



# The present and future of confidential microdata access

## Post-workshop report

**DRAGoN**

The Data Research, Access, and  
Governance **Network**

---

## Contents

Introduction .....	4
Section 1: Technology .....	7
Subtopic 1.1: Research data centres .....	7
Subtopic 1.2: Remote job servers and table servers.....	9
Subtopic 1.3: Other technology solutions to data access .....	11
Session 2: Statistical Disclosure Control .....	14
Subtopic 2.1: Input SDC .....	14
Subtopic 2.2: Output SDC .....	16
Subtopic 2.3: Synthetic data .....	19
Session 3: Organisation .....	22
Subtopic 3.1: Training.....	22
Subtopic 3.2: Access arrangements.....	24
Subtopic 3.3: FAIR, metadata, and sustainable management.....	27
Session 4: Societal context .....	31
Subtopic 4.1: Regulatory regimes .....	31
Subtopic 4.2: Public engagement.....	32
Subtopic 4.3: Ethics/ benefits and costs .....	34
Overarching findings.....	37
Goodbye scientific use files; hello synthetic data?.....	37
Co-creation of community governance models .....	37
Rise of the machines .....	38
Sustaining momentum .....	38
Road map for the future .....	39
Glossary .....	40

---

## Acknowledgements

We would like to thank the Advisory Board for their time, expertise and support for the workshop and subsequent report, for suggesting the pre-workshop briefing on which this report is based, and for their assistance with. The Advisory Board comprised

- Stefan Bender, Deutsche Bundesbank and INEXDA
- Aleksandra Bujnowska, Eurostat
- Taeke Gjalterna, UN Economic Commission for Europe
- Adam Harris, Australian Bureau of Statistics
- Tina Hotton, Statistics Canada
- Wim Kloek, Eurostat
- Steve McEachern, Australian Data Archive
- Eric Schulte Nordholt, Statistics Netherlands
- Pete Stokes, UK Office for National Statistics
- Steven Thomas, Statistics Canada
- Lynn Woolfrey, DataFirst South Africa

The report was compiled by Elizabeth Green and Felix Ritchie, with support from UWE students William Ashford and Pedro Ferrer Breda. Francesco Tava, of UWE Philosophy Department, wrote the introductory ethics section and reviewed the comments. Azeem Haroun and Juan Carlos Mondragon Quintana from the DRAGoN team provided additional support during the event.

All errors of omission, commission and interpretation remain ours.

For further information, contact: [elizabeth7.green@uwe.ac.uk](mailto:elizabeth7.green@uwe.ac.uk) or [dragon@uwe.ac.uk](mailto:dragon@uwe.ac.uk).

Website: [www.uwedragon.org](http://www.uwedragon.org)

---

# Introduction

In 2006 the UN Economic Commission for Europe/Conference of European Statisticians set up a task force on microdata access. The findings were published in 2007 as 'Managing statistical confidentiality and microdata access'<sup>1</sup>, sometimes known as the 'Trewin report' after the chair of the Task Force. 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 report aimed 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 vary 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<sup>2</sup>:

- *Technological delivery.* In 2006, systems such as the UK, Danish and Dutch RDCs were outliers, rather than the mainstream offering of a large number of NSIs. Other key technical developments have been in the development of sophisticated table and query servers, and the visualisation of data. Finally, the training of AI models provides challenges for the ethics of microdata use.
- *New data sources.* Few of the case studies in 2006 considered the growth in administrative or operational data, the demand to link data sources, and the pressure this would put on NSIs for both the construction of these datasets and to manage effectively the increased security risks.

---

<sup>1</sup> 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](https://www.unece.org/fileadmin/DAM/stats/publications/Managing_statistical_confidentiality_and_microdata_access.pdf)

<sup>2</sup> 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>

- 
- *Legal changes.* Regulation in 2006 was fundamentally data-centred. Simple distinctions were made between anonymous and personal data, with a high bar for non-data interventions to protect confidentiality. In contrast, modern regulation such as the GDPR recognises the spectrum of interventions available to NSI to ensure that ‘safe use’ is the outcome, rather than ‘safe data’.
  - *Statistical change.* Two major statistical developments of the last ten years have been synthetic data and differential privacy. Both existed in 2006 but were niche interests at the time; now most NSIs have considered whether they can be used to improve the range of microdata options. On the negative side, there are concerns over the security of distributed data, arising from increased computing power and the greater availability of corroborative data. Finally, techniques such as homomorphic encryption suggest new possibilities for secure distributed analysis
  - *Standards and principles.* The Trewin report asked case study authors to use a common template when responding, but this highlighted the variation in practice across countries. Since 2006, there have been substantial advances in the way data access discussions are framed, in the way risks and evidence are assessed, and in the way principles of use are understood.
  - *Low and middle-income countries (LMICs).* LMICs are notable absentees from the original report. Much of the research and practice data access has evolved in high-income countries (HICs). While NSIs in LMICs face some similar challenges in developing their data strategies, very little is known about how accepted good practice in HICs can effectively translate into LMIC practice.

To address the need to review and reflect on past and current practices, the DRAGoN team at the University of the West of England team ran a 5-day virtual workshop in July 2021 on ‘The present and future of microdata access’. With attendees from international organisations, central banks, statistical agencies, and academia, the expert workshop was intended to help to provide a potential roadmap for the development of confidential use across countries and organisations for the next decade.

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?

Each session consisted of different subtopics; for each subtopic, a brief overview by the chair was followed by a facilitated discussion considering, ideally,

- Is there a consensus on good practice?
- What lessons have we learnt (i.e. things not to do)? In particular, what did we learned from having to respond to Covid-19
- Are current practices sustainable (what happens if demand increases) and affordable?
- What are the lessons for implementation in LMICs?
- What are the lessons for international data sharing?
- What are the other main challenges for the next 10 years?

In practice, the nature of the discussion and the expertise of the attendees meant that the discussion often took directions that the attendees felt important. Where time allowed, recommended actions were developed at the end of each section. The sessions then reconvened in plenary to review subtopics and identify any cross-cutting themes. On the final day, all topics were reviewed and the attendees asked to identify next steps.

The structure of this report mirrors the format of the workshop. Each topic and subtopic is presented with the pre-conference briefing notes, followed by a summary of the discussion and recommendations arising. We finish with an overview of cross-cutting themes along with projections for the future of microdata access.

---

# Section 1: Technology

## Subtopic 1.1: Research data centres

At the turn of the century, researchers were often allowed to visit ‘Research Data Centres’, (RDCs) which were sites located in the data holders’ offices (or controlled and monitored by them). In return for physical travel and isolation in a secure environment, researchers were given access to the most detailed microdata. Many statistical organisations set up such facilities, often in partnership with academia, as in Canada. In 2002 the Danish NSI set up the first virtual RDC (vRDC), offering the same facilities like an on-site system but accessible from the desktop of researchers across Denmark. This was quickly taken up across much of Europe with the UK, the Netherlands, and Sweden being early adopters, followed by Italy, Finland, Slovenia, NORC (US), and Mexico in the third wave.

Most European NSIs, as well as the US, Canada, Mexico and Japan, now have or are planning some form of vRDC system. Most facilities use Windows-based ‘thin client’ software; the US Census Bureau uses Unix to achieve the same end. Most European government-run vRDCs cite the Danish/Dutch system as the model for their technology. The ‘virtual’ in vRDC relates to the way it can be used, not the way it is used in practice. For example, the ONS vRDC allowed access to users across the UK government network from 2004 but not beyond; only in 2019 were academics allowed to access it through their university systems. Some other vRDCs allow access across the university network, as does the Danish system. The Dutch and French systems are unrestricted, although eligibility requirements for international access, for example, are stringent.

The growth of vRDCs has led to several changes in perspective. First, the level of control has allowed the detail of data to be increased. Second, the efficiency gains from principles-based output SDC (see below) have encouraged training that emphasises engagement. Third, engagement is also encouraged because of the high cost of having untrustworthy users in the vRDC. Fourth, more control has increased confidence in allowing non-typical users (for example, private sector organisations) to get access to data. Finally, the unrestricted nature of the research carried out in a vRDC has led to the new field of output-based SDC targeted specifically at research.

---

## Workshop findings

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 the 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 resulting concerns around security and the geographical locations of servers. There was also a concern surrounding the demand for and volume of data available, and the rapid increase in demand for both remote access and 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 were 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 (cultural as well as technological), which caused participants to reflect whether RDCs are viable for LMICs.

An increase in RDC use by users can also place pressure on organizations that check requested outputs before release from the secure environment, especially if the process is not optimised<sup>3</sup>. 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.

---

<sup>3</sup> 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>



---

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 be 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; this solution is less relevant for LMICs as RDCs typically rely on a substantial cultural hinterland for implicit support.

## **Subtopic 1.2: Remote job servers and table servers**

Remote tabulation tools are designed to allow users to create their own tabulations of the data, rather than relying on the data owner's choice of tables or bespoke tabulations. As well as data tables, homologous tools can produce geographical images or time series. What distinguishes these tools is that the output statistic is a simple linear value (sum, mean, index) broken down by categories under the control of the user.

The value to the user of remote tabulation is twofold. First, it allows the user to have data presented in a useful form to his or her demands without needing to manipulate microdata. Second, it allows the data owner to present results from data that may be confidential and not suitable for release as microdata, but which nevertheless can produce secure tabulations. Confidentiality is ensured in one of two basic ways; restricting the input or restricting the output. To restrict the input, one option is to apply standard anonymisation techniques to the underlying microdata before analysis, so that the data is near enough non-confidential and can be safely tabulated without restriction. A slightly more flexible approach is to identify all acceptable combinations of key variables (allowing for differencing between possible tables) and then only allow any analysis on the confidential data which uses an acceptable combination, the 'hypercube' method. Recent developments in SDC such as cell-key encryption (see below) have increased the flexibility available to NSIs.

Remote job servers extend these principles of analysis at arm's length to allow users to run code and generate more complex statistical results. NSIs in the Netherlands, Canada, Australia and Norway offer such options, as do other organisations such as OpenSafely in the

---

UK (remote analysis on health records). Some data holders do not offer formal RJSs but do offer a service of uploading and running code on confidential data and returning the results. The longest-running such service is Lissy (Luxembourg Income Study), which has been offering remote analysis to researchers for some twenty years<sup>4</sup>.

Having all outputs generated by known code can make output checking easier compared to RDCs. Some RJSs ban all commands except those explicitly allowed, others allow all commands except those explicitly banned.

### **Workshop findings**

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 (unambiguous and strictly enforced) surrounding thresholds and the minimum number of observations. As outputs are not manually checked, administrators routinely check user activity/ released outputs to vet good and bad practices. Due to the need for 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 secure data). The tested code can then be executed in the secure environment and outputs checked and released.

There was also discussion about the need for streamlined definitions surrounding synthetic data and also anonymisation. There were conflicting views on whether data was anonymised

---

<sup>4</sup> <https://www.lisdatacenter.org/data-access/lissy/>

---

vs de-identified: if data is truly anonymized then there should be no risk to disclosure, but if the data is de-identified then disclosure is a potential risk.

It was noted that where organisations provided remote access to both RDCs and RJSs, there was a clear preference for RDCs amongst researchers. However, these do seem to serve slightly different markets.

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

### **Subtopic 1.3: Other technology solutions to data access**

Privacy-enhancing technologies (PET) is a term used to describe any technical method that protects the privacy of personal or sensitive information. Technologies considered under this definition range from the relatively simple, such as ad-blocking browser extensions, to complex encryption technologies used to secure communications.<sup>5</sup>

For this workshop session, we focus on PETs supporting the analytical use of data<sup>6</sup>. 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 or 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 currently under discussion or in development include

- *Homomorphic encryption* allows certain computations on encrypted data, generating an encrypted result which, when decrypted, matches the result of the same operations

---

<sup>5</sup> <https://cdei.blog.gov.uk/2021/02/09/privacy-enhancing-technologies-for-trustworthy-use-of-data/>

<sup>6</sup> This section is largely based on the Royal Society report <https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/privacy-enhancing-technologies-report.pdf> which discusses all the PETs in detail

---

performed before encryption (hence 'homomorphic'). Homomorphic encryption is extremely computationally expensive and has only so far been implemented in 'partial' forms.<sup>7</sup>

- *Trusted Execution Environment (TEE)*. A TEE is a secure area inside a processor where computations on sensitive data take place. As with other existing cryptographic technology, protecting secure keys in TEEs remains a challenge. It is necessary for particular to protect the system that generates secure crypto functions; that is, TEEs are secure, but the procedures that generate the TEEs might not be, especially on the cloud which is a shared environment.
- *Secure multi-party computation (MPC)* allows computation or analysis on combined data without the different parties revealing their private input. It may be used when two or more parties want to carry out analyses on their combined data but, for legal or other reasons, they cannot share data. Using MPC, parties send encrypted messages to each other and obtain the computation they want without revealing their input, and without the need for a trusted central authority. For example, in adding values across multiple data sources, each data holder adds the true value plus a large amount of noise; when the (noisy) total has been calculated, the noise can be removed safely. MPC can be a slow process, due in part to delays in communicating. Whilst secure multi-party computation has been applied in a limited number of 'products, research and development are ongoing and other applications are at a 'proof of concept stage.
- *Differential privacy (DP)*. DP is a method of systematically adding noise to any statistical outputs so that, when a result is released, it should not give much more information about a particular individual than if that individual had not been included in the dataset. DP mechanisms are designed to reduce the risk of revealing whether a specific individual or organisation is present in a dataset or output. DP can lead to a loss of utility from the statistics, especially on small datasets, rare events, or when dealing with highly skewed data)<sup>8</sup>. As a noise-addition method, it is susceptible to the standard attacks of repeatedly interrogating the data to remove 'average' noise. Nevertheless, it has become highly popular, particular with private sector management companies offering off-the-shelf DP packages.
- *Personal Data Stores (PDS)*<sup>9</sup> are systems that provide individuals with access and control over data about them, so that they can decide what information they want to share and with

---

<sup>7</sup> <https://eprint.iacr.org/2011/277.pdf>

<sup>8</sup> Bambauer J., Muralidhar K., & Sarathy R. (2013) Fool's gold: an illustrated critique of differential privacy. Vanderbilt J. Ent. & Tech. Law

<sup>9</sup> See <https://www.gov.uk/government/news/the-midata-vision-of-consumer-empowerment> for an example of PDS

---

whom. These systems are consumer-facing and aim to enable people to have more control over data. PDS enable a distributed system, where the data is stored and processed at the 'edge' of the system, rather than centralised. It is possible, for instance, to send machine learning algorithms to the data, rather than the data to the algorithms.

### **Workshop findings**

Workshop attendees noted that these were interesting developments but of very limited value for research use. Users only see their data which creates a sense of security, but there are few options for matching or exploring the database. It was noted that PETs (except DP) only resolve relatively simple linearizable problems, and suffer from the need for computational power. It was clear that the work is technically driven i.e. '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.

The discussion highlighted one crucial point: these technologies do not provide strong protection against output-based attacks (this includes DP because of its susceptibility to repeat attacks and the consequent need for it to operate in a restricted-query environment). 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:** these are not relevant at present for analytical use, perhaps more operational; still too experimental to be considered as a core option. Before they can be deployed in live environments, the output risk needs to be addressed.

---

# Session 2: Statistical Disclosure Control

## Subtopic 2.1: 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. This is a very large body of work going back fifty years. The body of literature includes both research pieces as well as general-purpose books and manuals for practitioners<sup>10</sup>.

Options for detail reduction are:

1. Removal of direct identifiers
2. Coarsening (such as reducing geographical detail; converting age to five-year bands)
3. Recoding (such as replacing higher earnings with “earnings above €100,000”)
4. Data swapping or other replacement techniques (for example, reallocating health conditions amongst individuals whilst maintaining the distribution of data)

Option (1) is usually a requirement on anyone collecting data: direct identifiers (such as name, or social security number) have little statistical value and a very high risk of disclosure. Options (2) and (3) are often carried out by researchers to reduce information content to the minimum necessary. Option (4) is a specialist function which only SDC professionals are likely to employ.

There are two general-purpose software tools for detail reduction, both open source and under continuous development. Mu-Argus is a stand-alone package, originally developed for

---

<sup>10</sup> In 2010 Eurostat commissioned a project to summarise perceived best practices. The final report was published as Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte Nordholt, E., Seri, G. and De Wolf, P-P. (2010). Handbook on Statistical Disclosure Control. ESSNet SDC, available at [https://ec.europa.eu/eurostat/cros/system/files/SDC\\_Handbook.pdf](https://ec.europa.eu/eurostat/cros/system/files/SDC_Handbook.pdf). An improved version of this report was published by Wiley as Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., Wolf, P.P. de, 2012, *Statistical Disclosure Control*, ISBN 978-1-119-97815-2. An example of a step-by-step guide for non-experts is <https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation.aspx>

---

use by national statistical institutes<sup>11</sup>. sdcMicro is an R package, intended to replicate mu-Argus but be called from within other packages<sup>12</sup>.

As well as carrying out detail reduction, the tools provide estimates of the risk associated with various protection measures. These are quantitative estimates based on risk models which can be feasibly calculated by a package that is designed to work without knowing the access context. This has caused some concern, that non-specialists might misinterpret models of relative theoretical risk as absolute objective measures. However, at present these tools are largely the preserve of professional data managers.

### **Workshop findings**

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 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. 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 also prove a way of de-identifying and/or assessing risk.

---

<sup>11</sup> <https://research.cbs.nl/casc/mu.htm>

<sup>12</sup> <http://www.ihsn.org/software/disclosure-control-toolbox>

---

## Subtopic 2.2: Output SDC

Output SDC refers to the application of SDC methods to potential publications after the analysis has been carried out, to guard against residual disclosure<sup>13</sup>. Statistical techniques aim to prevent individuals, households or enterprises from being identified in published information.

Because most output checking is carried out manually, operational considerations affect the disclosure rules. OSDC operating regimes are rules-based (RBOSDC: strict yes/no clearance processes), ad hoc checking (AHOSDC: rules, but not strict application), and principles-based (PBOSDC: guidelines rather than hard rules, and strict condition on how exceptions are handled)<sup>14</sup>. Statistical organisations are generally rules-based when it comes to producing official statistics, but there is more variation for analytical outputs. Most RDCs apply AHOSDC or PBOSDC, as RBOSDC is generally too constricting for research use. However, RJSs may apply RBOSDC as this means all processes can be automated. Automatic tools developed for output checking such as ACRO<sup>15</sup> also, rely on RBOSDC.

The ultimate aim of SDC is to maximise the usefulness of the outputs while minimising the risk of disclosure (or the perception of disclosure). This level of risk is never zero, so the aim is to reduce, to an acceptably low level, the possibility that confidential information is released. ‘Acceptably low’ is not clear: guidelines for output balance the risks to individuals of being identified in a publication and the risk to the public good of not being able to use statistical evidence (the ‘confidentiality’ and ‘usefulness’ problems). Consider the choice of a threshold

---

<sup>13</sup> Lowthian, P., & Ritchie, F. (2017). Ensuring the confidentiality of statistical outputs from the ADRN.

<sup>14</sup> 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>

<sup>15</sup> Green, E., Ritchie, F., & Smith, J. (2021). *Automatic Checking of Research Outputs (ACRO): A tool for dynamic disclosure checks*. *ESS Statistical Working Papers, 2021 Edition*, <https://doi.org/10.2785/75954>. Available from <https://uwe-repository.worktribe.com/output/7449840>. Code and support materials available from <https://github.com/eurostat/ACRO>.



---

count, universally used as a criterion for linear statistics such as frequencies, means or simple indexes:

- confidentiality: a low threshold increases the probability of disclosive cells being published, whereas a high threshold reduces this risk considerably
- usefulness: a low threshold allows most statistical findings to be published; a high threshold is likely to mean that some findings are not published

In RBOSDC, the threshold set is a compromise between these two. In PBOSDC the allowance for exceptions means that the threshold rule can target confidentiality; the usefulness is tackled through exceptions.

Not all decisions are statistical or operational. For example, one organisation uses a lower threshold for 'regular' outputs and a higher limit for outputs from 'sensitive' data. While the practical impact of this is unlikely to be significant, it sends a clear sign to both research and data holders that an extra degree of caution is being used for the 'sensitive' data.

Statistics can be classed as 'safe' or 'unsafe' (or sometimes 'low review' and 'high review', respectively). A 'safe statistic' has negligible disclosure risk due to its functional form (such as linear regression coefficients,  $R^2$ , or significance tests). These can be released for publication without further checks, except perhaps administrative ones (for example, in regressions, that  $N > K + 1$ ). In contrast, an 'unsafe statistic' is inherently problematic; it, therefore, needs to be reviewed for disclosure risk in the particular instance being presented. OSDC is therefore primarily concerned with (a) identifying statistics as 'safe' or 'unsafe' (b) devising methods for checking and dealing with the 'unsafe statistics'.

Most OSDC is concerned with frequency and magnitude tables. Tables of frequencies (the numbers of observations) can be problematic due to low counts or their unusual distribution:

- Identification disclosure is the act of identifying a specific person or unit in the data.
- Group (attribute) disclosure occurs when all respondents who have some feature also have some other feature.
- Within-group (attribute) disclosure occurs when there is one respondent in a single category with all other respondents in a different category.

Cells with 1 or 2 contributors are usually assumed to be disclosive, and so all textbooks use three as a minimum threshold for pedagogical purposes. However, in practice, most organisations use higher thresholds (5, 10 and 20 are popular) as this provides a 'margin of

---

error’ whilst not materially affecting outputs. Secondary disclosure (exploiting differences between tables, or between marginal totals, to reveal small numbers) is also a concern.

In magnitude tables, each cell value represents the sum (or average) of a value across all respondents belonging to that cell. As well as the problems of frequency tables, magnitude tables have the additional risk that the cell might be dominated by a small number of contributors, disclosing approximate information about those individuals.

There are tools for checking tables: sdcTable and tau-Argus are general-purpose open-source programmes. These were designed for the production of official statistics, and are used as such, but they have gained little traction in research environments with idiosyncratic outputs.

For tabular output, broad techniques for identifying and handling problems have remained much the same for many years; this is a very mature research area. Recent developments include differential privacy (see above) and the ‘cell key’ approach, a noise-addition method that offers security and flexibility at a cost of some consistency.<sup>16</sup> Unlike DP, the cell-key is not susceptible to multiple-query attacks as the same noise is added in a repeatable way irrespective of the table being created.

### **Workshop findings**

The discussions focused on the need for researchers to accept co-responsibility for output checking: the better the quality of the requested output, the 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 train 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

---

<sup>16</sup> <https://gss.civilservice.gov.uk/wp-content/uploads/2017/01/ExN-Disclosure-control-methodology-in-2021-Census-outputs-Spicer-Blanchard-Dove-ONS.docx>

---

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 (and ideally operational aspects), but not all organisations do training; there is still a very wide view on what can be expected from researchers.

## Subtopic 2.3: Synthetic data

Synthetic data is data created to replicate the structure of genuine data but without the confidentiality risk. Synthetic data can be

- Partially synthetic: some of the data in the synthetic data set are real; only some data (all the identifying variables, such as age, or detailed location) have been replaced by artificial data
- Fully synthetic data: all of the data in the dataset is constructed artificially from models.

Synthetic data does not mean that it is risk-free; if partially synthetic data retain identifiers, or if the synthetic data very closely resembles original data points (for example, reproducing outliers), then there may still be some need to protect the data.

Many synthetic datasets are created to improve coding efficiency: users develop code and use synthetic data (which replicates the structure of the source data) to iron out the bugs. This means users can be efficient when they get access to confidential data in secure environments as the secure-environment code needs less development. However, some synthetic datasets are designed to be used for research. As the synthesised data may not generate the same statistical findings as to the source data, so-called 'replication servers' allow users to submit code to be run on both synthetic and source data sets and provides a report on the results. finally, synthetic data is also used to run microsimulation, for example, to examine the effect of government policies. For microsimulation, the accuracy of the relationships is a key criterion (so that correct inferences in response to changes can be accurately modelled).

At its simplest, a synthetic dataset is created by reproducing distributions. Consider synthesising the observations on a single variable: it would be straightforward to generate random values which reproduce the moments of the distribution of that variable. However,

---

the synthesis of an entire dataset requires reproducing the relationships between variables as well as the distribution of each variable. In addition, there may be logical constraints: if gender and illness are both being synthesised, the code needs to ensure that “gender=female” cannot be associated with “cancer=testicular”, for example. If the data are highly skewed (as in the case of business data), it can be difficult to reproduce the extremes of the distribution without revealing characteristics of the original data. If multiple waves of data are to be synthesised (as in a repeated survey), then this introduces additional logical constraints such as  $age[t] = age[t-1]+1$ . These and similar problems mean that creating a synthetic dataset with plausible variable values and relationships is considerably harder than just making up numbers to reproduce univariate distributions. Practical applications are dominated by social data, which has relatively well-behaved distributions.

Most datasets are created by combining statistical models of the distribution with logical constraints. Variables are modelled sequentially, reflecting both expected dependencies in the data. This also keeps the computations manageable; modelling multiple relationships simultaneously is practically infeasible except for very simple data sets. Some recent researchers have begun applying machine learning methods, which so far seems to be more scalable but at a cost of lower utility.

Several software tools can automate the process. Two of the most popular are SynthPop<sup>17</sup> and SimPop<sup>18</sup>, both implemented as R packages.

### **Workshop findings**

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.

---

<sup>17</sup> <https://www.synthpop.org.uk/>

<sup>18</sup> Templ, M., Meindl, B., Kowarik, A., & Dupriez, O. (2017). Simulation of Synthetic Complex Data: The R Package simPop. *Journal of Statistical Software*, 79(10), 1 - 38. doi:<http://dx.doi.org/10.18637/jss.v079.i10>

---

The future involved the development of specialist knowledge and research eg AI techniques such as 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.

---

# Session 3: Organisation

## Subtopic 3.1: Training<sup>19</sup>

Should users be trained in the use and management of confidential data? When access was primarily through distributed data, user training was largely confined to giving users good practice instructions to read as part of the licence or data access agreement. There is little evidence to demonstrate that users read or remember this information, but data distributors build this into their security models by reducing detail.

With the growth of RDCs in the 21<sup>st</sup> century, it has become clear that the training of users can 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, user training and guidance can have a significant impact on the efficiency of output checking procedures<sup>20</sup>.

Historically the perspective has been that users of data are fundamentally untrustworthy – the ‘intruder’ model – despite the lack of evidence for malicious use amongst the research community. There is substantial evidence for researchers deliberately or accidentally failing to follow procedures. This has led, in recent years, to a move towards behavioural training in several countries, called the ‘human’ model<sup>21</sup>, or less polite names. This focuses on positive

---

<sup>19</sup> For a more detailed discussion, see section 4.4 and the appendix in Ritchie, F., & Green, E. (2016). Australian Department of Social Services Data Access Project: Final Report. ADSS, Canberra. <https://uwe-repository.worktribe.com/output/908255/departement-of-social-services-data-access-project-final-report>

<sup>20</sup> 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>

<sup>21</sup> For example, Eurostat’s *Self-Study material* <https://ec.europa.eu/eurostat/web/microdata/overview/self-study-material-for-microdata-users>

---

engagement with researchers to build a sense of community, rather than focusing on following specific rules<sup>22</sup>.

Even within RDCs, there is much disagreement about whether and how much training is necessary, whether it should be passive (such as reading online material) or active (face-to-face, interactive classes), and whether there should be tests or some form of certification. While face-to-face is often seen as the gold standard, it can be expensive and presents substantial difficulties in large countries.

The move to virtual training during the pandemic has forced organisations carrying out face-to-face training to redesign their training models. The evidence on this is not clear yet. However, in the UK, where a nationally accredited training scheme franchised to data holders moved online with notable variations in delivery style between partners, data on 3,500 trainees (2,100 face-to-face pre-pandemic; 1,400 online but interactive since March 2020) suggests that online interactive training can be as effective as face-to-face training, at least on a simple measure of pass rates. This has implications for countries that struggled with geographical distances pre-Covid.

Outside of RDCs, there is a growing interest in training users in confidential data management, ethics, data governance. Data holders are also training their staff in these issues. However, these tend to be associated with individual organisations (except in rare cases; for example, the Australian National Data Commissioner's government-wide data governance training). As a result, it is not clear if there is a consensus on what needs to be taught to whom, or what is the most effective teaching mode.

### **Workshop summary**

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, user training and

---

<sup>22</sup> Green, E., Ritchie, F., Newman, J., & Parker, T. (2017, September). Lessons learned in training 'safe users' of confidential data. Paper presented at UNECE/Eurostat work session on statistical data confidentiality - 2017, Skopje, FYR Macedonia

---

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. Some only train users in how to use the facility.

Covid-19 forced organisations to carry 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 that struggled with geographical distances.

Delegates felt that good practice meant integrating a community-building approach within the training course. This should include an 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 need to be appropriate to the organisation (as such LMICs materials might require tweaking). Training in AI and ML models will become the next challenge.

### **Subtopic 3.2: Access arrangements**

Data can be accessed in multiple ways, and often the same organisation makes data available in multiple ways. One common depiction is some form of 'data access spectrum' or 'continuum of access', developed independently in the UK and Canada in the 2000s and



subsequently adopted by several organisations. For example, for Statistics Canada the 'continuum of access' is<sup>23</sup>

	<b>Open Statistics</b> ← → <b>Restricted Data</b>					
Service	Statistics Canada Website	Depository Services Program (DSP)	Data Liberation Initiative	Custom Tabulations	Real Time Remote Access and Remote Job Submission	Research Data Centres

This is an intuitive representation of the trade-off between data confidentiality and flexibility in access arrangements. An alternative way of looking at access decisions is the 'Five Safes' which considers data governance as a combination of five separable but related dimensions<sup>24</sup>

Safe projects	Is this use of the data appropriate?
Safe people	Can the researchers be trusted to use it appropriately?
Safe data	Is there a disclosure risk in the data itself?
Safe settings	Does the access facility limit unauthorised use?
Safe outputs	Are the statistical results non-disclosive?

These are scales, not targets: the idea is that more 'safety' in one dimension can balance less control in another. For example, an RDC has great control over projects, people, settings and output, and so can hold very detailed ('unsafe') data whilst still maintain overall safe use. In contrast, data available for download under licence has some controls in place (users: sign

<sup>23</sup> Table 20.1 in Gray, S.V., and Hill, E. (2016) "The Academic Data Librarian Profession in Canada: History and Future Directions". Western Libraries Publications. Paper 49. <http://ir.lib.uwo.ca/wlpub/49>. This is derived from the Data Liberation Initiative 'Survival Guide' <https://www.statcan.gc.ca/eng/microdata/dli/training-events/dli-survival>

<sup>24</sup> Ritchie, F. (2017, September). The "Five Safes": A framework for planning, designing and evaluating data access solutions. Paper presented at Data for Policy 2017, London, UK. <https://uwe-repository.worktribe.com/output/880713/the-five-safes-a-framework-for-planning-designing-and-evaluating-data-access-solutions>

---

access agreements; are required to agree to storage conditions and no onward sharing, and so on) but these are weaker as the user is not under the direct control of the data holder. Accordingly, data detail is reduced to keep the overall risk manageable. Finally, data given an unrestricted release has no operational controls, and so protection is entirely vested in the anonymisation of the data.

The Five Safes and data spectrum can be seen as two different ways to look at the same issue. Good practice for the Five Safes states data should be seen as the ‘residual’ – reducing detail is what happens when other controls are not appropriate, feasible or desirable. The spectrum represents ‘data detail vs other controls’, and so it can help to clarify, for example, the need for the data and user expectations.

There are pros and cons for each data access route. Distributing fully anonymous data involves a high initial cost, but no future expenditure. Distributing partially anonymised data under licence allows data holders to give more detail to users; managing the risk involves both initial and ongoing costs, but there are large economies of scale. An RDC or RJS (Remote job server) involves expenditure on IT, not statistical solutions.

Users can make the same trade-off. Data distributed under licence remains perhaps the most important source of research data, as users highly value having data on their desktops. One uncertainty arising from Covid is the growth in remote access to restricted facilities. This may significantly change the balance between convenience and detail for researchers, but it is too early to tell.

Distributing data does face some challenges: increasing computing power, the greater use of administrative data as a source, social media as a source of intruder data, perhaps AI, all raise concerns about whether current de-identification and anonymisation practices are sustainable in the long run<sup>25</sup>. As there are significant cost and user advantages to the ‘treat once, then distribute’ model of distributed data, this is a concern for data holders.

---

<sup>25</sup> Government Data Quality Hub (2018) *Privacy and data confidentiality methods: a Data and Analysis Method Review (DAMR)*. National Statistician’s Quality Review. <https://gss.civilservice.gov.uk/policy-store/privacy-and-data-confidentiality-methods-a-national-statisticians-quality-review-nsqr/>

---

## Workshop summary

The Five Safes is the basis for structure across a lot of places; broadly, this is seen as good practice, but a '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 growth in computing power pose an 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 are 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 advances in computing each pose a risk for scientific use files. LMICs need further development in capacity and infrastructure to help support provisions for secure-use services and so are likely to rely on distributed data for the medium term.

## Subtopic 3.3: FAIR, metadata, and sustainable management

'Metadata' relates to the information supplied about data, to enable the data to be used and meaningful inferences drawn from it. Metadata can range from simple lists of variables and labels to complex descriptions of collection methodologies, notes on cleaning processes, or how to interpret similar terms in different datasets.

There are multiple metadata standards, as well as idiosyncratic systems used by individual organisations. Different communities and organizations have different goals that guide their

---

collection, usage, and sharing of data. For example, the Research Data Alliance's Metadata Standards Catalogue<sup>26</sup> lists 52 separate schemes, often but not always subject-specific.

With the rise of generally used and flexible metadata schemas (schema.org, DataCite, Dublin Core, and so on), datasets can nowadays be described in a flexible and generally understood way. Notwithstanding the multiplicity of schemes, metadata development has been fairly stable since the 2000s; most recent work has gone into the mapping and conversion of standards<sup>27</sup> or managing the integration of multiple schemes.

Information on how data are released, protected, controlled, and accessed is less well defined in the literature or practice. The low interest in this may be because metadata and metadata schemas tend to be producer-centred rather than user-centred. For a producer of data, compatibility and completeness of data descriptions is central to data management; the ability to describe one's access arrangements to others is not. In the ESS Quality Standards Framework, for example, metadata and archiving are distinct parts of one's dissemination strategy.

References to access in the metadata literature tend to refer to the specific technical problem of 'interoperability'; that is, ensuring that datasets can be readily integrated through automatic mechanisms. This tends to be interpreted in terms of record-level descriptors, rather than institutional arrangements, and research in this field is heavily influenced by IT security models. An exception is work being led by the INEXDA consortium on 'annodata' – defining a new metadata standard to allow data access mechanisms to be described in the same sort of systematic way as data items are described<sup>28</sup>.

---

<sup>26</sup> <https://rdamsc.bath.ac.uk/>

<sup>27</sup> Eg Hirwade M. (2011) A study of metadata standards. Library Hi Tech News v71 pp18-25. DOI 10.1108/07419051111184052

<sup>28</sup> S. Bender, J. Blaschke, H. Doll, A. Gordon, C. Hirsch, D. Hochfellner, J. Lane. (2019) The Annodata Framework: Putting FAIR data into practice. Deutsche Bundesbank Technical Report 2019-03

---

However, the major development in standards for the overall use of data has been the FAIR data principles: findability, accessibility, interoperability and reusability<sup>29</sup>. FAIR requires data providers to maximize the use-value of the data by making it easy to find and use with minimal human intervention. In the FAIR context, "accessible" does not mean "open" or "unrestricted", but those access paths are clearly defined. However, in line with the other literature, to date, this has been interpreted largely as a record-level requirement.

Reproducibility and replication are becoming an area of interest, and journals are increasingly demanding to see the provenance of research findings<sup>30</sup>. The high turnover of research in the pandemic has also raised some concerns about the quality of quick-release research findings. With confidential data being increasingly held behind firewalls, the assurance of research findings becomes more complex. A joint UK Data Archive/Office for National Statistics event in February 2020 brought together several interested parties from Europe and North America to examine ways forward, with a summary and reflection by the organisers posted on the web<sup>31</sup>.

The role of archiving in sustainable data management has not attracted much attention, except insofar as to relates to FAIR or reproducibility. There is a high degree of consensus on good data management and archiving amongst academic practitioners, with data archives, in particular, taking a leading role in advising researchers in good practice<sup>32</sup>. It is less clear if there is a consensus in government data collections, largely as government departments do not typically describe their data management arrangements in detail.

### **Workshop summary**

Delegates were clear that metadata should be publicly available, that DOI and annodata schema should be used, and that metadata has to be comparable. The DDI standard, although

---

<sup>29</sup> Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., & Bouwman, J., (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3.

<sup>30</sup> eg Vilhuber, Lars, (2019). "Report by the AEA Data Editor," *AEA Papers and Proceedings*, vol. 109, pp. 718-29.

<sup>31</sup> Louise Corti and Andrew Engeli: <https://blog.ukdataservice.ac.uk/loveyourcode2020/>

<sup>32</sup> For example, <https://www.ukdataservice.ac.uk/manage-data/handbook>

---

popular, was seen as being hard to reach, with secondary data often poorly documented. There was a trade-off between availability and meeting standards.

Several 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 or skills to use current ones. Metadata on access mechanisms would be helpful. An international MicroData Standard would be optimal but seems like a pipe dream.

---

# Session 4: Societal context

## Subtopic 4.1: Regulatory regimes

Broadly there are two types of regulatory regimes: rules-based and principles-based. In rules-based regulation, the system operates under a list of outlined requirements and rules; an example is a criminal law. Rules-based regulation is enforced by verification ("did you exceed the speed limit?"). Principles-based regulation identifies the goals of regulation, formulates rules as subordinate to those principles, and allows multiple ways to achieve those goals. Principles-based regulation is often managed by accreditation ("does the solution meet the standard expected?"). The advantage of principles-based regulation is its generalisability: it can be applied to any data management problem, such as designing internal administrative systems, setting up secure onsite facilities or releasing information on the internet. However, this advantage is also its main disadvantage: it is an approach to design but does not explicitly state a solution.

The rules-based model of regulation aims to specify in a binary manner what is allowed or not allowed. Under this model, the primary source of direction is the regulation itself. The principles-based approach instead focuses on what any system is trying to achieve, and then questions whether the system achieves those objectives. In a principles-based system, implementation decisions are primarily under the control of the implementor; regulation is there to specify the goals, and to identify what evidence should be presented that the goals have been achieved. The UK Digital Economy Act 2017 is an example of a principles-based approach: The Office for Statistics Regulation has to identify and approve accreditation processes for many of the different elements of data but does not need to follow a specific manifesto. Rules-based regulation works well when the terms can be unambiguously defined<sup>33</sup>.

---

<sup>33</sup> In Oceania, a movement called 'rules as code' has taken this to the logical extreme, and encourages programmers to build legal requirements directly into their data coding, as long as those rules can be clearly and unambiguously defined.

---

In recent years, the Five Safes model (see above) has become increasingly associated with the principles-based approach to regulation. There is an affinity between the two concepts. The Five Safes provides a framework for planning; the principles-based model provides a way of suggesting how that framework should be used, and how goals should be specified. Neither is specific on the actual implementation, but both provide a way that the effectiveness of any implementation can be measured. The popularity of principles-based regulation is that it seems to address some of the flaws of older legislation which struggled to provide adequate guidance.

### **Workshop summary**

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, concerns were 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.

## **Subtopic 4.2: Public attitudes and engagement**

Much of the academic literature concerning public attitudes towards data sharing comes from research regarding people's healthcare data. It is clear from the literature that people have an intuitive wariness of institutions that want to use and share their data. People are less wary of the healthcare industry and police, and people are warier of technology companies, insurance companies and the media.

Researchers acknowledge their duty to share data for the benefit of future generations. Ordinary members of society also express the benefit to future generations as their biggest reason for being willing to share their data. Some socio-demographic groups (particularly



---

those who feel marginalised or intimidated by the state) are less willing to share their data because of a historical distrust of certain types of institutions.

However, people are largely unaware of how data is already being used by the government, as well as academic researchers, charities, and commercial organisations. The more informed someone is of how their data will be used, the more willing they seem to be for their data to be shared. There appears to be a 'data trust deficit' whereby trust in institutions to use data appropriately is lower than trust in them in general. Privacy is the biggest issue for the public; safeguards are seen as essential.

These findings above are largely derived from UK analyses. Whilst some of these have resonance in other countries, there are also significant differences based upon historical and cultural factors. Citizens in continental Europe appear to be more confident that data use will be well-regulated/well-managed, compared to Anglo-Saxon countries. Countries with strongly federal structures (eg Germany, Australia) seem to place more trust in the sub-national governments. This may be a reflection that citizens are more likely to entrust data to individuals or organisations that they think they know.

### **Workshop summary**

'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 create 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 the public to gain trust in data use. It was noted that that focus groups tend to be more 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); these have been developed as the CARE principles (Collective benefit, Authority to control,

---

Responsibility and Ethics)<sup>34</sup>. 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, but better ‘storytelling’ is likely to be important for building trust.

### **Subtopic 4.3: Ethics/ benefits and costs**

Different regulatory regimes of data access are underpinned by alternative ethical approaches. Whilst the ruled-based regime reflects a **deontological framework** whereby what is permitted or forbidden is decided a priori based on a stable set of rules, a principles-based regime recalls a **consequentialist framework**, which, instead of focusing on the rightness or wrongness of the actions taken, looks rather at their impact on the future.<sup>35</sup> To assess such impact and therefore decide if an action is worth taking, we need to consider whether the consequences of this action align with a set of principles, which we deem valuable. In other words, from a consequentialist viewpoint, the ethical kernel is not the action itself (e.g. someone accessing or sharing sensitive personal data), but its principled consequences (e.g. the societal impact of that access/share).

The main challenge of a consequentialist framework consists of establishing a set of valid principles, which would allow us to assess the impact of our actions. In other words, what do we deem intrinsically valuable and therefore worth guiding our decision-making process in a specific context? **Privacy** has been often referred to as the cardinal principle of data access and

---

<sup>34</sup> <https://www.gida-global.org/care>

<sup>35</sup> See on this, for instance, Walter Sinnott-Armstrong, “Consequentialism”, *The Stanford Encyclopedia of Philosophy*, <https://plato.stanford.edu/archives/sum2019/entries/consequentialism/>.

---

governance to the point of leading scholars to question its hegemony.<sup>36</sup> There are many explanations for such emphasis on the value of privacy. Liberal societies give priority to individual values. This is not a sufficient reason, however, to prioritise privacy over health.

Data ethicists and stakeholders should discuss what values and public goods they consider crucial for applying a principles-based regime to microdata access, and what are the linkages and dependencies between such values and goods. Privacy should play a key but not exclusive role in this discussion. Other **collective principles**, which might determine the value of our data use, should also be considered.

### **Workshop summary**

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, priorities other 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 a 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

---

<sup>36</sup> See Tamar Sharon, “Blind-sided by privacy?”, *Ethics and Information Technology* 2020, <https://doi.org/10.1007/s10676-020-09547-x>

---

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.

---

# Overarching findings

The 5-day workshop attracted 130 attendees across 88 different organisations from 26 different countries. Discussions ranged from applied philosophical rhetoric questions to technological innovations in microdata access. This section outlines overarching findings which were cross-cutting across the different sessions; we then summarise what delegates thought would be in the pipeline for the future of microdata access. Finally, we outline recommendations for the next steps.

## **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.

## **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 the 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.

## **Rise of the machines**

With a lack of resources, we appear to be approaching a situation in which technological advances outpace current knowledge and practice of disclosure control. With the rise in computing power, 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 or devising new, more robust solutions

## **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, following on from the Christchurch earthquake. 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.

---

# Road map for the future

There was recognition that meeting and sharing ideas is in itself a good thing, and something that the data community would like to 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 take up?
- 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 ideas

As the conference was not organized under the formal authority of any group, there is as yet no mechanism to take these ideas forward. However, the conference will be discussed at the Eurostat/UNECE Expert Workshop in December 2021<sup>37</sup>; and number of the proposals are likely to be explored as part of the DRAGoN external engagement plan for 2022. The DRAGoN group also aims to support other groups/networks keen on taking aspects forward. Interested parties are encouraged to contact [dragon@uwe.ac.uk](mailto:dragon@uwe.ac.uk).

---

<sup>37</sup> A draft summary of this report and presentation can be found at the workshop web page <https://statswiki.unece.org/display/confid/Work+Session+on+Statistical+Data+Confidentiality+2021>

---

# Glossary

The Glossary was developed to aid discussion at the conference by having common meanings for terms, such as stating how ‘anonymous’ and ‘de-identified’ were to be used (as these differ substantially between countries). It is not intended to be definitive or replace other glossaries.

Term	Definition to be used in the workshop
Annodata	all information on the process for providing access to data, i.e. information about the set of legal requirements that must be considered when making data available for research and analysis.
Anonymous data	data that does not include sufficient detail to allow the data subject to be identified, under any reasonable conditions
Breach of confidentiality	the release of identified or de-identified data to an unauthorised system, environment or person; a breach of confidentiality may not mean a disclosure as it will depend on the circumstances
Breach of procedure	failure to follow appropriate operating procedures, irrespective of whether a breach of confidentiality occurs
Confidentialisation	the act of reducing the likelihood or harm of re-identification by reducing detail or perturbing the dataset
De-identified data	data which includes sufficient detail to allow the data subject to be identified, but only with effort and with less certainty (for example, a combination of gender, age, type of employer, salary range and disability status)
Differential privacy budget, or epsilon	quantitative measure of by how much the risk to an individual's privacy may increase, due to that individual's data included in the inputs to the algorithm.
Differentially private algorithms	population-level insights about a dataset to be derived, whilst limiting what can be learned about any individual in the dataset.
Distributed access	restricting the physical location of the data, but allowing users in other locations to carry out analysis and retain statistical results (but not microdata)



Distributed data	sending microdata to users under licence, to analyse on their machines
FAIR principles <sup>38</sup>	an acronym for Findability, Accessibility, Interoperability, and Reusability
Federated analytics	an approach for applying data science techniques by moving code to the data
Five safes <sup>39</sup>	The Five Safes is a framework for helping make decisions about making effective use of data that is confidential or sensitive. The Five Safes model also places <a href="#">statistical disclosure control (SDC)</a> in its proper context, as part of a system approach to data security. The Five Safes breaks down the decisions surrounding data access and use into five related but separate dimensions: safe projects, safe people, safe data, safe settings, safe outputs.
Fully homomorphic encryption (FHE)	encryption schemes where it is possible to compute any polynomial function on the data, which means both additions and multiplications.
Homomorphic encryption (HE):	a property that some encryption schemes have so that it is possible to compute encrypted data without deciphering it.
Identified data	Some data directly (not necessarily uniquely) relates to an individual respondent eg name, detailed address, social security number, Health service number, tax registration number etc
Input SDC	the application of SDC methods to raw data to reduce data risk before it is released to the users
Microdata	the individual unit records about a person or organisation, such as information collected from surveys or administrative data

<sup>38</sup> Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3: 160018.

<sup>39</sup> <http://www.fivesafes.org/>

Noise:	noise refers to a random alteration of data/values in a dataset so that the true data points (eg personal identifiers) are not as easy to identify.
Output SDC (OSDC)	the application of SDC methods to potential publications after the analysis has been carried out, to guard against the residual risk
Partial Homomorphic Encryption (PHE):	encryption supporting only additions or only multiplications (also referred to as additive homomorphic encryption and multiplicative homomorphic encryption).
Personal Data Stores (PDS)	are systems that provide individuals with access and control over data about them, so that they can decide what information they want to share and with whom.
Principles-based	A regulatory regime or operating model where 'principles' (what you are trying to achieve) are the basis for planning. Rules are designed to implement principles but can be changed if inconsistent. A principles-based system does not specify how a goal is to be achieved, only what the goal is. For example, a data protection regime could specify that the confidentiality of the individual is protected, but without specifying whether that occurs through anonymization or other methods.
Privacy-enhancing technologies (PET)	any technical method that protects the privacy of personal or sensitive information.
Public use file (PUF)	data file without restrictions on use or onward access
Raw data	the source data
Remote access	a system that allows users to 'see' and manipulate the source data
Remote job server (RJS)	a system allowing a range of complex analyses to be carried out, not just tabulations, without seeing the source data; a table server is a remote job server that has only one function
Research data centre (RDC)	a restricted access facility where users can manipulate the source data without restriction as if on their own computers; but the environment is made secure so that users cannot bring information into or take data out of the facility without approval, and additional services (such as internet access) are normally very restricted;

	typically provided by on-site access, where the facility is hosted on the organisation's premises
Rules-based	A regulatory regime or operating model where explicit rules are the basis for planning. The rules may specify how individuals and organisations should act, or (for example, defining 'anonymisation' and specifying what can be done with data that has or has not been anonymized).
Scientific use file (SUF)	the data file which retains some non-negligible confidentiality risk and so, therefore, has circulation restricted to authorised users for specific research purposes
Secure multi-party computation	multiple organisations collaborate to perform joint analysis on their collective data, without anyone organisation having to reveal their raw data
Secure use file (SecUF)	the data file which contains non-negligible confidential information therefore circulation and use is restricted to authorised users in controlled facilities
Sensitive data	data where release to an unauthorised person is likely to cause nonnegligible harm or distress to the data subject; for this report, we assume that all sensitive data is also confidential
Statistical disclosure control (SDC)	applying statistical measures to (a) determine if there is a substantive risk of unauthorised disclosure in a dataset or publication, and (b) make changes to the data or publication to reduce that risk
Synthetic data	generated data that can replace or augment sensitive source data
Table server	a system that allows users to generate their tables from the data flexibly, but without seeing the source data; a form of distributed access
Tokenisation:	obscuring a data item by replacement with a token, such as a regular expression, as part of a de-identification process. This is a reversible activity.
Trusted execution environments	code and data are protected in a processing environment that is isolated from a computer's main processor and memory.

---

Unauthorised disclosure	the unauthorised release of information about an identified data subject
Virtual RDC or Remote RDC (vRDC)	an RDC where technology is used to provide equivalent security to a physical site and to separate the RDC from the actual location of the data