# Designing a national model for data access

Felix Ritchie

**Abstract**

Access to detailed confidential microdata, often from official sources, is acknowledged to be one of the main constraints on the development of much industrial and labour economics. Recent developments across the world have improved access for some countries considerably, but progress is still patchy. Some of this delay is due to fear of the new; concerns over risk; worries over cost or feasibility; or just simply how to choose between a myriad of different solutions.

This paper considers the development of access to confidential microdata using the UK as an example. The UK Office for National Statistics (ONS) has developed a coherent framework for providing data: everything from internet access to secure labs can be considered within the same framework. The aim is to achieve a balance of cost, access and detail through a finite set of options, enabling both flexibility and economies of scale.

A central role is played by the ONS's remote access facility. This is due to be complemented in 2009 by an academic equivalent. This raises questions about the relevant level of scale for these operations, and the paper reports on a cost-benefit analysis carried out to try out ascertain both the value of remote access systems and efficient alternative methods of operation.

**Key words:**
Microdata; confidential data; remote access

**Corresponding author:**
Felix Ritchie
Microdata Analysis and User Support
Office for National Statistics
Newport, South Wales
felix.ritchie@ons.gsi.gov.uk

## 1. Introduction

Access to detailed confidential microdata, often from official sources, is acknowledged to be one of the main constraints on the development of much industrial and labour economics. Recent developments across the world have improved access for some countries considerably, but progress is still patchy. Some of this delay is due to fear of the new; concerns over risk; worries over cost or feasibility; or just simply how to choose between a myriad of different solutions.

The ground for making decisions has changed considerably over recent years. The value of microdata in its own right (as opposed to the construction of aggregates) is widely accepted (eg Trewin et al, 2007); legal positions are being changed or reviewed; and technological developments have made a much wider variety of solutions possible. An example is the rediscovery of the research data centre (RDC). Originally conceived as the only secure way to provide access to confidential data, RDCs fell out of favour as personal computers became available and distributed data became the norm. In recent years the development of thin-client technology has overcome the drawbacks of physically restricted sites, allowing the balance to shift from distributed data to distributed access.

In the UK, the Office for National Statistics (ONS) is the primary source of government microdata available for research. In recent years, ONS' approach to releasing this microdata has undergone a searching review in all areas: technological, procedural, legal, ethical, and financial. Increasingly, ONS has been refocusing on first principles – not just to ensure that access is appropriate, but as a way of developing sustainable solutions which can be applied consistently across a range of environments.

This paper sets out how a model for national data access may be defined, using ONS as an example. This model does not explain how ONS got to its present position – while some parts of the ONS system development were preceded by strong theoretical foundations (eg the Virtual Microdata Laboratory, or VML), others grew more organically (eg the 'data access

spectrum', or the VML SDC model) and it is only with hindsight that a coherent framework appears.

The aim of this paper is to use that hindsight to consider the issues surrounding the development of the framework for data access, illustrating it with the UK experience. A more extended discussion of the latter can be found in Ritchie (2009a).

## 2. Designing the principles

### 2.1 Definitions of principle

The basic data access problem is:

| NSI | <=> | users |
|---|---|---|
| *what have got?* | | *what do they want?* |

Much of the concern about providing access to data is couched in legal terms: this or that is not allowed by law. Law is the cornerstone of access. Diagrammatically, we can consider the decision making-process for data access:

| NSI | => | law | => | principles | => | technology | => | implementation | => | users |
|---|---|---|---|---|---|---|---|---|---|---|
| *what have we got?* | | *can we give access?* | | *what are trying to do?* | | *what options are available?* | | *how do we make this happen?* | | *what do they want?* |

This is a comforting view of the world. By ensuring that the legal basis for any access is considered first, it ensures that all subsequent actions are within the powers and competencies of the NSI.

This is also entirely the wrong way round, for a variety of reasons.

First, considering the legal position first restricts thinking to operating within the current framework. As that framework was typically defined some time ago and the situation being considered is novel, it is quite possible that the laws in place are inappropriate to data access problem; but an unwillingness to think outside the current legal framework may embed make-do practices.

Second, laws are not immutable; they are human constructs, and can be changed. This may be easy or hard; but it is certainly possible.

Third, laws are not always clear and unambiguous; otherwise there would be no need for lawyers and judges to interpret them.

Fourth, a law requiring, allowing or preventing a specific action may appear to limit solutions, but this may be because the wrong question is being asked.

Finally, not all ‘law’ is law. Many practices are the result of custom, which are so ingrained that they seem to be law.

This mutability of law is well illustrated by the UK experience in recent years. Academics were denied access to business data as this was restricted by statute to Civil Servants only. ONS had a number of responses to this. The first was to address the interpretation of the law; the lawyers consulted suggested a number of alternative ways of providing access. However, one of the most obvious was to change the question: not “can we give academics access to the data?” but “can we make academics Civil Servants for the purposes of research?”. On the purposes for which data could be made available, repeated enquiries showed that often ONS protocols were confused with law (at least by the data owners, if not by the legal team). Finally, having got a system which worked within the existing framework but not satisfactorily,

the law was changed to reflect changes in the statistical system since the last major round of lawmaking – and this included research data access.

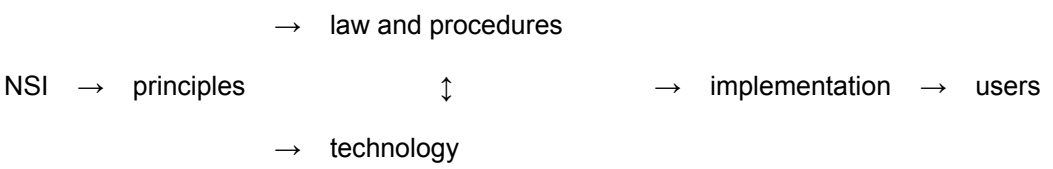
Crucially, the parts of the new law relating to research access were informed by the experiences of the previous years: in particular, the administrative burdens that needed to be overcome to achieve the enormous benefits of data access. In other words, had the ONS not already challenged received wisdom and pushed for a novel solution, then it would have been difficult to make the case for changing the law.

These considerations emphasise that the legal framework is the wrong place to start. Instead, data access models need to start from first principles: what are we trying to achieve?

Once we start designing a target outcome, the position of law becomes clearer: it is one of the constraints on the implementation, which may or may not be fixed. In this way, law plays a similar role to the technical implementation: we are limited by the technology currently available, but there may be different ways to apply it; and it can develop over time to reflect changing circumstances.

This view of the world also allows much more explicitly for the interplay between law and technology. For example, maintaining the ‘confidentiality’ of the data is a crucial part of any consideration, but it is not a simple matter of defining it in law. Nor is it a purely technological matter. Instead, as described below, an effective security model must take account of technological, psychological and legal protection mechanisms.

Hence, the decision-making process now takes a much more complex form:



This different perspective on the world is particularly important for international data sharing agreements. As law between countries is so variable (even within Europe, which shares a common legislative framework), points of agreement must lie on principle first. See Ritchie (2009b) for an extended discussion.

2.2 Principles and the legal basis for access in the UK

As noted above, access to data in the UK has grown organically. One of the changes was that a recognition of the inappropriateness of existing laws led to changes in the statutory framework. The ONS model is now in alignment with the 'principles-based' reasoning outlined above, and the following discussion describes how the new framework operates. nevertheless, it must be remembered that the ONS model did not did not grow that way; much of the alignment with the principles-based approach is due to a repeated assessment of the way things are done.

Research access to confidential data raises two distinct, but related, legal matters. First, there is the matter of privacy. Second, there is the matter of maintaining confidentiality.

The Human Rights Act (1998) gives citizens a conditional right to a private life. It is conditional, because this right can be interfered with where the interference is lawful, necessary and proportionate in a democratic society. Research access to confidential data is a further interference with the privacy of respondents beyond the original interference for the production of official statistics. Research access must therefore be judged separately as lawful, necessary and proportionate. Further, the Data Protection Act (1998) requires that processing personal data is fair, in that the respondent should be informed about onward disclosure of their personal data to researchers. ONS regulates the privacy impact of research data access by requiring that all research access arrangements are scrutinised and approved on a case-by-case basis by a senior panel of statisticians, taking account of the

recommendation of the source manager and ONS Legal Services. The Panel is called the Micro-data Release Panel, and it acts with the directly delegated responsibility of the National Statistician. Typically, failures in the right to privacy arise from faulty policy and/or decision making.

The UK Statistics and Registration Service Act (2007) has simplified the confidentiality legislation for ONS data. All ONS personal information sources are now protected by the same statutory prohibition on disclosure; Section 39.

Non-ONS sources of data are protected by the legislation under which they were originally collected, unless by Order they are transferred to ONS and fall under the above prohibition.

Many data sources are not collected under legislation. Often, the law of confidentiality will protect voluntary collections. ONS and any other public authority that collects data are bound by precedents set over the centuries by our Courts. The effect of these judgements is to impose a legal duty on ONS to treat citizen and enterprise private information in confidence, and any disclosure of such information may give rise to an action against ONS for a breach of confidence. Thus survey pledges effectively set down limits on the use of data because a Common Law duty of confidentiality is always present. Typically, the duty of confidentiality and/or statutory prohibitions on disclosure are threatened by failures of data handling.

In the past this led to a number of difficulties and inconsistencies. Legitimate research access to data could be blocked while administrative use of the data was permissible (in principle). Researchers wishing to access data often had to jump through many hoops justifying the legitimacy of their access – and these hoops were different for different datasets. Sharing data between government departments for research or administrative use was fraught with legal argument.

In 2007 the Statistics and Registration Services Act (SRSA) was passed. In recognition of the difficulties of supporting research under the existing legislation (and with increasing

awareness of the demand for and value of research use of ONS microdata), improving the research environment was one of the key aims of the SRSA.

The SRSA gave ONS a power to support research where to do so is consistent with its objectives to serve the public good. This removed one of the key barriers to research – the need for researchers to prove that their work was for the benefit of ONS, the Registrar General, or another government department. All that is now required is an easily established 'public good'..

The SRSA did not abolish any existing gateways for research, although many now are effectively redundant. However, it did create one new gateway – the principle of 'Approved Researcher' (AR). An AR is someone deemed 'fit and proper' to access confidential microdata for research, and who then potentially has access to all ONS data (with one exception). This enables all existing gateways to be brought together into a single process.

The exception is for cross-government administrative data. Under the SRSA, ONS can now receive administrative information from any other government department, subject to Parliamentary approval. However, access by non-ONS researchers is not allowed unless specifically agreed at the time of the approval.

The net impact is that, in theory at least, there are no limits on who can have access to which data for research purposes, as long as that research is in the public interest. This applies to individual researchers in government, academia, and the private sector.

The new powerful statutory regime under the SRSA means that decisions for access are founded primarily on a combination of assessments of data handling risks to confidentiality, and determinations of proportionate and necessary interference with privacy.

The assessments of data handling risk and privacy impact are resulting in an emerging and evolving data model.

## 3. Designing the data/risk model

Once principles of data access have been decided, the next step is to broadly outline how those can be fulfilled. This is an important stage between principles and implementation. The idea is to take a holistic view of data access, to study how multiple implementations could work together the achieve the principles in the most cost-effective way.

This section reviews this broad-brush planning; using extensive examples from ONS to illustrate how data and security models can be realised.

### 3.1 The data model

Data release can be seen as a ‘spectrum’ of access points balancing

- value of data
- ease of use
- disclosure risk

The aim is, for a given level of confidentiality, to maximise data use and convenience. There are no off-limits, no absolute prohibitions. Anonymised data is included in this spectrum, but so is fully identified linked data. There are of course trade-offs. The anonymised data may be available on a website; the fully identified data is only likely to be available under stringent conditions.

Having an infinite variety of options is not feasible, and so the NSI can only concentrate on a small number of solutions. The aim is to cover as much of the ‘use spectrum’ as cost-effectively as possible, while maintaining an effective level of risk. With the trade-off between ease of use and detail made explicit to the researcher (who is assumed to be uninterested in risk), the researcher can make informed decisions about which data solutions will meet his or

her own research needs and resources. From an NSI perspective, operating a limited set of options means that economies of scale can be achieved in the delivery of solutions.

The importance of this world-view is that there is an explicit acceptance that a one-size-fits-all policy is unlikely to be appropriate. There is also an implicit acceptance that the NSI might not have got the balance right. However, users wanting options outside the standard set will need to provide evidence of a more appropriate (and cost-effective) solution. In other words, the set of options is not fixed - nothing is ruled out - but the onus is on the researcher to demonstrate the value of alternative solutions.

For example, the ONS's secure data facility, the Virtual Microdata Laboratory (VML) is only accessible from government sites. For ONS the cost of providing direct access to the VML from the academic network is prohibitively expensive. However, the academic community has invested in a similar facility, the Secure Data Service (SDS), which from 2009 will be able to provide a very similar service with similar levels of security. The SDS has received approval, in principle, to be a host for ONS datasets. ONS supports the proposal because it raises the possibility of improving access significantly at negligible cost and with a manageable security risk. Thus an additional access point will appear in the spectrum, at the behest of users who were able to demonstrate a solution that met all three criteria.

As an example, the spectrum at the beginning 2009 looked like:

Table 1: The data access spectrum in the UK

| | More detail ↔ easier access | | | | | | |
|---|---|---|---|---|---|---|---|
| | **ONS survey managers** | **VML (ONS sites)** | **VML (govt. offices)** | **SDS*** | **Special arrange-ments** | **Licensed data archive** | **Internet** |
| **Census data** | Source data | Identifiable samples | | | Dedicated facilities | Anonymised microdata | Summary tables |
| **Enterprise** | Source | Identifiable | Identifiable | Unlinked | Secure | | Summary |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **data** | data | data | data | data | areas in other govt. offices | | tables |
| **Household data** | Source data | Identifiable data | Identifiable data | Identifiable data | | Anonymised microdata | Summary tables |

*SDS not fully operational yet. Table based upon likely content

The options are not ideal for every case, but the jump from one solution to another reflects data utility and patterns of research use. In practice, while researchers always want more detail more easily, the ONS model is generally seen as providing a reasonable set of options. The model is transparent and, on evidence to date, reasonably efficient.

The question naturally arises, how to decide what data goes in which box? The key to this is the recognition that technology does nto provide security on its own; human characteristics need to be brought into play. There is therefore a need for a security model as well as a data model.

### 3.2 The security model

A modern security model needs to starts by recognising that

- no solution is 100% safe
- different protection methods have different risks
- different risks do not necessarily mean greater risks
- different protection methods constrain operations in different ways
- protection methods can be independent yet complementary
- the complete risk profile needs to be assessed in any solution

In other words, security is a composite concept with complex relations between the component parts.

To turn this into a useable structure, ONS has identified five major components each of which has a part to play in guarding against either accidental or deliberate misuse:

| Criterion | Meaning | Guards against |
|---|---|---|
| Safe projects | The project has been reviewed to ensure that it has a valid research aim | Deliberate misuse |
| Safe people | The researchers can be trusted not to misuse their access | Deliberate misuse |
| Safe data | The data has been treated to limit disclosure risk | Deliberate misuse |
| Safe settings | A technical solution limits the options for misuse of data | Accidental/ Deliberate misuse |
| Safe outputs | Checking of outputs produced by researchers to reduce the chance of accidentally identifying respondents in statistical outputs | Accidental misuse |

Any solution (or set of solutions) can then be assessed against these five criteria. The idea should be that the security level for all is the same: negligible chance of confidential data being circulated. But the multi-part criteria allow this to be met in different ways for different levels of utility and convenience required by the user.

In terms of the data spectrum advanced above, the options would be assessed as follows:

| ***Safety criterion*** | ***VML*** | ***SDS*** *(provisional)* | ***Special arrangements*** | ***Licensed Data Archive*** | ***Internet*** |
|---|---|---|---|---|---|
| *People** | ONS-trained | ONS/SDS-trained ARs | ? | UK academics | Anyone |

| | | | | | |
|---|---|---|---|---|---|
| | ARs/ Civil Servants | | | | |
| *Projects* | Scrutiny by ONS | Scrutiny by ONS | Scrutiny by ONS | Self-certified academic projects | None |
| *Data (in theory)* | Any | Unidentified | N/A | Anonymised, almost no risk of identification | Anonymised, no risk of identification |
| *Settings* | Secure client from secure government office | Secure client from academic setting | ? | Use in academic setting only | None |
| *Outputs* | ONS staff checked | SDS staff checked, ONS guidelines | ? | No checking | No checking |

* AR = Approved Researcher

Thus for example, web data places its protection solely on safe data, because it is recognised that none of the other safety measures are applicable to users downloading from the internet. Data archive downloads are however controllable to some extent, and so it is possible to get a commitment from researchers to be 'safe'. On this basis the protection in the data can be relaxed somewhat, as it is counterbalanced by a decrease in the risk that individuals will misuse the data.

At the other end of the spectrum, the VML is designed to hold the most confidential data, and so safe data is not a relevant protection mechanism. Instead, all four other criteria are brought into play. The complementarity comes in a number of ways. For example, while the VML has

passed two external security reviews and an internal security audit, it is designed to facilitate use of data not limit it. It is possible that an experienced hacker could, eventually, find and exploit technical flaws in the system. However, the need to prove one's statistical bona fides before being granted access to the VML reduces this risk to an acceptable level, a fact acknowledged in the security reviews.

**4. Choosing the implementation**

It is far to say that there is a wide range of experience on most implementation modes, particularly for non-confidential data. Delivery of aggregates tables over the web (and the associated SDC methods used to ensure that the data is safe) is well established. Making data files non-disclosive is, again, a well-understood area. Modern data archives are well-established but still efficient and innovative.

For access to confidential microdata, there is more variation; but this is largely because this area has really developed in the last few years and is still going through major leaps. There are four real options:

- circulating confidential data to licensed users
- providing isolated RDCs with physically restricted access
- providing remote RDCs
- create remote job submissions systems

The first two are essentially uninteresting, as the only effective development has been on the transport media and processing power; there has been some discussion of 'watermarking' distributed datasets to track unauthorised distribution, but this is not widespread at present.

The major interest in recent years has been on remote RDCs, using thin-client technology (see Ritchie, 2008b). These have the advantage of providing the security of RDCs but without the need to physically be at sites. Moreover, because the security is provided by thin client

system, it is possible to tailor security within the design of the system (eg limiting access points, or types of users) very easily. One downside is that NSIs can feel they lose control over the researcher if they cannot physically see them. A number of alternative technologies have been identified to counter this: web cams, GPS-based security tags, whole-session recording, and so on. However, one of the aims of multi-stage security models, such as the VML model described above, is to change the focus from idealised technical solutions to more robust complementary methods.

There are a variety of implementations in the world: some based on Unix systems, some on Microsoft Windows, with a variety of operating systems and virtualisation tools. A software-based approach is highly amenable to alternative implementations: the same basic technology is used in the UK, Denmark, and the Netherlands; but one only provides access on government sites, one to universities only, and one over the internet but with security keys on computers.

Remote RDCs have also prompted a re-evaluation of the procedures needed to govern RDCs. This is one of the areas where, at time of writing, there are many reviews ongoing. Part of the reason for this is that the existence of remote RDCs suddenly raises many possibilities for international data sharing, and the need for effective, simple, transparent governance; see Ritchie (2009b) for a discussion and proposals for development of international standards as a first step.

Remote job submission technology, where a researcher sends in requests for statistical analyses which are answered automatically but which do not allow viewing of the data itself, has only been pursued by a small number of countries: currently Australia and New Zealand share a stable system; in Europe, ‘Lissy’ (access to European earnings data) has been running for some years and now services 50,000 jobs per annum; and in the US the Census Bureau has been experimenting with SDC models for remote job servers (eg Steele and Reznek, 2006).

Remote job submission has not reached the same level of technological consensus as the remote RDC. This is partly because it is less widespread; but it also partly because the technological implementation is often strongly determined by the disclosure control rules and the assumptions made about the researcher commitment.

In summary, the distribution of non-disclosive data is well understood. For confidential data, for remote RDCs (and RDCs) there is a wide variety of technical knowledge and experience, although theoretical development is still some way behind. For remote job submission, there is less experience and less consensus, although there are a number of successful examples in operation around the world. In addition to these, there are more traditional was of providing access to data such as licensing. In short, any NSI wishing to set up access to data, confidential or otherwise, now has numerous examples of good practice to follow in many different areas.

## 5. The ONS experience

A more detailed discussion of the ONS model is available as Rtichie (2009a).

### 5.1 Operational aspects

The VML holds business, health, social, Census and OGD data; in principle, all data. It is a thin-client system allowing users on any ONS site to access confidential data through a secure channel. Users are presented with a familiar Windows™ desktop, enabling them to work on the data as if it were stored locally, but without any data transmission taking place. This approach, of distributing access rather than distributing data, is increasingly recognised as best practice, and has been adopted across much of Europe and North America.

The VML became operational in January 2004, and has grown to become one of the most important research data resources in the UK, second only to the UK Data Archive and the Internet. Ritchie (2008a) describes the VML in detail and notes that demand from external

researchers has grown by around 50% every year[1]. The VML hosts a number of data sets for other government departments (OGDs); in several cases, the ability to partition the VML into isolated sub-areas have allowed OGDs to acquire, in effect, a dedicated secure research facility with administrative backup.

The VML has become increasingly important to ONS. As the implications of the SRSA duty to support research have been studied, the existence of a secure corporate universal delivery mechanism has allowed data managers to plan the release of data without needing to address details of implementation. For example, a policy decision was taken to release all social survey data at the most detailed level. The VML was identified as the default delivery mechanism, but not necessarily the best long-term solution. However the presence of a default solution allowed discussion to concentrate on principle.

At the same time as ONS reviewing its duties, the data sharing provisions of the SRSA have led to the acquisition of sensitive administrative data from other government departments which needs a high level of security.

Finally, other business areas have used the partibility of the VML to set up secure access areas to support their own research needs. The net effect of these changes, particularly the impact of the SRSA, is that VML usage is expected to double in 2009-10; and by the end of this period internal requirements for processing power and new datasets is likely to overtake external use.

In 2008 VML access points were set up in Glasgow and Belfast to explore the possibility of accessing the VML from other government departments. Following a successful pilot, it is likely that further access points will be set up around the country. These access points will all be based in the offices of central government departments: partly for technical reasons, and partly because of the additional physical security requirements in place at government offices.

[1] This does not include the ONS Longitudinal Study, of which the VML took over delivery in 2008 and which will be fully incorporated into operational procedures in 2009.

The requirement to visit a government office, even one close by, does impose an extra cost on academic researchers. To alleviate this, and in recognition of the exponential growth in demand for the VML, the UK's Economic and Social Research Council agreed to fund a pilot for a VML clone, housed on the academic network and supporting academic researchers. This is due to begin operations in Autumn 2009. While the long-term shape of the Secure Data Service (SDS) is unclear at the moment, it is expected that there will be a separating equilibrium for the VML and SDS based on different data sets, access points and user groups.

ONS has also been advising OGDs on setting up their own VML-style systems.

**5.2 Principles of access**

The VML was a radical departure from ONS practice on many fronts. The technological solution proposed was novel. There was much concern about the trustworthiness of external researchers. Equally, there was little perception of the value to ONS of providing a research facility. Disclosure control was the responsibility of people who created the data. Finally, there was a strong feeling that, if access was going to be granted, it should be on a familiar ONS model adjusted for external access, rather than being a tailored solution designed from the ground up.

As a result, the VML has always had a strong theoretical foundation. The VML team came together at the end of 2002, and the first papers in 2003 were on principles of research access, an early version of the VML security model (based upon concepts developed by the legal team) and its practical implementation, and a rationale for disclosure control to be managed by the VML team.

Subsequent iterations have refined the security model to the stage as described in section 2, and other papers have been written to address specific concerns. For example, Ritchie (2006) was a response to concerns about residual risks in VML operations. The paper argued that

the very existence of RDCs led to a certain amount of irreducible risk; trying to reduce this risk further was at best naïve and at worst counter-productive.

Much of the VML's conceptualising has been designed to simplify the administrative process. For example, since its inception the VML has had a compulsory training programme for all users. One of the aims of this training is to ensure that VML users buy in to the VML philosophy and management strategy, so that research is seen as partnership where both users and the VML team have the same incentives to make sure the VML runs efficiently. As a result, the VML operates one of the most cost-effective RDCs: in 2009-10 it will support around 120 projects (involving around 150 researchers making 1500-odd visits), plus a number of key ONS projects, for a staff cost of £400,000. This staff cost reflects a team of eight, but the VML can support this number of projects with as few as two people for short periods.

However, the main theoretical contribution of the VML to RDC management has been its development of statistical disclosure control (SDC) policies. In 2003, SDC for research outputs was expected to use the models established over many years for dealing with tabular outputs and for creating anonymised microdata sets. The VML team argued that the nature of the research environment made these rules inappropriate: it is easy to generate cases where simple threshold models, for example, both restrict research and fail to protect confidentiality. Instead, a tailored approach was needed, requiring a focus on principles, flexibility and rules-of-thumb, development and empowerment of RDC staff, and a classification of outputs based upon functional form. For details of the approach, see Ritchie (2009c); for the classification model, see Ritchie (2008a); a manual for researchers is available as VML (2009).

The VML approach to SDC causes some concern; it is seen as hard to implement, hard to train people in, hard to achieve consistency. However, the experience of the VML team is that RDC staff can be trained in a matter of weeks, and consistency is achieved partly by getting all staff to rotate training; a peer-review system is also being set up. On implementation, the VML team typically turns round clearance requests in less than a day, occasionally in

minutes; the difference in response times reflects the VML team's other work commitments, not the difficulty of clearance.

Whilst this model is only in use at the VML at present, it is being considered for adoption by the SDS, the UK Government Statistical Service, and European partners.

**5.3 Key concerns for the future**

Current key concerns are IT, access, and demand growth. Put simply, the VML has been a victim of its success and IT and other resources have struggled to keep pace. With hindsight, it is clear that demand projections (both within and outside ONS) have been too conservative; expectations of the growth in demand falling off have not been fulfilled: after almost six years of operations, demand is still expected to double this year. Hence, one of the lessons is that a capacity demands need to be reviewed early and often.

One other concern, for all of ONS, is the potential in the Statistics Act. The combination of the duty to support research, the publication of criteria for 'fit and proper' researchers, and the transparency of access rules, while all positive developments, does mean that ONS' policies and data release practices are more open to challenge than they have been before. The concern for ONS is that it might find itself receiving increasingly idiosyncratic requests for data for research.

This has made the need for a comprehensive data management policy more important. If ONS can demonstrate that all accessible data is available in some way, delivered in a form which is cost-effective for ONS, then the case for devising tailor made solutions needs to be very strong.

Other concerns about the proliferation of data types and alternative operating models have not been realised yet (or have been managed) but are still out there (see Ritchie, 2009a).

One perennial issue is what might be termed ‘fear of the new’. ONS is a government department, and by nature and design is risk-averse. However, with the playing field for data access changing so substantially in recent years, the assessment of risk has become much more complicated. Relative risk is still poorly understood. For example, in 2008 a potential security flaw in the VML was highlighted. The VML was closed immediately, pending an investigation. A failure to understand the nature of the risks that had been exposed led to a return to previous arrangements while the problems were evaluated: ONS temporarily replaced an almost (but not quite) watertight research facility with a system of much lower security.

Even with the VML in existence, there is a still a resistance to the security and data models. For a surprisingly large number, distributed access is still a poor substitute for distributed data, irrespective of the quality of that distributed access. Researchers who have access to the VML from their desktop still ask for local copies of data; and similar noises are being made about the SDS from the academic community.

These outcomes can be attributable to a ‘fear of the new’. There is a chance that more risky solutions may be adopted because decision makers are comfortable with something that had been used in the past – even if that past posed a higher risk. While impressions of risk are changing, there are still problems, particularly the identification of an appropriate reference frame. For example, under the new statistical law, the choice is no longer between releasing and not releasing data. The choice is between different methods of delivery. As a result it may be that one of the biggest risks for the future may be the under-utilisation of data if access methods are not appropriate to researcher needs.

### 5.4 Misplaced concerns

The previous subsection highlighted concerns and risks; however, it is also worth pointing out where concerns are relatively minor. For the VML the areas are staff, methodology, and law.

Staffing is a potential issue. However, a fast training time alleviates this. New VML staff are expected to provide user support two weeks after starting; by the end of the first month they should be clearing outputs without supervision. Much of this is due to the researcher training programme, encouraging researcher buy-in and simplifying disclosure control for both parties.

A supportive and knowledgeable user base is essential for keeping a very lean operation running. The VML team should operate on a team of nine; however, for a one month in 2009 it operated with just three members of staff (one senior manager, one junior researcher, one administrator) and for two weeks with just two people. Over this period key outputs (booking time, applications, clearances) were all maintained. The 'all hands to the pump' model is of course not sustainable, but it is an illustration of how paying attention to both the demand for and supply of user support can bring dividends.

Methodologically, the VML approach to RDC-specific SDC methods is proving robust for a wide range of datasets and operators, and has faced no serious challenges to their continued use. Opposition to the VML approach largely comes from the perceived difficulty of training RDC staff, and so far the experience of the VML has been an effective counter-argument. Note that this does not mean no mistakes have been made at the clearance stage. However, the VML system is designed so that the margin of error and additional checks and balances mean that there has been, as far as the VML team is aware, no illegal removal of data.

On the legal side, the new statistics law has proved flexible enough to provide reasonable responses to all needs. 'Reasonable' in this context means that both researchers and ONS are confident that the balance between access and confidentiality is fair, even if researchers do not get all the access they would like.

## 6. Evaluating the outcome

### 6.1 CEA versus CBA

Technology changes rapidly, and so does the environment within which data access solutions are defined. There may be opportunities to improve the IT or to adjust their procedures and processes to new conditions. A key component of good practice is for NSIs to periodically review and assess their operations.

Reviews can take two forms:

- cost-effectiveness analysis (CEA) reviews whether resources could be utilised better to either produce the same outputs for less money, or more outputs with the same
- cost-benefit analysis (CBA) questions whether the production of those outputs is a good use of resources *per se*.

CEA should be part of the regular process. It is a comparatively straightforward process to operate, as all the variables under review (cash charges, time, technological options etc) all have identifiable monetary values and a predictable effect on outputs.

CEA is often confused with CBA, but a true CBA is a very different beast. Instead of asking "how can we do this better", CBA asks "should we be doing this at all?". It is true that, for a given set of constraints, it is possible to see CEA and CBA as different sides of the same coin; for example, "do a CEA on this RDC" can be equated with "do a CBA on this RDC subject to the continuing provision of the services of the RDC". However, this obscures a fundamental difference in perspective: CEA is about effective delivery of outputs; CBA is about finding an appropriate level of outputs as well as the delivery of them.

The difference between CEA and CBA can be likened to the difference between a model with or without binding constraints on some of the variables:

CEA = f(inputs) subject to g(outputs)

CBA = f(inputs, outputs)

CBA raises many more problems: for example,

- identification of outputs: does ‘improved understanding of the world’ count as a valid output?
- valuation of outputs : what is the value of a research paper?
- scope of outputs: in the multiple-mode data access framework designed above, a CBA of an RDC should take into account that some of the RDC outputs could be produced through other mechanisms
- discontinuity in inputs: if inputs are only available in large discrete chunks, this complicates CEA; but it can lead to multiple outcomes in CBA where the outputs are being optimised as well
- additionality of operations: which outputs would be produced by some other mechanism if the system under review was not available, and which inputs would be needed to produce it?
- inter-dependency of inputs and outputs: some outputs are determined jointly with others, and some act directly as inputs for other outputs

As a result of these problems, CBA of data access has rarely been put into practice. For most access modes, either a preliminary investigation indicates that the benefits massively outweigh costs, or the NSI takes a policy decision that providing access in a particular way is one of its duties. Hence, little value is placed upon a full CBA, when a CEA is more appropriate:  see Stefan, Ray for examples).

One exception is in the creation of non-disclosive microdatasets for distribution. These are costly to create, and if NSIs get the confidentiality/utility balance wrong they can end up creating an unsafe or an unused datasets. Hence NSIs typically will carry out studies where the utility/confidentiality/cost trade-off is explicitly examined in the context of other ways of spending the money to support research. This is a CBA in practice, if not in name. Synthetic datasets, although still experimental, have a similar explicit cost/utility trade-off and so are also likely to undergo CBAs before going into production.

**6.2 CBA in the UK: assessment of the VML**

In the UK, the VML team is currently at the early stages of carrying out a full CBA of the RDC use for official microdata. This goes well beyond a CEA of how the VML is run most effectively: the aim is to review how both the VML and the academic equivalent, the Secure Data Service, should fit together to provide the maximum public benefit from the existence of one or two remote RDCs. to keep it manageable, the internal operations of the VML and the SDS are out of scope of this project; the primary aim is to decide the scale of operations and distribution of responsibilites, and how this will change ONS' implementation of solutions along the whole data access spectrum. In other words, this is a true CBA which challenges the whole purpose of RDCs.

As mentioned, this is still at its early stages, and is due to report in November 2009. However, there are two novel features of this analysis which can be mentioned here.

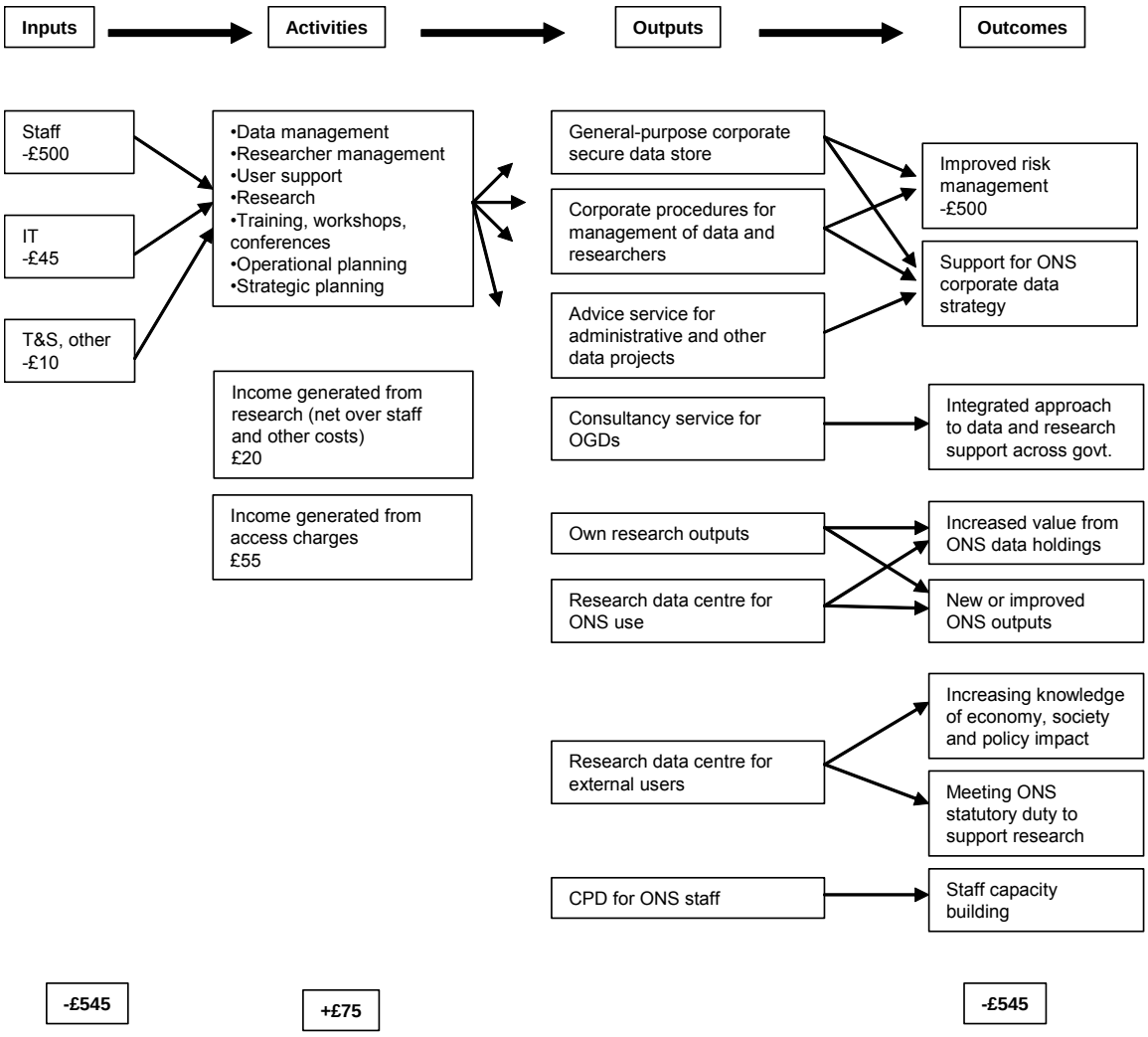
First, the analysis is in two stages: (a) is an RDC worthwhile (b) if so, should the UK have one, two or more, and if so, what would be the roles of each? This is novel because, historically, assessment of RDCs have tended to focus only on the specific implemtnation rather than the framework.

Second, the VML has borrowed a method from ONS' UK Centre for the Measurement of Government Activity (UKCeMGA) for its assessment of benefits. UKCeMGA has developed a model for assessing the productivity of government departments which tries to tackle directly the problem of valuing either incomparable or non-valuable items. This breaks down the production process into four components:

| Inputs | Activities | Outputs | Outcomes |
|---|---|---|---|
| What resources are used? | How are those resources manipulated? | What tangible outputs are produced | What are we trying to achieve? |

The idea is that, although only some of these can be measured financially, most can be measured in some way; and changes can then be identified in a structured way and productivity changes assessed. It also has the advantage of concentrating minds on the difference between tangible outputs and broad aims of any activity.

In the context of the VML, a first attempt to fill in this matrix (including projected costs from 2010-11) is




Note that, without needing to assess the physical activities and outputs, it is already possible to make some broad assessments of whether there is a net benefit of running the VML (or some similar system).

This is an ongoing project, due to report in November 2009. Interested readers are invited to contact the author for an update.

## 7. Summary

This paper has discussed a number of aspects of designing a research data access model which covers the full spectrum of possibilities. It has argued that the design of a system needs to start from principles: what is the NSI trying to achieve? Once this has been identified, a data model can be defined to address the level of security required. It is important that this model does not treat solutions in isolation: there is a spectrum of needs and opportunities, and a holistic approach to data access is needed to provide a set of valid alternatives.

Once the model and mode have been determined, implementation is a relatively simple part: there is enough experience around the world on efficient ways of using technology. However, the value of effective procedures meshed with technology is often forgotten; for a truly effective solution to data access both the human and technological components need to work together.

Finally, any system should be reviewed – regularly for cost-effectiveness, occasionally for cost-benefit. The latter has not been widely used except in assessment of the creation of non-disclosive or synthetic datasets. The UK has embarked on a full CBA of its RDC network, but the result of this are not due in until later in 2009.

## References

Desai T. and Ritchie F. (2009) “The role of researchers in effective data centre management”, presentation for UNECE Workshop on Confidentiality 2009, Bilbao, December

Ritchie, F. (2006) “Access to business microdata in the UK: dealing with the irreducible risks” in *Work session on statistical data confidentiality 2005*; UNECE/Eurostat; pp239-244

Ritchie, F. (2008a) “Disclosure detection in research environments in practice”, in *Work session on statistical data confidentiality 2007*; Eurostat; pp399-406

Ritchie, F (2008b) “Secure access to confidential microdata: four years of the Virtual Microdata Laboratory” in *Economic and Labour Market Review*; Office for National Statistics; May, pp 29-34

Ritchie, F (2009a) “UK release practices for official microdata”, mimeo, Office for National Statistics, forthcoming in *International Journal of Official Statistics*

Ritchie, F (2009b) *Designing an international framework for research access to confidential data*, mimeo, Office for National Statistics

Ritchie, F. (2009c) *Statistical disclosure control in a research environment*; mimeo, Office for National Statistics

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

VML (2009) *Default procedures for statistical disclosure detection and control: guide for researchers*; mimeo, Office for National Statistics