CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

1-3 December 2021, Poland

Microdata access where we are and where we need to go.

Elizabeth Green (University of the West of England)

elizabeth7.green@uwe.ac.uk

Abstract

In the summer of 2021, the University of the West of England in collaboration with UN Economic Commission for Europe, Eurostat, INXEDA and statistical organisations across the world will host a workshop to review lessons learnt over the past 15 years. The workshop will focus on the following themes surrounding microdata access: technology, statistical disclosure control, organisation management and societal context, the aim is within each theme consider; best practices, solutions for sustainability, affordability, international data sharing, and LMICs. This proposed paper will present the findings from this workshop and discuss proposed solutions particularly in terms of overcoming practical difficulties in defining data access strategies and systems.

.

Microdata access: where we are? Where we need to go?

Elizabeth Green[*] and Felix Ritchie[**]

[*] The University of the West of England, elizabeth7.green@uwe.ac.uk

[**] The University of the West of England, felix.ritchie@uwe.ac.uk

Abstract: In the summer of 2021, the University of the West of England, in collaboration with UN Economic Commission for Europe, Eurostat, INXEDA and statistical organisations across the world, hosted a workshop to review lessons learnt in microdata access over the past 15 years, and future directions. The workshop themes were technology, statistical disclosure control, organisation management and societal context, and within each theme the conference considered best practices, solutions for sustainability, affordability, international data sharing, and LMICs.

This paper presents the findings from this workshop and discuss proposed solutions particularly in terms of overcoming practical difficulties in defining data access strategies and systems.

1 Introduction

In 2006 the UN Economic Commission for Europe/Conference of European Statisticians set up a task force on access to confidential microdata**.** The findings were published in 2007 as 'Managing statistical confidentiality and microdata access', sometimes known as the 'Trewin report' after the chair(UNECE, 2007). This report reviewed microdata access practices by national statistical institutes (NSIs) across countries, presented country case studies, and provided a set of guidelines for NSIs to adopt. The aim of the guidelines was to create greater uniformity of confidentiality approaches by countries and improve statistical confidentiality processes within home countries. The guidelines were formulated as principles and acknowledged that precise arrangements for access to microdata varies from country to country. The report was an important step forward in highlighting the need for organisational transformation and the need for NSIs to move away from risk avoidance to risk management. An overarching theme from the report was the necessity to acknowledge the transference from confidentiality being perceived as a national issue to that of a global one.

Since 2006, the data landscape has changed considerably (Ritchie, 2021). The conference pre-reading (Green and Ritchie, 2021) highlighted, in particular:

- Technological delivery
- New data sources.
- Legal changes
- Statistical change
- Standards and principles
- The absence of perspectives from low and middle-income countries (LMICs)

To address the radical changes in landscape since 2006, the DRAGoN team at the University of the West of England team ran a virtual 5-day expert workshop on 'The present and future of microdata access'. Supported by the UN, Eurostat, central banks, and statistical agencies, the ambition of the workshop was to help to shape the next decade of confidential use across countries and organisations.

The primary aim of the workshop was to review lessons learned over the past 15 years, particularly in terms of overcoming practical difficulties in defining data access strategies and systems. The second aim was to examine and identify current good practice guidelines, reflecting both the range of access methods and the experiences/needs in different countries. Finally, the workshop also aimed to identify future opportunities for microdata access/use and potential risks to confidentiality. Throughout the workshop sessions we asked these two questions:

- What do we currently know and how do we share it?
- Where should we be looking ahead?

The 5-day workshop attracted 130 attendees across 88 different organisations from 26 different countries. Each of the four main topics was divided into three subtopics, with a summary on the final day. The structure of this paper mirrors the format of the workshop: a brief overview into the subtopic, a discussion of best practices, and areas for future developments are outlined along with a list of recommended actions detailed at the end of each section. The final section provides an overview of cross-cutting themes along with projections for the future of micro data access.

This is a very brief summary of a much longer conference report. This will be available, when completed, on the DRAGoN website www.uwedragon.org.

2 Microdata access technology

2.1 Research data centres

The discussions from this session predominately focused on the central role of RDCs in bringing together points of legislation, technological advances, research insights, data communications, and data security. The wide range of RDC activities alongside different contextual demands results in variations on how an RDC might be organized, level of access provided, method of access (virtual vs physical) etc. Variations were discussed surrounding the provision (and detail) of user training, forms of punishments, and use of financial charges to access data.

As digital solutions often develop faster than the societal and legal frameworks can adapt we begin to see new concerns. One particular issue raised was the move towards cloud computing and the role of security and understanding geographical locations of servers. There was also a concern surrounding the demand for data, volume of data available and the rapid increase in demand for remote access and also international

collaborations. Participants voiced concerns surrounding the current technical solutions which struggle to meet demands and needs (cloud solutions included). Concerns surrounding the viability of international collaborations was discussed; the main point of contention was the legal ambiguity between the different countries, but also a fear of data colonization and loss of autonomy surrounding the data use and application. Logistically it was also noted that RDCs were costly and required not only the technological set up but also the back-office infrastructure, which caused participants to reflect whether RDCs are viable for LMICs.

Increase in RDC use by users can also place pressure on organizations which check requested outputs prior to release from the secure environment, especially if the processed is not optimised (Alves and Ritchie, 2020). As such the time between the user accessing the microdata, conducting analysis, requesting output, and output being released has in the majority of places increased resulting in user frustration. The need for user expectation management was apparent. Reflection suggested the effectiveness of sanctions in modifying user behaviour was limited. There was a difference between those who simply trained users to use the facility, and those who actively promoted engagement and behaviour change; the latter was seen to good practice. There was a preference towards actively training users rather than reactively enforcing rules.

Key messages: there are concerns for long term sustainability; operational efficiency needs to be considered as an important design element; community based training is best practice and can help operations; less relevant for LMICs as RDCs typically rely on a substantial cultural hinterland for implicit support.

2.2 Remote job servers and table servers

Participants discussed the benefits of utilizing table builders via remote job servers as a form of managing disclosure control risk. Table builders are mainly automated and as such require hard rules surrounding thresholds and minimum number of observations. As outputs are not manually checked, administrators routinely check user activity/ released outputs to vet good and bad practice. Due to the reliance on remote job servers to have automated disclosure control integrated into the table builders the level of detail in the provided data is restricted.

Participants noted future concerns surrounding output attacks, particularly in the rise of sophisticated machine learning techniques and reverse engineering of data sets. There were also concerns about what data and outputs already exist in the public domain, and the lack of solutions for managing secondary disclosure.

Participants felt that the future area for remote job servers was the implementation of synthetic data sets. Researchers could test and develop code based on synthetic data (which holds the same properties as the secure data). The tested code can then be executed in the secure environment and outputs checked and released.

Key messages: becoming less of an outlier, with more practical examples to illustrate different choices made; can be very efficient; but always likely to be a second-choice preference compared to RDC

2.3 Other technology solutions to data access

This workshop session considered Privacy enhancing technologies (PETs) supporting the analytical use of data (CDEI, 2021). These use advanced computational techniques or hardware to allow the derivation of useful insights from data without requiring full data access (and the concomitant security risks and legal/ethical restrictions). PETs can therefore be seen as processing mechanisms, rather than the traditional cybersecurity embodied in RDCs and RJSs; hence, these traditional data management solutions are not usually included as PETs. PETs highlighted in the pre-reading and discussion included homomorphic encryption, trusted execution environments, secure multi-party computation, differential privacy, and personal data stores.

Users only see their own data which creates a sense of security, but there are few options for matching or exploring the database. It was noted that PETs (with the exception of DP) only resolve relatively simple linearizable problems, and suffer from the need for computational power. It was clear that the work is technically driven ie 'can tech do this?' rather than 'is tech the best way to do this?', but acknowledged that this is a necessary perspective to develop novel technologies.

Most importantly, these technologies including DP, do not provide strong protection against output-based attacks. There was concern that the overarching premise of guarantying security can result in overconfidence and less scrutiny of outputs.

One area these technologies could be valuable is for international sharing, but the technologies will need to be cost-effective; therefore, they may not be suitable for LMICs. Time and caution must be applied when implementing PETs: buy-in and user value identification is important alongside demonstrating value over other solutions.

Key messages: not really relevant for analytical use, perhaps more operational; still too experimental to be considered as a core option

3 Statistical Disclosure Control

3.1 Input SDC

Input statistical disclosure control (SDC) concerns the reduction in detail in a dataset so that the risks of a contributor to the dataset being identified are reduced to an acceptable level.

Discussions surrounding input SDC focused predominantly on the rise and development of machine learning and artificial intelligence techniques, These may prove a significant challenge, for example through reverse engineering of protective

transformations) to traditional anonymisation, which is highly labour- intensive. However, there are possible benefits too: ML may be a way to develop future risk profiles, and may help understand the transformations applied to historic data sets.

There was felt to be a need for better metadata and data standards universally, and the full application of the FAIR data standards. Finally, one area for consideration was whether some of the principles-based approaches developed for output SDC could be applicable here.

Key messages: input SDC is going to come under increased pressure from an arms race against computing power, AI and alternative databases; but AI might prove a way of de-identifying and/or assessing risk.

3.2 Output SDC

Output SDC (OSDC) refers to the application of SDC methods to potential publications after the analysis has been carried out, to guard against residual disclosure. Output checking is carried out manually, so operational considerations (rules-based, principles-based or ad hoc) affect the disclosure rules (Alves and Ritchie, 2020). Statistical organisations are generally rules-based when it comes to producing official statistics, but there is more variation for analytical outputs as rules-based is generally too constricting for research use; remote job servers are an exception.

The discussions focused on the need for researchers to accept co-responsibility for output checking: the better the quality of the requested output more efficient the system can become. However there was no clear consensus on how this could be achieved: it relies on trust and rapport between users and the output checkers, but only some data holders training their users on output SDC. It was noted that training not only assists the output checkers but can help inform the output checkers of new forms of analysis. With Covid-19, the pivot to online training was felt to be acceptable.

With the rise in microdata access and similarities in service structure and provision some organisations are considering allowing users to access services if they have completed safe data training elsewhere (which was of a similar standard). Another potential area for review was differential privacy with both the practical aspect and perceived level of utility considered in the conversations.

Key messages: organisations are concerned about the sustainability of a resource-intensive process; best practice suggests that researcher training should include OSDC, but not all organisations do training; there is still a very wide view on what can be expected from researchers.

3.3 Synthetic data

Synthetic data is data created to replicate the structure of genuine data but without the confidentiality risk. Synthetic data can be partially or fully synthetic. Synthetic data does not mean that it is risk-free, and there may still be some need to protect the data.

Delegates felt that, for datasets with few variables with limited possible values, synthetic data was a quick, cheap, easy solution to develop standard libraries. The most popular tools to automate the process are the R packages SynthPop, SimPop and RDV. These reflect consensus on 'basic' models of synthetic data.

The future involved development of specialist knowledge and research eg GAMS to allow synthetic datasets to become representative of real-world problems/ data; but additional complexity (eg more variables, or relationships such as family structures), creates difficulty in maintaining nuances without compromising the utility. Presently there are computational challenges as to whether this can be presently achieved.

Key messages: This is an area with great potential, if presently using basic datasets. Further work is required to address computation challenges and complex data sets.

4 Organisation

4.1 Training

Training of users is seen to achieve two objectives. First, training can improve the actual and perceived security of the facility. Second, training can be used to encourage positive behaviours which improve the efficiency of the operations. For example, the user training and guidance can have a significant impact on the efficiency of output checking procedures. Delegates discussed need, content, development and delivery. Not all organisations do training: some feel they are lacking expertise whilst others say it is not their remit.

Covid-19 forced organisations carrying out face-to-face training to redesign their models. The evidence on this is not clear yet, but in the UK a franchise structure suggested that online interactive training can be as effective as face-to-face training. This has implications for countries which struggled with geographical distances.

Delegates felt that good practice meant integrating a community building approach within the training course. This should include understanding role in the data community, understanding procedures and expectations. This helps researchers become cognisant of the wider picture and the potential impact of a data leak.

Challenges for training includes the transferability of examples for LMICs, particularly around assumption on the resource hinterland eg paper records and access arrangements. Further off, delegates raised the need for training in AI and ML models.

Key messages: Covid caused a change in training to virtual delivery, but experience suggests this can be as effective as face-to-face. The community based approach was seen as an integral and good practice element of training. Course materials needs to be appropriate to the organisation (as such LMICs materials might require tweaking). Training in AI and ML models will become the next challenge.

4.2 Access arrangements

The Five Safes is the basis for structure across a lot of places; broadly, this is seen as good practice, but safe person, in particular, is likely to be country/culture-specific. There is an awareness that data access should be seen as a management process, with applications and protocols viewed through an operational lens. This is particularly relevant as delegates expressed concern about sustainability (resources vs demand).

Covid19 has shifted a lot of base assumptions: will we return to the old ways?

ML and growths in computing pose extrinsic risk to distributed data – but does that mean we should only have public and secure-use files in future, no scientific-use ones? For HICs there was a feeling that this was the case, but LMICs ere clear that distributed data is likely to play a role in data dissemination for a long time.

LMIC delegates also noted that fear/lack of understanding can be a significant block to data access, due to a lack of cultural infrastructure on data governance.

Key messages: there are concerns for long term sustainability and operational efficiency. ML and advancing in computing poses a risk for scientific use files. LMICs need further development in capacity and infrastructure to help support provisions for a secure-use services and so are likely to rely on distributed data for the medium term.

4.3 FAIR, metadata, and sustainable management

Delegates were clear that metadata should be publicly available, that DOI and annodata schema should be used, and that metadata have to be comparable. The DDI standard, although popular, was seen as being hard to reach, with secondary data often poorly documented. There was a trade-off between availability and meeting standards.

A number of delegates associated with consortia (CESSDA, INEXDA, go FAIR) highlighted the potential value to be gained from exchanging experiences. This could include sharing knowledge of useful tools such as the World Bank's metadata editor. The world does not necessarily need more metadata tools, but better use of them.

Metadata can be vital for making sure that data is used, and should be part of any good practice dissemination programme. Funding contracts can be used to ensure that documentation is carried out. Engagement/training of researchers is also vital, to ensure that the metadata are useful (by using, reading, giving feedback).

Metadata around organisations and access (annodata) is important to encourage use. A list of repositories would be a useful start, and/or easy software to find them. An international MicroData Standard would be ideal, but hard to reach.

Key messages: Metadata should be widely available and well documented. The world does not need more tools, just the capacity to use current ones. Metadata on access mechanisms would be helpful. An international MicroData Standard would be optimal but seems like a pipe dream.

5 Societal context

5.1 Regulatory regimes

There are two types of regulatory regimes: rules-based and principles-based. Rules-based relies on specification, whilst principles-based focuses on alignment with goals. Rules-based compliance is managed by verification, principles-based via accreditation.

The UK Digital Economy Act 2017 is an example of a principles-based approach with the Office for Statistics Regulation having the authority to identify and approve accreditation processes. Training models are more likely to focus on a principles-based approach with a focus on one's role in the community. There are concerns about the scalability of a principles-based approach. The ICPSR is presently developing a 'passport' personal accreditation model. A universal agreement on terminology and definitions would be beneficial.

Delegates discussed the need to consider data colonization: if a passport model is developed, what implications does this have for eg indigenous datasets and LMIC data. We need to prioritize that these datasets are still held and controlled closely with the data owners and autonomy of data use remains with them.

Key messages: Universally agreed and consistently used terminology would be beneficial. Although principles based was seen as the preferred choice, there were concerns surrounding the sustainability and feasibility of this. Streamlining data access via a researcher passport could help cross-organisational studies. There are concerns about enforcing a single cultural model on indigenous and LMIC data use.

5.2 Public engagement

‘Stories not statistics’ matter; we need to focus on becoming trustworthy rather than assuming trust. Gaining trust can be complex; stories can be manipulated to created more problems (for example black-and-white arguments over whether opt-in or opt-out was best, or the claims of differential privacy supporters that it ‘guarantees privacy’). Often the wide array of opinions and information can make it difficult for public to gain trust in data use. It was noted that that focus groups tend to be supportive of data sharing than questionnaire respondents, and it was thought that this is due to the former being given more information and the chance to ask questions.

The issue of trust resonates with many marginalised groups who might have more fear about how their data is being used operationally (eg minority groups, indigenous peoples); this is sometimes known as the SCARE principle. Delegates considered who makes decisions about access to data: does public engagement include politicians? Can ethical groups act in the name of the public? And what about LMIC data collected by HICs, a common research pattern?

Finally, it was noted that explainable AI will be a substantial challenge for the future.

Key messages: Trust is complex and contextually sensitive to individual populations. How we communicate and engage with the public can be a double-edged sword.

5.3 Ethics/ benefits and costs

There was a consensus on best practices – but compliance is very low! Often the focus is on the public good/public attitude, but delegates asked whether there should be a greater role for institutions? Practices may not be sustainable as ethics is seen as something you 'do' at the project start – as time/money runs out, other priorities than checking against ethical standards may come to the fore.

In HICs, ethical approval processes are well established but not standard. Good practice says that the data subjects are the owners of the data, and data use should be for the common good. Data should be minimised to that necessary for the research or analysis (although it was recognised that different rules need to apply to archives and data repositories. Often, we hold LMICs to the same standard as developed countries, but they may be unable to meet these standards due to lack of resources or infrastructure.

It was also recognised that modern ethical review is much more about risk management, compared to the older risk avoidance strategies. Increasingly the argument is about the ethics of *not* using data i.e. a stronger awareness of the benefits missed by refusing to support access. Covid19 presented an excellent example, relevant to everyone, of how sharing data demonstrably saved lives.

Key messages: Focus on managing risk. Ethics needs to be considered throughout the project's lifecycle. Ethics assessments need to be tailored to the context and country- a one-fits-all size will not work especially in the context of LMICs.

6 Cross-cutting themes

In addition to the specific messages, a number of cross-cutting themes arose.

6.1 Goodbye scientific use files; hello synthetic data?

With the development of sophisticated synthetic data technology, one prediction from the workshop was that scientific use files will eventually become obsolete with synthetic data sets taking their place. Whilst this finding may be welcoming to many data providers, it comes with its own set of disadvantages. Concerns surrounding the loss of detail as synthetic data may wash out findings and trends in smaller populations, resulting in uneasiness that data sets could become ethnocentric focusing on white populations and losing details for marginalized populations. The use of synthetic data for decision and policymaking should be approached with caution. Synthetic data could provide an opportunity for researchers to test, develop and finalize code before sending it to a secure environment to be executed. As a tool for training and top-level insights, there is enormous potential for synthetic data sets.

6.2 Co-creation of community governance models

Successful data governance models are developed in tandem with public and data users. This is driven by the recognition of the need for better public engagement and public understanding of how microdata is being accessed and used, as well as tailoring process to user needs (as users ignoring rules is a key risk). A recent example from the UK was the resurgence of concern from the public about the use of GP records, causing a need to refocus on how the public engage with data governance. The perception of public data becoming a commodity has raised concerns at both a practical and a theoretical level. The issue of data colonization and the need to protect against exploitation is one of concern and the issue of indigenous data governance and sovereignty. HIC/LMIC co-development of training materials for data governance shows that the community model has high transferability, and supports developing capacity and ensuring microdata is retained in the country of origin.

6.3 Rise of the machines

With a lack of resources, we are encroaching towards a situation in which technological advances outpace current knowledge and practice of disclosure control. With the rise in machine learning, reverse engineering, and AI models there will be new concerns for output attacks and training in these models. However, these may also present an opportunity to assess risk better by mimicking real-life attacks.

6.4 Sustaining momentum

Covid has acted as a catalyst for action (overriding the typical defensive stance), and advances to data access practices can be attributed to this; but how do we maintain momentum in normal times? With previous natural disasters, an influx of action and transformation can be seen, partly due to the necessity to meet demand and partly due to extra resource provisions made available. Once normality begins to resume do the old processes and behaviours resume as well? NZ is perhaps the counter-example. The next challenge we face is continuing to use new practices, a difficult balancing act due to resourcing. The pandemic has forged new ways of working both nationally and internationally and maintaining momentum is essential for the future of data access.

7 Future actions

There was recognition that meeting and sharing ideas is in itself a good thing, and something data community good perhaps develop further. Several specific steps were suggested to ensure that the advance of data governance is well-founded and builds on good practice/consensus:

- One or more networks to share info and good practices on data management, ethics, and training

- A centralised group/ website to share info? Wiki/ linkedin/ launch event to help takeup?
- A support network/ mentors to help countries develop training, governance models etc, particularly for LMICs
- A webinar/lecture series on core concepts ("what is an RDC?" etc)
- More software solutions to help manage metadata and dataflow process, and a mechanism for sharing experience/advice
- A process for technical workers/coders/developers to discuss/share

We invite UNECE delegates to suggest ways/organisations to take these forward.

References

Alves, K., & Ritchie, F. (2020). Runners, repeaters, strangers and aliens: Operationalising efficient output disclosure control. Statistical Journal of the IAOS, 36(4), 1281-1293. https://doi.org/10.3233/SJI-200661

CDEI (2021). Privacy enhancing technologies for trustworthy use of data. https://cdei.blog.gov.uk/2021/02/09/privacy-enhancing-technologies-for-trustworthy-use-of-data

Green E. and Ritchie F. (2021) The future of microdata access; pre-workshop briefing. https://uwe.worktribe.com/record.jx?recordid=7918878

Ritchie, F. (2021). Microdata access and privacy: What have we learned over twenty years? *Journal of Privacy and Confidentiality*, 11(1), 1-8. https://doi.org/10.29012/jpc.766

UNECE (2007) *Managing Statistical Confidentiality & Microdata Access; Principles and Guidelines of Good Practice.* United Nations, Geneva. ISBN 13: 987-92-1-116959-1. https://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf