Journal of Privacy and Confidentiality
Vol. 11 (1) 2021

Perspectives

Submitted December 2020
Published January 2021

MICRODATA ACCESS AND PRIVACY: WHAT HAVE WE LEARNED OVER TWENTY YEARS?

FELIX RITCHIE*

* University of the West of England
e-mail address: Felix.ritchie@uwe.ac.uk

Abstract. Felix Ritchie reflects on lessons learned in twenty years of microdata access in the UK and Canada. Based on his contribution to the panel on "Privacy And Microdata Access: Two Worlds Colliding?" at the October 2020 Canadian Research Data Centre Network (CRDCN) conference celebrating the 20th anniversary of the network.

1. Introduction

Is there a conflict between microdata access and privacy? From the system designer's perspective, no. Access systems are designed to ensure that privacy is preserved. Concepts such as "privacy by design" encourage designers to consider and build in data protection measures at stage of the access system design. Each of the "five safes" (projects, people, settings, outputs, data; Ritchie, 2017*b*) used to frame data access decisions has a wealth of knowledge tools and support for the system designer to draw on.

From the privacy advocate's perspective, this reasoning confuses implementation with principle. There is a base level of privacy risk associated with the existence of the data. If the data are shared more widely, then there are a number of new vectors through which privacy could be breached: how the data are transferred to the user, what the user will use them for, whether the user can be trusted, how the transferred data be disposed of, whether inadvertent breaches could occur.... None of these can lower the risk to privacy; the need for access controls show that positive risks are associated with all these elements. Therefore, more access creates more risk to privacy.

This second perspective is correct. To argue that, because we are good at managing data access, there is no increase in risk, is clearly wrong. However, in the twenty years since the Canadian Research Data Centres (CRDCs) were first set up, there have been two

Key words and phrases: confidentiality; privacy; microdata access; evidence-based; default-open; user-centred.

Editorial note: This article is an edited version of the author's talk at the October 2020 Canadian Research Data Centre Network (CRDCN) conference. Information on the conference can be found at https://www.crdcn20.ca/crdcn20/program*. Articles in the Perspectives series reflect the author's opinions, and do not necessarily reflect the opinions of the journal's editorial board.*

www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.766

broad shifts in attitude, particularly in the public sector. The first recognizes there are much deeper questions of costs and benefits than the direct privacy risks and delivery costs of access/no access situations. The second change has been to move towards a 'managerial' approach to data access rather than a risk-avoidance. This is embodied in the 'EDRU' model (evidence-based, default-open, risk-managed, user-centered) that, implicitly or explicitly, underlies many modern data access arrangements.

We consider each of these in turn. We focus on the sharing of data within and by public sector agencies, for users across the economy. Sharing of commercial data raises different issues of cost and benefit that are outside the scope of this paper.

2. Private and public costs and benefits

When I started working in data access in the early 2000s, decisions were focused largely on the costs and benefits to the agency holding the data[1]. The key criteria were (1) the cost of providing access; (2) the risks of providing access, and (3) the public benefit of increased access.

The first two of these criteria were assessed in terms of the agency. Costs could include the initial set-up cost of providing a facility or anonymizing a dataset, and the ongoing costs of monitoring users in secure environments or managing end user licenses. The impact on the agency was assessed in terms of the risk of something going wrong, and the consequent financial and criminal penalties for the agency. The risk to the data subjects of having their privacy breached was subsumed into the penalties the agency would face. Alongside these direct risks, reputational risk also loomed large in the sight of the agencies, especially for the public sector, which is usually seen as one indistinguishable mass (Bhatta, 2003; Yang and Holzer, 2006) and so one data breach reflects on all of government. For statistical agencies, often relying on voluntary surveys, reputational risk may directly affect the organization's ability to collect data.

The public benefit of increased access is ideally characterized in terms of economic benefit and/or social welfare outcomes, not agency goals. In practice, "public benefit" is a very nebulous concept. Attempts to measure the value of data access (such as Diepeveen and Wdowin, 2020) demonstrate the difficulty, and even this is a recent development: the great majority of the papers cited in Diepeveen and Wdowin (2020) are from 2015 or later. As a result, public sector delivery goals were often framed in terms of broad aspirations or mission statements to make data available insofar as possible.

The agency considering whether to release data was therefore faced with direct costs; additional operational and reputational risks; and an unclear benefit that was unlikely to generate a direct return to the agency. In contrast, not releasing data involved no additional costs or risks; indeed, the foregone benefit could be paraded as a badge of honor: "we are protecting the data." When agencies made such cost-benefit evaluations of whether to release data, the cards were strongly stacked against release.

Moving forward twenty years, perceptions of costs and benefits have changed substantially. In terms of releasing data, agencies have increasingly recognized the direct benefits of sharing their data with different groups of users: policy insights, free and expert methodological input, data cleaning and improvement, engagement with user groups, even staff development.

[1]There is a substantial debate about how those who make decisions about data access should be referred to. For this paper we use 'data holder' as a practical description of the role of the agency.

When data are accessible, statistical agencies are more likely to publish analyses jointly with academic researchers.[2]

There is much more evidence about the wider public benefit of sharing data. Whilst specific valuations may be difficult, there is now a common agreement on the importance of shared data for policy-making. "Evidence-based policy-making" has become a prominent part of public administration, particularly during the Covid-19 pandemic.

Of course, identifying that value remains hard. Outside of medical research, there is a tenuous link between data access, research and policy outcomes. For example, the UK Low Pay Commission annually publishes a 300-page report outlining the research leading to that year's recommendations on the national minimum wage. Despite the centrality of the research to the policy recommendations, it remains almost impossible to gauge the value of any one research piece, any one dataset release, or even the research in its entirety. Nevertheless, the entire research corpus profoundly influences the lives of a large number of workers.

Understanding of costs has changed. There is greater recognition that some costs are investments which generate a return for the agency. For example, compulsory training for researchers can be used to build positive engagement and encourage co-production (Desai and Ritchie, 2010; Green et al., 2017). Most usefully for statistical agencies, training sessions have become a way to introduce researchers to the reputational concerns of the agencies, and the importance of "being seen to be safe." Thus a notional cost serves to ameliorate a major perceived risk.

The costs of "do not release" have also become more prominent. There are direct costs in terms of not having the information available to improve the efficiency of services. There are also opportunity costs: the benefits foregone by not releasing data. What is the impact on society of not identifying a new factor in rehabilitation, of not being able to scrutinize the efficacy of social support programs, of not being able to model commuting patterns? Again, the pandemic has demonstrated the direct, sometimes fatal, impact of not having the relevant information available at the right time.

In short, in 2020 data holders have a much better sense of the true costs, risks and benefit of data access. Privacy is still an important element, but the recognition that "do not release" is not a riskless or cost-free choice, at least as far as society is concerned, has changed the balance in favor of releasing data.

3. Managing data

The second development has been the realization that microdata access is primarily a series of operational problems. Multiple ways exist to achieve access goals, and there is now a substantial amount of evidence to allow data holders to make effective implementation choices. Most importantly, a change in attitude from "Should we…?" to "How do we…?" leads to very different approach to implementation. These changes have been collected in the "EDRU" approach: evidence-based, default-open, risk-managed, user-centred (Green and Ritchie, 2016).

[2]See, for example, outputs from the Economic Statistics Centre of Excellence www.escoe.ac.uk, accessed 2021-01-28

3.1. **Evidence-based.** Twenty years ago, most evidence around data access focused on two topics: the vulnerability of secure datasets to hacking, and the vulnerability of open datasets to re-identification. However, far more important than evidence was the use of worst-case scenarios for risk planning: what could possibly go wrong, and can we prevent it? The logic is that protecting against worst cases also protects against less serious cases. The problem with this approach is threefold.

First, these were typically not genuinely "worst-case scenarios" but instead "worst-case scenarios that we can usefully model." This is most obvious in software (such as `muArgus` and `sdcMicro`) developed to provide risk assessments, where the need for a general-purpose tool means practical compromises. There is a big difference between the two; genuine worst cases are highly idiosyncratic, and have no value for planning purposes (Ritchie, 2017*a*). Claiming something is protected against a worst-case scenario when it is not generates practical and reputational risks.

Second, the worst-case scenario approach assumes that there is a simple line between the more and less serious breaches of confidentiality: protect against the worst and you protect against the least serious. But the linear relationship is not linear; breaches of confidentiality originate in many places, and may occur directly as a consequence of other measures. For example, complex password requirements create a new risk of password recycling.

Third, modeling which is based upon extreme scenarios is likely to overprotect. This is fine if the only thing the agency is concerned about is the risk of confidentiality breach. But as noted above, agencies have become much more aware of the value of data access to themselves and the wider public. The costs of overprotection have become visible.

Today, there is a lot more evidence on all aspects of data management. Some of this evidence did not exist in 2000, such as the effectiveness of the CRDCs, or statistical disclosure rules for analytical outputs.

Other evidence existed but was not used. For example, consider how the perspective on user engagement has changed. Traditional models of data access treated users as potential criminals: data are valuable, therefore users will steal them if they can. This conceptualization, based on a simplistic economic theory of crime, was popular with data managers as it made user instruction easy: "don't do this or you'll go to jail." There was no evidence to support this approach, and the theories on which it was based have been largely debunked. Nowadays best practice is the opposite: identify common interests across all parties to the data transaction, and make the users of the data see themselves as partners in the same relationship. "We all want the same thing; we're all working together" is a much more effective way to make sure that users engage with security measures (Desai and Ritchie, 2010). This strategy, of course, does not protect against malicious misuse by authorized users, which is best dealt with by other measures. On the other hand, training has been shown to be effective against *deliberate, non-malicious* misuse, such as users ignoring rules they find burdensome—a far more likely occurrence.

More broadly, well-run facilities have stopped treating users as robots, and now think of them as humans (Eurostat, 2016). Humans are tyically self-motivated; make mistakes; avoid instructions they don't like or understand; but go out of their way to do the right thing if they are convinced that it is the right thing. This is old material for psychologists, with much evidence, but data holders are now building in these ideas when thinking about the access environment. What Green et al. (2017) described as the "well-meaning idiot" model is used not just for training but for system design.

Data access planners do have to consider unusual scenarios, and allow for unknown or unsubstantiated risks, as they have always done. What has changed is that this no longer drives thinking.

3.2. **Default-open.** Government departments everywhere are increasingly committed, more or less willingly, to promoting re-use of their data for public benefit. However, a mission statement should not be confused with the decision-making process, which is strongly driven by human psychology.

The traditional perspective on data access is defensive and "default-closed." Decisions are framed around the question "Can we make these data available?" The default assumption is that no data can be released if conditions cannot be met. The problem for society is that this places the entire burden of proof on the person wanting to release data, and none on the person wanting to stop it. As a result, data access is limited by the person who is hardest to convince—or the one least willing to take a decision (Ritchie, 2014). The old IBM adage can be rephrased as "no one was ever fired for saying they don't think the confidentiality issues have been addressed."

In recent years, organizations and countries have moved to a "default-open" perspective: "*how* do we make the data available?" The default assumption is that data access will go ahead unless confidentiality concerns cannot be overcome. The emphasis is on finding practical solutions, and so risks need to be clearly articulated; "something might go wrong. . . " no longer suffices.

In theory, there is no difference between the default-closed/default-open model: both should come to the same decision about access if the same tests of data protection are applied. However, as Ritchie (2014) demonstrates, the psychological endowments associated which each default position lead to very different outcomes.

Ritchie (2014) also argues that one reason for the preference for defensive models is confusion over the difference between objectives and constraints. In the last ten years or so, when faced with a gathering of interested data professionals, I often pose this question to the audience:"Maintaining confidentiality is our highest priority: agree or disagree?" Typically, 90% or more of the audience will say that they agree. But if this is truly what you believe, there is no issue about data access. Confidentiality can be maintained by locking data in a filing cabinet in a disused toilet in a cellar with no stairs, no lights, and a sign on the door saying *Beware of the leopard*[3] It also means that no value is derived from the data.

Maintaining confidentiality, or acting lawfully or ethically, is not an objective; it is a constraint. No organisation's list of objectives should need to state "we aim to obey the law." The objective is to use data; the constraint is to obey the law. This can be a surprisingly hard concept to get across, but the distinction is increasingly being realised.

3.3. **Risk-managed.** Twenty years ago we were focused on risk avoidance; now we focus on risk management. We recognize that it is not possible to remove all, or indeed even many, of the risks; moreover, as noted above, we have become better at recognizing the risks of not doing as well as the risks of doing. Accepting the existence of risks can be difficult for data holders: to articulate a risk entails a response to it. In the traditional approach, risks were dealt with by transferring responsibility as far as possible. The modern approach accepts that risks exist and the need to reduce both incidence and impact.

[3]This also apparently works for by-passes: see Adams (1979).

Consider the case of the UK in the 2000s, when it was reported in the press that a large supermarket was using Census microdata to send mail to individuals. This was not true: the supermarket used published Census tables to target promotions to different geographical areas. However, the narrative is more attractive than the truth, and the Census team had to expend effort rebutting the claims This reputational risk is impossible to prevent, but there are mitigation strategies, such as training agency staff and researchers to be precise in their language when talking to non-specialists; having media contingency plans; and ensuring that accurate information is easily findable. The alternative is to greatly reduce data access, but this is an over-reaction when there is no actual data risk. Hence, misrepresentation of data access arrangements becomes a necessary but manageable risk.

There is also a growing understanding that risk judgments are inherently subjective. Skinner (2012), for example, notes that even in the seemingly objective world of statistical disclosure control software, all important choices are subjective and under the control of the data holder. While authors such as Opperman (2018) have tried to re-introduce objective measures of risk, these have made little headway against the subjective trend. Ironically, recognition of subjectivity may be a direct result of the increased evidence about data risks.

A significant change has been the understanding that effective data protection systems need to consider a portfolio of control measures, such as the Five Safes: who is going to be using their data, how are they going to be using the data, how will results be released into the environment, what ethical checks are in place, and of course, should there be any protection in the data themselves? This allows us to think more holistically about data protection and build flexible solutions ttah recognize and can adapt to different circumstances. The Covid-19 crisis has illustrated how robust portfolio strategies can adapt to unexpected events: many national statistics institute were able to adapt their RDCs to home work and remote access with ease, whilst still maintaining very high levels of security.

3.4. **User-centred.** At the start of the century, data protection was highly data-centered: "we have this dataset, what can we do with it?" The current trend is to focus on the use case: "what is the use value of these data; why are we making them available?" If that basic question cannot be answered, there is no point in pursuing data access. But if a use case can be established, then building in the user perspective from the start brings several advantages. Users can be engaged more effectively; value can be realized more efficiently; and most importantly, we can reduce the security risks from potentially unwarranted assumptions about users (for example, that users pay much attention to legal sanctions).

In portfolio models, the rise of the user has been counterbalanced by the fall of data protection as a security measure. The point of data access is to allow use: once the "who," "why," and "how" of data use are established, the appropriate level of data detail to achieve that aim can be determined. Ritchie (2017*b*) explicitly argues that data anonymization be seen as the residual control.

4. Summary

The above discussion does present a slightly artificial view of events. Even in 2000, there were organizations that treated users as humans, understood the subjectivity of risk judgments, and did not confuse objectives with constraints. Equally, there are organizations now that still appear not to have learned anything over twenty years. Some of the same concerns still exist (statistical agency worries about response rates), some have shrunk (the risk of remote

working, post-Covid). Others have grown (the increased identifiability of administrative data).

Nevertheless, the broad change in attitudes and conceptual frameworks can be seen across organizations and countries in recent decades. It is not universal, and few organizations apply all the elements listed above; but the direction of travel has been fairly constant.

It is also reflected in the laws that are being passed around the world. Twenty years ago, we anonymized data; that was our main protection, and it was reflected in laws which distinguished simplistically between anonymous data and personal data. More recent legislation, such as the European General Data Protection Regulation, the UK Digital Economy Act or the Australian Data Availability and Transparency Bill, reflect the portfolio approach described above.

In summary, there is a conflict in theory between privacy and microdata access, but not in practice. Privacy is not the sole public benefit criterion, and the data holding agency's interests are not paramount. Much more importantly, we know a lot more about how to manage privacy: we know a lot more about risks, we are much better at evaluating them, and we are much better at dealing with them. And attitudes and laws are changing in response.

References

Adams, D. 1979. *The Hitch-hiker's Guide to the Galaxy.* Pan Books.

Bhatta, G. 2003. "Don't just do something, stand there! Revisiting the issue of risks in innovation in the public sector"." *The Innovation Journal*. http://innovation.cc/scholarly-style/bhatta-risks.pdf.

Desai, T., and F. Ritchie. 2010. "Effective researcher management"." Eurostat Work session on statistical data confidentiality. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2009/wp.15.e.pdf.

Diepeveen, S., and J. Wdowin. 2020. "The Value of Data: Policy Implications report – Accompanying Literature Review." Bennett Institute for Public Policy mimeo. https://www.bennettinstitute.cam.ac.uk/publications/value-data-accompanying-literature-review/.

Eurostat. 2016. "Self-study material for the users of Eurostat microdata sets." https://ec.europa.eu/eurostat/web/microdata/overview/self-study-material-for-microdata-users.

Green, E., and F. Ritchie. 2016. "Data Access Project: Final Report." Australian Department of Social Services. http://eprints.uwe.ac.uk/31874/. June.

Green, E., F. Ritchie, J. Newman, and T. Parker. 2017. "Lessons learned in training 'safe users' of confidential data." UNECE worksession on Statistical Data Confidentiality. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2017/7_lessons_learned_training.pdf.

Opperman, I., ed. 2018. *Privacy in Data Sharing: a guide for business and government.* Australian Computer Society.

Ritchie, F. 2014. "Access to sensitive data: satisfying objectives, not constraints"." *J. Official Statistics*, 30(3): 533–545,. https://doi.org/10.2478/jos-2014-0033.

Ritchie, F. 2017*a*. "Spontaneous recognition: an unnecessary control on data access?" European Central Bank ECB Statistical Papers 24, https://doi.org/http://www.ecb.europa.eu/pub/pdf/scpsps/ecb.sps24.en.pdf.

Ritchie, F. 2017*b*. "The 'Five Safes': a framework for planning, designing and evaluating data access solutions"." *Data For Policy Conference*. https://doi.org/10.5281/zenodo.897821.

Skinner, C. 2012. "Statistical disclosure risk: separating potential and harm." *International Statistical Review*, 80(3): 349–368. https://doi.org/10.1111/j.1751-5823.2012.00194.x.

Yang, K., and M. Holzer. 2006. "The Performance–Trust Link: Implications for Performance Measurement." *Public Administration Review*, 66(1): 114–126. https://doi.org/10.1111/j.1540-6210.2006.00560.x.