WP. 21
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (iv): Balancing data quality and data confidentiality

Sharing risks, sharing benefits: Data as a public good

Prepared by Felix Ritchie, ONS, United Kingdom and Richard Welpton, Secure Data Service, UK Data Archive, University of Essex, United Kingdom

Sharing risks, sharing benefits: data as a public good

Felix Ritchie[*] and Richard Welpton[**]

[*] Micro-data Analysis and User Support, Office for National Statistics, Newport, Gwent NP10 8XG felix.ritchie@ons.gsi.gov.uk
[**] Secure Data Service, UK Data Archive, rwelpton@essex.ac.uk

Abstract: Release of confidential data for research creates asymmetric costs and benefits for the National Statistical Institute (NSI) and the wider community. Considering access to publicly-owned data as a 'public good' (as defined by economists), the standard theory suggests that these asymmetries will lead to under-provision of data resources by an NSI acting in its own best interest. This paper uses the 'public goods' framework to demonstrate that sub-optimal data release is almost guaranteed by the relationship between the NSI, the rest of government and the research community.

However, public goods theory also has a well-established suite of techniques for addressing market failure. This paper applies these techniques to examine how society's preferences for data access can be met efficiently. In particular, we show the need to share, broadly and explicitly, risks and benefits; and we suggest practical steps to move in a positive direction.

1 Introduction

National Statistics Institutes (NSIs) are facing considerable pressure to make more data available for research use, particularly micro-data. This creates costs for the NSI: direct costs of data provision, and the indirect costs (for example, reputation) associated with the risk of unauthorised release of information. At the same time NSIs are largely isolated from the benefits of research use of data. This means that NSIs have relatively few incentives to consider a data access strategy; and in practice many historically have taken ‘no release’ as a default option for all but aggregate statistics.

This has consequences for society. Limiting data release may reduce risk for the NSI, but it also lowers the ability of society to make judgements based upon evidence. Even if data are released, they may be so constrained to be of limited value or even wrong (see e.g. Lane, 2007, for examples). However, the absence of useful research is rarely in the objective of the NSI.

In economics this problem is described in the context of ‘public good theory’. A public good is one where the provider of a good is typically unable to recover the costs of providing that good, because he is unable or unwilling to control access to that good. As a result, he will provide less of that good than society would require. The classical response to this from economic theory is to have an enlightened agent working on society's behalf and ensuring that the producer of the public good receives sufficient funds from society to produce an optimal output.

In the case of data access, the public good is research output from that data. The NSI cannot meaningfully control how those outputs are used and, crucially, valued; and so the NSI concentrates on its statutory tasks and gives less importance to the wider benefits to society from providing access to (confidential) data. As a result, data access is likely to be sub-optimal for society.

This may seem curious: as NSIs are government bodies, they surely are the ideal candidate to act as the 'enlightened agent'? In reality, this paper argues, NSIs are likely to act as any other producer who is required to focus on his internal objectives.

This is not a necessary outcome. This paper considers some practical changes to the relationship between the NSI, the rest of government, and the research community that can lead to improved outcomes. There is no magic formula, but there are steps which can be taken to improve the decision-making process. This discussion focuses on data released by NSIs but it is also applicable to other data owners.

The next section summarises public goods theory, and applies it to data access. Section 3 considers the problems with the model of the NSI as the 'enlightened agent'. Section 4 then proposes ways to reduce the imbalance in risk and reward. Because this is an area dominated by un-measurable effects and value judgements, trying to identify 'the' right level of data access seems an impossible task; and so the paper focuses on ways to move in useful directions. Section 5 concludes.

2 Public goods and data access

In economics, a 'public good' is defined by two key characteristics:

- *non-excludable*: that is, the good can be 'consumed' by all - there are no barriers preventing access
- *non-rival*: that is, consumption of the good by one individual does not exclude someone else from consuming that good

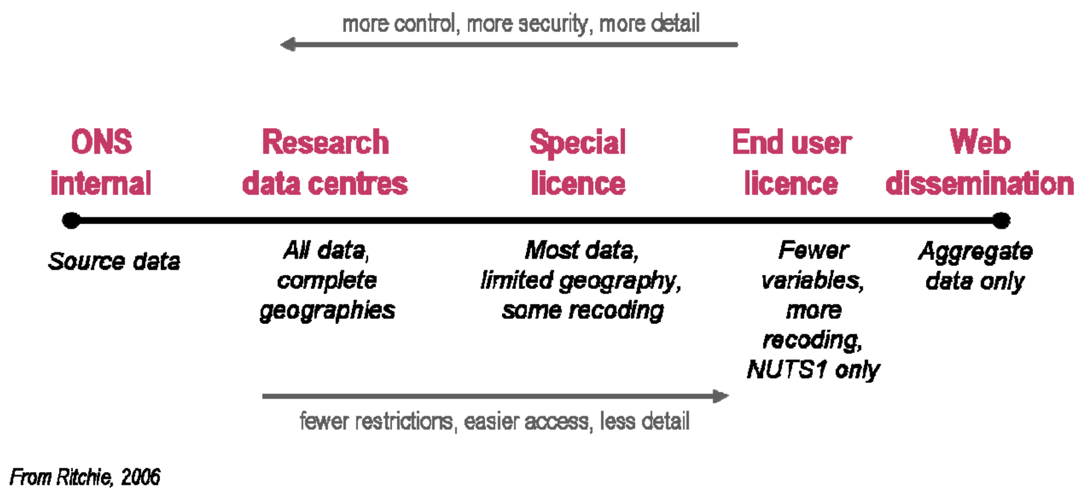
The classic example of a public good is defence - we are all protected by our armed forces, nobody is excluded, and my having an army to protect me does not stop you having that same army to protect you. On a smaller scale, consider someone planting flowers in the street in front of their house. The flowers are non-excludable – everyone walking past can see them – and non-rival – one passer-by enjoying the flowers does not reduce the number available for others to admire.

Private goods are excludable and rivalrous. If one earns enough income, one may be invited to purchase a Ferrari (it is *excluded* from those unwilling to pay), and in purchasing this good, another individual is denied (the *rival* must be content with another car).

The reason why public goods cause concern is because, left to private providers, they would be unable to gain sufficient funds to cover the cost of a socially optimal provision of the good. Consider the case of the flowers, above. It might well be that a bigger, brighter display would give greater pleasure to the community, at some small cost to the gardener. But why would the gardener put in the extra work? He cannot charge for looking at the flowers in the street; and so he makes a decision about which flowers to plant entirely in his own interests.

This economic framework can be applied to data access. Figure 1 offers a characterisation of data release in the UK, the 'Data Access Spectrum' (Ritchie, 2008). A number of access mechanisms are available, depending upon the confidentiality of the data.

Figure 1: Data Access Spectrum



Consider the case of the UK Office for National Statistics (ONS)[1]. Aggregated data can be downloaded freely from the website. These data are non-excludable and non-rivalrous - public goods.

For more detailed data, the problem becomes more complex. For example, the UK Data Archive supplies data under an 'End User Licence'. These are available to all UK academics. This electronic file is clearly non-rival as downloading does not affect others' ability to download; it is theoretically excludable but in practice these

[1] In this paper, we will use ONS and the UK Data Archive as examples; however, this is for illustration only and should not be taken to mean that these bodies are unique or exceptional in the cases mentioned.

data are available to all bona fide researcher through some mechanism. The UK Data Archive also supplies ‘Special Licence’ data, which can only be downloaded following commitments to certain security arrangements (such as non-networked PCs). These data are also non-rival but is more excludable.

There are also restricted datasets, accessible through a research data centre (RDC) at ONS or the UK Data Archive’s new Secure Data Service. Access is excludable, both directly by RDC managers, and because researchers may have insufficient funds to travel to ‘safe sites’. Access is also rival: if one researcher sits at a terminal, it prevents somebody else from doing so. Even if physical access is not limited (as at the Secure Data Service), researchers will be competing for finite computing resources.

So there seems to be a relationship between data detail (and confidentiality) and its ‘public’ nature; and between detail and the ability to fund it. There is no possibility of charging for data publicly available on the web. ONS therefore takes its decisions about what to put on the web based upon its internal view of its ‘public task’. For RDC access, however, ONS should be able to charge to cover the cost of provision of that service. The public-goods problem then appears to be restricted to uncontrolled data sources.

However, this concentration on access is a red herring. The value of data access to society is the output that it produces, whether it is a statistic taken from a web site on the current unemployment figure or a detailed analysis of the impact of health interventions. Almost all statistics and research are designed to go in the public domain, where they will be non-excludable and non-rival by design.

When considered in this way, the issue of the ‘right’ level of access becomes important for all data, not just the public data. Provision of data through controlled mechanisms can be provided and funded by NSIs and researchers making their own private decisions, but is this encouraging the ‘right’ level of research?

The issue of how much anonymised/aggregated data an NSI should provide is currently being widely debated. For this paper, this is of limited interest as the current trend internationally is to make as much data available as possible, with reasonable costs and funding being the limiting factor.

Of more interest is the case of restricted data where the NSI has identified that there is a non-negligible risk of disclosure from the data (identification of the observations). In these cases, the NSI has much more discretion to take unilateral decisions over data access. Part of the NSI’s ability to take independent decisions comes from the concentration of expertise on disclosure control in NSIs and associated bodies; however, an equally important source of power for the NSI is the view that the NSI is the only body competent to evaluate the risks and rewards of data access. This paper aims to probe that.

3 The provision of confidential data: risks and rewards

Identifying the public benefits of greater access to confidential micro-data is difficult: quantifying these benefits even more so. If a data owner is persuaded to provide access to their confidential data, what rewards do they and the public receive? Conversely, who bears the costs?

3.1 Benefits of providing data access

The most obvious beneficiary of data release is the researcher, particularly for restricted data. A culture of ‘publish or perish’ has created a bias whereby researchers’ primary objective is to publish in leading academic journals. The probability of this occurring is of course heightened by using more confidential ‘hard-to-reach’ data.

Unfortunately, explaining the results of these analyses, and the implications for public policy and society, tend to take a back seat. Often, researchers claim that it is in the public interest that their research is conducted, but often this means that it is in their own interest that the research can be carried out. There is very rarely a direct relationship between research and public policy, for example[2]. However, there clearly are benefits to society from improved access to data from the accumulated pool of knowledge. For example in the UK, the Low Pay Commission has a formal mandate to take account of research in its decisions on the level of the National Minimum Wage. Research might be contestable and contradictory; but the LPC makes recommendations on the total accumulated knowledge.

3.2 Costs of providing data access

These are two-fold. First, access to confidential data requires more administrative processes and may require sophisticated IT systems, or could involve significant expenditure on anonymisation. These costs are measurable and have a direct impact on the organisation’s resources.

The second type of cost is an indirect cost that is borne by the data owner in the event that the data is mis-used. This could include the loss of reputation for handling confidential data securely, declining response rates, and financial and legal sanctions. For an NSI, these consequences could be severe, so that even if the probability of mis-use is minimal the expected cost of the risk is still significant.

This indirect cost also falls on the researchers. A significant loss or mis-use of data may lead to their opportunities for future research being cut. This is, however, rarely on the radar of researchers.

[2] This is less true for harder sciences, where, for example, a clear refutation of a treatment or demonstration of an environmental impact might have an immediate effect on policy. In social science, for whom the release of NSI data is crucial, this is rarely the case.

3.3 Benefits of not providing data access

Most obviously, the benefit of not providing data access is that the costs mentioned above are not incurred. The data owner can also benefit from increased reputation as a result of taking a hard line on data release, although this is likely to be of limited value - beyond a certain point, the NSI may come to be seen as a malign influence on policymaking and overruled.

3.4 Costs of not providing data access

For society, the cost of not providing access to data is the loss of the ability to inform decisions with evidence; for example, evaluation of business support schemes requires the counterfactual provided by NSI business micro-data.

There is also a hidden cost to data owners, particularly NSIs. Secondary analysis of micro-data (where 'primary' analysis generates the headline aggregate statistics such as inflation, unemployment etc, and 'secondary' analyses are more detailed micro-data research projects) can also benefit them. Researchers, who are at the forefront of their fields, can play an important role in shaping survey design to ensure that NSIs are collecting data relevant to society. They can also provide feedback on quality of statistical data, particularly the multivariate characteristics which the NSI rarely has the ability to review. Finally, an educated class of data users is more likely to interpret NSI data intelligently. So a second type of cost is actually borne by the data owner when access to data is denied.

3.5 Will NSIs provide an optimal level of data access?

An optimal level of data access must be one where the balance between data confidentiality and ability to conduct research fairly, is accomplished. The problem in achieving this solution is that the risks of providing access and the costs of doing so are identifiable and measurable; and they fall on the data owner. If the data is misused by the researcher, the data owner will be blamed; if data is to be anonymised, it is the data owner's responsibility; if there is a need for access in a safe environment, the IT cost is probably borne by the data owner.

On the other hand, it is more difficult to assign a monetary value to the benefits of providing data access – some of these may only be realised far into the future. More importantly, it is difficult to identify

Data access therefore presents itself as a classic public good: the supplier faces a known cost structure; the benefits of access to statistics are non-excludable, non-rival and largely unquantifiable; and they are dispersed across time, space and users. Left to itself, then, the likelihood of an NSI providing an optimal amount of data access is small.

4 Solutions

In this section we propose three ways forward. None of them will guarantee the 'right' level of data release, but all are designed to improve the transparency of the decision-making process. By extension, that also makes the process more open to challenge.

4.1 The contract between researchers and data owners

Most researchers who access restricted data sign some form of contract with the data owner. Typically, this will specify the researcher's requirements to use the data securely. However, as noted in Desai and Ritchie (2009) and Ritchie and Welpton (2011), the widespread 'be grateful; be careful' approach to researcher agreements can be at best alienating and at worst counter-productive. If the researcher shares the NSI perspective on risk, the data owner is no longer shouldering so much of the burden and so may be prepared to make more data available. Making researchers responsible can persuade data owners to provide more access because they will not take all the blame should the worse happen.

However, in the context of public goods, there is an additional service the researcher can provide: to give the NSI some measure of the value of research. Researchers, as noted above, are typically using the data for their own benefit; however, measures of the value to society to this are available in terms of researcher grants, journal publications, government reports and so on. CNSTAT (2005, ch.5) describes a range of possible output measures.

A positive contract (formal or implicit) would encourage researchers to think about how the benefits of their research will impact society – and then feed that back to the NSI. For example, the Secure Data Service mandates that researchers inform them of publications and presentations. Researchers are also encouraged to complete Case Studies, with the aim of informing the wider society of their results.

But the key point is that the researchers should see publicising their work to data owners inherently beneficial, rather than something done to fulfil a contractual obligation. This implies that the NSI offers something additional to researchers. The most obvious route is an additional path to publication. For example, the IAB in Germany publishes a range of research compendia, topic summaries and monographs utilising the work of those accessing the various IAB data resources[3]. This increases researchers' visibility at negligible cost to them, while also providing the IAB with a handy theme–structured guide to the range of outputs produced and the user base.

The second route is to increase researchers' engagement with data creators within the NSI. This could be seen as a cost to the NSI, distracting the data producers from their key work. On the contrary, it is the authors' experience that the producers of data

[3] http://www.iab.de/en/publikationen.aspx

within the NSI generally welcome periodic interaction with experienced users, and will exploit this resource. For example, the E-Commerce Survey team at the ONS worked with researchers to ensure that the survey design reflected useful questions . For researchers, this contact is a way of feeling that their research has 'impact'.

4.2 The relationship between NSIs and the rest of government

The main value of research output comes from its use by government. There is no requirement on government to make good use of statistical data, and of course other parts of society use statistics to bring about change in society. However, government is the dominant recipient of research output, and we will make the assumption that having more information available is, on the whole, a positive benefit.

Access to data therefore provides benefits; not having access to that data creates some loss to society. Can we measure this so that the NSI can be 'paid' for its data access policies?

The difficulty, as noted earlier is that the value of research is difficult to assess. For example, the UK Treasury's periodic analyses of productivity increasingly draws on research on restricted micro-data. But the Treasury made policy decisions before micro-data was available, and there is no single 'killer' research paper which unequivocally provides a guide to policy.

However, the inability to quantify does not justify avoidance of any assessment. Assume, for argument's sake, that there only two government bodies: the NSI, which provides researchers with data, and the Treasury, which uses that research. The Treasury may not be able to say what the value of the research is, but it should be able to say whether the research it is getting is valuable and sufficient.

However, the Treasury is also aware that the NSI runs a risk by making restricted data available. That risk is partially borne by the Treasury too: a disastrous disclosure of data would affect the Treasury's ability to carry out further research.

Treasury and NSI therefore share a joint interest in both the use of the data and in the risk generated by that use. Both organisations may differ in their views of the relative component, but the point is that both explicitly acknowledge the contradictory tensions. The appropriate level of data release is a negotiated output, to which both parties agree.

How does this help data access? As noted above, one of the problems is that the risk taken by NSI is not valued externally, and the value of NSI data access is not fed back to the NSI. These are difficult to quantify; on the other hand, both organisations should be able to take a qualitative position on their own preferences. The crucial point is that both NSI and Treasury consider risk and benefit jointly and agree a position. If the two cannot agree, then the NSI can fall back on its unilateral position. But the fact that the Treasury is forced to recognise the risk, and by extension accept and <u>defend</u> any agreed position, can change the NSI's risk-reward balance.

This does not solve the public goods problem. It brings in some major customers, so that more of the 'output' of the NSI may be appropriately valued, but government is only part of society and will have its own interests. However, the point is that this is a *movement towards* a more societally beneficial outcome.

4.3 The NSI as an enlightened agent

In practice, an NSI is an organ of the government, and in theory should act as the enlightened agent. However, civil servants tend to be risk-averse (Ritchie, 2010), and the rotation of staff and consequent loss of knowledge can hamper efforts to develop a long-term working relationship between the NSI and the outside research community.

Instead, the use of a third party service provider (the Secure Data Service is an example) can act as an enlightened agent instead. Such independent RDCs, managed by researchers for researchers, can successfully provide intermediation between NSIs and researchers. For example, established processes and procedures can convince the NSI that confidential data access can be successfully managed. Researchers can be confident that they will receive the support they require to undertake their analysis.

These third-party providers shoulder much of the risk in providing access to data, transferring the risk away from the NSI, and transferring some of the risk to the researcher. This encourages the NSI to provide more confidential data for access by means of this RDC.

Most importantly, the duties of the third-party researchers is likely to be covered by a contract and, ideally, a risk assessment by the data owner, so that both parties know exactly what they are providing. This is what is missing in the relationship between NSIs and their clients in government.

5 Conclusion

Owners of confidential data are unlikely to provide a socially optimal level of access to restricted data. This is because the end product, research, is a public good, available to all without exclusion or restriction. Data owners therefore cannot enter into contracts with users of research outputs to ensure that they get sufficient recognition for the value of the product, or an appropriate recompense for the risks taken. In addition, the value of research outputs is largely un-measurable, whilst assessment of the risks involved in releasing data is entirely the province of the data owner.

However, there are some practical steps to be taken towards improving the outcome for society. The first is for researchers to provide some of the missing information about the value of outputs. The key is to design the engagement process such that this produces positive benefits for both parties; and there are examples from different countries where this occurs successfully.

The second step is for data owners to engage with the end users of research outputs, in government. The aim is to achieve a shared understanding of both risks and benefits; more importantly, for all parties to agree that the level of risk and reward is appropriate; most importantly, for users of research to be prepared to defend the risks taken by the NSI.

Interestingly, the one place this is currently carried out is in the provision of third party services. For example, organisations such as the Secure Data Service or the NORC Data Enclave have clearly defined levels of risk in the service agreements with data owners. Perhaps these could be a model for the relationship between NSIs, the rest of government, and the research community, in the future.

References

CNSTAT (2005) *"Expanding Access to Research Data: Reconciling Risks and Opportunities"*; National Research Council; Washington, DC

Desai T. and Ritchie F. (2010) *"Effective researcher management"*, in *Work session on statistical data confidentiality 2009*; Eurostat

Lane J. (2007) *"Optimizing the Use of Micro-data: An Overview of the Issues"*, Journal of Official Statistics, Vol.23, No.3, 2007. pp. 299–317

Ritchie F. (2008*) "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. (2010) *"Risk assessment for research access to sensitive microdata"*; presentation to John Deutsch Institute, Ontario/WDA 2010; May; http://jdi.econ.queensu.ca/sites/default/files/JDI - Access to business data v1.ppt

Ritchie F. and Welpton R (2011) *"Incentive compatibility in research data access"*, mimeo, ONS/UK Data Archive