WP. 29
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
(Geneva, Switzerland, 9-11 November 2005)

Topic (iv): Access to business microdata for analysis

ACCESS TO BUSINESS MICRODATA IN THE UK: DEALING WITH THE IRREDUCIBLE RISKS

Supporting Paper

Submitted by the Office for National Statistics, United Kingdom[1]

[1] Prepared by Felix Ritchie.

Access to business microdata in the UK: dealing with the irreducible risks

Felix Ritchie[1]

[1] Social and Economic Micro Analysis and Reporting Division, Office for National Statistics, Government Buildings, Cardiff Road, Newport, South Wales, NP10 8XG.

Abstract. The UK Office for National Statistics provides a thin-client remote laboratory service for secure research on confidential microdata. This technological solution is allied to tight procedural environment and compulsory training of researchers to achieve an effective mix of practicality and security. The result has been a ten-fold increase in the use of the ONS business data by external researchers, and a significant increase in ONS' in-house capabilities and projects.

This solution raises several potential problems: the irreducible person risk from providing access to identifiable data; a need for "intelligent" disclosure control policies; the management of off-site access to data; and the possibility of "unwanted" recreation of official data. This paper discusses how these have been addressed.

Keywords. Remote access, thin clients, disclosure control, confidentiality, identifiable data, research use of microdata

1 Introduction

Since January 2004 the UK Office for National Statistics (ONS) has been providing remote and on-site access to business microdata for research purposes in a controlled environment. This is achieved through thin-client technology and a tight procedural framework, within which researchers have complete access to identifiable (if not identified) data. This has allowed data to be used by researchers in academia, government and the private sector to produce analyses ranging from the impact of innovation on productivity to the calculation of labour cost adjustments for the health service.

Researchers are also allowed to link their own data with ONS data in the lab, and this has been taken up with enthusiasm by other government departments, often in

collaboration with academics. This has helped to keep the UK at the forefront of policy impact analysis and programme evaluation in Europe.

This solution raises several issues.

First, although the technical solution provides the highest protection from a technical attack, giving access to identifiable microdata creates an irreducible person risk which needs to be addressed through training programmes and credible penalties for misuse.

Second, the lack of restrictions on analysis means that disclosure control mechanisms need to be sufficiently flexible to cover an unknown variety of outcomes. As a result, automatic disclosure control methods are not feasible. Training in statistical disclosure becomes necessary for both researchers and lab managers, and this training has to be driven by example within broad principles rather than prescriptive.

Third, a remote access system potentially allows access to data from any location. Thus access is governed by perceptions of acceptable risk rather than technical feasibility; and these perceptions are more subject to criticism by outside bodies.

Fourth, the provision of complete datasets to researchers means that there is a risk of the ONS's own figures coming under attack from researchers using the same data.

In all these cases, there remains an element of risk which cannot be managed away. This paper describes how ONS has addressed these issues within a common corporate framework which is now being applied to non-business data.

2 Providing secure access to confidential data for research

2.1 Background

In 2003 ONS created the Business Data Linking branch (BDL) to address the issue of providing access to its business microdata for research purposes. This posed several significant problems, including the legality of access and the fitness of survey data for the purpose. For a detailed description of this, see Ritchie (2004).

BDL developed a solution based upon a four-part model of security

Aspect	Aim	Criteria
Safe projects	Projects will not embarrass ONS or damage ONS' ability to produce statistics	Projects must have a valid statistical purpose and be carried out by responsible researchers who take ultimate responsibility for inferences made
Safe people	Researchers can be trusted	Researchers must be from an approved research or government institution; there should be no conflict of interest
Safe settings	Deliberate and accidental removal of data is not possible	The lab environment must be inherently secure (that is, preventing the removal of data without action by BDL staff)
Safe outputs	Approved outputs do not contain any disclosure risk	Disclosure control methods are designed explicitly for the lab environment and outputs

This composite strategy is designed so that the elements reinforce one another. This model is now being adopted for other parts of ONS wishing to provide access to microdata for research[1].

2.2 The Virtual Microdata Laboratory (VML)

The concepts in the above table are familiar to many national statistical institutes (NSIs). One unusual feature is the technological solution implemented to secure the system electronically. This system has been in place in Denmark for some years, and is now being considered in test implementations in a few other countries.

The VML is a thin-client system; that is, researchers log on to a computer in a remote location which processes all requests centrally and returns information about the results. This is the way that Unix systems operate. Hence, no data travels over the network, save in the form of statistical results. In contrast, a "fat client" such as a PC downloads data over the network and processes it locally.

There are advantages and disadvantages of both systems. For the fat clients, the main advantage is that processing power is controlled locally, and hence a need for more resources can be satisfied relatively easily. However, the disadvantages are

[1] To clarify roles, ONS set up Business Data Linking (BDL); and BDL set up the Virtual Microdata Lab to provide the technical solution. The VML is now used by other areas in ONS as well as BDL, but the resource continues to be managed and developed by BDL. Hence references to ONS are to the overall policy; BDL, to how the business data setion operates; and VML to the underlying technology.

- Data travels over the network, which is slow and may be insecure
- Temporary files are stored locally
- Users may be able to store local copies of datasets
- PCs have to be individually configured for local security
- Fat clients with access to network drives cannot easily be made secure
- Individual machines need to have the sufficient processing power and analytical software.
- Processing power cannot be redistributed between machines

The thin client solution addresses all these problems. All processing is carried out on the central server; the security of the system is determined wholly by the security of that server, irrespective of the individual clients. Network traffic is minimised, and performance is almost identical at any location. The only software required at the user end is the thin client interface, which may be as simple as a web browser; all other software can be managed on the central server. Finally, central processing resources can be shifted more easily to demanding tasks. The major disadvantage of thin client systems is a dependence on the performance of the central server, which has been an issue with the VML at times.

2.3 Results: use of business microdata at ONS

Since January 2004, all BDL researchers have been using the lab. Because of the security and efficiency of the lab, BDL has been able to expand output enormously. In eighteen months, research output has expanded from roughly 10 projects and 15 accredited researchers to over 90 projects and over 150 researchers.

Whilst a large part of this is due to an increase in general academic research, there has been a significant increase in the use of the data by other government departments, either directly or indirectly through academics. There has been a significant movement towards more evidence-based policymaking in the UK, and the BDL has been able to support a large amount of this. Users have included the departments for industry, arts, defence, agriculture, overseas trade, and tax; the Treasury; the Bank of England; regional authorities; and training councils.

A large part of this government work is on programme evaluation (or policy impact analysis, as it is sometimes called). Recent projects have included the effectiveness of small business support programmes, export subsidies, tax changes, tourism promotion, and the National Minimum Wage, all supported by the relevant government departments.

Of more direct relevance to the ONS is the increasing use of the lab by internal staff. As well as investing in technology, the analytical capability of the economics departments at ONS has been expanding. These departments are now using the combination of technology and analytical skills to produce aggregate statistics; for example

- estimates of investment in intangibles
- an analysis of the impact of R&D capitalisation, including methodological recommendations and an experimental series
- productivity series consistent at the micro and macro level
- reconstruction of the main business register on a historical base to improve demography analysis

These are major long-term projects with significant implications for national accounts, only now made feasible by the combination of secure easy access to a wide variety of microdata sets and advanced econometric skills.

Finally, the VML has also provided a secure facility for other types of microdata: it currently houses Census and Labour Force Survey data at levels of detail not available on externally distributed datasets; and is being evaluated as a potential home for personal, medical and mortality data[2].

3 Issues

The ONS system of microdata access provides a flexible and secure system for the analysis of microdata. However, the consequences of this system are a series of risks which are relatively new to ONS. Not all the risks involved with providing such open access to microdata to can be managed away, and so ONS has had to re-evaluate and explicitly define a new series of risks.

3.1 Access to microdata and the irreducible person risk

The VML provides an extremely secure solution. The electronic removal of data from the VML by researchers is not practically possible. Independent verification of the technology and procedures (Echelon 2004) described the VML as meeting or exceeding best practice in almost every area; in some areas, such as disclosure

[2] Note that the VML is a last-resort solution for situations where the data cannot be released. ONS uses a variety of distribution channels (such as the UK Data Archive), and the default is to anonymise and release data if possible; running any kind of lab is expensive and inconvenient for all concerned. For example, social survey data is typically anonymised and distributed under licence, which has meant a much longer and wider tradition of analysis of social data than business data. Hence, there is no immediate pressure to place social data in the lab, as adequate and efficient solutions already exist.

control, BDL procedures far outstripped alternative methods across the UK government.

Nevertheless, providing access to identifiable microdata does give rise to an irreducible risk; that is, that a researcher could identify a company, and then remove information about the company through non-electronic means. This risk cannot be reduced for company data. First, company data cannot be anonymised effectively without destroying almost all information content. Second, even if all paper and writing materials were banned from the VML environment, it is not possible to stop a researcher remembering items of information.

As technological risk reduction is not possible beyond this point, ONS therefore concentrates on the "safe people" part of the security framework. In particular, researchers must come from "trusted" organisations (ie those where ONS is reasonably confident there is no conflict of interest), must themselves have a credible research background, and are made aware, through the BDL training programme, of the consequences of abusing the trust.

For this last point, the credibility of any sanction is important. Prior to 2002, researchers were brought in on "£1 contracts" of the type common in many countries. These were stopped, partly because they did not meet the spirit of the law, but also because they gave no credible grounds for sanctions in the case of a breach of trust: either the ONS could claim its £1 back, or the researcher could face a criminal prosecution. Neither solution is particularly credible. However, the new contracts tie researchers to their institutions, and enforce a form of collective responsibility. This means that, in the case of a breach of confidentiality, ONS will approach the researcher's institution who will be responsible for disciplinary action in all but the most serious of cases. This gives a much wider range of sanctions, and is also far more credible than the "all-or-nothing" nature of the £1 contracts. In addition, BDL is in discussions with key academic funding bodies about tying funding to the "trustworthiness" of the institution.

This has not resolved all problems. Some researchers reject the implication that they might not be trustworthy, whilst there is still some suspicion within government circles that academics might not show an appropriate awareness of the confidentiality of data. In addition, BDL has not yet resolved the position of private sector consultants; there the possibility of conflict of interest is felt to be too high in general to allow direct access to commercial data. Finally, there have been no agreements with international bodies, because of the difficulty of finding an effective legal framework. Nevertheless, the system as it stands seems to be generally accepted by most parties.

3.2 Disclosure control in a research environment

Statistical disclosure control (SDC) in a research environment is a fundamentally different animal to that used to produce aggregate tables or anonymised datasets. In both those cases, there is a finite set of outputs and a well-defined structure for the data.

In contrast, the purpose of a research environment is to combine data in innovative ways and produce a range of outputs which would not normally be generated by the statistical body. This means that

- Automatic disclosure control of most outputs is not possible
- SDC rules, whilst often extremely detailed, are rarely comprehensive and are almost always open to challenge for particular examples

For a more detailed discussion of the issue, see Ritchie (2005).

As manual checking of all outputs is required, and as it is difficult to define in advance acceptable outputs, ONS makes training in SDC compulsory for all researchers and staff involved in the lab. The training method is based on an understanding of principles, backed up by examples to illustrate particular issues; there is very little discussion of rules per se. For example, although the BDL threshold of a minimum 10 units for each cell is given, the main discussion of this rule illustrates cases where the rule can be adjusted or ignored, where it will be tightened, and what information researchers need to provide to BDL to allow them to make a meaningful judgement.

This requires some commitment on the part of both researchers and lab managers, as researchers are required to physically attend the training sessions. However, the response among researchers has generally been positive, with almost all taking the view that the time spent on the course has been productive. Certainly the experience of BDL before and after the introduction of the training course, and of other areas in ONS, is that this make a significant difference to the time taken to clear outputs.

Again, there are still unresolved questions. One is that, under the BDL model, there are few absolute rules as to what is allowed, and researchers have requested more clarity over acceptable outputs. In practice, once researchers become used to the data, the major problem for BDL is the quantity of output produced.

It has been argued that, by giving researchers a detailed insight into how disclosure control is carried out, the risk of them subverting SDC procedures is raised. At BDL, the view is taken that, should a researcher really want to remove output surreptitiously, no practical method of checking output is going to prevent this (even if it did, researchers could remove data in non-electronic ways, as described above). On the other hand, by involving researchers in the process of checking for outputs,

BDL encourages researchers to take a pro-active approach in avoiding problematic outputs. The results can be seen in that several of the examples BDL uses in its training have arisen from questions posed by researchers. Hence, overall BDL has decided that the benefits from more effective disclosure control outweigh the disadvantages of providing researchers with ways to avoid the control.

3.3 The possibilities of remote access: what is a "safe place"?

The VML is potentially accessible from any location connected to the internet or a phone line. However, at the moment, technically access is limited to ONS sites. An investigation is under way to put in place equipment to allow secure access across other government sites. In the longer term, it is possible to envisage secure sites being set up at universities, along the lines of the US Census Research Data Centres.

Whilst there is a technical element to ensuring the security of access, the major concern here of providing access on non-ONS sites is procedural. Off-site access to a research lab involves devolving responsibility for the physical security of the lab site to a third party; and because of the irreducible person risk noted in section 3.1, this means that some of the risk is also being devolved.

There are a variety of models for this. In the US, Census Bureau employees are stationed at each research centre, for direct supervision of researchers. This is costly but ensures the Census Bureau keeps the risk in-house. At the other end of the scale, Statistics Denmark, using the same technology as ONS, relies entirely on its "safe people" policy. Access is available from the desktop of any researcher, subject to both the researcher being approved and technological methods being in place to ensure that the researcher can only log on from an approved institution. This is a cheap, flexible and very secure option, save for the problem of ensuring that only approved users are physically in front of the terminal.

In preparation for extended access, the ONS site has begun providing lab access at its Southport office. This site was chosen as a test case because there is no local expertise in any of the datasets available in the lab. The facility is managed locally purely for access to researchers in a measure halfway between the US/Danish models. The local managers are responsible for escorting researchers on to the premises, and observing researchers to make sure there are no attempts to write down confidential data. A set of protocols for "safe places" and "safe kit" (that is, a standard working environment and technology for a remote lab) has been agreed.

This solution is not universally supported. In some areas it has been argued that devolving the responsibility of physically supervising researchers is increasing ONS'

risk unacceptably. Researchers, on the other hand, tend to view the restrictions to certain physical sites as an unwarranted limitation on research. This solution is more costly than the Danish solution; it is also potentially less secure than the US system, as non-specialist local managers may not understand whether confidential data is being removed or not. To address this point, BDL authorises the local manager to "remove first, question later"; that is, in the event of any suspicious activity, the local managers will err on the side of removing researchers from the lab. This semi-attended lab facility has been available since the summer of 2005, and so in the long term it remains to be seen whether this approach is appropriate.

3.4 Are ONS official statistics liable to attack?

One concern expressed with research access to microdata is the risk to the reputation of official statistics. If researchers are using the same data as that used to generate aggregate statistics, what happens if

- Researchers produce different aggregate statistics?
- Researchers discover significant errors in aggregate statistics or the underlying data?

This latter point has been accepted as a risk with potentially beneficial consequences to ONS. Although there is a risk of embarrassment of ONS, this is felt to be outweighed by the quality-control aspect of letting a large number of researchers stretch and twist data. BDL is currently formalising feedback arrangements so that queries and comments from researchers can be relayed to the data providers effectively.

The former problem is more subtle. "National Statistics" (ONS official non-experimental outputs) are generated by weighting survey and administrative data to reflect population characteristics. This is a complex process requiring large teams of people, and goes well beyond the simple weighting which researchers would typically use to produce descriptions of the data. Researchers' interest is in marginal analysis and sample description, rather than population totals. Hence, when researchers do produce figures which are comparable to "official" figures, they are unlikely to agree, and the reason for the difference may be quite technical. It is possible that ONS' reputation may suffer from a misinformed reading of outputs produced by different methods, particularly as the outputs of external researchers are not subject to the same stringent quality checks as ONS official statistics.

BDL took the decision not to restrict outputs, partly because it was difficult to police, but mainly because it was thought extremely unlikely that research papers would be compared to official statistics, even if the aggregated totals were similar. The readership for these publications is quite different, and most analysts reading technical papers would be familiar with the differences between aggregate and marginal analysis. The solution BDL has adopted is to explicitly exclude (by contract) research outputs from being described as National Statistics. A standard rubric is given to researchers to be added to the contracts. For internal ONS outputs, BDL staff papers state that the work is "based on research version of the ONS datasets, which may not exactly replicate National Statistics produced by the ONS". Although not without criticisms, the current position of BDL is to allow the policy to continue as there is no evidence yet that this is inappropriate or that there is a practical alternative – or that it is unnecessarily confusing readers.

4 Conclusion

ONS has spent some time and thought building a research facility for confidential microdata which is believed to offer an optimal combination of very high security, simplicity in operation and management, and flexibility in use. The lab and the procedures developed have proved extremely popular, and use of the data has rocketed with significant benefits to ONS, other parts of government, and academia.

However, even when the problems of providing access to data have been solved as far as possible, a number of significant risks remain. These can be seen to be irreducible risks; that is, an organisation has to decide whether the remaining risk is acceptable or not – there is no practical possibility of reducing the risk further without significantly affecting the operation of the solution. For example, it is necessary to trust researchers with identifiable data; otherwise, it is not possible to analyse business microdata except through costly and inefficient proxies.

In each of the four security aspects on which BDL and ONS base their access models, there is at least one irreducible risk: for safe people, the trust risk; for safe outputs, the disclosure risk; for safe settings, the location risk; and for safe projects, the reproduction problem. These are all contentious, and all are under continuous review. However, at least ONS does have a clear perspective on exactly how much risk is inherent in its microdata operation.

References

Echelon (2004), *Security Review of the ONS Microdata Laboratory: Part I*, Echelon Security Systems

Ritchie, FJ (2004). Business Data Linking: Recent UK Experience, *Austrian Journal of Statistics* v33:1/2

Ritchie, FJ (2005) *Statistical Disclosure Control in a Research Environment*, mimeo, Office for National Statistics, London