FEATURE

Felix Ritchie
Office for National Statistics

Secure access to confidential microdata: four years of the Virtual Microdata Laboratory

SUMMARY

This article explains how the Virtual Microdata Laboratory (VML) has become the Office for National Statistics' solution to providing access to sensitive microdata while maintaining the confidentiality and security of the data. In the four years since it was set up, the VML has gone from almost nothing to becoming a major resource for UK academic researchers. The VML has enabled both more detailed and wider research, and has influenced policy making at all levels. Looking ahead, the VML faces significant challenges and a bright future.

The Virtual Microdata Laboratory (VML) was first launched by the Office for National Statistics (ONS) in January 2004 to provide secure access to confidential business survey data for research purposes. From small beginnings, the VML has grown to become a key element of ONS's data access strategy. It is now, in practice, the default secure solution for data access across the Office and, increasingly, a major player in cross-government access, helping transform the research culture in ONS.

This article describes how and why the VML was designed and grew, where it is now, and the prospective opportunities that lie ahead.

The rationale for the VML

ONS collects a large amount of data on individuals and businesses and uses this to produce statistics about all aspects of the economy and society. Most of this information is presented in a highly aggregated form, to maintain the confidentiality of those supplying the data. However, the underlying microdata used to create these aggregate tables are a major research resource for the UK. Meeting the needs of researchers while strictly maintaining confidentiality is a priority for ONS.

For social data, this is addressed by 'anonymising' the data and placing the resulting data set in the UK Data Archive (see www.ukda.ac.uk) or, for files with less anonymisation, releasing the files under a 'special licence' to limited groups of researchers.

This is rarely feasible for health and business data because the distribution of characteristics for these types of data means that anonymisation tends to destroy all of the data of interest for statistical research purposes. Hence, such data are often only available, unidentified but without anonymisation, in research data centres (RDCs). Although ONS and the Economic and Social Research Council (ESRC) did set up an RDC for health data in the 1980s, there was no equivalent for business data. This is because RDCs tend to be costly, in time and money, for both researchers and data providers, and so of limited appeal.

In the late 1990s, a number of academic researchers starting looking into the business data collected by ONS. The ONS business data are unique within the UK in the scope and detail of their information and the potential for research is enormous. However, the legal status of the data made access difficult and the lack of prior record-level analysis meant that the data were often not organised usefully for research.

At the end of 2002, ONS began to formulate a considered strategy to resolve this. The first stage was to develop a security model (see **Box 1**) and to resolve legal issues surrounding access to data. A pilot VML using 'thin-client' technology (see **Box 2**) began operation in January 2004, and ONS had a generic solution for providing secure access to confidential microdata.

Conceptually, the VML is similar to a well-guarded physical laboratory where visitors are searched on entry and exit and allowed no communication with the outside world from inside the laboratory. This is how the balance between confidentiality and usage is managed. Because the locks on the 'virtual door' are strong, ONS can give access to sensitive data and allow more freedom to researchers inside the safe environment. The VML (and similar facilities in other countries) does this through software, allowing a much more flexible deployment; this is why the thin-client model has come to be seen as international best practice in recent years.

Growth and development

In this section, developments in four areas are considered: users, data, projects and outputs.

Users

In January 2004, the VML was providing access to a total of ten academic researchers, in four research groups, and no government or ONS researchers. Four years on, and the VML supports around ten users a day from all areas of government and academia, and has trained around 400 researchers in the use of confidential data.

Box 1

The VML security model

The VML security model recognises that no single solution can be expected to provide an absolute guarantee of security at a reasonable cost (see Ritchie 2006). Hence, the VML embraces a series of interlocking security controls for 'safe' access to confidential data:

- safe projects – access needs to be for a valid statistical purpose
- safe people – researchers can be trusted to use data appropriately and follow procedures
- safe data – the data itself are inherently non-disclosive
- safe settings – the technical controls surrounding access prevent the unauthorised removal of data
- safe outputs – the statistical results produced do not contain any disclosive results

For the VML, 'safe' data are included for completeness, but for planning purposes it is always assumed that the data are inherently unsafe.

Safe projects, safe people and safe settings are designed to protect the data from deliberate misuse; safe settings (again) and safe outputs are designed to prevent accidental releases of data. Hence, the VML security model is designed to ensure that there are overlapping controls for each identified risk.

Box 2

Thin-client research data centres

Thirty years ago, all access to computers was 'thin client': massively powerful central computers would do the processing for IT specialists working over a network. With the advent of PCs, having all the processing power needed on one's desktop became the norm. This was true for research data centres too: they involved bringing the researcher to the data, often in physically controlled spaces.

However, recent developments in technology, particularly for Windows™ computers, have caused thin-client computing to be re-evaluated. For RDCs in particular, there are significant advantages. First, the security of thin-client systems is far more easily controlled: for example, the VML has been using strong encryption for all data traffic since its inception, something which has only become a wider requirement for government IT systems in the last year. Second, thin-client systems means that data can be managed centrally, a great advantage when data are being linked and updated regularly. Third, thin clients mean that the user no longer needs to be physically close to the data store; researchers can access the VML from any ONS site without loss of performance.

When the VML was set up, Denmark was the only other country in Europe using this technology. In the four years since, use of this technology has grown considerably, and thin-client solutions are now widely considered best practice.

Figure 1 shows the number of researchers attending the VML training course, which is compulsory for all VML users. This can only give an impression of the cumulative use of the VML: the training course is not compulsory for researchers who do not need to visit the site, and the course is also used by the VML team as a way of demonstrating the security of procedures to potential data depositors. However, it is clear even from this rough approximation that both the number and range of users of the VML has expanded remarkably over the period. While academics still make up 55 per cent of the attendees, over 30 per cent are researchers in other government departments, at levels ranging from junior economist to Chief Statistician.

Figure 2 shows the daily usage. In terms of regular daily use of the VML, internal ONS demand has generally kept pace with that of external visitors, but in the last 18 months, ONS use has accelerated. It is difficult to establish exactly the amount of internal ONS usage, but a new monitoring system being introduced in 2008 will allow a much more accurate assessment of both internal and external use.

Data sources

The VML was originally set up to provide access to business data and relied heavily on early work by academic researchers[1] to create workable microdata sets from ONS's archives. Ritchie (2004) describes the early work of these researchers and the problems encountered in trying to make data collected for one purpose usable for another.

The most important of these early data sets was the Annual Respondents Database (see Barnes and Martin 2002 and Robjohns 2006). A longitudinal database of firm-level survey responses, it was constructed from the structural business surveys used to generate a substantial part of the UK's annual GDP and related National Accounts estimates. For the production sector this was available back to 1973 – although much of the pre-1996 data are still labelled 'unknown'. These data have now been extended to include the services sector, and responses to ONS business surveys on employees, R&D, e-commerce, capital expenditure, prices and so on. These are mostly available from the late 1990s onwards.

One crucial factor in the development of the business data sources has been ONS's Inter-Departmental Business Register

(IDBR). The IDBR covers 99 per cent of all non-governmental economic activity, and is used to provide the sampling frame for all ONS business surveys. As all surveys contain IDBR reference numbers, this enables, for example, information on a company's productivity from one survey to be linked to information on R&D from another survey, or to administrative and survey information from other sources, particularly other government departments.

This enables analyses to be carried out where collecting the data from a single source would have proved an intolerable burden on respondents, or where the data were collected for a different purpose. This is extremely useful in terms of extracting the most value from the data set: data can now be used repeatedly to address new questions without the requirement to collect further information.

The IDBR is a remarkable source of information whose analytical potential is still being discovered. One step forward is the construction by the VML team of the Business Structure Database, a firm-level longitudinal data set created from the IDBR which derives indicators for demographic events such as takeovers and mergers (see Hellebrandt and Davies 2007). Although only created in 2006 and moved into the regular research area in 2007, this is likely to increase its significance as more and more of the IDBR information is tied into it.

Although business data still accounts for 80 per cent of VML research, the VML has become the de facto secure data facility for ONS and increasingly holds non-business data. In most cases, the VML has been called upon when ONS wishes to allow research on a more confidential version of a data set that is already available in anonymised form. For example, the VML is used to provide access to more detailed census data than is available on CD (albeit still strongly anonymised) through the Controlled Access Microdata Samples (see www.statistics.gov.uk/census2001/sar_cams.asp).

Figure 1

Trained researchers, 2004 to 2007

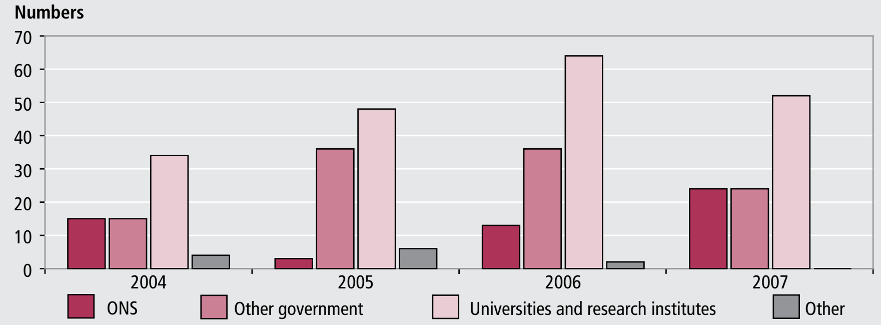


Figure 2

Days in the VML

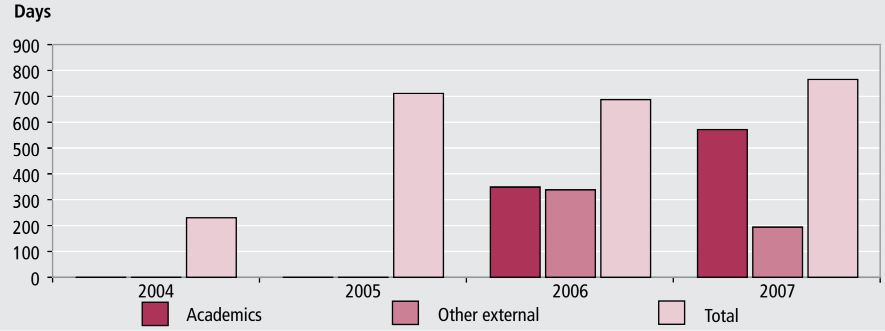


Table 1

VML projects, 2003 to 2007

	September 2003 to March 2004	April 2004 to March 2005	April 2005 to March 2006	April 2006 to March 2007	April 2007 to December 2007
New projects	20	27	42	76	45
Continuing projects	0	6	27	27	72
Total projects	20	33	69	103	117
Completed within year	14	6	42	31	28
Running at year end	6	27	27	72	89
Percentage continuing	*30*	*82*	*39*	*70*	*76*

Projects

As was noted, the VML started in 2004 with around six live research projects of external researchers. At the beginning of 2008, the VML had 89 live projects, of which 45 had commenced in the previous seven months.

Table 1 shows the growth in projects over the past four years. As well as more new projects, it is clear that much of the increase is due to projects being carried over from one year to the next and extended. However, this is partly a consequence of the general growth in projects. Generally, around three-quarters of projects in any one year (old and new) carry over into the next financial year.

The table includes work sponsored by other government departments (OGDs). The single biggest direct sponsor of VML research is the Department for Business, Enterprise and Regulatory Reform (BERR, previously the Department of Trade and Industry). Indirectly, the biggest sponsor is HM Treasury, particularly of productivity studies. Other departments commissioning academics or using the VML directly include UK Trade and Investment; the Health and Safety Executive; the Low Pay Commission; the Office of Fair Trading; the Department of Health; the Department for Work and Pensions; and the Department for Culture, Media and Sport.

This table does not include internal projects. ONS research staff use the VML as part of their regular business-as-usual activity, and so separate research projects are not identified. However, five broad work programmes and their start dates can be identified:

- the microeconomics of productivity (2003)
- methodological studies (2004)
- research in low pay and earnings data (2005)
- intangible investment (2005), and
- analysis of price data (2006)

The scope of projects has changed along with the data sources. Initially, all research was on the microeconomics of productivity, but this is no longer the main area of interest.

Table 2
Project theme ranks

Subject	Ranking by main theme only	Ranking by all themes	Ranking by all themes, weighted
Earnings	1	3	1
Employment and labour markets	2	3	2
Skills and productivity	3	1	3
Capital and investment	3	6	5
Globalisation, outsourcing, international	5	2	4
Energy and environment	5	7	7
Industry studies – manufacturing	7	7	8
R&D, innovation	8	3	6
Entrepreneurship	9	9	9
Business demography	9	10	10

Note:

Other themes also identified: programme evaluation, regional studies, macro-micro linkages, ICT and the new economy, finance, and service industry studies.

Table 2 shows the most popular themes for the 47 project applications received in April to December 2007, researchers selecting up to four topics from a list of 16. The themes are ranked by popularity as the major theme; by the inclusion in any of the four topics; and by the inclusion but where the first mentioned topic has a higher weight. Currently, the most popular issue concerns the UK labour market. This may be a temporary phenomenon: in late 2007, the VML team managed to link successfully the most popular earnings data source (Annual Survey of Hours and Earnings) to a widely used study of workplaces (Workplace Employment Relations Survey). Nevertheless, it is clear that a wide range of economic issues are being studied.

Outputs

One obvious measure of growth of the VML is the number of outputs produced by researchers. However, this is not an easy figure to assess.

VML researchers are isolated from the world outside; they have no access to email or the internet, and statistical results can only be released from the VML by VML staff. The VML operates a two-stage level of clearance. Researchers working in the VML may ask for results to be checked for confidentiality issues and released so that they may discuss with colleagues and write up results; these are called intermediate results. When results have been written up, researchers need to resubmit these results to the VML team where a tighter confidentiality regime is applied. Results approved here are given final clearance and can be released to the research community. These are not necessarily finished papers, but would also include, for example, a table to be included in a conference presentation.

Figure 3 and **Figure 4** show the number of intermediate and final results cleared by the VML team. Several features can be observed.

First, although there has been a general increase in outputs over time, this is less pronounced than the growth in projects or users. Partly this reflects the maturing level of research. As the VML data sets become better known, and researchers build more on earlier studies, there is less exploratory work and more analysis. Longer-term users tend to produce a lower volume of output. In addition, internal ONS staff using the VML rarely request outputs, as they have permanent access to the VML through their desktops. Outputs only tend to be collected when, for example, reports are being prepared for an external audience.

Second, the volume of outputs can vary considerably over the course of the year. This may be driven by particular events (such as a conference deadline) but often it reflects the academic year. Intermediate outputs, for example, tend to fall in August; final outputs tend to peak around June, possibly reflecting the start of the conference season.

Third, it is hard to predict in advance when the demand will arise. Although general patterns of demand can be drawn, in any one month there can be considerable variation in the need to clear outputs. As all outputs must be cleared by at least one member of the VML staff, this can make the allocation of VML resources more difficult.

Figure 3
Intermediate clearances

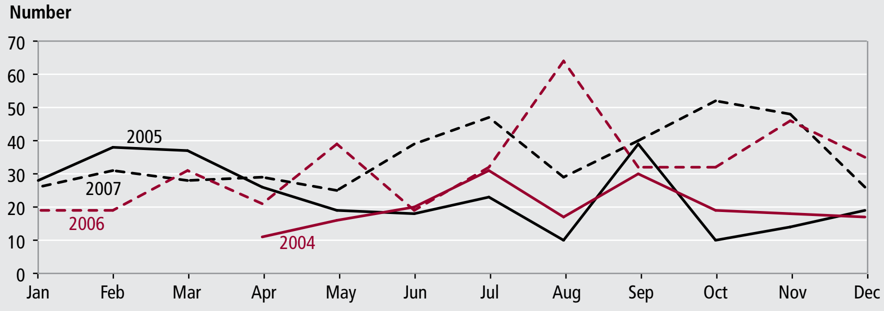
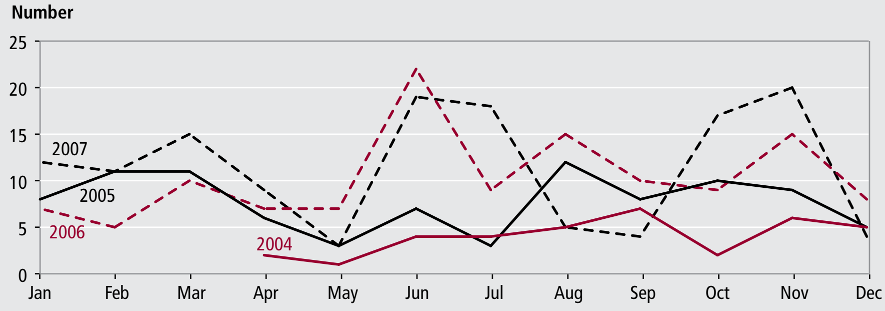


Figure 4
Final clearances



Impact

Part of the reason for the VML's success is that it was designed as a general-purpose data research facility. The following sections look at the impact on academia, government departments and ONS.

Academic impact

That there was a significant pent-up demand for the data resources of ONS is evident from academic output in recent years. For example:

- in 2004, HM Treasury's fifth Productivity Report contained macroeconomic analyses of the

UK economy, and international comparative studies, but relatively few microeconomic analyses of the UK economy. The sixth report in 2005 took evidence from a range of UK microeconomic studies, most based on VML research

- the two-yearly Comparative Analysis of Enterprise Microdata (CAED) conference, the main international gathering for microeconomic analysis of business data, was organised by ONS in 2003. Only six papers out of 78 were VML projects, with only ten relating to UK data; ONS presented a special session on procedural issues. ONS also hosted the conference in 2005: 26 papers out of 61 used ONS data, of which 21 were VML research projects; ONS itself presented six papers on technical subjects
- in 2007, over 40 per cent of research projects had been of sufficient quality to attract competitive ESRC funding. A conservative estimate of academic daily charge-out rates would suggest that VML research projects are currently worth at least £200,000 per year to the academic community
- the ESRC is currently in the process of setting up an academic equivalent of the VML, to be run in a similar manner but with access direct from universities, to provide an additional route for access to less restricted data

Of course, these examples reflect the skills of the UK academic community, as well as the concurrent development of ONS's integrated vision of processes for microdata access. However, the role played by the VML in facilitating this research has been considerable.

Government research

Government departments were legally allowed access to much of ONS's business data resources for statistical purposes, but made little use of this opportunity. There may be many reasons for this, but four stand out:

- there was little consideration of what could be done with the ONS data
- there was limited awareness of who could work with the data, in-house or with contractors
- there was no mechanism for advising OGDs on how their data could be combined with ONS data, and
- there was concern about perceptions that the data could be used for non-statistical purposes

Creating a virtuous circle is important, by supporting OGDs in both the commissioning of research and hands-on analysis. The team regularly visits OGDs and microdata researchers (economists, statisticians, social researchers) who are encouraged to discuss data queries with the VML team. Users can receive advice on the feasibility and practicality of data linking, allowing them to develop skills in ONS and other microdata. This leads to further questions about how data can be used, and the research community is strengthened. For example, BERR chairs a cross-government user group on research use of microdata, which grew out of the VML government user group.

At every stage along the way, the VML team helps to bring together expert academic researchers with policy analysts. Here, the early role of HM Treasury and BERR was crucial in leading by example, with support for research and a willingness to consider innovative uses of data.

The VML does now have a presence in government as a central contact point for microdata research. As well as organising workshops and conferences, the VML publicises other research events, circulates invitations to tender, and carries out a number of networking activities. Again, much of this is 'facilitating' rather than 'doing', but there is a clear demand for an expert unit to take on this role.

ONS

The VML has influenced many areas of ONS operations. Five examples are listed here.

First, the VML has helped increase ONS's visibility in, and contacts with, the academic community, so enhancing its own research. Economic analysts in ONS in particular have used these contacts to build up an enviable reputation for supporting innovative research; and they have taken a lead in a number of collaborative international research projects. For example, ONS is at the forefront of research into the capitalisation of R&D in National Accounts (see Galindo-Rueda 2007). The VML has contributed to this by, inter alia, arranging expert workshops.

Second, the ONS data collection units have been able to use both in-house and external expertise to study data in a range of new ways. Much of the impact comes from taking an analytical perspective on data, rather than the traditional population-estimation focus of ONS. For example:

- Ormerod and Ritchie (2007) showed how the level of the National Minimum Wage influences the accuracy of its measurement. As a result of this paper, new instructions to survey interviewers have been introduced
- a 2006 workshop on innovation and research showed that the two key ONS surveys in this area collected data from different parts of the same business; the sampling schemes for the surveys are now under review
- Ormerod (2007) highlighted inconsistencies in data on self-employment collected across three different surveys
- Hellebrandt and Davies (2007) investigated how standard National Accounts company classifications may be hiding the changing industrial structure of the UK
- Jenkins (2008) studied the possibility of linking census and earnings data; while only a feasibility study, this has far-reaching implications for the use of social data
- analysis-led reviews of data sources, typically undertaken at the request of the data managers, have informed reviews of variables, sampling frames, and forecasting

Third, the VML has affected ONS key outputs directly. The programme of work on how investment in intangibles (software, patents, branding) should be measured has already led to a revision of GDP estimates and a new experimental National Statistic (see Chamberlin, Clayton and Farooqui 2007 for a summary). This work was the result of a major project carried out by ONS's Economic Analysis Division in conjunction with academics and HM Treasury; the existence of the VML meant that the project could concentrate on the research and ignore issues of data collection, storage and management of external researchers.

Fourth, the VML has played a notable role in the development of ONS's data strategies. In recent years, ONS has been developing an access strategy for microdata which provides users with a range of options, tailored to the purpose, the user and the confidentiality of the data (see www.statistics.gov.uk/about/ns_ons/ons_microdata_releases.asp). The UK's integrated access strategy, developed in collaboration with the ESRC, has been identified by international bodies as an exemplar of how to effectively and safely support research on confidential data.

The VML is a key part of the data access spectrum, acting as the last link in the chain of possibilities.

Fifth, the VML was set up as an isolated system designed to meet unusual research requirements. As such, it has supported other ONS divisions, either as a short-term solution to a specific problem, or in a more methodical way. For example, the VML is now used by the Methodology Group for testing software and methods in a secure, isolated environment before authorisation for production use.

The last five years have seen a revival of interest in RDCs, using new technology to provide both better access to, and more security for, data. Along with technical developments, much of the practice of running RDCs has been under review. The VML was set up early on in this revival and took a leading role in discussions about the purpose and management of RDCs (see, for example, Ritchie 2007, 2008). As a result, ONS remains in the vanguard of international developments in this area, and was praised by the OECD in 2005 as 'one of the most innovative research efforts in the public sector across the 30 OECD member countries'.

The future

2008 will see the first major review of the VML since it was set up. This is driven by five main factors.

First, the VML has a plethora of information about how it could and should operate.

Second, since April 2008, the VML has been used to deliver the ONS Longitudinal Study (see www.statistics.gov.uk/about/data/methodology/specific/population/LS) and the VML team will become responsible for managing the user support (mainly provided by Celsius, a team of academics from the London School of Hygiene and Tropical Medicine; see www.celsius.lshtm.ac.uk). The ONS Longitudinal Study is a confidential data set containing census and health data, and is available for research under similar principles to those of the VML but with a different operating and funding model. It seems likely that there are many synergies to be exploited by combining the two research services. In the medium term, the aim is to provide a single seamless solution for all on-site access to confidential ONS data.

Third, ONS is piloting access to the VML from a small number of government offices around the country. This will address a criticism of the VML that researchers currently have to travel to ONS offices to use the VML. The pilot is due to be completed in summer 2008 and will report on the feasibility of allowing access from a wider group of offices, with the aim that 95 per cent of UK researchers should have less than one hour's travel to their local VML access centre. However, such a move would have financial, statistical and ethical implications. These need to be reviewed before any further development takes place.

Fourth, under an arrangement with the ESRC, academics engaged in research on their own account have the charges for accessing the VML paid directly by the ESRC. This arrangement runs out in March 2009, and so there is a need to review the funding model.

Finally, in April 2008 the Statistics and Registration Act came into force. This simplifies the legal framework for VML activities, but is likely to require some changes in operation.

Bringing these five elements together, in summer 2008 the VML will carry out a major review of its activities, in consultation with other parts of ONS, OGDs, the ESRC, IT specialists, and academia. The aim will be to put the VML on a secure footing for the next five to ten years by having:

- an overarching vision of how and why the VML exists
- a 'one-stop data shop' approach to supporting the research community
- best-practice security procedures
- flexible, efficient administrative procedures continuing the VML's tradition of being one of the most cost-effective RDCs in the world.

Overall, the prospects for the future of research into confidential data look bright.

Notes

1 Principally the Institute for Fiscal Studies, Queen Mary College London, the London School of Economics, and Newcastle University.

CONTACT

✉ elmr@ons.gsi.gov.uk

REFERENCES

Barnes M and Martin R (2002) 'Business Data Linking: an introduction', *Economic Trends* 581, pp 34–41.

Chamberlin G, Clayton T and Farooqui S (2007) 'New measures of UK private sector software investment', *Economic & Labour Market Review* 1(5), pp 17–28.

Davies R and Welpton R (2008) *Linking the Annual Survey of Hours and Earnings to the 2004 Workplace Employment Relations Survey.* WERS Information and Advice Service, Technical Paper No. 3. NIESR, London.

Galindo-Rueda F (2007) 'Developing an R&D satellite account for the UK: a preliminary analysis', *Economic & Labour Market Review* 1(12), pp 18–29.

Hellebrandt T and Davies R (2007) Some issues with enterprise-level industry classification: insights from the Business Structure Database, *Virtual Microdata Laboratory Data Brief*, No. 5 Spring, ONS, Newport.

Jenkins J (2008) 'Linking the Annual Survey of Hours and Earnings to the Census: a feasibility study', *Economic & Labour Market Review* 2(2), pp 37–41.

Ormerod C (2007) 'What is known about the numbers and 'earnings' of the self-employed?', *Economic & Labour Market Review* 1(7), pp 48–56.

Ormerod C and Ritchie F (2007) 'Issues in the measurement of low pay', paper for Work Pensions and Employment Group Conference 2007.

Ritchie F (2004) 'Business data linking: recent UK experience', *Austrian Journal of Statistics*, July, pp 89–97.

Ritchie F (2006) 'Access to business data: dealing with the irreducible risks', *Monographs in Official Statistics: Work session on statistical data confidentiality Geneva*, Eurostat, Luxembourg.

Ritchie F (2007) 'Statistical disclosure detection and control in a research environment', paper for Workshop on Data Access, Nuremberg.

Ritchie F (2008) 'Disclosure detection for research environments in practice', forthcoming in *Monographs in Official Statistics*, Eurostat.

Robjohns J (2006) 'ARD2: the new Annual Respondents Database', *Economic Trends* 630, pp 43–52.