Access to sensitive data: satisfying objectives rather than constraints

WISERD DATA RESOURCES

WISERD/WDR/007

Felix Ritchie

February 2012

Authors
Felix Ritchie

Address for Correspondence:
Trig Consulting Ltd
Cardiff

Email: felixritchie@trigconsulting.co.uk

WISERD Hub Contact:

Cardiff University
46 Park Place
Cardiff
CF10 3BB

Tel: 02920879338

Email: wiserd@cardiff.ac.uk

Abstract

The argument for access to sensitive unit-level data produced within government is usually framed in terms of risk, and the legal responsibility to maintain confidentiality, even where the government has a duty to provide data. This paper argues that the way the question is framed may be restricting the set of possibilities; and that the correct perspective needs to start from a more abstract plane, focusing on the data owner's principles and user needs.

Within this principles-based framework, the role of law and risk changes: they become enabling technologies, not setting the objectives but helping to define the solution. For law, this perspective allows for both more flexible solutions within the current legal framework, and a coherent way to assess changes to that framework. Similarly, changing the focus to treat risk as just another aspect of data management rather than some absolute standard gives the data owner the conceptual background within which to place the different ways to achieve an acceptable risk-utility balance.

Focusing on the objectives rather than the constraints also encourages the data owner to engage with users and build a case for data access which takes account of the wider needs of society.

Acknowledgements and disclaimer

1. Introduction

It is nowadays widely accepted that access to confidential or sensitive microdata collected by government is essential for the research needed to produce an evidence base for policy; see Trewin et al (2007) or Ritchie (2010) for a discussion. This data is usually collected either by statistical agencies to produce aggregates, or by government departments as part of their work. In both cases, use of the microdata directly allows the collecting body to leverage their investment in data collection at minimal additional cost.

General agreement on this principle is common; but principles can take a back seat when implementation is considered. In particular, the confidentiality of the data becomes paramount and access to data focuses on how that confidentiality can be maintained.

This paper argues that a wider perspective on the principles governing data access can help to improve the quality of decisions taken, as well as clarifying exactly what risks are being run and what the benefits are.

The next section proposes a perspective on access which emphasises the predominance of objectives over constraints. This leads to a model where law, technology and risk are all 'enablers': that is, they inform, and may constrain, decisions to be taken on how to implement an objective, but not the objective itself. The following three sections discuss these in more detail. Section 6 considers how this change of perspective can inform the debate on the 'public goods' problem of data access identified by Ritchie and Welpton (2012). The paper concludes by noting that the arguments advanced here run counter to the natural decision-making structures in government, and so an efficient system of confidential data access may need an active and engaged sponsor.

For simplicity, the paper throughout refers to the options of National Statistics Institutes (NSIs), who are generally the main or only holders of confidential survey data. However, it should be clear that the arguments apply to any owner of confidential data considering giving access to that data for research.

2. The framework principles

The usual decision-making process for giving access to confidential data can be framed as in Figure 1, which we will refer to as the 'constraint model'.

Figure 1: The 'Constraint Model'
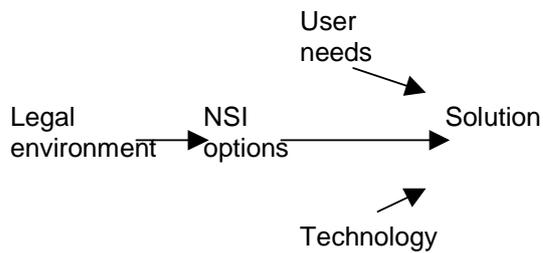
That is, organisations ask:

- what does the law say we can do?
- given that, what do we want to do?
- what technologies are available, and what are the needs of the user?
- what solution results?

The problem with the constraint model is the first step. Clearly, acting within the law is a requirement of any agency. The problem is that 'the law' is rarely a simple, unambiguous construct with only one possible outcome; any statement of law is an interpretation in relation to a specific set of circumstances. However, focusing on a particular interpretation constrains the set of solutions to a subset of outcomes, particularly if the circumstances surrounding the interpretation are not explicitly made known.

Consider the UK experience in 2003. The Office for National Statistics (ONS), the UK's NSI, was reviewing the options for giving academic researchers access to confidential business data. The initial legal opinion was that this was not possible: the Act governing such access strictly limited access to employees of the UK government.

This seems crystal clear, until it is considered that the question being asked is the implicit one, "can academics, in their own right, have access to business microdata?" This is a very specific and, as it turned out, very limiting question. An alternative question was put to the government legal advisors: "can academics become Civil Servants for the purposes and duration of their research?" There turned out to be several possible answers to this question, and of those a form of secondment was taken as the most workable. As a result of changing the perspective, an outcome, previously considered impossible, was achieved with a solution in keeping with both the spirit and letter of the law. For details, see Ritchie (2004, 2009).

The specific legal arrangements continued to evolve as different questions and circumstances came to light. ONS' legal services unit periodically reviewed the secondment arrangements, and the team providing access was required from time to time to amend its procedures to address potential areas for challenge. For example, the team was asked to develop clear guidelines for the 'fair and open competition' for access, and to specify the criteria for determining whether access contributed to ONS' 'benefit'.

The important lesson from this is that the attitude of the NSI determined the outcome; that is, whether access could be granted or not. Both the research team and the legal team shared the same aim, to see wider research use being made of confidential data, lawfully. The specifics of implementation were just that: specifics of implementation, not a universal statement of law.

This focus on objectives rather than implementation leads to a rather different framework for access, as displayed in Figure 2:

Figure 2: The 'Objective Model'
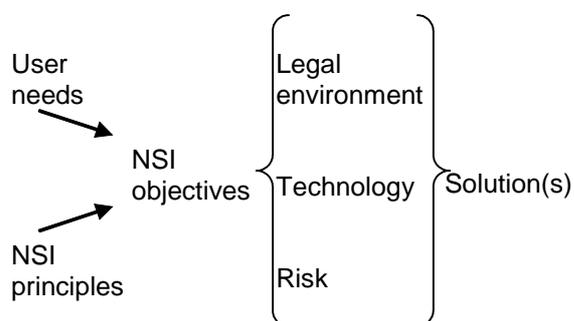
The questions now are, in order,

- What are our operating principles, and what do users want?
- What how do we turn this into a set of objectives?
- What legal and technological options are available? What risks arise and can be addressed?
- How do we employ these alternatives to meet the objectives in the best way?

This 'Objective Model' puts the aims of the NSI and the user at the start of the decision process. Law has the same status as technology: just as all implementations are limited by the existing technology, so they are constrained by the existing law. But technology and law

are both used to achieve the objectives; they do not count towards the definition of those objectives. Risk also becomes something to be evaluated and managed .

One of the implications of the Objective Model is that multiple legal and technological solutions may meet those objectives; there is no need to identify 'the' legal or technical solution. The solutions might also co-exist; the aim is to find the combination of solutions that meets the objectives best.

It could be argued that this is a largely semantic argument; that is, the real questions are always about implementation, and the same solution could be derived by individuals working from the different models. However, this misses the point: the two models imply a fundamentally different mindset, the difference between "what can we do?" and "what would we like to do?"[1]

The 'user need' also takes prime position. The logic for this is straightforward: if there is no user need (and the NSI should consider itself as one class of user), then the NSI objectives might be set but there is no necessity to fix a solution. User demands can be stereotyped (eg Ritchie and Welpton, 2011) as "give me all the data now, on my desktop, with no restrictions", but this is an exaggeration. Researchers are generally aware that not all tasks need all data, particularly as more detail typically involves more restrictions. As an example, the UK Data Archive provides many datasets in both anonymised and detailed files, with the latter having more access restrictions. A bona fide UK researcher would have little trouble getting access to either, but the usage of the anonymised files massively outstrips that of the restricted-access detailed files, indicating that users can make balanced judgements about costs and benefits.

The dichotomy characterised by the Constraint/Objective models may be unfair to individual NSIs, but in the author's experience, based on visits to numerous NSIs and discussions with colleauges, this state does persist in the real world. There are units within NSIs that consider user needs and then consider how to meet them; but the majority still seek to identify the legal framework and then assess which user needs can be accommodated within that framework. An even smaller minority are prepared to consider NSI objectives and user needs jointly without reference to legal limits on implementation.

---

[1] This is analagous to the primal/dual distinction in constrained optimisation. For example, maximising production subject to a budget constraint can produce the same specific outcome as minimising expenditure to achieve a level of production, for a given cost structure. However, the process by which this is achieved, the options available to the producer and the implicit cost of the constraints differ between the two processes, reflecting the objectives of the producer.

The next three sections discuss what is meant by being an 'enabler' in more detail.

3. Law as an enabler

Law as an enabler has three components: recognising that law is 'how', not 'why'; providing a rationale for and measure of changes in the law; and challenging embedded procedures which are often treated as law.

The example of the ONS above illustrated the first of these components: once the decision was taken to give access, the problem resolved itself into identifying and applying relevant legislation.

The second component is putting changes in law in the proper context. The Constraint Model suggests that changes in the law will change the NSI's objectives; how then do we know that changes in the law will lead to an improved set of objectives? In the Objective Model, NSI objectives are invariant to law; therefore, a test of the likely effectiveness of any new law is whether it improves the way the NSI meets its objectives.

Again taking the case of the UK, in 2007 the Statistics and Registration Act was passed. This formally gave ONS a function of supporting research for the public benefit, and provided a simple universal legal gateway for access to ONS microdata. This simplified enormously the process through which researchers gained access to the ONS Virtual Microdata Laboratory, and clarified the role of researchers' use of ONS data. ONS objectives were largely unaltered.

The third component is to clarify exactly what is law and what is merely established practice. In the context of the diagrams above, 'law' includes the NSI's procedures, which often go beyond the law into areas where the NSI feels it has an ethical or operational responsibility even if no legal responsibility exists. Fixed ways of working, particularly when in place for a long time, can also easily be confused with law. Even when procedures are explicitly recognised as NSI policy decisions, they can still be seen as immutable.

Under the Constraint Model, challenging established practice is hard. An objective which requires a change to established practice needs to be demonstrably preferable to another, more achievable, objective. This is difficult because the costs and benefits of objectives are often hard to quantify. But it also begs the question: why should objectives need to justify

their value by reference to specific implementations? Surely the implementation has to address the objective, not the other way around?

Under the Objective Model, established practice has to justify its existence using fair and proper criteria: cost, benefit, effectiveness, impact of disruption against the alternative solutions. These are also more easily quantifiable: the impact of a change on access rules on IT expenditure, for example, can be readily identified. Under the Objective Model cost-effective practice is what matters; the value of 'established practice' is only reduced costs of learning or change.

4. Technology as an enabler

Technology (meaning all the practical matters surrounding access to data) as an enabler is relatively self-explanatory. The technological options can be broadly grouped into six types (see Ritchie, 2011a, for a more detailed review):

- *Anonymisation* of the data: this is used for public files, such as those on the web
- *Licensing* of researchers, sometimes combined with a degree of anonymisation, is still the most common way for researchers to get access to microdata. Between countries there are huge differences in the degree of anonymisation.
- *Secure 'research data centres'* (RDCs), laboratory facilities at the NSI or the researcher's base. For many countries, this is still the only way to get access to detailed data
- *Remote access*, where 'virtual' RDCs allow users to manipulate data unhindered by geography. Again, there are differences in the degree of 'remoteness': the ONS' VML can only be accessed from central government departments, whereas the Dutch NSI's system and the NORC Data Enclave, using the same technology, allow secure access across the internet.
- *Remote job submission*, where users send statistical programmes to be run and get back results, are relatively uncommon but a number of NSIs have been exploiting web technologies to develop friendly interfaces
- *Synthetic data*, which has the same characteristics as the real data but has been imputed from statistical models. As it no longer refers to individuals, it poses no disclosure risk.

NSIs tend to be risk-averse and avoid new solutions, but in most of these areas a prospective data manager can draw on a wealth of international experience in implementation (synthetic data being the exception).

The everything-is-possible answer does not help planning. Ritchie (2011b) elaborates on the concept of the 'data access spectrum', whereby a finite variety of options, all assessed under the so-called 'VML Security Model' (see below), provide different combinations of accessibility and detail. The user can then determine which is most appropriate. There is also scope to expand, but this is based on a cost-benefit assessment. This model has been used both to classify existing operations and to justify the development of two new delivery mechanisms, a third–party remote access system and an improved off-site RDC model; see Ritchie (2011b).

5. Risk as an enabler

Risk as an enabler may seem less straightforward as a concept. NSIs sometimes state that they have an 'appetite for risk' which needs to be satisfied, irrespective of the implementation. This argument fits well with the Constraint Model: NSIs aver that have a legal duty to maintain confidentiality to some 'reasonable level' (that is, an acceptable level of risk), and that the requirement must be satisfied absolutely.

A popular representation of this argument is the risk-utility map, as in Figure 3. This is often used in discussing access to data (eg Lane and Schur, 2010).

Figure 3: Risk and utility
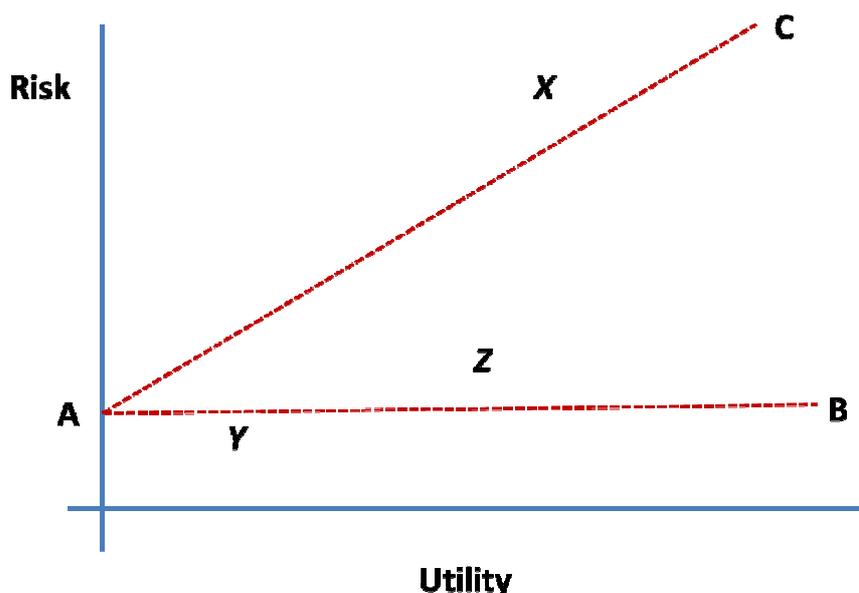
The statement that an organisation has 'a' level of acceptable risk implies a risk-utility frontier (R-UF) such as the line A-B, where the risk is an absolute value not to be exceeded.

For example, original data at point X is an unacceptably high level of risk. One way to reduce this risk is to reduce the utility of the data: recoding, top coding, record swapping, etc. This may be sufficient to reduce the level of risk to Y, but the utility of the data is low.

However, most of the literature draws the R-UF as an upwards sloping line, such as A-C[2]. In other words, it is generally accepted that risk is something to be balanced against utility. Some lower level of perturbation could produce point Z, where the risk is slightly higher but the utility approaches that of the raw data.

In practice, the risk profile of an organisation is a function of the solution it implements, not an abstract standard. As Ritchie (2011b) notes, all NSI data is initially at very high risk of disclosure. An organisation has many options to reduce this risk: by not releasing the data at all; by anonymising; by restricting access to particular sites and/or persons; by technical solutions; and so on.

For example, the risk of misuse of ONS Labour Force Survey data is reduced by

- employment policies (for raw data)
- technical measures, de-identification and researcher training (for RDC access)
- anonymisation (for Public Use data)
- licensing and limited anonymisation (for Scientific Use files)

ONS believes that all of these different approaches provide an acceptable level of risk – otherwise it would need to change its policies. In that limited sense, there is a corporate 'appetite' for risk. But clearly risk is being managed differently depending upon the need and the settings. It would be very difficult to say that the risk, in any objective sense, is the same across all access mechanisms, as assessing and comparing these risks is difficult. While there are methods for calculating residual risk in anonymised datasets, there is no relevant metric for the risk-reduction achieved by a researcher training programme or by a new firewall.

Risk-reduction measures also have different cost profiles. For example, anonymising a dataset or training researchers is an upfront investment; running a secure facility is an ongoing cost.

---

[2] Including NSIs; see for example, papers presented at the biennial UNECE workshop on confidentiality

Hence decisions about risk are choices, not constraints. An NSI can balance how costly a solution is against how well it meets it objectives and the needs of users. Risk is also tightly bound in with the other enablers: decisions on risk affect technological options and the legal basis, and vice versa.

6. Objectives, constraints and the public goods problem

As well as affecting the decision-making process, the Constraint/Objective model dichotomy has implications for addressing one of the key issues: how much research data should be made available?

Ritchie and Welpton (2012) argue that one reason why NSIs tend to focus on protecting data rather than maximising value is a 'public goods' problem arising from the unequal distribution of risks and benefits. The benefit from making confidential data available for research largely accrues to the wider public, but the risk of being blamed for something going wrong is typically borne by the NSI. For example, if a licensed user is sent a confidential dataset and loses it, the NSI may well get blamed for distributing the data however well-founded its distribution policy is. In contrast, if data is not released, or is only used by the NSI for its own purposes, then the NSI minimises risk; but the wider public loses the benefit of that data and runs an increased risk of bad decision-making.

In these circumstances it is rational for NSIs to take a cautious approach to data release, and consider their own priorities over the wider public benefit. The NSI's main function is to protect its interests: risk avoidance becomes the goal, a conservative legal stance appeals, and the Constraint Model predominates.

Ritchie and Welpton (2012) argue that one way to address the public-goods problem is to 'negotiate' the level of access with users; as part of that negotiation, issues of risk and responsibility are also addressed. This is entirely consistent with the Objective model , which puts agreement with users at the forefront of the decision-making process and views risk as something to be managed, not minimised. For the NSI to set its objectives, it needs to consult with users – and to get the buy-in necessary to ensure a collective responsibility for the data release policy. If that buy-in is not forthcoming, perhaps the NSI is justified in ignoring those user needs? Hence the Objective Model is consistent with the customer engagement necessary to avoid underprovision of data access from society's perspective.

7. Conclusion

This paper has dichotomised the decision-making process for data access into the 'Constraint Model' and 'Objective Model'. Whilst this is clearly an oversimplification, it nevertheless usefully illustrates some different approaches to setting the objectives and solving problems associated with data access. In doing so, this idealised worldview also suggests that NSIs may be missing opportunities for both their benefit and the wider public.

This paper has argued that access to the data collected by NSIs and similar bodies is often unnecessarily restricted. This is because NSIs decision-making processes tend to focus initially on what is allowed rather than what is desirable; and the incentives for NSIs do not encourage exploration of what is allowed.

An alternative perspective focuses on the NSI objectives, and uses this to address questions of constraints in implementation, rather than the other way round. In this perspective, law, NSI procedures, and technology all become 'enablers': options for or constraints on implementation which affect the delivery of objectives, but not the objectives themselves. Risk is also included as an enabler; the acceptable level of risk implied in a solution is a function of the objectives, not the other way round.

This perspective also provides a framework to bring users into the discussion on access principles, increasing the chance of community buy-in and reducing the NSI's incentives to implement an overly risk-averse release policy. The objective-based world-view opens up the NSI to wider and deeper engagement with users.

However, this strongly user-/objective-centric approach runs counter to the natural decision-making structures in government, which largely reflect the Constraint Model. An efficient system of confidential data access may therefore need active and engaged sponsors to have any realistic prospect of success.

References

Lane J. and Schur C. (2009) "Balancing access to health data and privacy: a review of the issues and approaches for the future", *Health Services Research* October v45:5.2 http://www.hsr.org/hsr/abstract.jsp?aid=45965869665

Ritchie F. (2004) "Business Data Linking – Recent UK experience", *Austrian Journal of Statistics* v33:1-2 pp89-97

Ritchie F. (2009) "UK Release Practices for Official Microdata", *Int. J. of Association of Official Statisticians*

Ritchie F. (2010a) *Use of confidential microdata for social and economic policy research: lessons from the UK experience*, mimeo, ONS/Motu working paper

Ritchie F. (2010b) *Risk assessment for research access to sensitive microdata*, presentation to 3rd Workshop on Data Access, May

Ritchie F. (2011a) *Methods for analytical access to confidential data*; paper prepared for OECD Working Group on Microdata

Ritchie F. (2011b) *Provision of ONS data for analysis: safe use not safe data*; mimeo, ONS

Ritchie F. and Welpton R. (2011) *Incentive compatibility and data security*, mimeo, ONS

Ritchie F. and Welpton R. (2012) "Data access as a public good" in *Work session on statistical data confidentiality 2011*, UNECE/Eurostat, forthcoming

Trewin D., Andersen A., Beridze T., Biggeri L., Fellegi I., Toczynski T., (2007) *Managing statistical confidentiality and microdata access: Principles and guidelines of good practice*; Geneva, UNECE /CES, 2007.