WP. 15
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Bilbao, Spain, 2-4 December 2009)

Topic (iii): Research data centres and virtual labs

EFFECTIVE RESEARCHER MANAGEMENT

Invited Paper

Prepared by Tanvi Desai (London School of Economics) and Felix Ritchie (Office for National Statistics), United Kingdom

Effective Researcher Management

Tanvi Desai* and Felix Ritchie**

* Research Laboratory, London School of Economics, London, UK WC2A2AE, t.desai@lse.ac.uk & Office for National Statistics, Newport, UK, NP10 8XG tanvi.desai@ons.gsi.gov.uk
** Office for National Statistics, Newport, UK, NP10 8XG, felix.ritchie@ons.gsi.gov.uk

Abstract: National Statistical Institutes [NSIs] are increasingly investigating new ways of providing access to confidential microdata for research purposes. These innovations are being driven by the requirement for NSIs to ensure the best possible return for their investments in data collection coupled with researchers' increasing demand for highly detailed microdata.

After a long period of decline as NSIs focused on confidentialising data to produce 'scientific use files' for circulation, Research Data Centres [RDCs], which allow researchers largely unrestricted freedom to work on highly detailed microdata within a secure environment, are making a comeback. The reasons for this include (a) the potential for remote access solutions which overcome most of the limitations of physical RDCs, (b) associated new models of working which have caused a revision of the confidentiality/utility tradeoff, and (c) increasing policy demands for analysis, such as local area studies, which can only be met by detailed microdata

These new ways of working which include an increased focus on 'customer engagement' and the effective use of resources within the public sector are leading to the realisation that the involvement of the researcher community is a key element of the success of any solution. Most obviously, the emerging field of output-based SDC requires the active engagement of researchers to be properly effective. If researchers are seen as an active part of the security model, as opposed to something for the data to be protected against, then both more efficient and more secure operating models can be devised.

This paper considers how the active engagement of researchers in the management of RDCs, and the design and implementation of data access systems in general can be used to improve data security. It also addresses a number of clichés espoused by both academics and NSIs and argues that, while there is some truth in these, a fair assessment of current risk is often eschewed in favour of simple judgments based on past practice. Finally we note that the quest for security based upon a technological perspective and a fundamentally negative view of human behaviour can lead to exactly the outcomes which RDCs designers are trying to avoid.

1 Introduction

Governments invest significant sums of money in data collection, and there is an increasing focus in the UK and internationally on ensuring the greatest possible return for these investments. A growing interest in data linking and the use of administrative sources require data access infrastructures in which data can be accessed at levels of detail that are considered highly sensitive. It is only by creating these infrastructures that the available microdata resources can be fully exploited.

However, there is a concern that infrastructure developments for confidential data access undertaken by the data owners may be unduly risk-averse. This is because data protection measures are often simply translated from environments where there is little interaction with data users, and so the aim of the data owner is to embed as much protection in the infrastructure (data and contracts) as possible.

This paper will argue that it is not just infrastructure that can have a significant impact on data security and research quality, but also the approach the data provider takes to researcher management. Researcher management is the key to an efficient and effective security model, and should be built in to the model from the beginning. This can have an impact on short term research projects, but can also bring long term benefits for the NSI in terms of an improved culture and understanding of data security among researchers.

This paper focuses primarily on the provision of access by NSIs through RDCs, but the issues described here have a wider application. Beyond NSIs, building an effective relationship with researchers is arguably even more important for organisations without the statutory penalties that NSIs can call on. The benefits of active researcher engagement highlighted here may also be realised in other dissemination contexts, not just RDCs.

2 Data and technology management versus researcher management

2.1 Traditional views of risk

While the value of confidential data as a tool for policy relevant research is well established (e.g. Trewin et al, 2007), the release of confidential microdata has been traditionally dominated by a focus on risk rather than reward. Until recently, NSIs have primarily been concerned with protecting data from researchers, by making sources difficult if not near impossible to misuse.

This has lead to NSIs taking full responsibility for data security and a focus on 'intruder scenarios'. These scenarios are typically 'worst-case', and while there has been some recognition (e.g. Mackey and Elliot, 2009) that this may not be an appropriate way to assess risk, few practical alternatives have been explored.

2.2 Real world risks

In spite of this focus on 'worst case scenarios' there is no evidence, to the authors' knowledge, of academic researchers maliciously misusing data.

The UK Office for National Statistics [ONS] has a long history of releasing microdata for policy relevant research. The UK Data Archive has been providing access to anonymised ONS microdata through End User Licenses for over 40 years, and in recent years the service has been expanded to include Special License files at a more detailed level. From 1991 to 2002, researchers in a small number of universities around the UK were given access to the NESPD, an anonymised but identifiable dataset on earnings of some 400,000 individuals. With the addition of the ONS RDC, the Virtual Microdata Laboratory [VML], in 2004, researchers in the UK have access to a large collection of highly detailed business, social and health data (see Ritchie, 2008b, for a description of the VML).

During the time that the UK has been making microdata available there has been no evidence of malicious misuse of confidential data[1]. There have, of course, been cases of inappropriate, careless or selfish behaviour: professors giving passwords to research assistants, researchers analysing data on a home PC even when expressly forbidden by their license, data being transferred to unprotected zip drives or USB sticks, and researchers making notes from protected information on the screen.

Other countries experience similar problems. In one country a researcher failed to apply simple disclosure control rules, allowing a journalist to identify personal information; in another, a group of academics systematically abused a loophole in a remote job submission system to download, over a period of months, an entire dataset.

In these cases, there was no intention to misuse the data for non-statistical purposes; in the last example, the academics merely disliked the inconvenience of using the remote job system.

[1] It is not possible to say definitively that there has been no malicious misuse of the data, as there might have been successful attempts to mislead data owners. However, all breaches of confidentiality known to the authors concern attempts to circumvent processes, not to identify individuals.

One reason for this relaxed attitude to data security is that academics do not see themselves as a risk. Typical, if not verbatim, responses familiar to the authors include

- on giving access to unauthorised users: 'but they are working with me; I trust them.'
- on storing data locally: 'it's my computer; no-one else can use it'; 'I've always worked like this, and I've never lost anything.'
- on transferring data inappropriately: 'If I can't take my data with me to the conference I won't get my paper finished'
- on following procedures: 'this is just ticking boxes; it doesn't make any difference'
- overall: 'you can trust me';

The underlying assumptions are that academics are trustworthy, data are probably non-disclosive anyway, and the NSI is oversensitive and doesn't understand research.

By taking full responsibility for data security while viewing and treating researchers as a risk, NSIs have contributed to a culture where data users do not traditionally think of themselves as responsible for data security. Users rely on the files they are accessing to be safe, and have an expectation that NSIs will not release sensitive data, leading them to argue that any conditions placed on the use of the data that they might find inconvenient are just unreasonable demands that do not add to risk management.

Limited communication between data users and data providers over the principles of data access has created a situation where there is a mutual lack of understanding of the aims, needs and the working practices of both sides. This has led to distrust on both sides, making cooperation more difficult, and impacting on a data provider's ability to manage and implement change without opposition from data users.

Most seriously, a lack of trust impacts on data security. Researchers who do not understand the aims and responsibilities of the data provider, or why security mechanisms are in place, are more likely to view the mechanisms as an indication of the NSI's lack of interest in their needs and working practices, and thus are more likely to attempt to subvert security. (Desai, 2004).

3 Researcher management: the active approach

This paper argues that shifting the emphasis from one of 'data management' where the NSI takes responsibility for data security and researchers are viewed as a risk, to

'researcher management' where both data users and providers are responsible for the confidentiality of outputs and researchers are viewed as collaborators, leads to a more efficient trade off between data security and data utility.

	Data management	Researcher management
NSI	Data owners Data suppliers	Data custodian Primary analysts
Researchers	Data users Data risk	Data custodian Secondary analysts

3.1 What is 'active researcher management'?

A lot can be achieved just by careful management of the language used when communicating with researchers. A vocabulary which suggests a cooperative rather than an adversarial view can go a long way towards fostering understanding between parties, for example:

	Data Management: *researcher as risk*	Researcher Management: *researcher as collaborator*
Explaining security policy	'we're doing this to protect the data' (from you)	'doing this allows us to supply you with more detailed data'
Limiting quantity of results	'you must limit your output to reduce the chance of disclosure'	'limit your output because we have finite resources; people who produce good output get their results back quicker'

In the 'researcher management' method, researcher training is vital. Before accessing the VML, researchers are trained to understand not only their responsibilities in relation to confidential data but also the basis for those responsibilities: for example, legislation, licensing conditions and ONS' aims and priorities. An interactive component of the course then provides potential data users with the opportunity to explore examples of statistical output and to assess whether they would be considered disclosive. Finally, researchers are introduced to the support team; the role of the VML team (including what support is provided and where researchers are expected to manage on their own); and the rationale for various administrative processes.

Physical attendance at a training course is a precondition for access to the VML. The training programme not only underpins the cooperative approach to researcher management, allowing data users and data providers to meet and develop trust, but it also contributes significantly to resource management within the VML. The researchers are encouraged to ask questions to ensure that there is a common understanding of purposes and constraints. All this is designed to enable support staff to work more efficiently by avoiding confusion and unnecessary involvement in discussions about appropriate output.

3.2 Example: engaging researchers in SDC

It is in an NSI's interest to choose a disclosure control method which makes the most efficient use of resources possible. As it is impossible to determine all the manipulations and analyses that a researcher will subject microdata to, an NSI is left with a choice between imposing a set of inflexible rules or else implementing a more flexible system.

Inflexible rules that are designed to cover all eventualities will inevitably lead to data that is not actually disclosive being refused release due to the necessity of concentrating on worst case scenarios. This leaves researchers unsatisfied, and reduces the utility of the data. With the cost of data collection, and the acknowledged benefit to society of policy based socio-economic research, it is in the interest of NSIs to ensure that data resources are fully exploited for research purposes.

The flexible method adopted by the VML, focusing on outputs, is an example of how the training programme generates both efficient and secure operations. In the VML training researchers are told that outputs are either 'safe' or 'unsafe' (Ritchie, 2008a). A 'safe' output, such as regression coefficients, would normally be released unless the NSI makes a case for why it should not be; an 'unsafe' output, such as a table, would not be released unless the researcher can demonstrate to the NSI that it is non-disclosive. The training then gives researchers guidelines on how to make 'unsafe' outputs non-disclosive; it also suggests (but does not prescribe) how output should be structured in order for it to be easily understood and quickly passed by the support team.

This system gives researchers a clear view of their responsibilities; it encourages production of 'safe' outputs; and it demonstrates to researchers that there is no 'them and us' when it comes to disclosure control. Although the NSI always has the final say if a dispute arises, disclosure control is a co-operative process with one or other party putting in more effort depending upon the type of output.

The training enables the VML team to make very clear to researchers what is and what is not acceptable as output. It therefore reduces any unrealistic expectations researchers may have in relation to the data they can remove from the secure

premises, avoiding disappointment and wasted effort on both sides, and increasing the chance of high-quality output.

3.3 Benefits of a 'researcher management' approach

There are a number of benefits to employing a 'researcher management' approach instead of a 'data management' approach to data access. These are largely inter-related and include:

- increased communication
- increased understanding
- increased cooperation
- better change management
- better data security
- better research
- more efficient use of NSI resources

Though most NSIs undertake consultations of varying intensity with users when collecting data, there has traditionally been very little consultation in relation to data access procedures and infrastructure. This has led to a disconnection between data users and providers which has fostered a feeling of mistrust.

The increased communication involved in a 'researcher management' approach to data security has enabled the VML to significantly increase researcher understanding of NSI goals and responsibilities, the conditions under which the ONS is able to release data and the consequences of misuse. At the same time ONS has been able to gain a greater understanding of researchers working practices, and research and data priorities. The ONS has been able to use this information to tailor a system that is as appropriate to researchers working practices as is possible within the constraints of data security. This system and the increased levels of communication, has built trust between the ONS and researchers, and decreased the incentives for subverting data security. This model of 'open innovation' to stimulate behaviour change is being recommended as a way of improving services while reducing costs (see e.g. Bunt and Harris, 2009).

In addition increased trust and communication leads to researchers being more cooperative and more willing to accept service interruptions, and modifications to systems and procedures, thus making change management far easier for NSI staff.

The flexible, cooperative approach to disclosure control practised by the VML fosters trust and understanding between data users and providers, and allows the fullest possible exploitation of data resources. This technique also allows an NSI to make use of researchers' statistical expertise to gain understanding of the analysis

techniques used and whether they might lead to the release of disclosive data. This gives the NSI access to free statistical consultancy, and the input necessary to improve SDC methods for the future. As well as increasing the statistical skills and understanding of SDC staff.

One of the most significant benefits of the 'researcher management' approach is that training and involving researchers in statistical disclosure control promotes a culture of understanding data security. With increased access to sensitive microdata a significant cultural shift is needed in the academic research community away from one in which data security is traditionally seen as the responsibility of the data provider to one where researchers feel accountability for the safety of their data resources.

There are costs associated with a 'researcher management' approach, primarily in terms of the initial training programme, and on-going support and communication, which is far more involved than it would be for end user license, or special license files. However, when the costs are compared with the amount spent on the development of anonymised license files, they do not seem so onerous, particularly when the additional benefits are taken into account.

4 Engaged researchers and the wider security model

The full security model used by the VML is "safe projects, safe people, safe settings, safe outputs" (Ritchie, 2008b). The above discussions suggest that the engagement of researchers is an essential component of an effective security model, but this is clearly not the only component.

However trustworthy researchers are, problems can still arise from accidental releases of data. For example, IT equipment gets circulated which may not have been cleaned properly; or, researchers accessing files from a secure store may not be aware that programs such as SAS and STATA store temporary files locally, and so may not take appropriate precautions with their machines.

On the NSI side, habits can form that may impact on security. For example, if a researcher only produces very safe outputs for a long period and then switches to a project which generates more risky outputs, will the researcher's more risky outputs come under less scrutiny because of past good practice?

Finally, RDCs are becoming popular because they offer an unprecedented level of access to data, but it is not clear that researchers always need that level of access. In the case of the UK, the time and money expended to visit the VML, and the ready

availability of less detailed versions of the same data source through the UK Data Archive, acts as a form of filter.

Researcher management therefore is only one component of the security model, but it is a central one and complementary to the others. For example, in designing the IT model, assuming that the user might be foolish and selfish but is generally well-disposed towards the RDC makes system design much easier than if the 'worst-case scenario' is applied. However, just assuming that users are well-intentioned and taking no steps to bring that about that may be equally indefensible.

5 Conclusion

By taking full responsibility for data security while viewing and treating researchers as a risk, NSIs have contributed to a culture where data users do not traditionally think of themselves as responsible for data security. They rely on the files they are accessing to be safe, and have an expectation that NSIs will not release sensitive data, allowing them to argue that any conditions placed on the use of the data that they might find inconvenient are just unreasonable demands that do not add to risk management. This paper has argued that making active engagement of researchers should be part of any security model.

'Active engagement' goes beyond setting up contracts and making sure researchers know their legal responsibilities. It requires making researchers partners in secure effective access – and making sure that they understand this.

Effective engagement of researchers is not the only basis for secure access to microdata, but it is possibly the most important element. Without a positive engagement from researchers, protecting data from accidental and deliberate misuse becomes more difficult and costly.

Moreover, the active engagement of researchers brings other benefits. An involved researcher is more willing to exchange in debate, more willing to accept change, and is more tolerant of NSI processes. Most importantly, he or she is more likely to bring ideas to the NSI. Getting researchers involved in the security of the RDC improves the efficiency and security of the RDC, provides free statistical consultancy, increased feedback on data quality issues, and an element of training for RDC staff; Moreover, developing a culture of engagement is the only way in which NSIs can affect a change in researcher attitudes to data security in the long term, by improving researchers' understanding of their responsibility in relation to confidential data.

References

Bunt L. and Harris M. (2009) *The Human Factor: how transforming healthcare to involve the public can save money and save lives*; London; NESTA

Desai, T. (2004). *Providing remote access to data: the academic perspective*. UN(2004).

Mackey, E.C and Elliot, M.J. (2009). *An application of game theory to understanding statistical disclosure events*. UNECE http://www.unece.org/stats/documents/ece/ces/ge.46/2009/wp.40.e.pdf

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

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