pp. 109 – 115, European Commission.

Casciano, C., Ichim, D. and Corallo, L. (2011). Sampling as a way to reduce risk and create a Public Use File maintaining weighted totals. In: Joint UNECE - Eurostat work session on statistical data confidentiality, Tarragona, Spain, November, 2011.

Drechsler, J. and Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets, Computational Statistics and Data Analysis, 55, 3232 - 3243.

Elliott, M. and Purdam, K. (2007). A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymized Records. Environment and Planning A 39, 1101–1118.Foschi, F., Casciano, C., Ichim, D. and Franconi, L. (2012). Designing Multiple Releases from the Small and Medium Enterprises Survey, In J. Domingo Ferrer and I. Tinnirello (eds) Proceeding of PSD2012, Vol. 7556, Lecture Notes in Computer Science, pp 200-215. Springer, Berlin/Heidelberg.

Fuller, W.A. (1993). Masking Procedures for Microdata Disclosure Limitation. Journal of Official Statistics, Vol.9, No.2, 1993. pp. 383–406.

Hawala, S. (2008). "Producing Partially Synthetic Data to Avoid Disclosure," Proceedings of the Section on Government Statistics, American Statistical Association.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E. and De Wolf, P. (2012). *Statistical Disclosure Control*, Wiley series in Survey Methodology, Wiley & Sons, Chichester, UK.

Ichim, D. and Franconi, L. (2010). Strategies to achieve sdc harmonisation at European level: multiple countries, multiple files, multiple surveys. In Privacy in Statistical Databases, PSD 2010 (eds. Domingo-Ferrer J. and Magkos E.), vol. 6344 of Lecture Notes in Computer Science, pp. 284–296. Springer, Berlin/Heidelberg.

Ichim, D., Franconi, L. and Corallo, L.(2012). Rilascio di file ad uso pubblico a livello internazionale – opzioni metodologiche e condizioni di utilizzo. Istat internal report.

Little, R. (1993). Statistical analysis of masked data. Journal of Official Statistics, vol. 9, pp. 407–426.

Kennickell, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson, eds., Record Linkage Techniques, 1997, 248-267. Washington, D.C.: National Academy Press.

Kennickell, A. and Lane, J. (2006). "Measuring the Impact of Data Protection Techniques on Data Utility: Evidence from the Survey of Consumer Finances," in Berlin: Springer Verlag, pp. 291-303.

Lochner, K., Bartee, S., Wheatcroft, G. and Cox, C. (2008). A practical approach to balancing data confidentiality and research needs: the NHIS linked mortality files. In Privacy in Statistical Databases, PSD 2008. Domingo-Ferrer J. and Saygın Y. (eds.), vol. 5262 of Lecture Notes in Computer Science, pp. 90–99. Springer, Berlin/Heidelberg.

Rubin, D. (1993). Satisfying confidentiality constraints through use of synthetic multiply-imputed microdata. *Journal of Official Statistics*, vol. 9, pp. 461–468.

Trottini, M., Franconi, L. and Polettini, S. (2006). Italian Household Expenditure Survey: A Proposal for Data Dissemination, in: Privacy in Statistical Databases 2006, (eds.) J. Domingo-Ferrer, L. Franconi, LNCS 4302, Springer, Berlin pp. 318−333.

Winkler, W. E. (2007). "Examples of Easy-to-implement Widely Used Masking Methods for which Analytic Properties are not Justified," http://www.census.gov/srd/papers/pdf/rrs2007-21.pdf

Zayatz, L. (2005). Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update. Research Report Series (Statistics #2005-06), Statistical Research Division, U.S. Census Bureau, Washington, D.C.

CHAPTER 10. LICENSING PUBLIC USE FILES[27]

by Brian Negin

1. Introduction

278. Discussions on international collaboration on microdata access generally focus on access to confidential microdata by researchers. Certain research projects do indeed require very detailed microdata, which could potentially allow for the identification of individual units. For many studies, however, less detailed microdata files satisfy the researchers' needs, in particular non-confidential microdata files commonly referred to as Public Use Files (PUFs) made available by the NSOs of some countries. In terms of international access to microdata, PUFs are an ideal vehicle since, in principle, they do not contain confidential information and access to them is, or should be, simple.

279. Yet, there exists confusion regarding what constitutes a PUF and under what conditions it can be provided to the public. The purpose of this document is to make some order out of this, from a legal point of view, with the goal of offering guidelines that could reduce barriers to international access of these microdata sets.

280. As explained in the *Glossary*, a PUF definition requires two parts: a description of the "product" being offered, and the terms under which it is made available.

281. PUF datasets can be provided to researchers by different means. Files can be made available on a website, in a manner that allows individuals to download them at their own convenience. PUF datasets can also be provided by NSOs to individual researchers on a physical medium, such as a disk, or the NSO can send them to individual researchers by email. In some countries, a social science data archive may curate, maintain, and disseminate the PUF. The channel by which the dataset is provided is not material to our consideration whether or not it is a Public Use File. However, the different ways in which the datasets are provided may impact on how they are licensed (if they are licensed at all).

2. Legal and methodological considerations regarding confidentiality

282. A Public Use File must be anonymised to the extent that it is not considered to contain confidential data according to the law and methodology applicable to the NSO producing it. National laws are not always clear on what legal standard should apply -- and even when there is a legal definition, applying that definition to a particular dataset from a methodological point of view may be less than clear.

283. Following are some examples:

i) Article 3, Paragraph 7, of Regulation (EC) No 223/2009 of the European Parliament and of the Council: *Confidential data means data which allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information. To determine whether a statistical unit is identifiable, account shall be taken of all relevant means that might reasonably be used by a third party to identify the statistical unit.*

[27] A version of this Chapter completed with Annexes is available on OLIS as document STD/CSTAT/MICRO(2013)9.

ii) Article 19 of Regulation (EC) No 223/2009 of the European Parliament and of the Council:
Data on individual statistical units may be disseminated in the form of a public use file consisting of anonymised records which have been prepared in such a way that the statistical unit cannot be identified, either directly or indirectly, when account is taken of all relevant means that might reasonably be used by a third party.

iii) Draft definition of "data subject" in the proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), 25 January 2012 (COM(2012) 11 final):
'data subject' means an identified natural person or a natural person who can be identified directly or indirectly, by means reasonably likely to be used by the controller or by any other natural or legal person, in particular by reference to the identification number, location data, online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that person;

iv) Draft definition of "data subject" by the Committee on Civil Liberties, Justice and Home Affairs of the European Parliament, in its Draft Report on the proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), 17 December 2012 (2012/0011(COD)):
'data subject' means an identified natural person or a natural person who can be identified or singled out directly, alone or in combination with associated data or indirectly, by means reasonably likely to be used by the controller or by any other natural or legal person, in particular by reference a unique identifier, location data, online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity or sexual orientation of that person;

v) Section 39(3) of the United Kingdom's Statistics and Registration Service Act 2007, regarding "personal information":
For the purpose of subsection (2) information identifies a particular person if the identity of that person -
(a) is specified in the information
(b) can be deduced from the information
(c) can be deduced from the information taken together with any other published information

vi) Section 502 (4) of the United States Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA):
The term "identifiable form" means any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means.

vii) Section 17(1)(b) of Canada's Statistics Act:
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 an identifiable individual person, business or organization.

viii) Clause 7 (1) of Australia's Statistics Determination 1983, relates to confidential information as:
information relating to a person, other than information of a personal or domestic nature which is likely to enable the identification of that person

ix) Section 308(d) of the United States Public Health Service Act:
….. such information may not be published or released in other form if the particular establishment of person supplying the information or described in it is identifiable ……

284. The above sampling shows legislative language on a spectrum from absolutes to nuance. At one end of the scale, with language that seems quite absolute, one has: the United Kingdom, where "identification ***can be deduced*** from the information together with any published information"; and Canada, referring to identification "in such a manner ***that it is possible*** from the disclosure to relate the particulars …to an identified individual." At the more nuanced end of the scale, one has: Regulation (EC) 223/2009, "To determine whether a statistical unit is identifiable, account shall be taken of ***all relevant means that might reasonably*** be used by a third party to identify the statistical unit"; and the DGPR drafts, "by means ***reasonably likely to be used****"*.

285. Regardless of the stance taken by a legislative body, it is proposed that from a methodological point of view, there is no possible means of determining if identification can be deduced without taking into account "all relevant means that might be reasonably be used" by the public. These means include, but are not limited to, other microdata sets that could be linked to the PUF files, data on individuals available through commercial data mining sites, data on individuals publicly available on social communication sites, such as Facebook and Twitter.

286. From the above, one may also deduce that the confidentiality of a Public Use File could be compromised by the use of means outside of the spectrum of "all relevant means that might be reasonably used". The degree of this risk varies from file to file, depending on a variety of factors, including the aggressiveness of statistical disclosure control methods used to anonymise the data. Each NSO must strike a balance between the usefulness of the file for research and the degree of anonymisation applied, taking into account the risk of identification of an individual unit.

287. Due to the risks involved, some NSOs do not publish or make available Public Use Files. Others may make them available for educational use (such as the DESTATIS CAMPUS files). Some might make them available subject to limitations on attempts at identifying the individual -- such as a note in the National Health Interview Survey warning that "any effort to determine the identity of any reported case is prohibited by this law" -- even if the validity of such a warning is doubtful, as shall be discussed below.

288. In conclusion, an NSO could have legitimate reasons for imposing limitations on the use of Public Use Files as regards attempts to identify individual units, though it is suggested that such limitations are virtually unenforceable, as shall be discussed below. Guidelines and Best Practices for SDC of PUFs could assist in harmonizing levels of disclosure risk, leading to reduced reliance on user obligations meant to protect the identity of individual units in the dataset.

289. Confidentiality of the data might not be the only concern of an NSO. Public Use Files come under the category of "Protected Works" (as compilations of data or databases) according to international copyright conventions and national law. The next section of this report will focus on the legal status of PUF datasets from this standpoint -- and the implications arising from the conclusions.

3. Intellectual property rights in Public Use Files

290. Two international conventions set the baseline for copyright protection for compilations/databases. PUF files come under the category of compilations/databases. Pursuant to local legislation in line with these conventions, copyright protection extends to the selection and arrangement of datasets. This is known as "thin protection" because it does not apply to the content of the datasets -- i.e. the data itself.

291. Article 5 of the WIPO Copyright Treaty (1996) (WCT) states:

> Compilations of data or other material, in any form, which by reason of the selection or arrangement of their contents constitute intellectual creations, are protected as such. This protection does not extend to the data or the material itself and is without prejudice to any copyright subsisting in the data or material contained in the compilation.

292. Article 20 of the World Trade Organization Agreement on Trade Related Aspects of Intellectual Property Rights (TRIPS)[28] states:

> Compilations of data or other material, whether in machine readable or other form, which by reason of the selection or arrangement of their contents constitute intellectual creations shall be protected as such. Such protection, which shall not extend to the data or material itself, shall be without prejudice to any copyright subsisting in the data or material itself.

3.1 The standing and purpose of international copyright conventions

293. A discussion about the consequences of inclusion of Public Use Files in the category of "protected works" under the WCT and TRIPS requires first some understanding of the international copyright conventions that set the baseline for this protection.

294. Underlying both the WCT and TRIPS is the Berne Convention for the Protection of Literary and Artistic Works.[29] Created in 1886, this is the primary copyright treaty in the world, which both the WCT and TRIPS reference themselves to. 166 nations are signatory to the Berne Convention, including all OECD member countries. The Berne Convention is administered by the World Intellectual Property Organization (WIPO), which also administers other intellectual property conventions, including the WCT.

295. Both the WCT and TRIPS base themselves on the Berne Convention and then add to it, each in its own way.[30] The two sections referred to above regarding compilations of data do not appear in the Berne Convention, but according to both WCT and TRIPS, they are subject to the terms of the Berne Convention. Therefore, one must understand what baseline the Berne Convention sets.

296. The Berne Convention establishes the basic principles of copyright protection that member states are required to legislate in their own national legislation. In addition, it establishes the principle of national protection -- a protected work shall enjoy copyright protection under the laws of each and every member state. For example, a work created in the United Kingdom will enjoy the protection of French copyright law. Below is the language of this protection, from Article 5 of the Berne Convention.

---

28 The TRIPS agreement is Annex 1c of the Marrakesh Agreement establishing the World Trade Organization. TRIPS gateway: http://www.wto.org/english/tratop_e/trips_e/trips_e.htm. TRIPS text: http://www.wto.org/english/docs_e/legal_e/27-trips.pdf

29 http://www.wipo.int/treaties/en/ip/berne/trtdocs_wo001.html

30 Article 3 of the WCT: "Contracting Parties shall apply mutatis mutandis the provision of Articles 2 to 6 of the Berne Convention in respect of the protection provided for in this Treaty." Note 4 of the WCT states: *Agreed statement concerning Article 5:* The scope of protection for compilations of data (databases) under Article 5 of this Treaty, read with Article 2, is consistent with Article 2 of the Berne Convention and on a par with the relevant provisions of the TRIPS Agreement.
Article 9 of TRIPS: "1. Members shall comply with Articles 1 through 21 of the Berne Convention (1971) and the Appendix thereto. However, Members shall not have rights or obligations under this Agreement in respect of the rights conferred under Article 6*bis* of that Convention or of the rights derived therefrom. 2. Copyright protection shall extend to expressions and not to ideas, procedures, methods of operation or mathematical concepts as such".

(1) Authors shall enjoy, in respect of works for which they are protected under this Convention, in countries of the Union other than the country of origin, the rights which their respective laws do now or may hereafter grant to their nationals, as well as the rights specially granted by this Convention.
(2) The enjoyment and the exercise of these rights shall not be subject to any formality; such enjoyment and such exercise shall be independent of the existence of protection in the country of origin of the work. Consequently, apart from the provisions of this Convention, the extent of protection, as well as the means of redress afforded to the author to protect his rights, shall be governed exclusively by the laws of the country where protection is claimed.
(3) Protection in the country of origin is governed by domestic law. However, when the author is not a national of the country of origin of the work for which he is protected under this Convention, he shall enjoy in that country the same rights as national authors.
(4) The country of origin shall be considered to be:
(a) in the case of works first published in a country of the Union, that country; in the case of works published simultaneously in several countries of the Union which grant different terms of protection, the country whose legislation grants the shortest term of protection;
(b) in the case of works published simultaneously in a country outside the Union and in a country of the Union, the latter country;
(c) in the case of unpublished works or of works first published in a country outside the Union, without simultaneous publication in a country of the Union, the country of the Union of which the author is a national, provided that:
(i) when these are cinematographic works the maker of which has his headquarters or his habitual residence in a country of the Union, the country of origin shall be that country, and
(ii) when these are works of architecture erected in a country of the Union or other artistic works incorporated in a building or other structure located in a country of the Union, the country of origin shall be that country.

297. TRIPS establishes the same principle of national treatment in its Article 3:

1. Each Member shall accord to the nationals of other Members treatment no less favourable than that it accords to its own nationals with regard to the protection of intellectual property, subject to the exceptions already provided in, respectively, the Paris Convention (1967), the Berne Convention (1971), the Rome Convention or the Treaty on Intellectual Property in Respect of Integrated Circuits. In respect of performers, producers of phonograms and broadcasting organizations, this obligation only applies in respect of the rights provided under this Agreement. Any Member availing itself of the possibilities provided in Article 6 of the Berne Convention (1971) or paragraph 1(b) of Article 16 of the Rome Convention shall make a notification as foreseen in those provisions to the Council for TRIPS.

2. Members may avail themselves of the exceptions permitted under paragraph 1 in relation to judicial and administrative procedures, including the designation of an address for service or the appointment of an agent within the jurisdiction of a Member, only where such exceptions are necessary to secure compliance with laws and regulations which are not inconsistent with the provisions of this Agreement and where such practices are not applied in a manner which would constitute a disguised restriction on trade.

298. "National treatment" is of importance to international collaboration on microdata access. It means that copyright protection is available regarding any dataset protected under national law. That dataset shall enjoy copyright protection throughout the world, in all the member states signatory to the Berne Convention. The nature of this protection is the topic of the following section of this Chapter.

299. International conventions on copyright do not, in and of themselves, have legal force as laws. They require that the signatories to them legislate in their national laws the terms of the conventions, subject to the exceptions noted in each convention. Therefore, the conventions serve as a baseline for understanding the principles underlying international copyright law, while the exact manner of implementation must be examined at the national level. By way of example, and pertinent to the following discussion of copyright in Public Use Files, the United States Federal Government does not have copyright in works it has created. Title 17 USC Section 105[31] states:

> Copyright protection under this title is not available for any work of the United States Government, but the United States Government is not precluded from receiving and holding copyrights transferred to it by assignment, bequest, or otherwise.

300. However, it is argued in the Copyright Law Revision (House Report 94-1476)[32], p. 59, that such protection does not extend beyond the borders of the United States:

> The prohibition on copyright protection for United States Government works is not intended to have any effect on protection of these works abroad. Works of the governments of most other countries are copyrighted. There are no valid policy reasons for denying such protection to United States Government works in foreign countries, or for precluding the Government from making licenses for the use of its works abroad.

301. Regardless of the position of the United States on extraterritorial copyright protection in Federal Government works, it is noted that Public Use Microdata Sets (PUMS) made available to the public via Federal Government agency websites, are distributed freely, without any claims to copyright in them.[33] These files can be accessed from anywhere in the world -- so for all practical purposes, it would seem that there is no contention on the part of the US Government that it exercises copyright control over these files.

3.2 Copyright in a dataset

302. A "protected work", such as a database, acquires copyright protection upon its creation. No formalities are required. No registration is required. Upon creation, the "author" (or another individual or entity designated by the national copyright law or by agreement) receives a set of exclusive rights in the work, subject to limitations and exceptions.

303. One such limitation is the term of protection. Article 7 of the Berne Convention sets a minimum term of copyright protection as the life of the author and fifty years after his death.

304. However, copyright in works created in the course of employment (including by government employees) are generally deemed, under national laws, to belong to the employer, unless otherwise agreed upon. Copyright in PUF files created by employees of an NSO will belong to the NSO, an entity that is not a natural person.

305. Article 12 of TRIPS addresses copyright terms for non-natural persons.

> Whenever the term of protection of a work, other than a photographic work or a work of applied art, is calculated on a basis other than the life of a natural person, such term shall be no less than 50

---

31 http://www.law.cornell.edu/uscode/text/17/105

32 http://en.wikisource.org/wiki/Copyright_Law_Revision_(House_Report_No._94-1476)

33 A table listing PUMS in the United States, with links to the sites, can be found in *Chapter 9*.

> years from the end of the calendar year of authorized publication, or, failing such authorized publication within 50 years from the making of the work, 50 years from the end of the calendar year of making.

306. Therefore, the copyright in a PUF file will expire no sooner than 50 years from the end of the calendar year of its authorized publication, as determined in national legislation.

307. The set of exclusive rights in the work is known as a property right -- or intellectual property right. Like all rights in property, the rights of the copyright owner are valid towards all others to the extent determined under each country's national law. The exercise of the right is not dependant on a contract or agreement between the copyright owner and anyone else.

308. One of the several limited exclusive rights under copyright, relevant to the use of datasets, is the right of reproduction (copying) of the work.

309. Article 9(1) of the Berne Convention sets the baseline for the right of reproduction. It states:

> (1) Authors of literary and artistic works protected by this Convention shall have the exclusive right of authorizing the reproduction of these works, in any manner or form.

3.3 Allowing use of copyright protected datasets: licenses and agreements

310. The use of a protected digital work on a computer involves copying of it, for example from its source to the user's computer and from memory storage on the computer to RAM. It has long been held that such use necessarily implicates the copyright owner's exclusive right of reproduction in the work.[34] Therefore, use of that protected work on a computer requires authorization from the copyright owner.

311. Copyright owners grant others the right to reproduce their works by way of "license". A license determines to what extent rights reserved to the copyright owner may also be exercised by others. Generic copyright licenses are made readily available today via Creative Commons.[35]

312. A license, such as the CC license mentioned above, might be expressed as part of a contractual agreement between the copyright owner and another party or parties. Such agreement would include, in addition to the copyright license, obligations undertaken by the other party that are outside the realm of copyright protection. Such undertakings could include the other party's payment of money for the use of the protected work, an undertaking to maintain the confidentiality of the data in a dataset, or an undertaking to refrain from data linking and attempting to identify individual units in the dataset. An example of such a license agreement is provided by Statistics Canada.[36]

---

34 *MAI Systems Corp. vs Peak Computer, Inc*., 991 F.2d 511, 518 (9th Cir. 1993); Working Group on Intellectual Property Rights of the U.S. Information Infrastructure Task Force, "Intellectual Property and the National Information Infrastructure", http://www.uspto.gov/web/offices/com/doc/ipnii/ipnii.pdf, pp. 65 - 66; *Religious Technology Center vs. Netcom On-Line Communication Services, Inc*. 907 F. Supp 1361, (N.D. Cal., 1995), http://www.law.cornell.edu/copyright/cases/907_FSupp_1361.htm

35 http://creativecommons.org/licenses/by/3.0/legalcode

36 Statistics Canada, License Agreement for Public Use Microdata Files, http://search1.odesi.ca/documentation/CCHS_2012/Microdata_license_agreement.pdf

313. While copyright licenses rarely appear alone, outside of license agreements, one must be aware that there are significant differences between "pure" licenses and licenses secured within agreements that must be taken into consideration when providing licensed datasets to the public.

314. Licenses can grant use of a copyrighted work to everyone simply by stating the terms of the license in reference to the work. Agreements require that both parties express their agreement to the terms of the agreement. This poses challenges, especially when enabling the creation of such agreements in an online environment.

315. License agreements that are buried in the terms of use of an NSI website and that do not require an action on the part of an end user expressing agreement to its terms, may not meet the minimum requirement for the creation of a binding contract. Two examples of such license agreements can be found on the Statistics Canada website[37] and the ONS website[38].

316. The CC license referred to above is based on a system intended to bring the license terms to the knowledge of the end user when a work is provided to the public on the Internet. There are three levels to this system. The first is a notice that the work is protected. This could be through embedding machine readable language in the dataset using Creative Commons Rights Expression language[39] or a human readable link and notice, such as Some rights reserved. Clicking on the link opens a new page with a "human readable summary of the legal code", linked to the legal code of the full license. A disclaimer on the summary page declares that the summary is not the license, only the legal code of the full license constitutes the license. The legal code of the full license declares, among other things, in full capital letters, as follows:

> BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. TO THE EXTENT THIS LICENSE MAY BE CONSIDERED TO BE A CONTRACT, THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

317. While the CC system is arguably more effective in bringing the license terms, including the terms of the agreement, to the attention of the end user, one might still question the efficacy of the system in creating a binding contractual agreement between the licensor and the licensee. At the same time, the terms of the CC licenses are almost wholly copyright oriented, meaning that their terms relate to the permitted exercise by the licensee of the licenser's exclusive rights in the licensed work. The CC licenses also contain non-copyright related statements intended to protect the licensor against legal claims based on implied warranty and to limit the extent of the licensor's legal liability. To the extent these are binding only upon agreement by the licensee, their binding nature will depend on the efficacy of the mechanism for creating that agreement.

318. License agreements that require that an end user click an "I agree" button before access to a dataset is allowed would create a binding agreement. However, without knowing the identity of the person who agreed to the terms, enforcement of the agreement might well be impossible.

319. Licenses can apply to anyone, anywhere. Agreements only bind the party's to them. Thus, the agreement based terms of a "copyright license" (such as the undertaking to preserve confidentiality or to

---

37 Statistics Canada Open Licence Agreement, http://www.statcan.gc.ca/eng/reference/copyright-droit-auteur-eng

38 Non-Commercial Government Licence for public sector information, http://www.nationalarchives.gov.uk/doc/non-commercial-government-licence/

39 http://wiki.creativecommons.org/Ccrel

refrain from data linking) will apply personally to the user. If the user should transfer the dataset to another user who is not party to the agreement, that undertaking will not bind the additional user. However, the dataset will continue to remain protected under copyright.

320. Licenses are valid only during the term of copyright protection. If protection in a dataset expires after 50 years of publication, then the dataset moves into the public domain, and use of the dataset can be made freely by anyone. Contracts are binding for the term set in them -- which can be forever. Thus, when the copyright term expires, one could have the agreement terms continuing while the copyright protection ceases. The dataset would move into the public domain, usable by all. On the other hand, someone holding the dataset by dint of a license agreement might still be bound to protect the confidentiality of the file or to refrain from data linking if that obligation has been stipulated in the agreement. This is not a desirable situation, which could be mitigated if the license agreement includes a condition by which the user undertakes to not transfer the dataset to anyone else.

321. Licenses can be enforced in all countries signatory to the Berne Convention and TRIPS, under the principle of "national treatment". However, this is relevant only as regards infringement of the copyright owner's rights in the dataset, for example by violation of the license terms pertaining to a dataset. Agreements are enforceable under contract law by the courts of the country designated in the agreement and according to the national law of the country designated in the agreement. Thus, the enforcement of the contractual elements of a license agreement might be limited to laws and in the courts of the country of origin of the dataset, while action against copyright infringement could take place wherever the infringement occurred, according the copyright law of the country in which it occurred.

322. In countries where no copyright exists in the dataset being released (such as in the United States regarding federal agency datasets), there is no option for licensing datasets - either under pure licenses or under license agreements. However, the absence of copyright in the dataset does not preclude the option of binding the end user to an agreement by which he/she undertakes to refrain from activities that could compromise the confidentiality of the dataset. One such example is the “Data User Agreement” for the access to public use datasets provided by the National Center for Health Statistics in the United States. The agreement only relates to the protection of the confidentiality of the data and the permitted use of the datasets (for statistical analysis and reporting only).

323. This agreement is accessed through a link on the "home page" of access to the public use files on the CDC website.[40] One is not required to read the terms of the agreement before accessing the datasets available to the public. This "agreement" may or may not have binding effect on the user.

324. From the above, it is possible to summarize as follows.

- Copyright licenses that do not include any element of agreement, protect only the exclusive right of the copyright owner. They cannot include undertakings by the end user to refrain from data linking or from attempts to identify individual units. Therefore, a dataset made available under such a copyright license must ensure that the disclosure risk of the dataset is extremely small, approaching zero.
- Such copyright licenses can be enforced internationally under the principle of "national treatment" under the Berne and TRIPS conventions. This means that infringement of copyright in country B relating to a dataset released under license in country A, can be enforced in country B under country B's copyright laws.
- Since copyright licenses do not require the agreement of the end user, no process is required to affect such agreement and to preserve a record of that agreement.

[40] http://www.cdc.gov/nchs/data_access/ftp_data.htm ; http://www.cdc.gov/nchs/data_access/restrictions.htm

- Copyright licenses have a life span concurrent with the copyright protection afforded by the relevant national law -- no shorter than 50 years from the time of publication.
- License agreements allow inclusion of obligations on the part of the end user, such as those that are intended to enhance protection of confidentiality of the data. License agreements might also include additional terms the NSI deems important, such as attribution to the NSI as the source of the data should research be published based on the dataset and disclaimers meant to protect the NSI against liability.
- License agreements, because they are agreements, require some act on the part of the end user to indicate his/her agreement to its terms. To the extent that the (non-copyright related) terms under the agreement are important to the NSI, positive identification of the end user might also necessary. License agreements expressed somewhere on a web page and that do not require the end user to explicitly express his/her agreement as a condition to accessing the dataset, might not be legally enforceable as contracts.
- License agreements can stipulate what national law should apply, and which nations courts should have jurisdiction, in the case of a breach of the agreement terms.
- License agreements can contain non-copyright related obligations that do not expire, unlike copyright licenses that are subject to copyright term limits.

4. Representations, warranties and disclaimers

325. Copyright license agreements will often include statements by the licensor intended regarding representations, warranties and disclaimer. Following are some examples from the attached license agreements.

i) Sections 5 and 6 of the CC license state respectively:

Representations, Warranties and Disclaimer
UNLESS OTHERWISE MUTUALLY AGREED TO BY THE PARTIES IN WRITING, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTIBILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO SUCH EXCLUSION MAY NOT APPLY TO YOU.

Limitation on Liability.
EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ii) Section 10 of the Statistics Canada PUF license states:

The Collection is provided ‘as-is,’ and Statistics Canada makes no warranty, either express or implied, including but not limited to, warranties of merchantability and fitness for a particular purpose.

iii) The Statistics Canada Open License Agreement states:

No Warranty and no Liability
The Information is licensed 'as is', and Statistics Canada makes no representations or warranties whatsoever with respect to the Information, whether express or implied, in relation to the Information and

expressly disclaims any implied warranty of merchantability or fitness for a particular purpose of the Information.
Statistics Canada or any of its Ministers, officials, servants, employees, agents, successors and assigns shall not be liable for any errors or omissions in the Information and shall not, under any circumstances, be liable for any direct, indirect, special, incidental, consequential, or other loss, injury or damage, however caused, that you may suffer at any time by reason of your possession, access to or use of the Information or arising out of the exercise of your rights or the fulfilment of your obligations under this agreement.

iv) The UK Open Government License states:

No warranty
The Information is licensed 'as is' and the Information Provider excludes all representations, warranties, obligations and liabilities in relation to the Information to the maximum extent permitted by law.
The Information Provider is not liable for any errors or omissions in the Information and shall not be liable for any loss, injury or damage of any kind caused by its use. The Information Provider does not guarantee the continued supply of the Information.

326. France's INSEE has a very brief license for its PUF files (referred to later in this paper), which includes the following disclaimer (translated from the French):

> By express agreement, in all cases, no warranty, explicit or implied, is given by INSEE, either for direct or indirect damage, commercial, financial or any other cause. The use made of the file is under the sole responsibility of the user, especially as regards the results obtained from them.

327. Such disclaimers are meant to dispel any claims for implied or explicit warranty in the product being provided to the public under consumer protection or contract laws. They could also be intended to protect the licensor against claims in tort for negligence regarding nature of the product being offered. A notice such as "use at your own risk" would serve such a purpose.

328. Such disclaimers can be included in the terms of a license agreement or as part of a license not based on agreement. They can be placed conspicuously on a website at the point of access to a dataset as a free standing notice.

329. The efficacy of such disclaimers is dependent on the national law. Disclaimers that are part of contractual agreements could be subject to laws regulating unfair contracts or contract terms, or consumer protection laws. Notices not based in contract would be judged in the context of tort or other applicable law.

5. Creation of agreements and identification of users

330. National laws determine when a binding agreement between parties is created. For the purpose of this paper, the terms "agreement" and "contract" are synonymous. However, there might be countries that distinguish between them both as regards the necessary conditions for their creation as well as regarding their content. Common law countries might require "consideration" as necessary condition for creating a binding contract, in which case the contract must indicate what consideration has been given.

331. As noted above, two issues are of relevance in the context of this paper regarding the creation of Internet based agreements. The first is the necessity for some positive evidence that the end user has agreed to the terms of the agreement. Browser based agreements within a website that do not require the user to click on an "I agree" button or statement, are of questionable legal value.

332. NSOs must also consider whether or not the identity of the user, agreeing to the terms of use of the dataset, is important. Online registration to use a dataset may or may not provide accurate information about the person registering. The most succinct statement of this reality was made 20 years ago in Peter Steiner's famous cartoon published by The New Yorker on July 5, 1993, captioned: "On the Internet nobody knows you're a dog."

333. It is suggested that the more concerned the NSO is about protecting the confidentiality of the data in a PUF dataset, the more vigorous it should be about ensuring that the agreements it makes with the end users are legally binding and enforceable against well-defined individuals.

334. An institutional PUF provides a good means of establishing legal responsibility by an identified entity – the institution licensing the dataset for its members. The institution must enforce confidentiality restrictions on its own end users, and bears responsibility for breaches of the contract/license terms. This method also allows for one point of licensing while allowing an indeterminate number of researchers to access and use the datasets.

6. License terms not read or agreed to by the user

335. License terms based entirely on copyright (meaning – the terms of the license do not require the agreement of the end user) might be included on an NSO's website in a manner that does not ensure or guarantee that the end user accesses and reads those terms. This should not invalidate the terms of the license.

336. By default, a database protected under copyright may not be used without permission of the copyright owner. By placing a copyrighted dataset on the Internet for access, the copyright owner is granting use of that dataset to others. The extent that others may make use of the dataset is governed by the license terms relating to that dataset, whether or not a person has read those terms. The licensor would be prevented ("estopped" in legal terms) from enforcing copyright in the dataset in contradiction to the conditions that were stated in the license. At the same time, it is in the copyright owner's interest to make it clear to end users that the dataset is protected by copyright, particularly if the copyright owner is concerned about how the dataset is to be used. For example, if the copyright owner wants to restrict use of the dataset to personal use only and prohibit commercial exploitation or redistribution of the dataset to others – this should be explicitly brought to the attention of the end user in plain and understandable language.

7. Licensing issues for PUF datasets - with examples from NSOs

337. As noted earlier in this paper, there is no international standard for determining when a dataset is safe for use as a Public Use File. Assessing the level of disclosure risk is as much an art as it is a science. There are many unknown variables to take into consideration. One can assume that there is always some risk of identification of an individual unit -- but that risk could vary from infinitely small to something tolerable which does not quite turn the dataset into confidential by the definitions referred to above.

338. Therefore, the manner in which a PUF dataset is licensed, if at all, is a function of the NSO's assessment of the disclosure risk involved as well as other legal and other factors that might be peculiar to an individual country -- such as the lack of copyright in US federal works (including databases and datasets).

339. Where the NSO is convinced that the risk of identification of an individual unit is so remote as to make it effectively zero, it might consider making the PUF available:

- With minimum license terms (such as INSEE's) and without any terms requiring an agreement
- Under a Creative Commons license, or a government open license agreement such as in Canada and the United Kingdom. PUF files of this nature could be made freely available online for downloading. It is recommended that before allowing the download of the dataset, that the terms of the license are presented to the user, and that code in the dataset link directly to the CC license. CC provides the necessary code and support for such linking.

340. Where the NSI is concerned about disclosure risk and wishes to constrain users by terms of use that require them to refrain from certain activity -- such as attempting to identify individual units or linking the dataset to other datasets or giving the dataset to an unauthorized individual -- then a more aggressive means of creating a license agreement (or an agreement sans license) should be considered.

341. The quandary here is less in the terms of will be included in the license agreement (*i.e.* permitted usage, obligations, etc.) than how the agreement will be created. Since the NSI would be making such an agreement because it is concerned about confidentiality, then it might follow that the NSI would also want to know who it has made an agreement with so that any breaches of the agreement can be enforced.

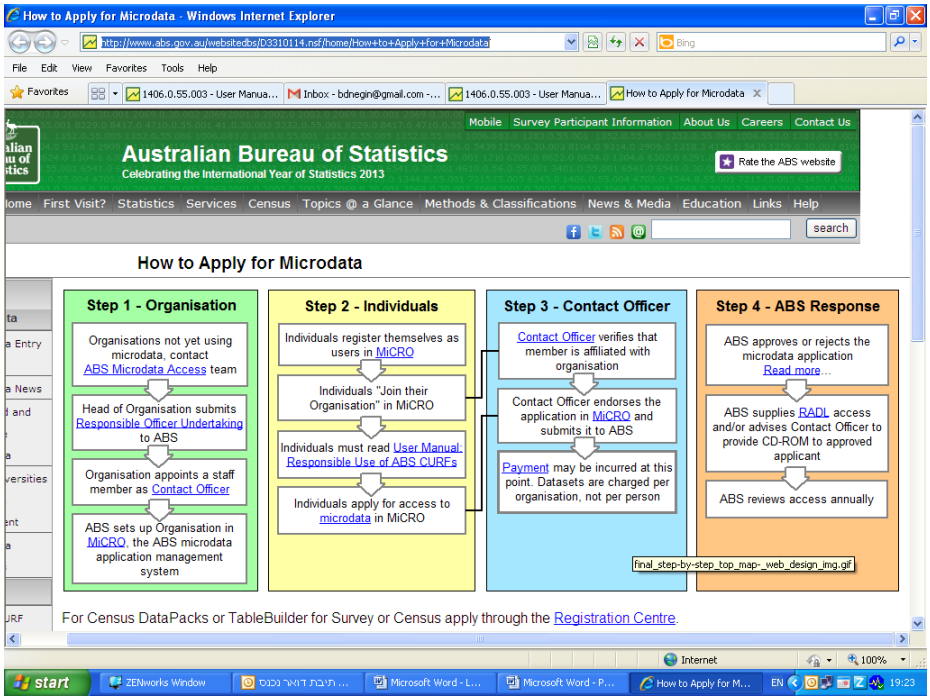
342. One way to deal with this is as Statistics Canada has done. Public Use Files are made available to institutions of higher education under the Data Liberation Initiative, whereby the free use of the datasets can be made by recognized members (students, faculty) of the educational institution.[41] The PUF datasets are provided to the subscribing institutions under a license agreement, which contains terms of use similar to those in Statistics Canada's PUF license agreement for individual end-users. The obligation to ensure use of the datasets in accordance with the license agreement terms is upon the institution signing the agreement. The license agreement requires that the contracting institution bring the terms of usage to the attention of the institutional users. Sanctions for breach of the terms of the license include discontinuation of the service.

343. The Australian Bureau of Statistics (ABS) provides researchers with CURF datasets via their institutions. According to the ABS website, CURF datasets are "Confidentialised Unit Record Files" covering responses to ABS surveys containing very detailed data while also protecting the confidentiality of individuals. The datasets are made available via institutional affiliation, and each request is assessed to ensure that the use intended is for a legitimate statistical purpose. Regarding the statistical purpose, ABS website states:

41 The Data Liberation Initiative, http://www.statcan.gc.ca/dli-idd/dli-idd-eng.htm

> "The statement of statistical purpose [in the application] need not be lengthy but should show that there is a clear objective to the analysis and demonstrate that access to the CURF is essential to be able to undertake that analysis. Statements such as 'academic research' or 'policy research' without further detail are not in themselves adequate descriptions of statistical purpose, and will not be approved by the ABS."

344. ABS does not license its CURFs -- it provides them at an institutional level subject to an undertaking by a "responsible officer". The application process for use of a CURF is not simple, and is intended to ensure the identity of the end user, the end user's institutional affiliation, and the purpose of the research.



345. Should the CURF datasets, under the above circumstances, be considered PUF files or scientific use files? On one hand, ABS states that the datasets are "confidentialised" -- though it is not clear from that statement if the datasets themselves are non-confidential. It would appear that the vetting process – approving specific research requests of specific researchers within the institution, would preclude including CURF datasets in the category of institutional PUFs.

346. France offers online through INSEE what would be appear to be "pure" PUF datasets without any restrictions of use, under the following license terms (Microdata INSEE):

> Files and documentation are owned by INSEE. They can be downloaded free of charge and the data contained in the files can be reused, including for commercial purposes without a license and without payment of royalties.
>
> By express agreement, in all cases, no warranty, explicit or implied, is given by INSEE, either for direct or indirect damage, commercial, financial or any other cause. The use made of the file is under the sole responsibility of the user, especially as regards the results obtained from them.
>
> INSEE provides no service of any kind whatsoever, including consulting, apart from the documentation provided on the files themselves.

The files are typically offered in dBase format and in Beyond 20/20 ®.

347. Italy through Istat, now offers Public Use Files, downloadable via the Internet. Their new microdata web site provides a gateway to microdata access, including public use mIcro.STAT files. These are described as follows:

> Public use mIcro.STAT files are collections of elementary data downloaded directly from the Istat website. The mIcro.STAT files are developed for some surveys starting from the relative file for research purposes, as a subsample. They contain a lower level of detail in comparison with files for research purposes.
>
> Because of the sampling, the calculations performed on the file mIcro.STAT may lead to results in some extent differing from those published. To acquire such files it is necessary to register at the area of the Istat website dedicated to them and to accept the terms of use.

348. The above examples illustrate the wide variety of options open to an NSO or other national statistical agency for providing datasets to the public -- as PUFs, PUMS or under any other name one might wish to use. In general, it is proposed that the manner of dissemination is determined, first and foremost, by limitations under national law. Adding to the heterogeneous nature of national laws is the fact that there are no agreed upon standards as to when a dataset is so safe it can be released to the public without any limitations. While some countries do not seem to have a problem with this (the United States and France), others seem to be more hesitant to allow microdata files to be used by the public without some type of limitation (Australia, Canada, Italy). Then there are those countries that simply do not produce Public Use Files due to disclosure risk, for example the Netherlands and the Scandinavian countries.

349. The above examples also illustrate that some countries (United States and France) make their PUF datasets available at no cost and without restrictions regarding use and re-use. Such datasets could be considered "open data" in accordance with the EU initiative on "Open Data".[42] However, even PUF

---

42 http://ec.europa.eu/digital-agenda/en/open-data-0. Below is a summary of the Directive 2003/98/EC that provides the legal framework for the open data initiative.

- Charges for re-use have to be limited at a ceiling calculated on the basis of actual costs. Public sector bodies need to calculate charges per re-user in a way so that the total income from charging does not exceed the costs incurred to produce and disseminate the information, together with a reasonable return on investment.
  Public sector bodies are encouraged to apply lower charges or to apply no charges at all. On request, public sector bodies must indicate the method used to calculate charges.
- Conditions for re-use shall be non-discriminatory for comparable categories of re-use.
- Prohibition of cross-subsidies: If public sector bodies re-use their own documents to offer added-value information services in competition with other re-users, equal charges and other conditions must apply to all of them.
- Prohibition of exclusive arrangements: Public sector bodies may not enter into exclusive arrangements with individual re-users, excluding others. Such exclusive rights may only be authorised in exceptional circumstances if they are necessary to provide services in the public interest.
- Charges and other conditions for re-use have to be pre-established and published. If a request for re-use is refused, the grounds for refusal and the means of redress need to be explained.
- Requests for re-use shall be processed within a specific timeframe (20 days for standard cases).
- Licences should not unnecessarily restrict possibilities for re-use or be used to restrict competition. Member States are encouraged to use standard licences in digital format

datasets with restrictive licensing, where appropriate, could be considered "Open Data" under the EU initiative if they meet all the other criteria pertinent to the initiative. Article 8 of Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information states:

Licences

1. Public sector bodies may allow for re-use of documents without conditions or may impose conditions, where appropriate through a licence, dealing with relevant issues. These conditions shall not unnecessarily restrict possibilities for re-use and shall not be used to restrict competition.
2. In Member States where licences are used, Member States shall ensure that standard licences for the re-use of public sector documents, which can be adapted to meet particular licence applications, are available in digital format and can be processed electronically. Member States shall encourage all public sector bodies to use the standard licences

350. PUF files are capable of providing a tool for international collaboration on the use of microdata among those countries that make such datasets available to the public. The concluding section of this paper summarises recommendations in this respect.

8. Recommendations and Conclusions

351. Guidelines and Best Practices for Statistical Data Control of PUFs should be developed to assist in harmonising levels of disclosure risk, leading to reduced reliance on user obligations meant to protect the identity of individual units in the dataset.

352. Ideally, PUF datasets should be provided to the public only when the risk of identification of an individual unit approaches zero in the estimation of the NSO producing them, in order to reduce reliance on user obligations.

353. Such datasets should be made available as "open data" under the guidelines of Directive 2003/98/EC. The datasets should be made available to anyone, anywhere, without restrictions regarding how the dataset will be used or re-used, and without terms by which the end user agrees to refrain from attempting to identify individual units. In countries that hold copyright in the datasets and wish to enforce the copyright, minimum license terms should suffice, such as a CC license appropriate to the needs of the NSO. Such licenses do not require the agreement of the end user, and the CC licensing system is meant to ensure that the license terms are brought to the attention of the end user. These licenses are also enforceable internationally under international copyright treaties. Datasets should be provided without limitations regarding use or re-use. If they are provided subject to license terms, such terms should not unnecessarily restrict possibilities for re-use and should not restrict competition.

354. License terms, as well as license agreements, should not be buried within an NSO's website where the chance of them being read is extremely small. License agreements on a web site should only be used when the end user is required to provide positive proof of agreement – such as by clicking on a statement attesting to that agreement.

355. Countries that do not hold copyright in the dataset, or for whom copyright is simply not an issue, should make the datasets available free of license terms (such as for the United States PUMS), or subject solely to disclaimers and wavers (such as in France).

356. Should an NSO have concerns regarding the disclosure risk of a dataset, even when the dataset is considered non-confidential by national standards, it should consider at least three options.

- The first is to refrain from making the dataset available as a PUF.

- The second option is to make the dataset available under a license agreement by which the end user undertakes to refrain from attempting to identify individual units. While such an agreement could be an anonymous "click through" agreement on a website, its anonymous nature would undermine the purpose of the agreement – the ability to enforce it should it be breached. Ideally, should an NSO wish to prohibit attempts at identifying individual units in a PUF, it should require positive identification of the end user in addition to a positive manner of ascertaining agreement to the contractual terms. It is recommended that such an agreement be used only for end users who reside in the country providing the dataset, as international enforcement would be very difficult. The NSO should strive to keep the license agreement terms within the parameters of Article 8 of Directive 2003/98/EC as well as to ensure that the additional parameters of "open data" set out the Directive are followed.

- A third option is institutional licensing, such as under Statistic Canada's DLI. The institution would undertake contractual obligations to protect the integrity of the PUF datasets and to provide access to them only to its members. Institutional licensing would also provide an option for international access to PUF datasets, especially when the NSO providing the dataset wishes to impose conditions of use on the end users, such as obligations to refrain from attempts to identify individual units. In the event of a breach of contractual terms by the institution (or its end users), the NSO would recourse to more than one sanction option – such as refraining from providing datasets in the future or civil law suits for breach of contract or copyright infringement. Contracts could also contain clauses for agreed damages and other sanctions that do not require court action.

357. In conclusion, there is bad news and good news. The bad news is that there is no simple solution for providing PUF datasets to the public, particularly due to concern for the risk identification of individual units in spite of the dataset being considered "non-confidential" under national law. The good news is that there are a variety of options available to NSOs for making such non-confidential datasets available to researchers as Public Use Files in accordance with national legislation and the disclosure risk associated with such files. It is hoped that the examples brought in this paper, in conjunction with the legal analysis can serve to guide NSOs in pursuing the dissemination of PUF datasets – especially in the international sphere -- for the furtherance of academic and scientific research.

CHAPTER 11. MICRODATA EXCHANGE AND THE CHALLENGES OF OPEN DATA AND TRANSPARENCY

by Paul Jackson

Introduction

358. This Chapter is based on the experiences of the Office for National Statistics, but is not a statement of the policy of that office or of the United Kingdom government. It discusses some of the policy and legislative initiatives for Open Data in the context of microdata in national statistics institutes. The drivers and solutions for Open Data in microdata are distinct from those for the exchange of confidential microdata, but challenges in either activity cannot be addressed without an understanding of them both.

359. For the purposes of this report, "Open Data" are data (datasets) that are:

- *accessible* to anyone and everyone, ideally via the internet,
- in a *digital machine readable* format that allows interoperation with other data, and
- available at *reproduction cost or less,*
- and are *free from restrictions on use and re-use*.

360. Thus for the purpose of discussion of Open Data for this Expert Group and its reports, the relevant microdata are the individual records of direct observations and facts obtained through instruments of statistical inquiry (such as surveys), registers, or administrative sources. The data are modified before dissemination only where necessary to address obligations of confidentiality to the subjects of the observations. Such data might be called Open microData. The purpose of the paper is to explore expectations that micro-data exchange should increasingly be based on Open Data principles, and the issues that arise for producers of official statistics.

Drivers for Open Data

361. The drivers for development of microdata as Open Data include the adoption of scientific principles, support for democracy, stimulus for social and economic growth, and response to legislation and public policy.

Scientific principles

362. Open Data supports the status of statistics as a scientific discipline.

363. The Fundamental Principles of Official Statistics[43] require the methods of dissemination for official statistics data to be decided according to *scientific principles*. Users' interpretation of the data is facilitated when they are presented according to *scientific standards*.

364. What are these scientific principles and scientific standards for the dissemination and presentation of data? The Fundamental Principles suggest they pre-exist and are familiar.

[43] http://unstats.un.org/unsd/methods/statorg/FP-English.htm

365. The Royal Society published its report "Science as an Open Enterprise" in June 2012[44]. The report reflects on science as a self-correcting process. Theories can be independently corroborated, invalidated or improved. Findings, and the supporting data, are presented to the widest possible audience for further development.

366. Producers of official statistics should challenge themselves as to whether their published statistics - their assertions of patterns in observations and facts - can be independently corroborated, invalidated, or improved. Open Data allows this scientific principle to act on official statistics. Published methodology allows abstract challenges only, unless the data as used by the producer of the statistic is also available to another independent party. This can be achieved to some extent through channels of controlled access such as statistical peer review, or through research access to data. However, maximum transparency is achieved when no restrictions, selections or pressures (whether real or imaginary) are brought to bear upon the independent scrutiny of the statistics.

367. Public trust and confidence may suffer if the data that underpin scientific information and statistics are not freely available along with the methods used. For example, public trust in the publications of the Climate Research Unit of the University of East Anglia was undermined when some data and methods were withheld from the public domain. Two inquires were held, and the Government's response was to recommend the full disclosure of the raw data along with the necessary computer programmes and methodologies to replicate the results.

> "The disclosure of raw data and sufficient details of the computer programmes is paramount in encouraging people to question science in the conventional way, challenging existing work, enabling validation of it and coming forward with new hypotheses".
>
> - Government Response to the Science and Technology Committee's First Report of Session 2010-12[45]

Indispensable element for democracy

368. The *UN Fundamental Principles of Official Statistics* position official statistics data as "...an indispensable element in the information system of a democratic society."

369. The Open Government Partnership is an initiative to make governments better. It was founded in September 2011 with 8 government members, and is now expanded to 55 government members. The Partnership declaration includes a commitment to increase the availability of information about government activities. The partners are expected to:

> "...proactively provide high-value information, including raw data, in a timely manner, in formats that the public can easily locate, understand and use, and in formats that facilitate reuse."
>
> - Open Government Declaration, Open Government Partnership[46]

---

44 http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf

45 http://www.publications.parliament.uk/pa/cm201012/cmselect/cmsctech/496/496.pdf

46 http://www.opengovpartnership.org/open-government-declaration

370. Implementing the *UN Fundamental Principles*, most national statistical systems will incorporate statutory objectives that include reference to support for the scrutiny of public policy. The primary objective of the UK Statistics Authority and its executive office ONS reads:

> "The Board is to have the objective of promoting and safeguarding the production and publication of official statistics that serve the public good. [T]he reference to public good includes in particular informing the public about social and economic matters, and assisting in the development and evaluation of public policy."
>
> - Statistics and Registration Service Act 2007[47]

Open Data for economic and social growth

371. A 2011 study conducted for the European Commission estimated that the direct and indirect economic gains when public sector information is open for re-use are in the order of 140billion euros. [48]

372. The *OECD Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information* (2008) makes several relevant recommendations for Open Data in statistical microdata, in order to "increase economic and social benefits in particular through more efficient distribution [of information], enhanced innovation and development of new uses". [49]

373. Member countries are recommended to adopt its principles, which include these of relevance to statistical microdata:

- *Openness.* Maximising availability, with open access as the default rule.
- *Transparent conditions for re-use*. Non-discriminatory conditions, eliminating exclusive arrangements and removing unnecessary restrictions.
- *Asset lists*. Inventories of Open Data, published online.
- *Quality and integrity*. Use of best methods in data preparation and protection from misrepresentation,
- *Pricing*. Free of charge, or cost recovery only
- *International.* Facilitation of cross-border use and interoperability

374. These recommendations provide a framework for NSOs to assess their adoption of Open Data principles in microdata.

Legislation and public policy for Open Data

375. Legislation and policy for Open Data take two basic forms. First, those laws and public policies that oblige the public sector to *push* Open Data. And second, those laws and public policies that entitle users to *pull* Open Data from the public sector.

---

47 http://www.legislation.gov.uk/ukpga/2007/18/section/7

48 Vickery G. (2011), Review of recent studies on PSI re-use and related market developments, http://ec.europa.eu/information_society/policy/psi/index_en.htm

49 http://www.oecd.org/internet/interneteconomy/40826024.pdf

'Push' legislation and policy

376. Many NSOs are subject to provisions in law and policy that establish an expectation that they will push Open Data to the public under a regulatory regime.

377. For example, in the European Union the public sector information market was first regulated by the Directive 2003/98/EC on the re-use of public sector information.[50] The Directive is now under revision to strengthen its provisions and to respond to the digital world. The Directive does not, strictly speaking, oblige the production of public use documents, but provides a positive regulatory framework for adoption in the legislation of member states, with an assumption that public sector information should be Open within reasonable limitations. For example, the Directive enshrines the principle that the total income from supplying and allowing re-use of data should not exceed the cost of collection, production and dissemination and a reasonable return on investment.

378. Implementation of the Directive has been achieved in all member states, and some have gone beyond its requirements in both national law and national policy. For example, the United Kingdom has adopted national Public Data Principles binding on the public sector that include an obligation to actively encourage the use and re-use of departments' public data. [51]

379. The public sector copyright rules in the United Kingdom ("Crown Copyright") have been modified to establish a 'push' principle. The default copyright licence for the UK public sector is the Open Government Licence (OGL).[52] The OGL is a positive, proactive, 'push' licence:

> "[The government] grants you a worldwide, royalty-free, perpetual, non-exclusive licence to use the Information... You are free to copy, publish, distribute and transmit the Information; adapt the Information; exploit the Information commercially for example by combining it with other Information, or by including it in your own product or application."
>
> - UK Open Government Licence

380. A UK public body that wishes to obtain an exception to marginal cost recovery must seek accreditation to the Information Fair Trader Scheme, submit a business case, and be subject to the scrutiny of the Office for Public Sector Information.[53]

381. In the future 'push' policies and structures can be expected to develop, often in response to popular campaigns.[54] The UK is now launching the Open Data Institute as a public/private partnership.[55] It is established expressly to push Open Data out of the public sector:

> "...a collaboration between our leading businesses and entrepreneurs, universities and researchers, government and civil society to unlock enterprise and social value from the vast amount of Open Government Data now being made accessible."
>
> - About the ODI, www.theODI.com

---

50 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32003L0098:EN:NOT

51 http://www.data.gov.uk/library/public-data-principles

52 http://www.nationalarchives.gov.uk/doc/open-government-licence/

53 http://www.nationalarchives.gov.uk/information-management/ifts.htm

54 http://www.freeourdata.org.uk

55 http://www.theodi.org/people/nrs

'Pull' legislation and policy

382. Many NSOs are also subject to examples of laws and policies that empower the public to pull data from the public sector. Examples include the Environmental Information Regulations, and the Freedom of Information Acts of many countries. It is usually the case that statistical data are subject to the pull power of these laws and policies, with exemptions only where it can be shown that the data are confidential or scheduled for future publication in the form requested. The impact of such laws on statistical data can be significant:

- 'Pull' laws typically have a time limit for compliance. The NSS in the United Kingdom must respond with either the data requested or an explanation of why an exemption applies within 20 days of the request.
- In combination with the Open Government Licence (or equivalent), limitations and conditions on use and re-use cannot be imposed.
- The applicant typically has a right to appeal a decision to refuse or to comply on in part to the request for data. The appeal is typically heard by an Information Tribunal or other such non-statistical authority. Whilst the statistical office may present its evidence for why the data should not be disclosed (for example, for reasons of confidentiality), the decision lies with the Information Tribunal.
- Typically, the application for data is 'purpose blind'. The applicant does not need to explain why the data are request, nor what their intended uses are for the data.
- Often there is no exemption for data that are of poor quality.
- The release of data under 'pull' legislation is often of a precedent-setting nature.
- Unless handled carefully, release of data under 'pull' legislation can look like privileged access.
- Release of datasets can undermine confidentiality through 'mosaic attack'.
- Disclosure control issues are often non-obvious to the uninitiated.

383. The push and pull laws and policies are now enabling information entrepreneurs, from financial analysts to App designers, to invest in confidence in data exploitation. The industry resulting exerts its own pull on the public sector data owners.

Challenges of producing Open (micro)Data

384. How are NSOs responding to the challenges of Open Data? With their statistics, we might argue the response is very good. Fundamental practices in statistical presentation and dissemination are by their nature compatible with Open Data principles. However, where the challenges of Open Data are applied to the 'raw' microdata of NSOs, the response is currently less convincing. Typically, a statistical office will assign their data assets into three categories:

- "Statistics" for publication or for use in public administration - usually aggregate data or visualisations.

- "Research datasets" for statistical research uses under controlled conditions - usually microdata and usually marked 'confidential'.
- "Production data", being the microdata or aggregate data within statistical production systems and which is not expected to be used for any purpose other than the derivation of statistics and the derivation of research datasets.

385. The concept of 'public use files' is familiar, but current practice is usually to assign them to the "research datasets" category. Usually, that is the correct classification, because most public use files cannot be called Open Data, in that there are terms and conditions of use. Neither NSOs nor the users of public use files currently have the expectation that such files could be used to replicate the statistics produced by the NSO - not least because the NSO is not using the public use file themselves to produce the statistics. Thus the scientific principle of independent corroboration, verification, or improvement of the statistics is not achieved through public use files in the manner in which they are currently produced.

386. It is not common current practice to recognise Open Data as a category for microdata and assign information assets to that category for dissemination under those conditions. This may be because of the many challenges presented by producing Open microData

Confidentiality

387. Confidentiality issues are the most frequently cited reason for not producing Open microData. Very often it will be an entirely legitimate reason. However, it is common practice for an NSO to consider all its unpublished data to be confidential data either by default, or by virtue of its status as 'data held for the purpose of official statistics'. This practice is unlikely to survive the push or pull legislation and policies for Open Data. First, it should be recognised in the policies of the NSO that some data assets are inherently less likely to raise confidentiality issues than others - for example, public sector budget and expenditure data, or prices data, may be identified microdata but not confidential due to the information being already available in the public domain. Second, a more systematic and critical analysis of whether a data asset is truly disclosive or not should be established in NSOs. This may involve so-called 'penetration tests', whereby a trusted party is provided with a candidate dataset and allowed to see whether, under a reasonable test, any private and confidential information about identified individuals can be discovered from the data. If such personal information can be discovered, the NSO has obtained independent evidence to present to an Information Tribunal or equivalent, helping them to sustain a challenge against them withholding the data from an applicant. If such personal information cannot be discovered through the penetration test, the NSO may have a candidate Open microDataset.

388. Data Protection Supervisors recognise the difficulty of distinguishing between personal and non-personal information. The Article 29 Working Party Opinion on the concept of personal data helps NSOs with the task of identifying what is, and what is not, personal data.[56] The UK Information Commissioner has recently published a Code of Practice on anonymisation of personal data and managing data protection risk, with the help of Office for National Statistics and many others.[57]

---

[56] The Data Protection Working Party is an independent European advisory body on data protection and privacy. Its tasks are described in Article 30 of Directive 95/46/EC and Article 15 of Directive 2002/58/EC, http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf

[57] http://www.ico.gov.uk/news/latest_news/2012/~/media/documents/library/Data_Protection/Practical_application/anonymisation_code.ashx

Data quality and reputation

389. Of all the factors that affect public confidence in official statistics, data quality may be the most important. A reputation for good quality statistics is hard won, and easily lost. It is entirely understandable that Open Data, and especially Open microData, is seen as a threat to a reputation for good quality statistics. The threat may arise from two scenarios. First, the data may well be of poor quality. If the data are 'pulled', they may be released before quality assurance and before additional input data comes in to the NSO. Second, the data may be of good quality, but may be interpreted differently or even wrongly by the other party. The NSO may have to explain the differences and justify them to a public that may-be prefer the non-government party to be 'right'. These issues are to be dealt with by excellent metadata, a readiness to comment on the quality of the other party's analysis, and a campaign of education for the public on matters of quality and analysis. The cultural issue can be addressed by reference to good scientific practices – good quality statistical methods have nothing to fear from the reuse of the underlying data in an Open Data world.

Dissemination standards

390. The standards for dissemination of Open microData are demanding. Few NSOs will achieve the expected standards in the short term. The UK's public sector is challenged by the 5 star Open Data expectation.[58] In particular, the use of non-proprietary formats and URIs for all Open microData releases is a challenge, and may not even be welcomed by current users.

Using websites designed for other purposes

391. Compared to aggregate statistics, Open microData may be very large in volume and may have a format that is incompatible with existing dissemination channels such as NSO websites. Reengineering NSOs websites for Open Data may be a low priority, especially if there are hard limitations such as bandwidth, file size, and proprietary format restrictions. A solution is to adopt alternative dissemination channels designed specifically for this challenge. For example, the United Kingdom has established www.data.gov.uk as a shared service available to all public sector institutions. It should also be born in mind that, as Open Data, restrictions on government information security standards are not applicable, allowing private sector dissemination channels to be used. Google Public Data is just one example, currently hosting Open Data from OECD, Eurostat, the IMF, the World Bank, DeSTATIS, and the US Census Bureau. Although Google Public Data is primarily an aggregator of data released through other channels, it could be used as a channel of first release of Open Data.

Authenticity and attribution

392. NSOs are the attributed original source of Open microData, but legitimately may have concerns about modifications to the content of the data that affect the products of analysis but which are unrecorded and/or unexplained by users. In other words, the authenticity of the data is lost, but the attribution of 'Source: NSO' persists. The solution is the use of 'persistent identifiers' by the producer of the Open Data. Every Open Data asset should be associated with a unique Direct Object Identifier/ Uniform Resource Name, in combination with a permanent Universal Resource Locator and a unique citation for electronic publications using the ISO 690-2 standard. This will allow the NSO, and users, to readily identify and retrieve the original and uncorrupted source data.

[58] http://5stardata.info

Metadata

393. The metadata standard most relevant for national statistics is SDMX, but this is optimised for aggregate statistical data rather than Open microData. If the NSO does not have expertise in metadata standards for Open microData it might seek assistance to pass this obstacle from partners such as data archives, in particular the members of the Consortium of European Social Science Data Archives (CESSDA).

Compliance with Code of Practice undertakings for scheduled releases and equality of access, versus obligations in Open Data standards for timelines.

394. NSOs will want to ensure that statistics scheduled for future release, according to the good practice for predictable and pre-announced release dates, are not undermined by similar statistics derived from re-use of Open microData. Where Open microData are pushed, this is easily achieved by proper scheduling of the two releases. However, where Open microData are pulled, there may be conflict with release schedules. A solution might be to bring forward the scheduled official release where possible, or to proactively explain the potential availability of unauthentic and unauthorised statistics from other sources in advance of the official publication.

395. Where Open microData are pulled from NSOs it is important to preserve the principle of equality of access. This is an important statistical principle, but it is an important Open Data principle too. If Open Data or Open microData are provided as (for example) attachments in an email to an applicant, this will give the appearance of privileged access. The use of the data may appear as a 'scoop' for the user. NSOs should establish a disclosure log within their website, used to present ad-hoc releases of Open Data that may have been pulled by a particular user but is clearly simultaneously available to all.[59] The user who pulled the data should be provided with a link to the URL only when the data resource is available to everyone.

Allocation of resources

396. The resource use profile of Open microData is different to research use files. Initial resource costs are high, despite the raw information already existing, by definition. The analysis of disclosure risk, the preparation of metadata, the assembly into an open format, etc, are all up-front costs. By definition, the income from Open Data cannot exceed costs plus a reasonable return on investment into their production. However, once produced, the ongoing costs of Open microData are negligible, especially if the dissemination channel is a shared resource or a cost to another party. In contrast, the resource use profile for research datasets is low initially, as the data are in effect unchanged from their status in the production environment. The resource costs for research datasets are high in the maintenance of a secure research environment, the accreditation of researchers and research projects, and the checking of outputs. NSI budgets are, typically, oriented towards meeting ongoing costs and easier to allocate, even if in aggregate over a number of years the burden on resources of research datasets is greater than the investment in Open microData of equivalent value to society.

Response to the challenges

397. This chapter discusses the drivers and challenges to the exchange of microdata under the emerging standards of Open Data.

[59] http://www.ons.gov.uk/ons/about-ons/what-we-do/FOI/foi-requests/index.html

398. The benefits of achieving the Open Data standard for official microdata are clearly very substantial. Microdata as Open Data allows the scientific principles of corroboration, validation, and improvement of the official statistics derived from the same sources. Also, Open microData

- can be exchanged without costly and bureaucratic administrative obstacles;

- can be used for any purpose, including those never envisaged when they were produced;

- allow NSOs and third parties from all sectors to cooperate and collaborate fully on a shared information resource;

- are an additional information asset category for a NSO, making the NSO an important and (hopefully) valued partner in the public sector for the future of modern Information Societies; and

- encourages the development of information entrepreneurs, fostering economic and social growth.

399. The obstacles and challenges are equally substantial:

- Confidentiality risks, and concepts, have to be addressed.

- Logistical issues arise, presenting challenge to the business architecture of NSOs.

- Authenticity and identification of assets must be addressed.

- The expected standards for Open Data for official microdata are high.

Recommendations

400. It is proposed that NSOs:

1. Adopt Open Data standards for their routine statistical production. This ensures the NSO becomes familiar with Open Data challenges before the particular challenges of Open microData are tackled.
2. Include a category of Open microData in their information asset registers.
3. Collaborate to spread the costs of developing methodologies for creating Open microData.
4. Explore alternative dissemination channels for Open microData if existing architecture is unsuitable.
5. Work closely with their national data protection supervisor on anonymisation standards and the concept of personal data.
6. Consider in advance how obligations under Codes of Practice for statistics can be upheld when Open microData are produced.
7. Use the skills and experience of data curators, computer scientists, knowledge and information management professionals, data archives, and national libraries, to assist with the preparation and dissemination of Open microData.
8. Prepare a communication strategy with key statistical users, and the public, to address novel Open Data products.

PART IV. MAKING MICRODATA ACCESS OUR BUSINESS

CHAPTER 12. MICRODATA ACCESS AND THE GSBPM

by Heather Dryburgh, Tuulikki Sillajõe and Tomaz Smrekar

The challenges: Information management for international access to microdata[60]

401. There are many legal and technical challenges for National Statistical Institutes (NSI) seeking to set up a way to provide access to microdata to international researchers. Once the legal questions have been dealt with, the remaining challenges can be addressed through well-defined and executed procedures and processes, a management information system, data repository and technical controls of access permissions and data. These process flow systems ensure that access is controlled to reflect approvals and that only non-confidential outputs are released from the NSI.

402. This Chapter will focus on the process flows and information management requirements for microdata access. It will then recommend some changes to the Generic Statistical Business Process Model (GSBPM) that will better capture the processes required in order for NSOs to provide access to microdata by researchers (Figure 12.1) [61]. The recommendations will include descriptions of the planning, development, implementation and archiving of microdata access and its outputs as part of the survey life cycle.

403. Finally, three general recommendations are made regarding microdata access provision by NSOs:

a. NSOs should use the revised GSBPM descriptions to guide them in ensuring access to their microdata is planned from the beginning stages of the survey life cycle, and all costing is included in the planning stages.

b. NSOs should allow researchers access to their microdata by one or more of the access types as a first step.

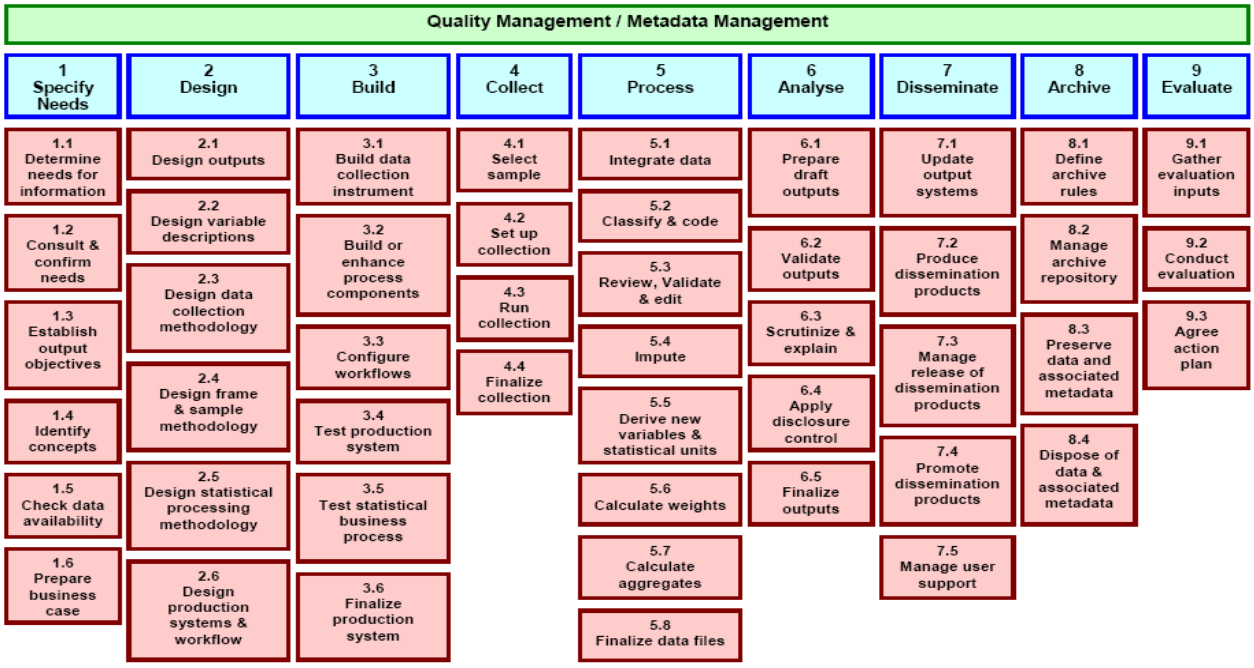
c. NSOs should work toward an embedded access model where multiple access modes are available to meet the needs of different users, and where international access is made possible through at least one of the available access types.

---

60 Work for this Chapter was done on the basis of Version 4.0 of the GSBPM. After the completion of the Chapter, a new Version 5.0 has become available. While Version 5.0 does not reflect all the changes proposed in this Chapter, it integrates references to microdata in most of the phases of the GSBPM model as suggested in this Chapter.
http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0

61 "The GSBPM should therefore be seen as a flexible tool to describe and define the set of business processes needed to produce official statistics. The GSBPM is intended to apply to all activities undertaken by producers of official statistics, at both the national and international levels, which result in data outputs. It is designed to be independent of the data source, so it can be used for the description and quality assessment of processes based on surveys, censuses, administrative records, and other non-statistical or mixed sources"; see page 2, paragraphs 3 and 4,
http://www1.unece.org/stat/platform/download/attachments/8683538/GSBPM+Final.pdf?version=1&modificationDate=1241066597110

Figure 12.1. Generic Statistical Business Process Model, Version 4.0



Types of microdata access

404. Once an NSO has determined what is legally possible in the provision of access to confidential microdata – having established a legal framework for access to microdata - there are four main types of microdata access that could be adopted.. The choice depends on the requirements of the researchers and the legal requirements of the NSO.

405. The current types of access that are most frequently found in NSOs of OECD countries include secure centre access, remote access, remote execution, and direct access to different levels of anonymized microdata, such as Scientific Use Files and Public Use Files:

- Anonymized data include public use files and scientific use files. Researchers can have access to and see the microdata, which are less detailed than the full confidential file, available anywhere and at no or low cost to the researcher.
- Confidential data are accessible through the other three remaining access modes. In the remote access option and in secure centres, researchers see the confidential data, however all outputs to be removed from the NSO's servers are vetted for confidentiality before being released to the researcher. The cost of this kind of access is higher as the vetting is manually conducted for each researcher.
- Confidential data are also accessible through remote execution. In this case, researchers can submit program code remotely from any computer (in an automated remote execution process), or to NSO staff to run the program for them, but researchers do not see the microdata. The cost of an automated program is moderate after it is developed. When running the program and vetting the output is done through a manual process it can be costly.

The Maturity Model of Access Types

406. NSOs should consider their current status of providing access to their microdata, and their plan to reach the ultimate goal of a fully embedded Microdata Access program. At the early Initiating stage an NSO will be providing access to microdata by sending approved researchers Scientific Use Files. This activity, while potentially desirable to some researchers, is only viable in this early stage when demand is small and NSOs can accept the risk of a few files that are no longer held on the secure NSO servers. In the Heroic stage, the NSO develops the legal framework for expanded access to microdata for research purposes and commits to at least one of the other access types described above. A fully Embedded microdata access model is identified by multiple modes of access, including options that will allow for cross-border access to the NSO's microdata holdings.

Process flow

407. Thinking of the process flow in a simplified way as input, process (access and research) and output, the following table shows the important differences between the processes for each of these four types of access.

Table 12.1. Typical process flow by Access Type (although there are differences in different NSIs)

	Remote Execution	Remote access	Secure Centre	Direct access to Anonymized Datafiles
Input	Slightly masked files stored in NSO, controlled by password, metadata on official website	Slightly masked anonymized files, stored in secure servers, controlled by file permission, metadata on secure server	Detailed anonymized files, stored in secure research centres, controlled by file permission, metadata on secure server	PUFs are greatly masked files, stored in NSI and distributed externally, end-use license, controlled by password, searchable metadata coded to DDI standard. SUFs have less masking and require legal agreement.
Process for access	Real time, anywhere, Internet submission, email notification, automatic confidentiality vetting, or manual submission that is run and output vetted by NSO staff	Real time, anywhere, Internet access, email notification, manual or automatic confidentiality vetting	Restricted access in secure centres, peer and institutional review process, paper approval documents and signed contract, manual confidentiality vetting	Public access to microdata in education libraries, as a data repository for other organizations and available individually free of charge. SUFs provided under legal agreement only for single researcher use.
Process for research	Pre-scan, scanned code run on microdata, confidentiality rules applied, post-scan, returned to	Access permissions automated (LDAP), researcher in secure centre outside NSO, restricted access to approved files, output submitted to NSO	Access permissions automated (LDAP), researcher in secure centre, restricted access to approved files, output submitted	End-use license (no malicious use), Research on microdata file as per researcher needs.

	researcher (automated or manual)	employee for confidentiality vetting	to NSO employee for confidentiality vetting	
Output	Tables, means, medians, percentiles and rations are available by automated process, any statistical output if run manually vetting automated using controlled rounding (for example), or vetted manually	Tables, means, medians, percentiles and ratios are available, vetted manually	Any statistical output from detailed microdata, vetted manually	Any statistical output from masked microdata
Benefits to Researchers	Customized output without having to travel to NSO	Can see microdata without having to travel to NSI, so data quality assessment possible. Customized output, but have to meet NSO's security requirements	Full access to the microdata and all quality assessment and customization possible, but have to travel to safe centre	Full access to the microdata on researcher's computer, but microdata masked so detailed work limited
Benefits to NSI	Automation possible, which allows more control and efficiency. Data remain at NSO for greatest security. Researchers do not see microdata.	Data remain at NSO for greatest security, however, researchers do see microdata, but infrastructure and staffing costs can be high.	Cost effective as researchers work autonomously, but staffing costs can be high.	Reduction in the number of researchers requiring more costly access options.

Information management for international access

408. In the process flow there are a number of different kinds of information, including the microdata files, the metadata, the project administration documents (approvals, contracts, oaths, security clearances and other paperwork), the code or syntax, working files, temporary files, reports and outputs. All of this information must be managed carefully for all access, but particularly for access by international researchers. The following list summarizes the information at each stage of the process flow:

a. *Input:* Data holdings (Anonymization, storage, metadata, security, permissions)
b. *Process for access:* Approvals, project process flow, permissions and controls
c. *Process for research:* Data processing, temporary files, metadata use
d. *Output:* Confidentiality vetting, some indicator of data quality, residual risk, archiving

409. Process flows for secure and successful access programs vary according to the aims of the access program and the requirements of researchers. Many concerns around international access can be resolved with careful planning and implementation of procedures and process flow design. For this reason, the Expert Group is recommending the explicit inclusion into the GSBPM of elements of the process flow for microdata access.

Recommended changes to the GSBPM

410. At the second meeting of the OECD CSTAT Expert Group for International Collaboration on Microdata Access in December 2012, a sub-group was established to address the concern that the GSBPM did not adequately capture microdata access activities of NSOs. The sub-group was tasked with fitting

access to microdata into the vision of the High-Level Group for the Modernisation of Statistical Production and Services (HLG) [originally called High-Level Group for Strategic Developments in Business Architecture in Statistics (HLG – BAS)].

411. Since then, the sub-group consulted with the UNECE (Mr Steven Vale and Ms Thérèse Lalor), the Chair of the METIS Steering Group (Ms Alice Born, Canada), and with the members of the OECD Expert Group to determine what recommendations would come from the Expert Group and where, in addition to CSTAT, that recommendation should be sent for consideration. Finally, the team also reviewed the Generic Longitudinal Business Process Model (GLBPM), the Data Documentation Initiative (DDI) as well as other contextual documents on the GSBPM.

412. Based on the work conducted, the OECD CSTAT Expert Group for International Collaboration on Microdata Access recommends the following changes to ensure that the process flow for microdata access is included in the GSBPM. Recommended changes are all to the descriptions (see changes in bold red), not to the high level elements of the model. Note that the Expert Group identified as very important to the development of a mature Microdata Access program, the establishment of a legal framework for such access. As this is an overarching process – i.e.; it requires doing once before statistical programs can plan, develop and implement an access strategy for their specific program – it is not included in the GSBPM, but should be a key part of the early development under the broad rubric of "Legal Frameworks."

Recommendations[62]**:**

1 SPECIFY NEEDS

This phase is triggered when a need for new statistics is identified, or feedback about current statistics initiates a review. It determines whether there is a presently unmet demand, externally and / or internally, for the identified statistics and whether the statistical organization can produce them.

In this phase the organization:

- determines the need for the statistics;
- confirms, in more detail, the statistical needs of the stakeholders;
- establishes the high level objectives of the statistical outputs;
- identifies the relevant concepts and variables for which data are required;
- checks if current collections and / or methodologies can meet these needs, and
- prepares the business case to get approval to produce the statistics.

This phase is broken down into five sub-processes. These are generally sequential, from left to right, but can also occur in parallel, and can be iterative. The sub-processes are:

1.1 Determine needs for information

This sub-process includes the initial investigation and identification of what ***microdata,*** statistics ***and metadata*** are needed and what is needed of the ***microdata,*** statistics ***and metadata***. It also includes consideration of practice amongst other (national and international) statistical organizations producing similar data, and in particular the methods used by those organizations.

62 The following reproduces the text of Version 4.0 of the GSBPM. Suggested changes appear in Italics.

1.2 Consult and confirm needs

This sub-process focuses on consulting with the stakeholders and confirming in detail the need for the ***microdata,*** statistics ***and metadata***. A good understanding of user needs is required so that the statistical organization knows not only what it is expected to deliver, but also when, how, and, perhaps most importantly, why. For second and subsequent iterations of this phase, the main focus will be on determining whether previously identified needs have changed. This detailed understanding of user needs is the critical part of this sub-process.

1.3 Establish output objectives

This sub-process identifies the statistical outputs that are required to meet the user needs identified in sub-process 1.2 (Consult and confirm needs). It includes agreeing the suitability of the proposed outputs and their quality measures with users.

1.4 Identify concepts

This sub-process clarifies the required concepts to be measured by the business process from the point of view of the user. At this stage the concepts identified may not align with existing statistical standards. This alignment, and the choice or definition of the statistical concepts and variables to be used, takes place in sub-process 2.2.

1.5 Check data availability

This sub-process checks whether current data sources could meet user requirements, and the conditions under which they would be available, including any restrictions on their use. An assessment of possible alternatives would normally include research into potential administrative data sources and their methodologies, to determine whether they would be suitable for use for statistical purposes. When existing sources have been assessed, a strategy for filling any remaining gaps in the data requirement is prepared. This sub-process also includes a more general assessment of the legal framework in which data would be collected and used, and may therefore identify proposals for changes to existing legislation or the introduction of a new legal framework.

1.6 Prepare business case

This sub-process documents the findings of the other subprocesses in this phase in the form a business case to get approval to implement the new or modified statistical business process. Such a business case would typically also include:

- A description of the "As-Is" business process (if it already exists), with information on how the current ***microdata,*** statistics ***and metadata*** are produced, highlighting any inefficiencies and issues to be addressed;
- The proposed "To-Be" solution, detailing how the statistical business process will be developed to produce the new or revised ***microdata,*** statistics ***and metadata***;
- An assessment of costs and benefits, as well as any external constraints.

2 DESIGN

This phase describes the development and design activities, and any associated practical research work needed to define the statistical outputs, concepts, methodologies, collection instruments and operational

processes. For statistical outputs produced on a regular basis, this phase usually occurs for the first iteration, and whenever improvement actions are identified in phase 9 (Evaluate) of a previous iteration.

This phase is broken down into six sub-processes, which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

2.1 Design outputs

This sub-process contains the detailed design of the statistical outputs to be produced, including the related development work and preparation of the systems and tools used in phase 7 (Disseminate). Outputs should be designed, wherever possible, to follow existing standards, so inputs to this process may include metadata from similar or previous collections, international standards, and information about practices in other statistical organizations from subprocess 1.1 (Determine needs for information). ***Establishing confidentiality vetting rules for outputs and the procedures for access to microdata are key parts of this sub-process.***

2.2 Design variable descriptions

This sub-process defines the statistical variables to be collected via the data collection instrument, as well as any other variables that will be derived from them in sub-process 5.5 (Derive new variables and statistical units), and any classifications that will be used. It is expected that existing national and international standards will be followed wherever possible. This sub-process may need to run in parallel with sub-process 2.3 (Design data collection methodology), as the definition of the variables to be collected, and the choice of data collection instrument may be inter-dependent to some degree. Preparation of metadata descriptions of collected and derived variables and classifications is a necessary precondition for subsequent phases.

2.3 Design data collection methodology

This sub-process determines the most appropriate data collection method(s) and instrument(s). The actual activities in this subprocess will vary according to the type of collection instruments required, which can include computer assisted interviewing, paper questionnaires, administrative data interfaces and data integration techniques. This sub-process includes the design of questions and response templates (in conjunction with the variables and classifications designed in subprocess 2.2 (Design variable descriptions)). It also includes the design of any formal agreements relating to data supply, such as memoranda of understanding, and confirmation of the legal basis for the data collection. This sub-process is enabled by tools such as question libraries (to facilitate the reuse of questions and related attributes), questionnaire tools (to enable the quick and easy compilation of questions into formats suitable for cognitive testing) and agreement templates (to help standardize terms and conditions). This sub-process also includes the design of process-specific provider management systems.

2.4 Design frame and sample methodology

This sub-process identifies and specifies the population of interest, defines a sampling frame (and, where necessary, the register from which it is derived), and determines the most appropriate sampling criteria and methodology (which could include complete enumeration). Common sources are administrative and statistical registers, censuses and sample surveys. This sub-process describes how these sources can be combined if needed. Analysis of whether the frame covers the target population should be performed. A sampling plan should be made: The actual sample is created sub-process 4.1 (Select sample), using the methodology, specified in this sub-process.

2.5 Design statistical processing methodology

This sub-process designs the statistical processing methodology to be applied during phase 5 (Process), and Phase 6 (Analyse). This can include specification of routines for coding, editing, imputing, estimating, integrating, validating and finalizing data sets.

2.6 Design production systems and workflow

This sub-process determines the workflow from data collection to archiving, taking an overview of all the processes required within the whole statistical production process, and ensuring that they fit together efficiently with no gaps or redundancies. Various systems and databases are needed throughout the process. A general principle is to reuse processes and technology across many statistical business processes, so existing systems and databases should be examined first, to determine whether they are fit for purpose for this specific process, then, if any gaps are identified, new solutions should be designed. This sub-process also considers how staff will interact with systems, and who will be responsible for what and when.

3 BUILD

This phase builds and tests the production systems to the point where they are ready for use in the "live" environment. For statistical outputs produced on a regular basis, this phase usually occurs for the first iteration, and following a review or a change in methodology, rather than for every iteration. It is broken down into six sub-processes, which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

3.1 Build data collection instrument

This sub-process describes the activities to build the collection instruments to be used during the phase 4 (Collect). The collection instrument is generated or built based on the design specifications created during phase 2 (Design). A collection may use one or more modes to receive the data, e.g. personal or telephone interviews; paper, electronic or web questionnaires; SDMX hubs. Collection instruments may also be data extraction routines used to gather data from existing statistical or administrative data sets. This sub-process includes preparing and testing the contents and functioning of that instrument (e.g. testing the questions in a questionnaire). It is recommended to consider the direct connection of collection instruments to the statistical metadata system, so that metadata can be more easily captured in the collection phase. Connection of metadata and data at the point of capture can save work in later phases. Capturing the metrics of data collection (paradata) is also an important consideration in this sub-process.

3.2 Build or enhance process components

This sub-process describes the activities to build and test new and enhance existing software components needed for the business process, as designed in Phase 2 (Design). Components may include dashboard functions and features, ***micro***data***, statistics and metadata*** repositories, transformation tools, workflow framework components, ***microdata access systems, provider and*** metadata management tools.

3.3 Configure workflows

This sub-process configures the workflow, systems and transformations used within the statistical business processes, from data collection, right through to archiving the final statistical outputs. It ensures that the workflow specified in sub-process 2.6 (Design production systems and workflow) works in practice.

3.4 Test production systems

This sub-process is concerned with the testing of computer systems and tools. It includes technical testing and sign-off of new programs and routines, as well as confirmation that existing routines from other statistical business processes are suitable for use in this case. Whilst part of this activity concerning the testing of individual components could logically be linked with sub-process 3.2 (Build or enhance process components), this sub-process also includes testing of interactions between components, and ensuring that the production system works as a coherent set of components.

3.5 Test statistical business process

This sub-process describes the activities to manage a field test or pilot of the statistical business process. Typically it includes a small scale data collection, to test collection instruments, followed by processing and analysis of the collected data, to ensure the statistical business process performs as expected. Following the pilot, it may be necessary to go back to a previous step and make adjustments to instruments, systems or components. For a major statistical business process, e.g. a population census, there may be several iterations until the process is working satisfactorily.

3.6 Finalize production systems

This sub-process includes the activities to put the process, including workflow systems, modified and newly-built components into production ready for use by business areas. The activities include:

- producing documentation about the process components, including technical documentation and user manuals
- training the business users on how to operate the process
- moving the process components into the production environment, and ensuring they work as expected in that environment (this activity may also be part of sub-process 3.4 (Test production systems)).

4 COLLECT

This phase collects all necessary data, using different collection modes (including extractions from administrative and statistical registers and databases), and loads them into the appropriate data environment. It does not include any transformations of collected data, as these are all done in phase 5 (Process). For statistical outputs produced regularly, this phase occurs in each iteration.

The Collect phase is broken down into four sub-processes, which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. These subprocesses are:

4.1 Select sample

This sub-process establishes the frame and selects the sample for this iteration of the collection, as specified in sub-process 2.4 (Design frame and sample methodology). It also includes the coordination of samples between instances of the same statistical business process (for example to manage overlap or rotation), and between different processes using a common frame or register (for example to manage overlap or to spread response burden). Quality assurance, approval and maintenance of the frame and the selected sample are also undertaken in this sub-process, though maintenance of underlying registers, from which frames for several statistical business processes are drawn, is treated as a separate business process. The sampling aspect of this sub-process is not usually relevant for processes based entirely on the use of

pre-existing data sources (e.g. administrative data) as such processes generally create frames from the available data and then follow a census approach.

4.2 Set up collection

This sub-process ensures that the people, processes and technology are ready to collect data, in all modes as designed. It takes place over a period of time, as it includes the strategy, planning and training activities in preparation for the specific instance of the statistical business process. Where the process is repeated regularly, some (or all) of these activities may not be explicitly required for each iteration. For one-off and new processes, these activities can be lengthy.

This sub-process includes:

- preparing a collection strategy
- training collection staff
- ensuring collection resources are available e.g. laptops
- configuring collection systems to request and receive the data;
- ensuring the security of data to be collected;
- preparing collection instruments (e.g. printing questionnaires, pre-filling them with existing data, loading questionnaires and data onto interviewers' computers etc.).

4.3 Run collection

This sub-process is where the collection is implemented, with the different collection instruments being used to collect the data. It includes the initial contact with providers and any subsequent follow-up or reminder actions. It records when and how providers were contacted, and whether they have responded. This sub-process also includes the management of the providers involved in the current collection, ensuring that the relationship between the statistical organization and data providers remains positive, and recording and responding to comments, queries and complaints. For administrative data, this process is brief: the provider is either contacted to send the data, or sends it as scheduled. When the collection meets its targets (usually based on response rates) the collection is closed and a report on the collection is produced.

4.4 Finalize collection

This sub-process includes loading the collected data and metadata into a suitable electronic environment for further processing in phase 5 (Process). It may include automatic data take-on, for example using optical character recognition tools to extract data from paper questionnaires, or converting the formats of data files received from other organizations. In cases where there is a physical data collection instrument, such as a paper questionnaire, which is not needed for further processing, this sub-process manages the archiving of that material in conformance with the principles established in phase 8 (Archive).

5 PROCESS

This phase describes the cleaning of data records and their preparation for analysis. It is made up of sub-processes that check, clean, and transform the collected data, and may be repeated several times. For statistical outputs produced regularly, this phase occurs in each iteration. The sub-processes in this phase can apply to data from both statistical and non-statistical sources (with the possible exception of sub-process 5.6 (Calculate weights), which is usually specific to survey data).

The "Process" and "Analyse" phases can be iterative and parallel. Analysis can reveal a broader understanding of the data, which might make it apparent that additional processing is needed. Activities within the "Process" and "Analyse" phases may commence before the "Collect" phase is completed. This enables the compilation of provisional results where timeliness is an important concern for users, and increases the time available for analysis. The key difference between these phases is that "Process" concerns transformations of microdata, whereas "Analyse" concerns the further treatment of statistical aggregates.

This phase is broken down into eight sub-processes, which may be sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

5.1 Integrate data

This sub-process integrates data from one or more sources. The input data can be from a mixture of external or internal data sources, and a variety of collection modes, including extracts of administrative data. The result is a harmonized data set. Data integration typically includes:

- matching / record linkage routines, with the aim of linking data from different sources, where those data refer to the same unit;
- prioritizing, when two or more sources contain data for the same variable (with potentially different values).

Data integration may take place at any point in this phase, before or after any of the other sub-processes. There may also be several instances of data integration in any statistical business process. Following integration, depending on data protection requirements, data may be anonymized, that is stripped of identifiers such as name and address, to help to protect confidentiality.

5.2 Classify & code

This sub-process classifies and codes the input data. For example automatic (or clerical) coding routines may assign numeric codes to text responses according to a pre-determined classification scheme.

5.3 Review, validate and edit

This sub-process applies to collected micro-data, and looks at each record to try to identify (and where necessary correct) potential problems, errors and discrepancies such as outliers, item non-response and miscoding. It can also be referred to as input data validation. It may be run iteratively, validating data against predefined edit rules, usually in a set order. It may apply automatic edits, or raise alerts for manual inspection and correction of the data. Reviewing, validating and editing can apply to unit records both from surveys and administrative sources, before and after integration. In certain cases, imputation (sub-process 5.3) may be used as a form of editing.

5.4 Impute

Where data are missing or unreliable, estimates may be imputed, often using a rule-based approach. Specific steps typically include:

- the identification of potential errors and gaps;
- the selection of data to include or exclude from imputation routines;
- imputation using one or more pre-defined methods e.g. "hot-deck" or "cold-deck";
- writing the imputed data back to the data set, and flagging them as imputed;

- the production of metadata on the imputation process.

5.5 Derive new variables and statistical units

This sub-process derives (values for) variables and statistical units that are not explicitly provided in the collection, but are needed to deliver the required outputs. It derives new variables by applying arithmetic formulae to one or more of the variables that are already present in the dataset. This may need to be iterative, as some derived variables may themselves be based on other derived variables. It is therefore important to ensure that variables are derived in the correct order. New statistical units may be derived by aggregating or splitting data for collection units, or by various other estimation methods. Examples include deriving households where the collection units are persons, or enterprises where the collection units are legal units.

5.6 Calculate weights

This sub process creates and applies weights for unit data records according to the methodology created in sub-process 2.5 (Design statistical processing methodology). These weights can be used to "gross-up" sample survey results to make them representative of the target population, or to adjust for non-response in total enumerations.

5.7 Calculate aggregates

This sub process creates aggregate data and population totals from micro-data. It includes summing data for records sharing certain characteristics, determining measures of average and dispersion, and applying weights from sub-process 5.6 to sample survey data to derive population totals.

5.8 Finalize data files

This sub-process brings together the results of the other subprocesses in this phase and results in a data file (usually of macro-data), which is used as the input to phase 6 (Analyse). Sometimes this may be an intermediate rather than a final file, particularly for business processes where there are strong time pressures, and a requirement to produce both preliminary and final estimates.

6 ANALYSE

In this phase, statistics are produced, examined in detail and made ready for dissemination. This phase includes the sub-processes and activities that enable statistical analysts to understand the statistics produced. For statistical outputs produced regularly, this phase occurs in every iteration. The Analyse phase and sub-processes are generic for all statistical outputs, regardless of how the data were sourced.

The Analyse phase is broken down into five sub-processes, which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. The subprocesses are:

6.1 Prepare draft outputs

This sub-process is where the data collected are transformed into statistical outputs. It includes the production of additional measurements such as indices, trends or seasonally adjusted series, as well as the recording of quality characteristics.

6.2 Validate outputs

This sub-process is where statisticians validate the quality of the outputs produced, in accordance with a general quality framework and with expectations. This sub-process also includes activities involved with the gathering of intelligence, with the cumulative effect of building up a body of knowledge about a specific statistical domain. This knowledge is then applied to the current collection, in the current environment, to identify any divergence from expectations and to allow informed analyses. Validation activities can include:

- checking that the population coverage and response rates are as required;
- comparing the statistics with previous cycles (if applicable);
- confronting the statistics against other relevant data (both internal and external);
- investigating inconsistencies in the statistics;
- performing macro editing;
- validating the statistics against expectations and domain intelligence.

6.3 Scrutinize and explain

This sub-process is where the in-depth understanding of the outputs is gained by statisticians. They use that understanding to scrutinize and explain the statistics produced for this cycle by assessing how well the statistics reflect their initial expectations, viewing the statistics from all perspectives using different tools and media, and carrying out in-depth statistical analyses.

6.4 Apply disclosure control

This sub-process ensures that the data (and metadata) to be disseminated ***or outputs released to external researchers accessing microdata*** do not breach the appropriate rules on confidentiality. This may include checks for primary and secondary disclosure, as well as the application of data suppression or perturbation techniques.

6.5 Finalize outputs

This sub-process ensures the statistics and associated information are fit for purpose and reach the required quality level, and are thus ready for use. It includes:

- completing consistency checks;
- determining the level of release, and applying caveats;
- collating supporting information, including interpretation, briefings, measures of uncertainty and any other necessary metadata;
- producing the supporting internal documents;
- pre-release discussion with appropriate internal subject matter experts;
- approving the statistical content for release.

7 DISSEMINATE

This phase manages the release of the statistical products to customers. For statistical outputs produced regularly, this phase occurs in each iteration. It is made up of five sub-processes, which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

7.1 Update output systems[63]

This sub-process manages the update of systems where data and metadata are stored for dissemination purposes, including:

- formatting data and metadata ready to be put into output databases;
- loading data and metadata into output databases;
- ensuring data are linked to the relevant metadata.

7.2 Produce dissemination products

This sub-process produces the products, as previously designed (in sub-process 2.1), to meet user needs. The products can take many forms including printed publications, press releases and web sites. Typical steps include:

- preparing the product components (explanatory text, tables, charts etc.);
- assembling the components into products;
- editing the products and checking that they meet publication standards.

7.3 Manage release of dissemination products

This sub-process ensures that all elements for the release are in place including managing the timing of the release. It includes briefings for specific groups such as the press or ministers, as well as the arrangements for any pre-release embargoes. It also includes the provision of products to subscribers ***as well as opening up and closing access (permissions) on the start and completion of projects requiring access to microdata by external researchers, and releasing researcher's output that has been vetted for confidentiality. Publication of the results of these microdata research projects, or providing reference to them is also a part of this sub-process.***

7.4 Promote dissemination products

Whilst marketing in general can be considered to be an over-arching process, this sub-process concerns the active promotion of the statistical products produced in a specific statistical business process, to help them reach the widest possible audience. It includes the use of customer relationship management tools, to better target potential users of the products, as well as the use of tools including web sites, wikis and blogs to facilitate the process of communicating statistical information to users.

7.5 Manage user support

This sub-process ensures that customer queries are recorded, and that responses are provided within agreed deadlines. ***Requests for microdata access require putting in place a contract and a confidentiality statement with the researcher.*** These queries ***and microdata research requests*** should be regularly reviewed to provide an input to the over-arching quality management process, as they can indicate new or changing user needs.

8 ARCHIVE

This phase manages the archiving and disposal of statistical data and metadata. Given the reduced costs of data storage, it is possible that the archiving strategy adopted by a statistical organization does not include provision for disposal, so the final sub-process may not be relevant for all statistical business processes. In

[63] Formatting, loading and linking of metadata should preferably mostly take place in earlier phases, but this sub-process includes a check that all of the necessary metadata are in place ready for dissemination.

other cases, disposal may be limited to intermediate files from previous iterations, rather than disseminated data.

For statistical outputs produced regularly, archiving occurs in each iteration, however defining the archiving rules is likely to occur less regularly. This phase is made up of four sub-processes, which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

8.1 Define archive rules[64]

This sub-process is where the archiving rules for the statistical data and metadata resulting from a statistical business process are determined. ***In the case of giving access to microdata for scientific purposes, archiving of the results of the research projects or reference to them is a part of this sub-process.*** The requirement to archive intermediate outputs such as the sample file, the raw data from the collect phase, and the results of the various stages of the process and analyse phases should also be considered. The archive rules for a specific statistical business process may be fully or partly dependent on the more general archiving policy of the statistical organization, or, for national organizations, on standards applied across the government sector. The rules should include consideration of the medium and location of the archive, as well as the requirement for keeping duplicate copies. They should also consider the conditions (if any) under which data and metadata should be disposed of.

8.2 Manage archive repository

This sub-process concerns the management of one or more archive repositories. These may be databases, or may be physical locations where copies of data or metadata are stored. It includes:

- maintaining catalogues of data and metadata archives, with sufficient information to ensure that individual data or metadata sets can be easily retrieved;
- testing retrieval processes;
- periodic checking of the integrity of archived data and metadata;
- upgrading software-specific archive formats when software changes.

This sub-process may cover a specific statistical business process or a group of processes, depending on the degree of standardization within the organization. Ultimately it may even be considered to be an over-arching process if organization-wide standards are put in place.

8.3 Preserve data and associated metadata

This sub-process is where the ***micro***data***, statistics*** and metadata from a specific statistical business process are archived. ***In the case of giving access to microdata for scientific purposes, archiving of the results of the research projects or reference to them is a part of this sub-process.*** It includes:

- identifying ***micro***data***, statistics,*** ~~***and***~~ metadata ***and research projects*** for archiving in line with the rules defined in 8.1;
- formatting those ***micro***data***, statistics,*** ~~***and***~~ *metadata* ***and research projects*** for the repository;
- loading or transferring ***micro***data***, statistics,*** ~~***and***~~ metadata ***and research projects*** to the repository;
- cataloguing the archived ***micro***data***, statistics,*** ~~***and***~~ metadata ***and research projects***;
- verifying that the ***micro***data***, statistics,*** ~~***and***~~ metadata ***and research projects*** have been successfully archived.

64 This sub-process is logically strongly linked to Phase 2 – Design, at least for the first iteration of a statistical business process.

- ***storing the microdata accessed by external researchers and their intermediate results in the form they were made available to external researchers for an agreed-upon period.***

8.4 Dispose of data and associated metadata

This sub-process is where the data and metadata from a specific statistical business process are disposed of. It includes;

- identifying data and metadata for disposal, in line with the rules defined in 8.1;
- disposal of those data and metadata;
- recording that those data and metadata have been disposed of.

9 EVALUATE

This phase manages the evaluation of a specific instance of a statistical business process, as opposed to the more general over-arching process of statistical quality management described in Section VI. It logically takes place at the end of the instance of the process, but relies on inputs gathered throughout the different phases. For statistical outputs produced regularly, evaluation should, at least in theory occur for each iteration, determining whether future iterations should take place, and if so, whether any improvements should be implemented. However, in some cases, particularly for regular and well established statistical business processes, evaluation may not be formally carried out for each iteration. In such cases, this phase can be seen as providing the decision as to whether the next iteration should start from phase 1 (Specify needs) or from some later phase (often phase 4 (Collect)).

This phase is made up of three sub-processes, which are generally sequential, from left to right, but which can overlap to some extent in practice. These sub-processes are:

9.1 Gather evaluation inputs

Evaluation material can be produced in any other phase or sub-process. It may take many forms, including feedback from users, process metadata, system metrics and staff suggestions. Reports of progress against an action plan agreed during a previous iteration may also form an input to evaluations of subsequent iterations. This sub-process gathers all of these inputs, and makes them available for the person or team producing the evaluation.

9.2 Conduct evaluation

This sub-process analyzes the evaluation inputs and synthesizes them into an evaluation report. The resulting report should note any quality issues specific to this iteration of the statistical business process, and should make recommendations for changes if appropriate. These recommendations can cover changes to any phase or sub-process for future iterations of the process, or can suggest that the process is not repeated.

9.3 Agree action plan

This sub-process brings together the necessary decision making power to form and agree an action plan based on the evaluation report. It should also include consideration of a mechanism for monitoring the impact of those actions, which may, in turn, provide an input to evaluations of future iterations of the process.

CHAPTER 13. PROCESS FLOW AND COSTS OF A DATA ACCESS SERVICE[65]

by Leo Engberts

1. Introduction

413. Many National Statistical Offices (NSOs) provide, under strict conditions, access to confidential data for scientific research and only to obtain statistical output; they offer a Data Access Service. Some of these Services allow access to researchers across the borders. In particular, Remote Access Centres limit travel and accommodation costs. Ideally, an International Research Centre, where access to microdata from multiple countries would be provided, would need a Remote Access Network; the development of such a network should benefit from the experience that national Services have with Remote Access Centres.

414. The organisation of an International Research Centre should encompass several dimensions: a transparent overview of the available microdata databases, IT facilities, appropriate legal rules and a cost model. The cost issue is particularly relevant: often researchers and research institutions consider that data detained by NSOs are for free, as they have already been gathered for other purposes and the cost of their collection has been borne. However, a Data Service Access can only function if a number of elements are reunited, *i.e.* a system for data protection, modern IT facilities and a smoothly organised administration; all this requires considerable time and money.

415. Indeed, the overall budget of most NSOs does not include the maintenance of a Data Access Service. In order to support a reflection on the benefits and challenges of Data Access Services, this Chapter provides information on the costs of establishing and managing a Data Access Service, based on the experience of Statistics Netherlands.

2. Necessary elements for a Data Access Service

416. A Data Access Service requires several elements. It is useful to list them, also because detailing the workflow for providing data access allows to show the breakdown of costs. The information provided in the following section refers to the Data Access Service of Statistics Netherlands. http://www.cbs.nl/NR/rdonlyres/50625EDE-3274-4D7C-B19B-5E5D0F239E2F/0/131112dienstencatalogusosra2014eng.pdf

Information on the website

A webpage presenting the Service should be available on the website of the NSO, with an explanation of the institutes/researchers that are eligible to work with the microdata and information on the accreditation process (how to get access to the micro data needed). http://www.cbs.nl/en-GB/menu/informatie/beleid/zelf-onderzoeken/default.htm

A price list of the Service

The researcher should know in advance approximately the costs for an On-Site or Remote Access project. Data are free, but running a Service and its facilities has a cost that should be appreciated by the researcher (see example in Box 13.1).

[65] A version of this Chapter completed with Annexes is available on OLIS as document STD/CSTAT/MICRO(2013)5.

Box 12.1. Example: Prices for microdata services at CBS, as of 1 January 2011

ON SITE
Description Price (in €)
Workstation[1] *per half day:* 50
Administrative costs *once only for each project:* 163
User configuration *once only for each user:* 185
Outputcheck[2] *per project per check:* 92
Use of SAS[3] *per project per half day:* 40
Use of Ox[3] *per project per half day:* 21
Use of Gauss[3] *per project per half day:* 26
Use of MLWin[3] *per project per half day:* 17

REMOTE ACCESS
Description Price (in €)
Installation of a workstation *once only for each workstation:* 490
Workstation subscription[1] *per workstation per month:* 525
Administrative costs *once only for each project:* 163
User configuration o*nce only for each user:* 185
Fingerprint enrolment location[4]/at CBS *once only per person:* 490 / 92
Smartcard *once only per person:* 11
User subscription *per user per month:* 57
Outputcheck[2] *per project per check:* 185
Use of SAS[5] *per project per month:* 556
Use of Ox[5] *per project per month:* 46
Use of Gauss[5] *per project per month:* 57
Use of MLWin[5] *per project per month:* 36
Workstation at CBS *per half day:* 50

1. Use of SPSS and Stata is included in the workstation subscription costs.

2. The first four output checks for each project are free of charge.

3. Use of these software packages is assigned per project for the full duration of each project.

4. For each visit to the user's location, € 490 will be charged, irrespective of the number of tasks performed (workstation installation, fingerprint enrolment of more than one user).

5. This fee is charged for a minimum of three months, or the duration of the project if this is less than three months. Every quarter, use of these packages can be purchased in one deal for all current projects. The tariff is twice the monthly tariff concerned per month in the given quarter, irrespective of the number of projects the software is used for.

6. This workstation can only be used incidentally by researchers if no workstation is available at their own institution. Remote access does not provide for long-term workstation rent.

REMOTE EXECUTION
Description Price (in €)
Running of a SPSS or Stata-script *per script:* 100
Running of a SAS-script *per script:* 140
Running of an Ox-script *per script:* 121
Running of a Gauss-script *per script:* 127

Preparation of DATASETS
In addition to the costs specified above for use of the facilities, costs will also be charged for preparation and documentation of datasets. Datasets listed in our catalogue have a standard price. For each custom-made dataset a price is set based on the amount of work involved to prepare and document it.
Datasets
Description Price (in €)
Dataset in catalogue *per dataset:* 928

A standardised and transparent catalogue of available micro data bases

Researchers need clear information on the available microdata databases. Standardised catalogues of the national Services are therefore necessary. These catalogues should obtain also documentation reports and meta-information of the microdata databases. Because of a simultaneous use of the corresponding databases of different countries, the design (and thus the documentation) ought to be standardised.
http://www.cbs.nl/en-GB/menu/informatie/beleid/catalogi/default.htm

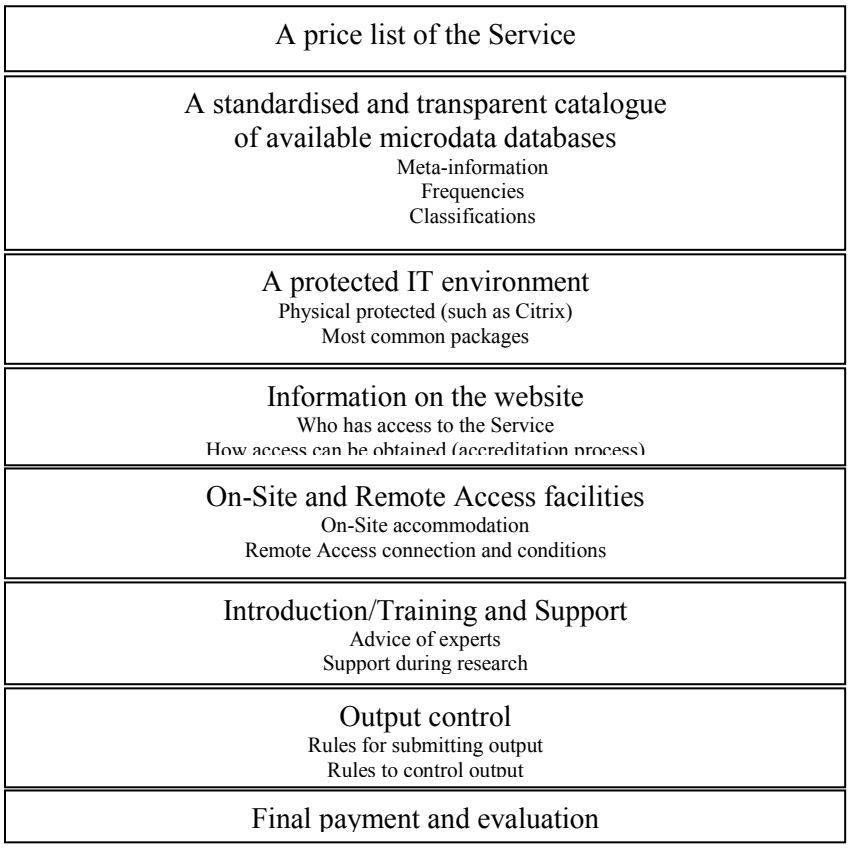
On-Site and Remote Access facilities

A Microdata Service could provide an accommodation for working On-Site. Except for some local facilities, supplementary actions beside the usual paperwork are not needed. A Remote Access facility needs physically requirements of the room and the hardware and software. A contract with the researcher/ research institution involved is recommended.

Output Control and Evaluation

Despite the contract which holds the researcher responsible for not revealing personal or enterprise information in the output produced, the Service will control the output before this is set free for the researcher. For an efficient control of the output, some rules must be set respected by the researcher in submitting his/her final output. International rules for the security of output are respected. Finally, an evaluation of every project is done for the benefit of both the researcher and the Service.

Figure 12.2. Facilities of a NSO-Service for On-Site and Remote Access



3. Type of Costs

417. The facilities and activities mentioned before can roughly be divided in three type of costs: i)°Facilities, ii) Personnel and iii) Information technologies.

Facilities

418. In the first place there is a need for computers as components of the Remote Access Network. If the RAN-computers are owned by the connected institutions there are no costs for an International Research Centre.

Personnel

419. This is typically the most expensive item. Staffs are involved in several facilities of the Service:

- *A standardized and transparent catalogue of available microdata databases:* a catalogue should also hold meta-information, frequencies and classifications of the microdata databases. The main task of this catalogue belongs to the National Services: an International Research Centre has no detailed information at its disposal. Nevertheless these costs cannot be ignored and will belong to the total costs of a successful IRC. The costs for the production and the maintenance of uniform database documentation for the development of a catalogue will be a huge effort.
- *A protected IT environment:* Establishing a protected environment is a demanding and costly task. Maintaining the environment, including software, is less time-consuming but is a continuing engagement.
- *On-Site and Remote Access facilities:* Mainly the Remote Access facility will cost a single effort to install the necessary software and the biometric device. Support for addressing problems and malfunction has to be calculated in the total cost.
- *Introduction/Training and Support:* Introductory interviews are easily accommodated by national Data Service Centres. These interviews are less of an option for an International Research Centre, as they would involve greater costs and problems (*e.g.* the costs for travel and accommodation, lack of experts on the different microdata databases). Support during the research will always be necessary and also expensive.
- *Output control:* This is a time-consuming task. Although there are standards for producing output and international standards for checking output, it still will involve much effort. Instead of a total control, check-ups can be made by sample; however, this is only possible if all members of the IRC agree.
- *Final payment and evaluation:* The very existence of an IRC depends on funding. An IRC engages time and resources. Finally, the evaluation of every project benefits both the researcher and the IRC; in fact a periodical summary of projects can be useful to improve the Service. Once a clear evaluation template is established, evaluation does not require substantive resources.

Information technologies

420. Two elements are relevant:

- *Information on the website*: the design and maintenance of a site involves considerable costs for IT-personnel time.

- *A protected IT environment:* Building a protected environment requires a great effort. The maintenance, including running software, is less time-consuming but involves a continuous effort. Finally, the costs of software licenses must not be underestimated.

421. The following figures are based on a Service that has been functioning in the Netherlands for over 13 years with many satisfied users, especially since the Remote Access service started in 2006.

Box 12.2. Example of costs of a national Data Access Centre: the Data Service Centre of Statistics Netherlands

In 2011, about 200 projects were active on Remote Access, involving 350 researchers. Around 100 projects took place base on On-Site facilities, involving 200 researchers. Remote Access projects lasted longer (from some weeks up to over a year), involving an average of four researchers. On-Site projects had an average of two researchers per project and lasted one to two weeks. The Service has over 800 registered users.

Facilities

There are more or less 100 Remote Access computers connected with the Data Service Centre of Statistics Netherlands. Another eight Remote Access computers are situated abroad. Also, 12 on-Site computers are available in two different offices.

Personnel involved

Full-time account managers (introduction, accounts, support)
Seven staffs for documentation
Staffs to support Remote Access (and On-Site) (also contracts)
35-40 staffs are available, not on a full-time basis, to check the output, in addition to two full-time staffs.

IT

IT-maintenance and support: one full-time staff
Servers, about 24 blade servers: 24 TB of user data (OS/RA)
Storage (NAS, 4 TB)
2,5 TB for Central repository
Software:

- Network components
- Authorization hardware and software
- Citrix-application
- Applications for research: MSOffice, SAS, Acrobat Reader, MLWin Ox, WinZip L, Gauss, StatTransfer 'R', GIS, STATA, Blaise, WinZip

Total cost and average cost

1. TOTAL COST	**€ 1.980.000**
Personnel	
Management	€ 150.000
Account management	€ 400.000
Administrative staff	€ 500.000
Documentation Centre	€ 650.000
ICT	***€ 80.000***
Licenses	***€ 200.000***
2. AVERAGE COST	
Per project	**€ 6.500**
Per user	**€ 3.500**
Account management	€ 365
Administrators	€ 910
Documentation	€ 1.180
Licenses	€ 365

4. Security Aspects of Data Access Services

422. Most Data Access Services have to deal with the same aspects in order to protect the data, develop technical (IT) solutions for hardware and software, provide facilities and manage the use and users of the Service. The "flow" of the security system could be described as follows:

a. The national law
b. The selection of institutes and researchers who get access
c. A project proposal
d. An introduction/training
e. A project contract
f. A security declaration
g. A facility contract
h. A security device
i. A well protected IT-environment and IT-facilities
j. Output control

The limitation of security

423. A system that can fully guarantee that no detail of a person or business would be revealed at any time does not exist. It is possible however to diminish the risk of revealing to an *acceptable* risk level.

424. Then there is 'trust'. For those aspects which cannot be completely secure, as the view recognition on the screen, one has to trust the integrity of the researcher. Besides, in many countries universities have rules for a careful use of sensitive data. If there is no trust in the researchers, the Service should be stopped.

425. The danger of misuse of confidential data is mainly related to the image of the NSO, and the International Research Centre eventually set. The whole existence of these institutions is based on people's trust in the protection of their privacy. From that perspective, reputation is more important than law and punishment.

Aspects of Security

The national law

426. Most countries have a national Law on Statistics and/or a Law on Protecting Private Information. Most of these laws do not contain sections which forbid giving access to confidential data for scientific research. Neither these laws give details about how the access should be organised. Mostly the law emphasise that data will be well protected and that the Director (-General) of the Institute is responsible for it. The explanatory notes of the laws sometimes give some leads for the design of a Service. Nevertheless, the National Law typically establishes the boundaries to the Service which cannot be ignored.

The selection of institutes and researchers who get access

427. This is one of the aspects present in most Laws on Statistics: restrictions of the institutions which are allowed to get access to confidential microdata for scientific research. Sometimes non-academic research institutions such as "Planning bureaus" cannot get access; access for other research institutions differs per country.

A project proposal

428. For security reasons, most NSOs with a Service want to give access to microdata databases which are necessary for a project. So a project proposal is required to get access. This forces the researcher to think about the research design and helps him/her in selecting the right databases. For the NSO itself, a proposal is interesting to observe what external researchers do with the microdata.

Introduction/Training

429. Not all Services have an introductory interview before the start of a project. The advantage of such an introductory session is that a research proposal can be fine-tuned because the experts from the NSO would know the databases involved better than an external researcher. Also, in a personal interview it can be made clear to the researcher that he/she would be working with very sensitive data. The disadvantage of having an introductory session is that it takes time for the account manager of the Service and the researcher. Besides, the researcher would need to travel for a face-to-face meeting, which can be difficult when the researcher is located far away and intends to work from a long distance (Remote Access). Some institutions provide training for the use of their Service; other only provider on-line training.

A project contract

430. For every project, a contract is made with the university/institution, holding the name of the project, the name of the professor/employee responsible for the project and the names of the researcher(s). (researchers involved ought to have a contract with the university/institution). Also, the databases concerned and the costs for the project are mentioned in the contract.

A security declaration

431. To make it possible to hold the researcher and the institution legal responsible in case data of persons or enterprises are disclosed, every researcher has to sign a security declaration. Also the responsible Dean or Director has to sign the declaration.

A facility contract

432. If a Remote Access facility will be used for a project, a separate contract for this facility will be necessary. This RA-connection has no direct link with the project: the facility can be used for different projects, for a longer period. Someone of the 'higher management' must be responsible for the required local facilities.

A security device

433. Depending on the design of the security model, different devices for the control of researchers are needed. Some devices give 'access on a distance', *i.e.* the Service can observe what the researcher is doing (Team Viewer). Other devices control the identity of the researcher: these are biometric devices (fingerprint- or iris-scanners).

A well protected IT-environment and IT-facilities

434. The IT environment where the sensitive microdata are obtained and can be used needs to be optimally protected. To that purpose, there are some well-known applications, such as Citrix. For an efficient use and analysis of the data the most common programmes and facilities should to be available.

Output control

435. The researcher is responsible to make sure that the output he/she want to take with him/her does not disclose protected information. In light of the serious consequence that information disclosure could have, most Services choose nevertheless to perform an output control, integral or random.

436. Some Services have rules concerning the way the output produced by the researcher should be presented to the Service.

5. Concluding remarks

437. The costs of establishing and managing an International Research Centre will depend on the choices that are made about the facilities, the security measures and the overall design of the Service. The costs are expected to be substantial anyhow. It is an important task for the party responsible for the development and establishment of an International Research Centre to make a proper financial plan.

CHAPTER 14. NEW MICRODATA DISSEMINATION SYSTEMS

by Luisa Franconi and Daniela Ichim

Introduction

438. The impact of technological advances on economic systems and societies, new priorities of governments and the increased volume and diversification of sources of data freely available on the web result in new needs concerning statistical information. National Statistical Offices (NSOs) and statistical agencies are one of the key sources of statistical data and information in every country. Originally created to serve government's needs, they are now expected to provide a wide range of products and services to a broad spectrum of users. These users differ greatly in their information needs and in their ability to manage statistical information. In this context, several challenges are faced by NSOs: with regard to the release of microdata, the challenges concern i) the production of an increased number of consistent and relevant microdata products, on more topics, more and more spatially detailed, and ii) the development of strategies and dissemination tools which will be able to supply information to different users.

439. From a marketing management perspective, to achieve the maximum return on their (and public) investment, statistical agencies should plan, develop and align their products and services to the needs and expectations of key user groups, and match the content and attributes of products to the corresponding statistical literacy and related competencies. In practical terms, this translates into making sure that the information provided incorporates manifold rhetorical levels, makes use of diverse graphical and descriptive appearances, is delivered through several different devices, and is organized so as to respond to the questions and concerns that different user groups may have. This approach adopted for classical statistical products should be applied also to microdata access.

440. Despite the current unfriendly and not-immediate way in which microdata are provided by NSOs, more and more students, public administrations, private and public institutes request access to them, giving rise to the phenomenon of massive requests for Public Use Files (PUFs). If microdata access has to become a key service for statistical agencies, then organisational competences, management processes, cultural norms, legislative frameworks and microdata ownership need to be reconsidered (McMillan, 2010).

441. This document highlights general strategic issues to be addressed in the renovation of dissemination systems and the production of PUFs.

New dissemination systems

442. The definition of dissemination strategies influences improvement on each dissemination activity: characterising dissemination policies; designing products and services, editing and presentation; outlining breakdowns; disseminating statistical information on the website; promoting products and marketing the modes and channels for the actual “release” to users. However, if microdata access has to become one of the key services for statistical agencies, far more than mere dissemination strategies is needed. As the input phase of official statistics, *i.e.* data capturing, has recently been transformed by company-centric communication approaches (see, for example, Marske and Stempowski, 2008), the output phase should follow a similar renovation and put the user at the centre of microdata access programmes.

This change is inspired by simple principles: unique point for finding information, microdata products and services tailored to user demands, services facilitating microdata usage.

443. The design of a user-centric dissemination system implies the definition of governance, an infrastructure, a legal framework, together with access policies, diversification and flexibility of microdata products and services, relationships with users, etc. If such an approach is pursued in many countries, economies of scale could be reached during setting up and maintenance, links could be easily implemented between systems and international access could be favoured.

Governance

444. *Current experiences*. In Italy, the vision for a user centric approach to microdata access sees a group of institutions that shape connected data clearinghouse initiatives and related infrastructure that will provide the broadest possible access to public funded microdata. The governance of this operation is led by official statistics: Istat will be the pivot of a joint venture that aims at developing a network of data archives that will manage microdata stemming from public funding. Future developments foresee Istat as the hub for access to microdata from the central institutions (the ministries) of the Italian National Statistical System. Other countries have chosen similar approaches: for example, Réseau Quetelet in France and UK Data Archive in the United Kingdom together with their partner organisations provide several services for microdata.

445. The governance of the dissemination is the key issue to be addressed: products and tools can be developed, systems can be designed and built, architectures can be adapted, but a sound governance of this evolution is essential for efficiently managing access to official statistics microdata.

Infrastructure

446. In terms of infrastructure, the new dissemination systems aim at developing a coherent and single retrieving system of microdata products or services, with a web catalogue of all microdata available and accessible to users either directly through a download or via requests for further specialised services (remote execution, remote access, etc.).

447. This vision of a single access point via the web for services related to national microdata access is supported for example by the United Kingdom, and is envisaged at European level by the Council of European Social Science Data Archives (CESSDA). Under the auspices of the FP7 project Data without Boundaries, 28 institutions are working for the coordination of existing infrastructures for access to official micro data in Europe.[66] Also, the International Household Survey Network (IHSN) helps statistical offices in developing countries to improve the availability, accessibility and quality of survey data in their nations and provides a single point of access for microdata for all such countries.

Legal framework

448. In Europe, the revision of EC Reg. 831/2002, see Chapter 8 of this report, on microdata access leads to improvements not only for microdata research access but also in the area of PUFs. In any case, a legal framework that sustains such architecture is crucial.

[66] See www.dwbproject.org, 7th Framework Programme for research and technological development (FP7) of the European Union.

Standardised metadata

449. A "user-centric" system allows for a quick and clear picture of microdata really available and under which conditions. User-friendly searching protocols need to be developed to help users in finding the microdata they need. Additionally, standardised metadata protocols, which are a milestone of any dissemination system, need to be carefully selected to guide data interpretation.

Relationship with users

450. The creation of user-centric dissemination systems where PUFs are key products will modify the relationship with users and help the diffusion of statistical literacy: statistics, analytics, data analysis, computer programming are all essential skills when it comes to microdata analysis. A culture of data analysis should be understood in its simplest form by common citizens, should be clearly taught to students and should be a core competency in public administration.

451. To that purpose, the presence of powerful tools to analyse vast data stores or statistical microdata from surveys is not sufficient. Statistical literacy (*i.e.* how to choose suitable methods, apply them correctly and understand and interpret results) should be constantly promoted and supported by NSOs both inside their national statistical systems (government agencies, public administrations) and outside (undergraduates, master and PhD students, junior researchers).

452. For example, in response to the increasing need for quantitative literacy coming from the society and public administrations, Istat established in 2011 the Advanced School for Statistics and Socio-Economic Analyses (Scuola Superiore di Statistica e Analisi Economica, SAES). The school offers training programs on advanced survey techniques and statistical methodologies, bespoke training, traineeships, promotion of statistical literacy. The aim is twofold: increase statistical knowledge and analysis inside the public administration, and train future users of official statistics. Co-operation initiatives will pave the way for changes in the relationship between producers and users of official statistics. Users will not just be data analysts but will be more and more called for an active contribution toward the improvement of official statistics.

Standardisation as a driver for developing new dissemination systems

453. In recent years official statistics have been confronted with both accelerating change in the society and resource constraints that led chief statistician to question the whole architecture of the statistical production process. In Europe, this is addressed in the *Communication from the Commission to the European Parliament and the Council on the production of EU statistics: a vision for the next decade (COM 404/2009)*. At the worldwide level, the initiative of the High-Level Group for the Modernisation of Statistical Production and Services (HLG) (previously called High-Level Group on Business Architecture in Statistics HLG-BAS) has strengthened the need for standardisation and industrialisation of official statistics as it seeks to reuse and share methods, components, processes and data repositories and adopt a shared “plug-and-play” modular component architecture with the aim of increasing efficiency.

454. The implementation of such an architecture can be made possible only by sharing common standards represented by, on the one hand, the Generic Statistical Business Process Model (GSBPM) and, on the other hand, the Generic Statistical Information Model (GSIM). The former is necessary to define common components of the statistical process; the latter is a reference framework of information objects which enables generic descriptions of the definition, management, and use of data and metadata throughout the statistical production process. Together, they represent the starting point to define a common language and to set up an integrated production system.

455. In Europe, a strategic task force, the Sponsorship on Standardisation, has been set up to advise the European Statistical System on how to pursue standardisation and integration (see Braaksma, et al. 2013). At Istat, the Stat2015 project aims at the standardisation and industrialisation of production processes based on re-use and on the adoption of a model founded on shared services in a service-oriented architecture (SOA) framework (Falorsi et al. 2013).

456. According to the vision launched by the HLG, validated microdata stem from a standardised and harmonised process and are accompanied by appropriate semantics that can be used unambiguously across and between different implementations. Further standardised estimation procedures lead to the classical official statistics products such as indicators, aggregates, graphs, small area estimates, etc. The validated microdata are, in the GSBPM, singled out in a specific sub-section (5.8) to state the importance of the achievement: validated data are, themselves, the output of a standardised process. Moreover, by adopting the perspective of a statistical process which is governed by semantics and is truly metadata-driven and thanks to the use of standardised protocols that link the data to the metadata since their capture (*e.g.* DDI and/or SDMX), validated microdata already feature the high level of quality necessary to be considered a product on its own right.

457. Under the vision of HLG, the development of metadata for microdata to be released will not be a demanding task as it will come as a by-product of the statistical production process.

PUFs as a way to increase impartiality, transparency and public trust

458. Dissemination and communication of "value added" analytic products require making specific choices about what data or patterns to show, what findings to interpret, what implications to discuss, but also what to bypass, ignore or downplay. By increasing impartiality and neutrality of official statistics, microdata-based products definitely contribute to increase the transparency of the NSOs and, most importantly, their credibility and public trust. Indeed, only access to microdata allows users to "clone" the NSOs estimates, perform analyses and comparisons, thus contributing also to a continuous innovation and customisation of the statistical system to the society information needs.

459. The high standards of quality, as well as strict ethical and professional principles followed by the NSOs in the production of statistical information should encourage more NSOs to open their microdata banks, in full compliance with confidentiality laws. Transparency, impartiality and neutrality can be increased by adopting microdata as a product on its own right and by distributing microdata in a more open manner like it is the case of PUFs. Transparency and public trust have however requirements: on the one hand, to adopt sound statistical methodology for the statistical disclosure control (SDC) process, and on the other clearly report the methods used and their statistical properties. The metadata released together with a microdata file should contain information on the methods used to limit disclosure, which statistics are maintained, which inferences could be affected by these methods and to what extent.

Public use files and the philosophy of re-use

460. Public use files are samples of individual data having well defined characteristics: they are freely available on the web (*i.e.* users do not need to sign any access agreement), they allow users to make sensible inferences on the phenomena for which the data were collected, usually under no restrictions/conditions on their use. Usually, sub-sampling technique are used for the production of PUFs, together with other protection methods, as they reduce the risk of disclosure by increasing the uncertainty on the number of population units sharing the same score on identifying variables (Hundepool et al. 2012).

461. Other approaches could be followed under the umbrella of the "re-use" philosophy. For example Istat re-used two ingredients for PUF production: i) an already existing product, namely microdata files for

scientific research (MFR); and ii) strong sampling competences inside the organisation. In order to gain efficiency in the production of PUFs, a new methodological solution was developed based on sub-sampling from the corresponding MFR (Foschi et al., 2012). The PUF and the MFR therefore share the same structure. Such hierarchical structure of the two data sets greatly simplifies assessment of the disclosure risk and information loss associated with the anonymisation procedure, and preserves the hierarchical detail as well as the internal consistency of the records. Moreover, the hierarchical structure between the two files allows for a reduction in the cost of preparation, therefore increasing efficiency.

Reasons for Public use files: democracy and statistical literacy

462. The development of a public use file that shares the same details, quality and complexity of the corresponding file for research purposes is based on the principles of democracy of access and right to research as well as the certainty that only by allowing students to be trained on complex official data statistical literacy will increase in a country.

463. The interest in developing proper PUFs instead of teaching files, which contain very few variables, basic information, very few observations and present a very simplified structure, stem from the observation that teaching file are valid merely to apply methods/formulas. On the contrary, there is a need to teach how to practice on the representativeness of a real survey microdata, to teach how to make reasoning out of data and, more importantly, how to extract knowledge from data.

464. If a PUF satisfies some predefined quality standards, it would positively contribute to the diffusion of statistical literacy. The value added of a PUF is a straightforward consequence of its quality standards defined as the ability to simulate real applications. The file dimension, expressed as number of records and number of variables, provides a first quality indicator. Moreover, since the data production process and data quality are not extensively discussed in statistical lectures, any PUF could contribute to the reduction of this gap. At the same time, a large number of variables would favour the development of a critical reasoning on the variables meaning, their operational definition, the surveyed phenomenon, etc. The precision and accuracy of the estimates that could be derived using the PUF would significantly improve the conceptual learning. Table 1 shows which PUF characteristics suit the process of knowledge extraction.

465. Finally, under the communication activities related to the launch of PUFs, it has to be stressed the need for a clear license for use as well as the development of tools and services for analysing microdata products. Microdata offer the greatest possible flexibility when analysing a phenomenon. At the same time, microdata are not user-friendly and do not offer the possibility to get immediate results. Only the design and development of adequate tools and services for microdata analysis would improve their usability.

Table 14.1. Characteristics of PUFs and MFRs and their suitability in the process of knowledge extraction

PUFs	Knowledge extraction	MFRs
	Problem definition – hypothesis formulation	
✓	Data reading (format, documentation, classifications)	✓
✓	Initial data analysis	✓
✓	Use of statistical methods (modelling, cluster analysis, etc.)	✓
✓	Interpretation of the results	✓
✓	Audit of the results	✓
✓	Comparisons	✓
✓	Consequences	✓
✓	Hypothesis re-definition	✓
✓	Report and presentation of the results	✓
✓	Development of statistical and non-statistical application	✓
	Development of social and economic theories	✓
	Policy decision making	✓

Recommendations

466. In terms of maturity model an NSO that has not addressed the issue of developing or collaborating in the creation of a user-centric dissemination system indicates a low level of maturity. NSO with a high level of maturity has identified a way to reach the creation of such a system and has a clear model of governance for it.

467. In light of the discussion above, it is proposed that NSOs should:

- Make efforts inside their national statistical systems in order to collaborate with other public administrations and institutions for the coordination in the release of microdata products.
- Collaborate to promote centre of expertise in the domain of SDC methodologies to increase the release of PUF and access to microdata in general.
- Address the issue of the governance of microdata dissemination systems in order to favour national and international access.
- Promote the re-use approach also in the area of PUF development.
- Foster the development of new, integrated and coherent dissemination systems.

REFERENCES

Braaksma, B., Colasanti, C., Falorsi, P.D., Kloek, W., Martinez Vidal, M. and Museux, J.M. (2013), Standardisation in the European Statistical System, NTTS Conferences on New Techniques and Technologies in Statistics, Brussels, 5-7 March, 2013, available at: http://www.NTTS2013.eu

Falorsi, P.D., Barcaroli, G., Fasano, A. and Mignolli, N. (2013), A Business Architecture framework for industrialisation and standardisation in a modern National Statistical Institute. NTTS Conferences on New Techniques and Technologies in Statistics, Brussels, 5-7 March, 2013, available at: www.NTTS2013.eu.

Foschi, F., Casciano, C., Ichim, D. and Franconi, L. (2012), Designing Multiple Releases from the Small and Medium Enterprises Survey, In J. Domingo Ferrer and I. Tinnirello (eds) Proceeding of PSD2012, Vol. 7556, Lecture Notes in Computer Science, pp 200-215. Springer, Berlin/Heidelberg.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K. and De Wolf, P.P. (2012), *Statistical Disclosure Control*. Wiley.

MacMillan, P. (2010), Unlocking Government: How Data Will Transform Democracy, Commonwealth Innovation, Vol. 16, 2, pp. 13-17.

Marske R. and Stempowski D.M. (2008), Company-centric Communication approaches for Business Survey Response Management. Proceedings of Statistics Canada Symposium 2008. Data collection Challenges, Achievements and New Directions.

CHAPTER 15. CASE STUDY: RELEASING THE VALUE OF ADMINISTRATIVE SOURCES

by Claus-Göram Hjelm, Ivan Thaulow and Johan-Kristian Tønder

Administrative data in statistics and research

468. All national statistical offices (NSOs) have a duty to produce official statistics with the highest possible quality, and to balance the need for quality against the available resources.

469. In the Nordic countries[67], detailed administrative data are produced as a result of administrative processes mainly defined by laws. Thus huge amount of administrative data are produced each day for other purposes than statistics and research. In the Nordic countries these administrative data are provided free of charge to the NSOs. Exploiting administrative data for the production of statistics is often a most cost-effective way to build statistical databases rather than establishing special data collections for statistical purposes only. Instead, the resources are better used for corrections and supplements of the administrative data.

470. Statistical registers based on administrative sources include all relevant units, not only a sub-sample of the population, and often more variables than censuses and other total enumerations (Box 1). This potentially improves statistical production compared to traditional data collection. Since administrative sources are regularly updated, statistical registers, statistics and research based on these sources will be timelier than statistics and research based on surveys and censuses. Indeed, the re-use of administrative data for statistical purposes is one of the key elements identified by the European Commission to make the production of European statistics more efficient.[68]

Box 15.1. Administrative registers, private registers and statistical registers

Administrative registers are registers primarily used in administrative information systems. In practice, the information is a result of the production of goods and services in public or private institutions and companies. In the Nordic countries most administrative registers used for statistical purposes are country-wide registers operated by the state or jointly by local authorities. However, private registers are also used in statistics, for instance registers operated by insurance companies and employer organizations.

Statistical registers are created by processing data from administrative registers. Statistical registers could be based on a single administrative register, but they are more frequently based on combined data from several administrative sources.

523. Researchers too are interested in detailed data that are well-documented and of high quality. For that reason, researchers often prefer to receive administrative microdata from NSOs instead of asking the responsible institution for such data. A main condition for the utilisation of these data for research is in fact

67 For the purpose of this chapter, Nordic countries include: Denmark, Norway and Sweden.

68 Communication from the Commission to the European Parliament and the Council on the production method of EU statistics: a vision for the next decade, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0404:FIN:EN:PDF

the availability of metadata, especially as concerns the description of data collection methods, accuracy, changes in data over time, such as changes of laws etc.

524. In the Nordic countries, the national NSOs have established services for access to microdata for authorised research institutes. Nordic NSOs provide researchers access to tailor-made sets of data compiled from various registers and data sources. Depending on the country and researchers' needs, data are either anonymous or de-identified, i.e. data where names, social security numbers and business registration numbers have been removed or replaced by project specific identification numbers. However, as some research projects require a large amount of variables and details, the identification of particular units can still be possible.

525. When deciding on the number of units and details of the data files to be delivered, the NSOs have to use the "need to know"-principle, balancing the researchers need for flexibility in their analytical work against the need for protecting official microdata. This discussion is, of course, very important in cases where researchers want to link their own data to microdata from NSOs, or when researcher ask for "sensitive information", such as crime records or mental diseases.

526. In addition to national researchers affiliated with institutions such as universities and hospitals, other users of statistics might want to have access to microdata. Also, by remote access techniques, it is already possible in some countries to provide central and local planning institutions, policy makers and the media with a service of microdata access; other groups of users are expected to be covered by the service, including researchers outside the national boundaries.

A model for the progressive introduction of the re-use of administrative microdata and their access to researchers

527. Establishing statistics based on administrative data may be realised in three steps. *The first step* is to make an agreement between NSOs and one or two owners of administrative registers. The NSOs receive administrative microdata, prepare the data for statistics and produce official statistics from the prepared microdata. NSOs may also produce, upon demand by the owner of the original data, statistics based on the treated microdata; however, these treated microdata remain under the responsibility of the NSOs in the same way as microdata from surveys and censuses.

528. After getting some experience with administrative microdata from one or two owners, *the second step* must be to convince the authority responsible for official statistics on the positive effects that the NSO's re-use of administrative microdata has on the quality and costs of official statistics, research and evidence-based policy. The statistical authority would then persuade all the concerned parties about the benefits for the whole government of allowing the transmission of administrative microdata to the NSO. The NSO may subsequently extend the range of administrative microdata accessible to researchers, by making agreements with other institutions. The agreements should clearly specify that the NSO has the possibility to give researchers access to prepared microdata without previously asking permission of the owner of the original data. If the agreements have no such provision, the NSO has to ask the owner's permission before providing microdata access to researchers, and this for each research project.

529. Some institutions and owners of administrative microdata are not willing to make an agreement with the NSO for different reasons. The two most used argument are: "The microdata are too sensitive to be given to the NSO", and "Statistics on our microdata must be produced and presented by our own experts." The first one may reflect a sense of mistrust in the NSO. The second motivation goes against the principle of independence of official statistics.

530. *The third step* of the process of producing statistics from administrative data is to have a Statistics Act that gives the NSO the right to re-use administrative data from all nationwide registers owned by government and local institutions (Box 2). The Statistics Act must give the NSO the control of the microdata they prepare from administrative microdata, both for producing statistics and giving access to researchers. It is possible that, even after setting such an Act, for some time some institutions would continue to refuse to provide their microdata to the NSO.

Box 15.2.The Statistics Act in Nordic countries

The legislation provides a key foundation for the use of administrative data sources for research and statistical purposes. National legislation must reflect the broadly held view that it makes good sense to take advantage of using administrative data sources rather than to re-collect data for statistical purposes.

All Nordic countries have a national Statistics Act that gives the NSO access to administrative data on the unit level with identification number, and that allows the NSO to link them with other registers and sources for statistical purposes. Furthermore, the Statistics Act has a detailed specification of data protection. The legislation of the administrative registers gives, with some exceptions, the possibility for the administrative institutions to provide microdata access to research institutions, directly or from the NSO.

Challenges arising for international exchange of microdata

531. A country's legislation should provide the key foundation for the re-use of data from administrative sources for research and statistical purposes. Indeed, national legislation must reflect the broadly held view that it makes good sense to take advantage of using administrative data sources rather than to re-collect data for statistical purposes. As it already occurs in Nordic countries, a national Statistics Act should give the NSO access to administrative data on the unit level with identification number, and should allow the NSO to use identification numbers to link administrative data with other registers and sources for statistical purposes, if necessary. Furthermore, the Statistics Act should contain precise provisions for the protection of microdata, both inside the NSO and when access to microdata is given to external users. Alternatively, legislation regulating the administrative registers could allow the administrative institutions responsible for the registers to provide microdata access to research institutions, directly or from the NSO.

532. The national legislation would presumably apply to all institutions within the country and regulate access to microdata and sanctions in case of breaches. The problem remains of research institutions in other countries; the uncertainty concerning the applicability of sanctions to such institutions could result in the denial of access to national microdata. As far as remote access can satisfy a researcher's needs, this may be a safe and effective way to solve the problem, because researchers are only allowed to send tables and statistical calculations to their home computer and not the detailed microdata. In the Nordic countries where remote access is used, the files with tables that are sent to the researchers are routinely checked by the NSOs, in order to ensure that no microdata are leaving the NSO server. If remote access is not available, de-identified microdata may be transferred to the research institution (the solution used in Norway today), or the researcher may have access to de-identified microdata from a "research laboratory" inside the NSO (the solution in Denmark before they got remote access). In both cases the access to microdata are restricted to some specified research project appointed between the NSO and the research institute.

533. Finally, building a "Circle of Trust" (Chapter 2) would significantly facilitate a country's provision of administrative data to researchers from other countries. Still, with many national owners of administrative data, it might be difficult to establish a "Circle of Trust" involving different institutions in

different countries. A more feasible alternative would be to distribute microdata from administrative sources and registers through a "Circle of Trust" encompassing NSOs with admission to such data for statistical purposes; an example of implementation is presented in Chapter 7.

Recommendations

534. It is recommended that:

a. NSOs integrate administrative data as input for their production of statistics.

b. NSOs provide access to microdata based on administrative and other sources to researchers.

c. The national Statistics Act gives the NSO, as an independent institution, the right to re-use administrative data for official statistics as well as permission to give access to microdata for research purposes.

d. Legislation on data protection, both general and the special for statistics and research, is implemented in a way that persuades the public opinion on the benefits of using administrative data in statistics and research.

e. Comprehensive documentation becomes accessible to researchers. Specific documentation on important changes in the statistical registers and administrative sources over time is critical for the usability of the registers for research purposes, since researchers very often use time series in their analytical work.

CHAPTER 16. CASE STUDY: NEW PRACTICES IN MICRODATA ACCESS IN ESTONIA

by Tuulikki Sillajõe

Introduction

535. The aim of this chapter is to share the Estonian experience in the adoption of new practices in microdata exchange, supported by the experience in developing new statistical law. It consists in a description of the policy and implications of innovation in microdata access; only aspects related to the dissemination of microdata are addressed, while other changes in the Estonian law on official statistics are not covered.

536. The State Statistical Central Bureau of the Republic of Estonia was established in 1921 and operated in accordance with its statutes. The first act governing the State Statistical Central Bureau was not adopted until 1938. During the Soviet era, there was no law on official statistics, but a Statistics Act was adopted again in 1990. This was a law typical of the transition period and had to be revised in 1997 according to needs of an independent state.

Previous legislation and practice

537. The Official Statistics Act of 1997 was broadly relevant. Its major shortcoming was that it included only 14 sections. Thus, it was also brief about the use of microdata for scientific purposes. The Act stipulated that only data allowing the identification of a natural person could be disseminated for scientific purposes without the person's consent. In addition there was a government regulation „Procedure for Transmission of Data That Permit Identification of the Data Subject without the Consent of the Data Subject for the Purposes of Scientific Research". The data of economic entities were not allowed to be disseminated for scientific purposes without the entity's consent. At the same time, no distinctions were made based on the source of the microdata. Thus, the use of data of natural persons obtained from both surveys and administrative registers was equally allowed, as long as used for scientific purposes.

538. The research institution requiring microdata had to present a simple application to Statistics Estonia, specifying the purpose of using these microdata. The Director General had to answer the request within two weeks and was not allowed to refuse without giving a reason. Microdata were disseminated on the basis of an individual agreement concluded between Statistics Estonia and the research institution. The content of these agreements was not standardised.

539. In practice, there were very few agreements with some major universities. Each agreement had several appendixes concluded over a relatively long time period for the use of different datasets. However, legally, these could not be considered framework agreements.

540. Microdata of the Labour Force Survey, the Household Budget Survey, the Social Survey, the Time Use Survey, etc. were disseminated under these circumstances. Microdata that allowed indirect identification (with light disclosure control applied) were transmitted on CD-ROMs, and always free of charge. Highly sensitive data was used for scientific purposes in the safe centre of Statistics Estonia. These practices were not publicly known or even communicated, but it were accepted and appreciated by researchers who happened to know about them and were thus able to use the service.

Steps of development

541. Statistics Estonia operates within the area of government of the Ministry of Finance. In 2008 the Ministry of Finance initiated and started to lead an inter-institutional working group for re-drafting the Official Statistics Act. The group included representatives of different ministries, the national central bank and Statistics Estonia as well as researchers.

542. In order to regulate the dissemination of microdata for research purposes more precisely, the representatives of Statistics Estonia included in the inter-institutional working group studied the corresponding practice of other countries. Based on information gained at international meetings, at conferences, from the websites of statistical offices, but also from the UN/ECE/CES handbook "Managing Statistical Confidentiality & Microdata Access: Principles and Guidelines of Good practice"[69], the relevant chapter of the Act was drafted. It was decided that the detailed procedures will be described in a separate document to be approved by the Director General of Statistics Estonia. Also, certain institutions were chosen to be visited in order to study their experience and practices more carefully.

543. In 2010 the representatives of the Ministry of Finance, the Ministry of Education and Research, the Bank of Estonia and Statistics Estonia made study visits to four institutions: Luxembourg Income Study, Statistics Finland, Statistics Denmark and Statistics Netherlands. The working group decided to be realistic in the introduction of a new policy and procedure for dissemination of microdata for scientific purposes.

544. During the study visits it became clear that it would be possible to take the necessary documentation (e.g. application form, oath of confidentiality, agreement with research institution, etc.) used by other statistical offices, modify it and shortly introduce it at Statistics Estonia. It also appeared that the LISSY software, developed and used by the staff of Luxembourg Income Study, was suitable for the provision of remote execution service to Statistics Estonia's customers.

Current legislation and practice

545. In August 2010, the fourth law on official statistics in Estonia came into force. By the end of 2010, the Procedure for dissemination of confidential data for scientific purposes had also been approved. So, the new policy of microdata dissemination had been implemented.

546. Dissemination of microdata is currently governed by sections 35–38 of the Official Statistics Act[70] and the Procedure for dissemination of confidential data for scientific purposes[71].

547. All categories of microdata collected for the production of official statistics (social survey data, census data and business data) are available for scientific purposes without the consent of the person, i.e. microdata of both natural persons and all kinds of economic entities. Also, data derived from administrative records and other databases may be disseminated for scientific purposes.

548. Data permitting direct identification and data permitting indirect identification of a statistical unit are both allowed to be disseminated for scientific purposes.

---

69 Managing Statistical Confidentiality & Microdata Access: Principles and Guidelines of Good practice. (2007). UN. http://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

70 Official Statistics Act. http://www.legaltext.ee/et/andmebaas/tekst.asp?loc=text&dok=XXXXX42K1&keel=en&pg=1&ptyyp=RT&tyyp=X&query=riikliku+statistika

71 Procedure for dissemination of confidential data for scientific purposes. http://www.stat.ee/dokumendid/51669

549. In case of Statistics Estonia, the following means for access to microdata are available:

- Remote Access: a service whereby a researcher performs the analysis and can immediately see the answer on the screen
- Remote Execution: a service whereby users can use microdata for making statistical analyses in a manner which precludes the users' direct access to microdata, i.e. a researcher submits a query and receives the output later over the Internet
- Safe Centres: are located in the offices of Statistics Estonia in Tallinn and Tartu.

  In case of these three modes of access, the results are sent back to the researcher via computer networks after checking for confidentiality

- Scientific Use File on CD-ROM or via FTP-server: microdata to which methods of statistical disclosure control have been applied, in order to reduce the risk of identification of the statistical unit to the appropriate level in accordance with current best practice.
- Public Use Files on the web: are published on the website of Statistics Estonia and do not allow any direct or indirect identification of a statistical unit, because these data files are prepared by applying statistical disclosure control methods
- Order For Information: the staff of Statistics Estonia produces tailor-made information for a particular use

550. Microdata are allowed to be used for scientific purposes only by legal persons or agencies, but not by freelance natural persons. Pursuant to the Estonian law, "a research and development institution" is an institution specified in section 3 of the Organisation of Research and Development Act, or a university or another establishment providing higher education of a foreign state or a research institution of a foreign state, or an institution listed in the relevant Decision of the European Commission. Pursuant to section 3 of the Organisation of Research and Development Act, a research and development institution is an institution in case of which: the principal activity is carrying out basic research, applied research or development; the activity accompanying principal activity is to spread knowledge through teaching, publication or technology transfer; the results of the principal activity financed from the state budget funds (which do not involve intellectual property rights) are public information; the membership includes the research staff necessary for carrying out the principal activity.

551. Students pursuing a Master's or Doctor's degree are also considered researchers. The same rules apply to domestic and foreign research institutions.

552. A legal person or agency in need of confidential data for scientific purposes must submit a written application to Statistics Estonia. The application must set out the following information: name of legal person or agency; registration code of legal person or agency; title of the research; objective of the research; name of the statistical action or a list of data necessary for the research; a list of data which the applicant has obtained from other sources and which the applicant wishes to link with the data applied for; a list of persons wishing to use the relevant data during the research (given and surname, personal identification code, email address); in case the use of personal data is involved, a confirmation issued by the Data Protection Inspectorate to prove that the applied organisational, physical and information technology related security measures are sufficient and, if an ethics committee has been founded, also the opinion of such committee; in case the use of sensitive personal data is involved, a confirmation issued by the Data Protection Inspectorate proving that the processing of sensitive personal data has been registered.

553. Statistics Estonia has the obligation to consider each application separately. Applications for the dissemination of confidential data for scientific purposes are reviewed by the Confidentiality Council

according to their order of arrival. The Confidentiality Council considers the substance of the application and decides whether the confidential data can be used for scientific purposes, whereas the decision must be made within ten working days from the receipt of all documents necessary for evaluation of the application.

554. The Confidentiality Council consists of the following public servants of Statistics Estonia: the Deputy Director General, the Head of Methodology Department, the Head of Population and Social Statistics Department, the Head of Enterprise Statistics Department, the Head of Price and Wages Statistics Department, the Head of Agricultural Statistics Department, the Head of Information and Marketing Service of the Marketing and Dissemination Department, the person responsible for personal data protection, and the lawyer.

555. The Confidentiality Council meets once a week, if an application has been received. Decisions taken by the Confidentiality Council shall be confirmed by a directive of the Director General.

556. The Confidentiality Council shall consider every dataset requested by the applicant, taking into account two aspects: 1) risk of identifying a statistical unit; 2) the impact that the identification of a statistical unit may have (sensitivity of data). If the identification risk and sensitivity are rated as high, confidential data can be used only in a safe centre, by remote access or remote execution. If both aspects are rated as low, the requested data can also be delivered on removable devices.

557. Statistics Estonia may refuse to disseminate microdata only if:

- it is not convinced that the data will be used only for scientific purposes;
- the applicant wants to use personal data but the conditions provided for in the Personal Data Protection Act are not fulfilled;
- the applicant has previously violated the terms and conditions of an agreement entered into with Statistics Estonia and, in the opinion of the producer of official statistics, the applicant has not implemented sufficient measures to prevent violation of the terms and conditions of an agreement in the future.

558. Before the dissemination of confidential personal data for scientific purposes, Statistics Estonia shall enter into an agreement with the user of microdata stating the purpose of the research, the persons entitled to use the transmitted data for research, the procedure for processing and transmission of data and the obligation to ensure the organisational, physical and information technology related protection of data, and conditions for the destruction of data after completion of the research.

559. All users of microdata mentioned in an agreement shall sign a written confidentiality agreement.

560. The use of microdata is free of charge. A user shall only pay for the tailoring of the data for their needs and purposes (*i.e.* linking, matching, etc.): the cost is 100 euros for the first set of data per year and 50 euros for every subsequent dataset.

561. Penalties for the violation of relevant regulations are stipulated in section 40 of the Official Statistics Act. The unlawful dissemination of data which have been collected during the production of official statistics or which enable the identification of a respondent, or the use of data for other than statistical purposes is punishable by a fine of up to 200 fine units (800 euros). The same act, if committed by a legal person, is punishable by a fine of up to 3,200 euros.

562. Since the adoption of the current policy on 19 October 2010, Statistics Estonia has received and processed 63 applications (as of 12 December 2012) leading to the conclusion of 35 agreements on remote

access or use of safe centre, and 12 agreements on ftp-services. In case of seven applications, an agreement is being prepared. Five applications have been cancelled and four has been rejected. Two of the agreements have been concluded between Statistics Estonia and the OECD on access to confidential data using remote access.

Conclusions and lessons learnt

563. As a result of public debate, the new law on official statistics is quite detailed, including the regulations concerning microdata. The Official Statistics Act includes 62 sections, instead of the fourteen of the previous version.

564. There is no point in reinventing the wheel. This means that by using the best practices already available in statistical offices, a relevant and up-to-date policy for access to microdata can be implemented relatively easily and quickly.

565. Statistics Estonia has observed that supply creates demand; also, a clear and quick procedure enhances access to microdata.

566. The new policy is one of the most liberal ones in the world; it was created in close cooperation with the main stakeholders; the microdata collected for production of official statistics are, in practice, more easily accessible for scientific institutions; and in two years Statistics Estonia has only four times refused to allow access to microdata, because the data were clearly asked for non-scientific purposes. Despite all these considerations, the policy change has been perceived as negative by many researchers, and Statistics Estonia has still gained a somewhat negative image as unfriendly to research. Therefore, communication with stakeholders and interest groups is an ongoing challenge.

CHAPTER 17. CASE STUDY: CONFIDENTIALITY ON THE FLY IN AUSTRALIA

by Melissa Gare and C. Chien

Introduction

567. Vast amounts of microdata are collected by agencies from Censuses, surveys and administrative sources. Such microdata can be used in the development and evaluation of policy for the benefit, or utility, of society. The demand for gaining greater access to such microdata has continued to grow since the 2003 Conference of European Statisticians (CES) session on the topic of confidentiality and microdata access.

568. The ABS mission is to "assist and encourage informed decision making, research and discussion within governments and the community, by leading a high quality, objective and responsive national statistical service". The ABS, like many other National Statistical Organisations (NSOs), has needed to 'rethink' microdata access, faced with the challenge of balancing the trade-off between legal obligations to ensure that the likelihood of disclosing information about a particular person or organisation is unlikely, with releasing more detailed microdata for informed decision making, research and discussion to benefit society.

569. Managing the risk of disclosure is commonly referred to as Statistical Disclosure Control (SDC). Even after removing personal identifying information, such as name and address, from the microdata the risk of disclosure remains (see for example Willenborg and de Waal, 2001). Given the increasing amount of available data via administrative and linked sources, facilitated through technological advances, the risk of disclosure for microdata is arguably ever increasing.

570. There are a number of software products for confidentiality that are currently available. The majority of these are designed to be used by NSOs to confidentialise data before release. These can be either for tabular data (for example, Tau-Argus and sdcTable) or for microdata (for example, the Special Uniques Detection Algorithm and sdcMicro). Many NSOs have also developed their own tailored processes and software specific to their legislative requirements.

571. The ABS has progressively developed a range of methods, processes and applications to produce a range of products to provide access to its statistical data including the release of statistic tabular data to the ABS's Web site, the release of customised tabulations through an information referral service and through the analysis of pre-confidentialised microdata files, referred to as Confidentialised Unit Record Files (CURFs). CURFs are made available either in the form of a Basic CURF for analysts to use in their own research environment, or as an Expanded CURF which is accessible through submitting queries remotely in the ABS Remote Access Data laboratory (RADL), or by visiting one of the on-site ABS Data Laboratories (ABSDL). Typically, approved users submit queries in SAS, STATA or SPSS. In the case of RADL, the results of the queries are automatically checked and cleared outputs made available to users via their desktops. Additionally, a sample of queries is selected on an ongoing basis for manual inspection. For the ABSDL all outputs to be removed from the laboratory are cleared manually. The level of pre-confidentialisation applied to each CURF file is dependent on the method of dissemination, from the heavily confidentialised Basic CURF to the more detailed, less confidentialised Expanded CURFs. Considerable amounts of staff resources and time are required to produce a CURF in either form.

572. In response to a number of drivers of change, including the increased complexity of datasets where traditional approaches are likely to break down, the ABS has recently invested in the development of a suite of research applications that enable registered analysts to submit queries via the internet that are executed against the underlying microdata and confidentialised in real-time "on the fly" with

confidentialised output returned to the analyst. Internationally, such applications are commonly referred to as Remote Analysis Servers (RASs). They provide users with control over the particular outputs they want to extract from a dataset. The challenge for the statistical offices is to provide SDC for the different possible outputs.

Drivers for change

573. Many national statistical offices, including the ABS, are changing the way in which microdata are disseminated to researchers. Key drivers for this change include:

- increasing demands for better, more flexible and more timely access to detailed microdata;
- enhancing user experience through the provision of user friendly, menu-driven interfaces that are not reliant on users having statistical programming skills;
- reducing costs associated with existing manual and resource intensive approaches to disseminating microdata, such as the process of creating CURFs;
- increasing the timeliness of access to microdata (ABS CURFs are currently available up to 6 months later than static tabulations are disseminated);
- mitigating the increasing risk of disclosure as a result of increased computing power (both hardware and software), the increased volume of outputs disseminated and the increased accessibility of large datasets;
- facilitating analysis of emerging sources of data (including transactional, administrative and integrated data) where traditional approaches to SDC are not sufficient to mitigate the increased identification risk;
- a growing recognition that not all essential statistical assets are held by the National Statistical Organisations, hence the need to develop methods and infrastructure that can be utilised by other organisations; and
- the changing model of the typical data analyst to organisations looking for increased accessibility of outputs through more efficient machine to machine querying (such as through the use of SDMX Web services).

574. These drivers for change have led many NSOs, including the ABS, to commence development of sophisticated Remote Analysis Servers.

ABS Remote Analysis Servers - confidentiality on the fly

575. The suite of Remote analysis servers at the ABS enable the ABS to make the full detail of the dataset available while minimising the loss of utility through the application of SDC tailored to each specific output. From a user's perspective, they have control over the particular outputs they want to extract from a dataset. This is a fundamental shift in the process, from the traditional paradigm where the ABS would decide all the outputs that would be released to one where users can specify what they require when and as they need it.

576. There is a real risk of disclosure from the dissemination of tabular and analytical outputs that needs to be mitigated. A number of papers have been published on this subject, including proposed approaches to manage the disclosure risk. In respect to analysis output see Gomatam et. al (2005), Bleninger et. al (2011) and Sparks et. al (2008) and in respect to tabular output see Shlomo (2007). The goal of this literature is to protect against data attacks, which involves an analyst using output from an analysis server, including graphics and model diagnostics, to reconstruct attributes for one or more records

which, if successful, could be used to attempt identification. The challenge for the ABS is to provide SDC for the different possible outputs "on the fly".

577. A very simplistic model for a remote analysis server is:

A. The agency makes available a microdata file for researchers which is held securely by the agency. The sensitive micro-data is typically not observable to the analyst.

B. An analyst submits a query, via the internet, to the agencies analysis server which is processed against the sensitive micro-data.

C. The statistical output (e.g. regression coefficients or tabulation) from the query is modified by a tailored confidentiality method specific to that analysis for the purpose of SDC.

D. The analysis server sends the modified output, via the internet to the analyst. Some outputs may be restricted on the basis that they could allow an analyst to reconstruct the attributes of an arbitrary record.

578. The next generation remote analysis server at the ABS comprises three applications – Census TableBuilder, ABS TableBuilder and ABS DataAnalyser. Each of these is described in the following paragraphs. At no time does the analyst see the underlying microdata.

579. For the 2006 Census of Population and Housing, the ABS jointly developed **Census TableBuilder** with Space-Time Research Pty Ltd**.** This is a web-based product that is available to analysts outside the ABS. Census TableBuilder incorporates a perturbation method (Fraser and Wooton (2005)) for automatically protecting tables of Census count data. This method was designed to mitigate disclosure risks from requests for similar tables, repeated requests for identical tables, and repeated requests for the same table cell within different tables. The method was intended to allow greater access to data for subpopulations, and to enable the development of web-based systems allowing users to define their own tables. All tabular output from the 2006 Census is protected using the same method, including tables created by ABS staff for publications. For this reason, there are internal systems for applying the method as well as the web-based Census TableBuilder for analysts.

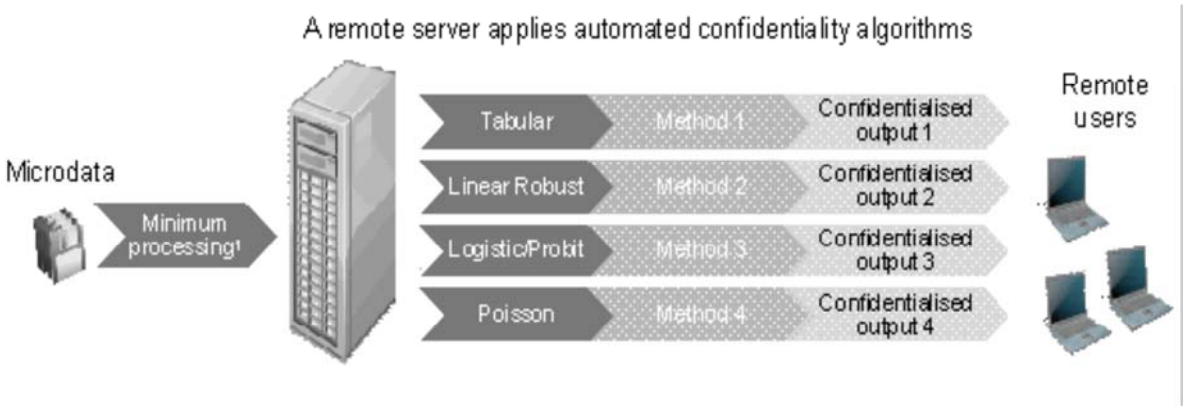
580. Expanding from this basis, the ABS has continued to develop a range of customised confidentiality algorithms and protections that run on the fly and minimise utility loss, while ensuring that the outputs produced are not likely to enable the identification of a respondent. Each confidentiality method is tailored to each specific output, reducing the level of confidentialisation required. In comparison, the existing CURF approaches require heavier confidentiality to protect against the full range of potential outputs that could be produced.

581. **ABS TableBuilder**, the successor of Census TableBuilder, also developed jointly with Space Time Research Pty Ltd, incorporates dynamic confidentiality routines for weighted survey data that have expanded beyond population counts to key summary statistics from magnitude data (such as custom ranges, totals, means, medians and quantiles). In addition to the perturbation, a number of in-built protections/restrictions have also been incorporated. These include restricting combinations of data items from being tabulated together, restricting the output of sparse tables where there are a large number of small cells, and requiring a minimum population size for the calculation of medians and quantiles. ABS TableBuilder also facilitates machine to machine querying of the microdata via an SDMX web service.

582. **ABS DataAnalyser** has been developed to facilitate exploratory data analysis and regression modelling. The ABS DataAnalyser is a secure menu based system for conducting statistical analysis through a remote user interface. The system enables users to remotely estimate the parameters of the statistical models fitted to ABS data while protecting the confidentiality of providers. All statistical outputs

that can be viewed by the user are automatically confidentialised using various disclosure control methods, including the perturbation of the estimating equation. This perturbation alone is not enough and a set of restrictions and protections against specific attacks have also been incorporated in the system, including confidentialised graphical displays to assist users to diagnose model goodness of fit. A summary of the perturbation approach and additional protections is described in Chipperfield, Gare and Yu (2011). The initial release of ABS DataAnalyser enables users to undertake data transformation and manipulation, tabulation, exploratory data analysis and Linear Robust, Logistic, Probit, Poisson or Multinomial modelling. The system is currently being trialled with users and a full production release is planned for 3rd quarter 2013. An illustration of ABS Data Analyser is provides below (Figure 1).

Figure 17.1. ABS DataAnalyser incorporating on the fly confidentiality



583. Advantages of the ABS approach to remote analysis servers include:

- the analysis is undertaken on the real microdata, retaining complex relationships in the data;
- statistical output is modified to a degree that is tailored specifically to the type of analysis being undertaken as well as the level to minimise information loss;
- once the server is set up, it can process multiple analyses in real time;
- all submitted programs can be logged and audited and if an attempt at disclosure is identified the analyst's access to the server can be revoked; and
- the point and click menu interface means that the user requires little training and does not have to learn a new software language.

584. Disadvantages of the ABS approach include:

- the analyst is restricted to use only analysis techniques, data transformations and manipulations supported by the server;
- analysis through the remote servers may take longer than if the micro-data were available on the analyst's personal computer; and
- the substantial investment of time and money required to develop confidentialisation software routines for each new analytical functionality.

International approaches to providing access to microdata

585. In terms of access to microdata, National Statistical Organisations have taken very different approaches, largely driven by the differing legislative requirements. A number of NSOs release public use files. These files are heavily confidentialised for general use.

586. NSOs also make extensive use of Research Data Centres, similar to the ABSDL, for the analysis of detailed microdata. Disadvantages of these approaches are that the outputs removed from these centres must generally be checked manually. This is resource intensive, leading to restrictions on the number of researchers that access can be provided too. The ABS seeks to provide a solution to facilitate as wide access as possible.

587. Some research centres are on-site while others are utilising technology advances to create a virtual research centre that can be accessed from dumb terminals installed in other organisations. The level of detail accessed in also varies greatly depending on the legislation of the NSO. In some cases authorised or "trusted" researches are provided with the same access to the detailed microdata as the employees of the NSO. The ABS legislation prevents this in Australia.

588. In comparison, the ABS's web-based systems are designed for external users to access remotely. The approach does not require extensive confidentialisation of the microdata prior to analysis, but utilises confidentiality methods that are applied in real-time. In recognition of the advantages of remote analysis servers that incorporate on the fly confidentiality, a number of NSOs have commenced research and development programs. Two fairly advanced developments worth noting are Morpheus (Höninger (2011)) developed by the State Statistical Institute Berlin-Brandenburg and the Microdata Analysis System (Lucero et. al (2011)) under development by the US Census Bureau.

Future directions and opportunities for international collaboration

589. The ABS remote analysis servers offer the advantage of providing high quality outputs derived from microdata files, as well as the convenience of access by users, without compromising the confidentiality of the data. Many challenges still remain.

590. Current research is focusing on assessing the existing approaches and their efficacy for linked datasets, which pose a higher disclosure risk. Future research work will focus on the development of on the fly confidentiality methods for the dissemination of business and longitudinal datasets via remote analysis servers and the provision of utility loss measures to researchers (Marley and Leaver 2011).

591. The ABS is also very interested in exploring the potential of synthetic microdata, even if only to allow analysts to test and review their models prior to running those on the real data held securely within the ABS DataAnalyser.

592. The ABS would be interested to work with the international statistical community to carry out research to address the challenges from these datasets. The ABS would also be willing to provide the statistical methods and algorithms behind those methods to the international statistical community if they want to incorporate these methods in their applications.

REFERENCES

Bleninger, P., Drechsler, J. and Ronning, G. 2011, 'Remote Data Access and the Risk of Disclosure from Linear Regression: An Empirical Study', *Privacy in Statistical Databases*, ***Springer.*** See http://www.idescat.cat/sort/sortspecial2011/DataPrivacy.1.bleninger-etal.pdf

Chipperfield, J., Gare, M. & Yu, F. 2011, '*Providing access to microdata for statistical purposes - experiences of the Australian Bureau of Statistics with Remote Analysis Servers'*, paper presented to the Statistics Canada 2011 Methodology Symposium, Ottawa, Canada, 1-4 November.

Fraser, B. & Wooton, J. 2005, 'A proposed method for confidentialising tabular output to protect against differencing', paper presented to the Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva, Switzerland, 9-11 November.

Gomatam, S., Karr, A. F., Reiter, J. P., & Sanil, A. P. 2005, 'Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk –Utility Framework for Remote Access Analysis Servers', *Statistical Science*, 20, pp.163-177.

Höninger, J. 2011, 'An Innovative Approach to Remote Data Access', 58th International Statistical Institute World Statistics Congress, Dublin, Ireland, 21-26 Aug 2011

Lucero, J., Zayatz, L., Singh, L., You, J., DePersio, M. and Freiman, M. 2011, 'The Current Stage of the Microdata Analysis System at the U.S. Census Bureau', 58th International Statistical Institute World Statistics Congress, Dublin, Ireland, 21-26 Aug 2011

Marley, J. & Leaver, V. 2011, 'A method for confidentialising user-defined tables: statistical properties and a risk-utility analysis', paper presented to the International Statistical Institute session, Dublin, Republic of Ireland, 22-26 August.

Shlomo, N. 2007, 'Statistical disclosure control methods for census frequency tables', International Statistical Review, 75, (2), 199-217.

Sparks, R., Carter, C. Donnelly, J., O'Keefe, C.M., Duncan, J., Keighley, T. and McAullay, D. (2008), 'Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-Preserving AnalyticsTM', *Computer Methods and Programs in Biomedicine* 91, pp. 208-222.

Willenborg. L. and de Waal, T. 2001, *'Elements of Disclosure Control'*, *Lecture Notes in Statistics*, Vol 155, ISBN 978-0-387-95121-8, Springer.