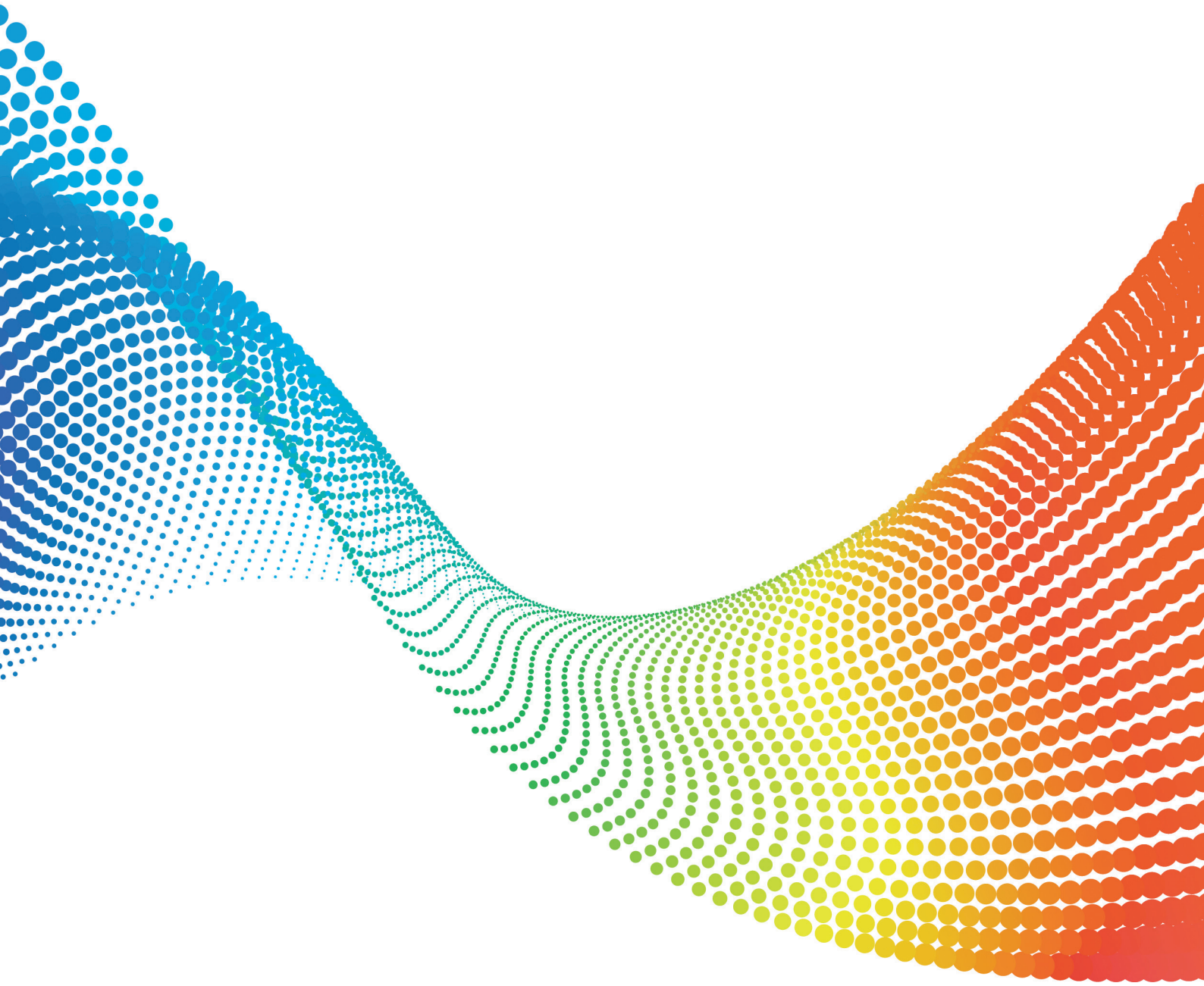


UK data forum

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

HMRC Datalab: Engaging with the External Research Community

Richard Welpton and
Melanie Wright
UK Data Archive,
University of Essex

Acknowledgements

We thank the following individuals for their kind help and input into this work: Aliya Saied and Katherine Pottinger, HMRC; Susan Cozzolino and Mus Ahmet, UK Data Archive; Chuck Humphrey, Canadian RDC Network; Ray Lambert, Department for Business, Innovation and Skills (BIS); Tony Clayton and Peter Evans, Intellectual Property Office (IPO); Martin Weale, formally of National Institute of Economic and Social Research (NIESR); Helen Simpson, University of Bristol; Stuart Adams, Institute for Fiscal Studies (IFS); Michael Devereux, University of Oxford; Tanvi Desai, London School of Economics (LSE); Annarosa Pesole, Imperial College; Paul Jackson, Office for National Statistics (ONS).

We kindly acknowledge financial support from the ESRC for this project.

Foreword

A key element within the *UK Strategy for Data Resources for Social and Economic Research 2009-2012*[1] (the 'National Data Strategy') is the promotion of initiatives and actions which are designed to 'improve research access to potentially disclosive microdata records on people and organisations whilst protecting confidentiality via the use of new data licensing and access procedures' (National Data Strategy, p.5).

Throughout 2010 the Economic and Social Research Council (ESRC) and Her Majesty's Revenue and Customs (HMRC) have been engaged in discussions seeking to develop closer links between the research community and the policy interests of HMRC. As a part of these discussions, HMRC requested ESRC to give consideration to ways in which access to data held within the HMRC Datalab could be promoted across a wide research community.

This report was commissioned in response to this request. I would like to thank Melanie Wright, the director of the Secure Data Service at the UK Data Archive, and Richard Welpton, Senior Data and Support Services Officer at the Secure Data Service for the effort they have put into this. HMRC wishes to express its gratitude to the ESRC for conducting this useful piece of work and will give careful consideration to the recommendations.

Peter Elias

March 2011

[1] **Elias, P.** (2009) 'UK Strategy for Data Resources for Social and Economic Research 2009-2012', Swindon: Economic and Social Research Council, Swindon http://www.esrc.ac.uk/_images/NDS_publication_Sep09_tcm8-5155.pdf

Contents

1.0 Introduction

Data are collected for many administrative purposes. For example, the UK Office for National Statistics (ONS) collects data to produce aggregate economic and population statistics. Her Majesty's Revenue and Customs (HMRC), the UK tax department, collects tax returns from individuals and companies to assess their tax liability.

The resources that are annually ploughed into gathering these 'administrative' data are matched only by the demand to analyse them. If "data needs are driven by research and policy interests" (ESRC National Data Strategy), then providing access to such administrative data will go a long way to contributing towards successful research outcomes that often feed into policy.

Concerns about protecting the confidentiality of such data are often used to prevent access to these data for legitimate research purposes. But the momentum to balance access and security is summarised in the 26th recital of EC Regulation No 223/2009:

"The research community should enjoy wider access to confidential data.....Access to confidential data by researchers for scientific purposes should therefore be improved without compromising the high level of protection that confidential statistical data require"[1].

Furthermore, in a climate of economic uncertainty, and where there is competition for resources, Government Departments are under increasing pressure to maximise the returns on the resources they invest in data collection by exploiting administrative data to the full. This, and the need to reduce the burden of survey data collection on themselves and the respondents, builds a compelling case for opening up access to administrative micro-data to the research community.

Relatively little research utilises rich sources of administrative data, often because there is a lack of awareness that the data exist, and/or because secure mechanisms for accessing data usually considered 'sensitive' have not been established. Yet because research and policy-making go hand-in-hand, and play an important role in "mapping the causes and consequences of change in this complex and dynamic world" (ESRC Strategic Plan 2009-2014), it is becoming more important to make the research and policy community aware of the benefits of accessing these data.

But if any resources are to be invested in making administrative data available, for example via Research Data Centres (RDCs), then there should also be a clear objective to promote access to these data and access to the RDC. Upon successfully creating access to administrative data, promotion to ensure their research potential is realised is important.

HMRC and the ESRC share the "research and policy driven" approach to data access, identified in the National Data Strategy. In addition, a considerable 'research agenda' has emerged at the Department following many recent policy initiatives, and there is recognition that opening up data to the research community will have a positive impact on this agenda, as well as widening research.

---

1 Regulation (EC) No 223/2009 of the European Parliament and of the Council

To this end, a Research Data Centre (RDC) similar to the Office for National Statistics Virtual Micro-data Laboratory (VML) is now being established at HMRC (known as the HMRC Datalab). A pilot project, evaluating corporation tax data, has been completed by researchers at Oxford University. HMRC aim to launch their Datalab in Spring 2011.

This report focuses on three key areas that are important for HMRC to consider:

- Whether there is scope for access to HMRC data to be more widely distributed
- The benefits of liaising with similar services
- How to promote the data, and access to the data

We recognise the important achievement of HMRC in establishing its RDC and making available data for research, especially given the confidential nature of the data and general public unease about data security. The first part of this report therefore considers whether the data that HMRC will make available could also be more widely accessed through RDCs such as the VML, or the new ESRC-funded Secure Data Service.

HMRC already realise the benefits of liaising with similar services throughout the UK and internationally. The example of the efficient model applied by the VML has been instrumental in the design of their own data enclave. But what other lessons can the new RDC learn?

In addition, other Government Departments regularly collect data, driven by policy needs. Are there benefits of liaising with other Government Departments? The second part of this report examines the benefits of liaising with 'sister' services and other Government Departments, and how HMRC could effectively engage with them.

Developing a strong communications strategy will be an important step to generate demand for the new facility at HMRC by the research community. There are many options for promoting the service: a strategy that ensures all methods of communication work coherently is important for establishing a service identity.

Drawing from the experience of data facilities in the UK such as the VML and the Secure Data Service, and around the world, the third part of this report recommends activities that can be undertaken by HMRC to encourage take-up of the facility by the research community, and maximise the returns on the investment made by HMRC in their RDC.

2.0 HMRC and its place in the data world

2.1 HMRC data and access

This part of the report describes the data that HMRC will make available to the research community, as well as arrangements for accessing the data by academic researchers. We consider how this setting compares with other data services in the UK. We also assess the scope for interaction with other services such as the Secure Data Service or VML.

2.1.1 What is the HMRC Datalab?

HMRC Datalab is a secure environment that researchers can visit to analyse confidential HMRC micro-data. It is located at Bush House in Central London. Researchers are provided with a computer account, which contains the micro-data they have applied to use. Researchers undertake all of their analysis using a computer terminal. Their research outputs are saved to their account, and are released to them by email from a member of the Datalab staff.

The decision was made to locate the Datalab at this less-high-profile building to minimise the amount of time researchers have to await their security clearance (high profile buildings, such as HMRC Headquarters in Whitehall, would have resulted in lengthy security clearance times for researchers).

Researchers arrange to visit the Datalab in advance, and are signed in by Front Desk Security. They are escorted to the terminals by the Datalab staff.

The Datalab has two terminals at present, and these are located in an isolated room within view of the Datalab Manager. The number of terminals can be quickly expanded if there were sufficient demand from researchers. Support for the data and using the system is provided by Datalab and IT staff.

2.1.2 What data are available?

An initial pilot dataset consisting of corporation tax data from the CT600 tax return has been made available for research. At the moment, these data may only be used in isolation (i.e. the observations in the dataset do not contain reference numbers that would allow a researcher to link these data to observations from other sources).

There may be scope for extending the catalogue of data available to researchers. It is intended that trade statistics data will shortly be made available, allowing researchers to analyse trade undertaken by firms in the UK and abroad. Self-assessment tax data may also be made available in due course, and possibly Child Benefit and Tax Credit data in the future.

These micro-data are observations of individuals or firms. HMRC will remove all direct identifiers (such as names and addresses) from the data before researchers are given access. This will prevent researchers from associating observations to individuals and/or firms directly (it should be noted that indirect identification may still be possible, and for this reason, analytical results that researchers wish to remove from the Datalab will be screened, please see section 2.1.2 below).

The research community has expressed hope that it will be possible to combine data from two or more sources. This would benefit HMRC/HMT (Treasury) because the number of research possibilities would increase considerably, as the additional variables from combining sources amount of research that can be conducted. In particularly, there is considerable interest from researchers hoping to combine HMRC data with business data available from the ONS (for example, productivity data that is accessible at the ONS VML and the Secure Data Service).

Documentation

Researchers benefit enormously from well-documented data. The experience of the UK Data Archive concludes that good quality documentation can lead to better quality research. HMRC would also benefit from reducing the costs of supporting researchers (and there are also reputational benefits too).

Staff at HMRC are currently arranging for documentation to be made available for researchers. Decisions about how to make documentation available, and the appropriate meta-data standards that should be adopted, have yet to be made. HMRC are to seek advice from the UK Data Archive shortly.

2.1.3 What setting?

Legal Framework

No legislation exists that specifically provides researchers with a right to access HMRC data if there is a 'public benefit' reason for doing so (as the Statistics and Registration Services Act 2007 does). However, researchers may access HMRC Datalab with permission of the Micro-data Release Panel (see below), and will be treated as any other Civil Servant employed by HMRC. This means that they will be subject to the terms of the Commissioners for Revenue and Customs Act 2005.

Non-compliance with the terms of this legislation can lead to prosecution (for example, in the event that a researcher discloses information about an individual or company during the course of their access to the data). HMRC will shortly include information about governance in the training that they provide to researchers, which is required before they can access the Datalab.

Applying for Access

By Spring 2011, HMRC will provide researchers with information to apply to access HMRC data on a website. The website will also allow researchers to download application forms, which they can submit to the Datalab manager. Documentation, and in the future meta-data, will also be made available through this website.

A Micro-data Release Panel (MRP) has been established to consider applications from researchers to use the data. HMRC Datalab will be unique among data enclaves because the MRP have specified that applications to use the service should broadly reflect the research interests (and policy agendas) of HMRC and HMT. Researchers who apply to access data in RDCs such as the VML or Secure Data Service must simply prove that their research will deliver a 'public benefit' to society.

Researchers must successfully demonstrate to the MRP that the research they propose to undertake will benefit these Departments. An example of recent HMRC research interests is provided below, although applications in other areas that may also benefit HMRC/HMT are encouraged too (these examples are merely a guide and not intended to be restrictive):

1. **Improved understanding and modelling of behavioural impacts on tax policies and administration**

   Within the context of tax policy, this is an exciting new area of analysis, drawing upon research from developments in the relatively new discipline of behavioural economics. It would be interesting to analyse the effects of policies which create incentives for changing the behaviour of individuals.

   In addition, the design of new surveys to collect data which could be used to model consumer behaviour represents new potential for research and policy design. Research that would feed into such an outcome is highly desirable.

2. **Optimising economic and organisational efficiency in the design and administration of tax policies and operations**

   This area of research is concerned with examining the perceived administrative burden of taxes on individuals and companies, and in particular, examining the trade-off between marginal tax yields and collection costs.

3. **Developing better approaches to forecasting and modelling of the effects of tax structures and tax reforms**

   Research in this area is concerned with developing models which forecast accurate responses to changes in taxation policy. In particular, there is demand for research which develops micro-simulation models to forecast the behavioural and/or distributional impacts of tax credits and/or to improve distributional analysis of indirect tax credits.

4. **Comparative learning from international approaches to tax policy and administration**

   This strand of research will focus on differences and similarities of international tax regimes, including how other administrations use evidence and analysis to design tax policy.

In terms of promoting the service, there are at least two reasons why HMRC should consider providing information about its 'research agenda' to assist researchers who wish to apply to use the service:

- It will avoid confusion among researchers about whether they can apply to use the data

- It will save the MRP from wading through and rejecting applications which the Panel does not deem suitable

Upon successfully applying to use the Datalab, researchers will be required to attend a training event. It is envisaged that this will be similar to the VML and Secure Data Service training programme, in which researchers pass through the following modules:

- Governance of the data (including legal responsibilities)
- Statistical Disclosure Control (see 'Outputs' below)
- How to use the service

There will be no fee for researchers who access the Datalab because HMRC are keen to encourage researchers to use the facility: the resulting knowledge 'spillovers' can be absorbed by HMRC and HMT (that is, the benefits of the research feed into the policy-making process of these Departments). It is felt that an access charge would deter applications to use the facility, particularly from less experienced researchers (such as junior research assistants or supervised PhD students), or those requiring to travel to the Datalab. This feature should have some prominence in the promotional activities of the Datalab.

On the other hand, HMRC will not provide research funding for researchers wishing to use the facility. Researchers should obtain funding from sources such as the ESRC.

Outputs

Although direct identifiers (such as names and addresses) have been removed from the data, there is still a possibility that individual observations can be identified indirectly. This depends on the level of detail (the number of variables which 'describe' observations). For example, a company could be identified using a combination of turnover, employment and industrial activity. But these are also essential variables to undertake analysis of business data.

Statistical Disclosure Control of outputs (SDC), achieves a balance between allowing access to potentially identifying variables, and protecting the anonymity of observations. Providing that researchers can access sensitive data securely, they can produce and receive outputs, subject to an SDC screening by staff to ensure that the results in the materials cannot be used to identify an observation.

The MRP has established rules concerning statistical disclosure control (SDC). Researchers will be provided with hands-on training to make them aware of these SDC rules.

However, a large body of international research on SDC practices recently culminated in the production of an ESSnet (European Statistical System Network) report to Eurostat ('Guidelines for the checking of output based on micro-data research'). The work draws upon the experiences of data enclaves in the UK, Europe, and North America, who apply SDC techniques to outputs on a day-to-day basis. These guidelines have been deployed by the VML since its inception in 2004, and will also be used by the Secure Data Service.

We believe that HMRC would benefit by learning from the experience of these data enclaves when establishing its SDC requirements. The ESSnet report recommends tried-and-tested SDC rules for output checking.

A common framework for SDC of outputs across services is highly desirable, especially because researchers are likely to apply for access to data to more than one service. Variations in the SDC of outputs among services are likely to lead to the unequal treatment of researchers. This could have reputational consequences for the service and the provision of data access, which we are keen to avoid.

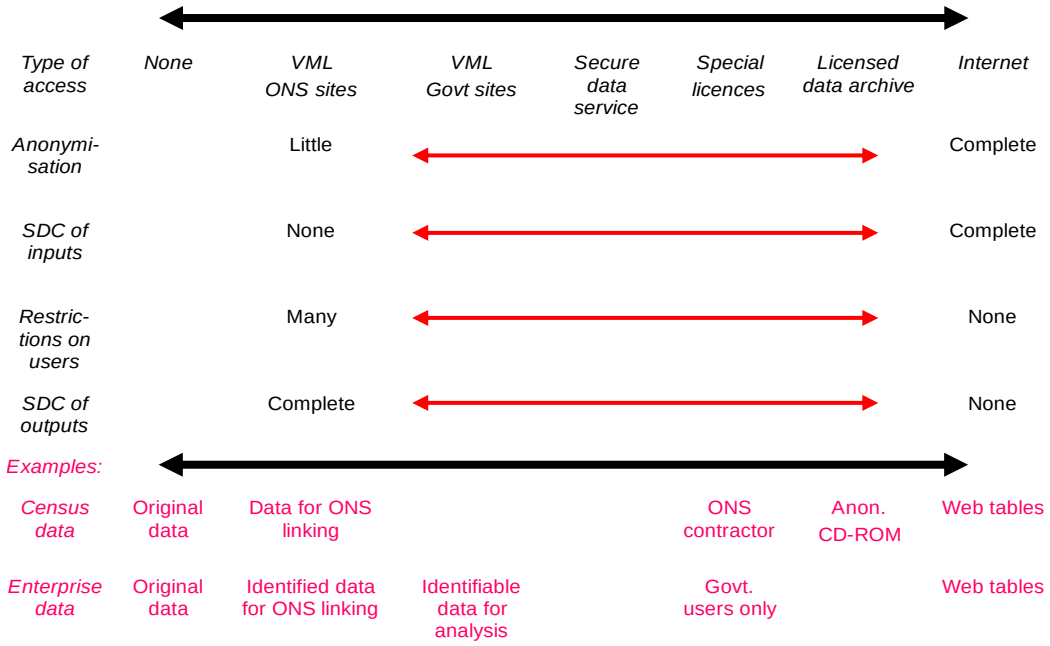
2.2 HMRC and Existing Data Infrastructure

This section briefly describes the data infrastructure scene in the UK (the international picture is similar). We then describe the VML (the template for HMRC Datalab) and the Secure Data Service. Both services provide access to confidential micro-data collected by the ONS and other sources. This setting provides the background by which we can assess where HMRC Datalab fits in.

2.2.1 The Data Access Scene in the UK

Ritchie (2010) usefully depicts the UK data infrastructure in the diagram below:

Figure 1: Data Access in the UK



There are a number of options for accessing data, depending on the sensitivity and/or confidentiality of the data. As we move across the diagram from left to right, researchers are required to meet fewer conditions for accessing data, and the data contain fewer variables (less detail). The data therefore become less sensitive/identifiable as fewer access conditions need to be met by researchers.

Original (or 'source') data, containing the names and addresses of individual survey respondents, are never released to researchers. At the other end of the spectrum, fully-anonymised, aggregated data may be made unconditionally available. For example, the ONS,

and Other Government Departments, make aggregated statistics available online and downloadable from the internet.

In between these poles, conditions restrict the distribution of these data in different ways, according to the additional detail made available.

For example, the UK Data Archive allows researchers to download the majority of their data collections after they have accepted the conditions of an End-User License (EUL). In this license the user agrees to cite the data appropriately, to neither attempt to identify individuals/households nor claim to have done so, and to provide the Data Archive with information about publications arising from their research using the data. They also give the Data Archive permission to share information about their data usage with the data owners.

Data which fall under EUL tend to have less detailed geography (e.g. a variable for Government Office Region rather than ward or postcode), have less detail (for example respondent age aggregated to 10-year ranges rather than year of birth) and potentially disclosive outliers are removed (such as households of greater than 8 or 10 members, incomes above a certain level, etc).

The UK Data Archive also allows access to more detailed data via a Special License. Researchers must provide more detail about their proposed use of the data, and their applications for access are approved on a case-by-case basis by an appropriate decision-maker within the data owners' organisation. In addition, researchers agree to additional access conditions governing the setting in which the data are used, and agree to abide by 'good data housekeeping' principles outlined in the Data Archive micro-data handling guide, for example, agreeing to destroy data after their project is completed. These data contain more detailed variables, such as lower level geographies, more detailed age, more detailed classification codes, etc.

RDCs: secure access to confidential data

A Research Data Centre (RDC) is the last port of call for accessing data before the release of data is prohibited. The ONS VML (at Government sites) and UK Data Archive Secure Data Service are examples of RDCs that provide access to potentially identifiable economic and social data.

While source data is not provided, detailed data is available which could indirectly identify survey respondents. These data may be useful for research. For example, business data is difficult to usefully anonymise. Researchers undertake analyses that often require variables including turnover, employment and industrial activity – yet these variables could be used indirectly to reveal the identity of an individual firm.

RDCs now play the role of 'intermediaries' between data owners, who want to ensure that the confidentiality of data is maintained, and researchers, who want to access data for research purposes. They present an ideal solution to data owners, who welcome access by researchers for legitimate research purposes, but are concerned about the security and confidentiality of the data. RDCs achieve the correct balance between security and access because they can place strict security arrangements upon researchers when they use the facility.

Heavy penalties (such as custodial sentences in the examples of the VML and Secure Data Service) can be imposed upon researchers who abuse their access rights. The Secure Data Service is also the first RDC that has arranged with a Research Council to withdraw funding for the research and their institution as an additional punishment against a researcher who has abused their data access rights.

On the other hand, researchers can access the data they require. RDCs also have experience of managing researchers, and staff often have research backgrounds themselves. A lot of the security concerns become negligible because RDC staff understand how researchers work (see 2.2.4 Security Arrangements for more detail).

Researchers produce their outputs within the secure confines of the RDC. Any outputs they wish to receive are subjected to Statistical Disclosure Control, which ensures that outputs cannot be used to identify survey respondents.

Finally, researchers are allowed to combine sources of data together within the secure confines of an RDC (as currently is the case at the VML and the Secure Data Service). Observations in ONS business survey data contain unique but anonymous identification references. It is thus possible to link one source of business data with another using this reference.

The ability to combine sources of data together is extremely valuable for researchers, because much more research can be conducted. For example, by linking observations from an ONS R&D survey to ONS productivity data, a researcher can research the effects of R&D on productivity. This would not be possible without the ability to combine sources.

It should be noted that there are concerns with regard to statistical confidentiality when sources are combined, because more detail about a single observation becomes available. For this reason, linking sources is also confined to the secure environment of an RDC.

2.2.2 ONS Virtual Micro-data Laboratory

The Virtual Micro-data Laboratory (VML) was established in 2004 to allow access to researchers interested in analysing productivity data. This initiative was prompted by HMT, because the Department desired to improve its understanding about the productivity of the UK manufacturing sector.

Since then, demand for the service has grown exponentially, following deposits of a large number of ONS business, earnings and social survey data. The VML is also a facility for research used by internal ONS staff. A large deposit of confidential administrative data from HMRC, Department for Work and Pensions (DWP), UK Borders Agency, Higher Education Statistics Agency (HESA)and the Department for Children, Schools and Families (DCSF), was made in 2009, with the intention of assisting the ONS to improve its migration and population statistics.

Services such as the VML are emerging as a mechanism for distributing access to detailed data securely – eventually this type of access will replace the use of CDs and other media for transferring detailed data. It is now widely accepted that the use of removable media such as these is simply not plausible any more – there is too much risk that a CD could go missing.

Secure storage in a central location with remote access using Virtual Private Network (VPN) technology (as used by the military and banking sectors) is much safer, and far more reassuring to all who hold a stake in the data.

This highlights the distinction that should be made when thinking about the use of detailed, sensitive and potentially disclosive (i.e. identifiable) data: between distributing data and distributing access to data. The former may cause security concerns about where data are stored and who has access to them. The latter is concerned simply about access – storage can be managed centrally and thus securely. So the question is about who is given access to data – which in the event of security concerns, can easily be removed.

The VML is internationally recognised as a model of best-practice. It has been emulated across the UK (such as the Secure Data Service), and world-wide (American and European data enclaves have followed the principles that underpin the VML). HMRC Datalab has adopted the VML model as its preferred choice for distributing access to its data.

2.2.3 UK Data Archive Secure Data Service

In December 2009, the UK Data Archive launched the (pilot) Secure Data Service, (with funding from the ESRC). This is a similar service to that provided by the VML, except that academic researchers are able to access data at their university, either at their desktop or at an institution 'safe room'. Like the VML, the Secure Data Service uses VPN technology to ensure that data cannot be removed. The Secure Data Service will be officially launched in March 2011.

The Secure Data Service builds upon the service provided by the VML because it is accessible to researchers who are not conveniently located near to an ONS office (currently the only locations where the VML can be accessed).

The Secure Data Service can also provide the benefits of UK Data Archive support, which has assisted researchers with their data needs for more than forty years. Support includes online documentation, applications to use the service and in general, a dedicated support infrastructure.

2.2.4 Security Arrangements

RDCs such as the VML and Secure Data Service are robust technical solutions which prevent data from 'escaping'.

In particular, the Secure Data Service is ISO 27001 accredited and CHECK (a product of CESG, the Information Assurance arm of GCHQ) compliant. This means that the service meets standards in:

- Security policy
- Organisation of information security
- Asset management
- Physical and environmental security
- Access control
- Information security incident management

- Human resource security
- Information systems acquisitions, development and maintenance
- Business continuity management
- Communications and operational management
- Penetration testing

However, IT security is in fact a relatively minor element of service security. Far more important is ensuring that users act properly and in accordance with the terms and conditions of using the service.

Both the VML and Secure Data Service invest time in 'active research management' (ARM). More information is available in Desai and Ritchie (2009). The approach emphasises that good researcher management is developed by fostering a close working relationship between the RDC and the researcher. Researchers will understand the risks of using the data, and behave in a way that does not endanger their future access. As such, RDCs allow risk-averse data owners to manage risk sensibly, rather than avoid it.

The experience of the VML suggests that, given the opportunity, researchers will happily engage with the RDC/data providers. For the RDC's part, it should return this engagement. ARM forms part of the security of the VML and the Secure Data Service, because both services strive to 'make people safe'. Removing barriers between researchers and data providers ('them and us') is important for security. If researchers find a procedure troublesome, they are more likely to speak to the RDC to find a solution, rather than 'break the rules'. In addition, this method shares the risk of providing data between researchers and data providers. Other benefits of ARM include:

- More efficient management (such as communication, understanding of procedures, cooperation, and use of RDC resources)
- Better change management (easier to implement new solutions)
- Better research (better engagement will lead to 'repeat custom' by researchers, and the improved reputation of the service, leading to more applications to conduct more research).

2.2.5 Where does HMRC fit in?

Initially, HMRC will provide access to firm-level Corporation Tax data. As discussed previously, these data cannot be usefully anonymised. From Figure 1, it is clear within the present infrastructure that access may only be possible from an RDC.

HMRC Datalab sits parallel to the VML and Secure Data Service in terms of the type of data that is currently offered (confidential firm-level data). In terms of security, there are differences between the three services. These are shown in Table 1 below, adapted from Ritchie (2010):

Table 1: RDC Specifications

	No protection				Strong Protection
Safe…..	0	1	2	3	4
Projects	Administrative processes only	Check researcher background	Check for statistical outputs only	Review by support officers able to critically assess feasibility and need for data	Review by Micro-data Release Panel to critically assess impact of outcomes
People (1) Knowledge	Administrative processes only	Check researcher background	Written assent to conditions of access	Passive training	Interactive training
People (2) Incentives	No effective sanctions	Procedural sanctions only	Mix of civil, criminal or procedural sanctions	Civil, criminal and procedural sanctions	Civil, criminal, procedural and institutional sanctions
Data	No data protection	Removal of direct identifiers	Identification within RDC environment unlikely	Identification outside RDC unlikely	Public use micro-data
Setting (1) Access	No restrictions	Access only from limited sites with no supervision	Access from secure networks with no supervision	Access from secure networks with occasional supervision	Access from secure networks with continual supervision
Setting (2) Networks	No restrictions on data transfer (downloadable from internet)	No internet access, data sent to researcher on encrypted removable media	No access to other parts of network. Data saved locally on terminal within secure room. No access to internet, no mobile phones, no access to printers.	Data provided on central server. Access by secure remote connection. No access to other parts of network, no internet, use of thin clients in secure room, no mobiles, no access to printers	Data provided on central server. Access by secure remote connection. ISO 27001 compliant, penetration tests on servers, unsuccessful attempts by 'ethical hacker' to 'break' the system.
Outputs	No checks.	Random checks.	Random plus targeted partial checking.	Full checking except for 'experienced' researchers.	Full checking.

When these requirements are applied, the following comparisons can be illustrated:

Safe…..	VML	SECURE DATA SERVICE	HMRC
Projects	4	4	4
People (1) Knowledge	4	4	4
People (2) Incentives	3	4	3
Data	2	2	2
Setting (1) Access	4	2 or 4 *	4
Setting (2) Networks	3	4	2
Outputs	4	4	4

* Secure Data Service can provide unsupervised access via desktop, or supervised access in an institution 'safe room', depending on the requirements of the data owner.

We believe that HMRC have made the correct assessment by providing access to their data in the secure environment of the Datalab (the data scores relatively low mark due to its sensitivity – the fact that the data are observations of businesses make them liable for indirect identification).

We also note that, within the broad range of administrative data collected by Government, there are varying degrees of sensitivity: some tax data can lie towards the most sensitive end of this spectrum. This can account for the extra care and seemingly slower approach towards decision-making and roll-out of access than other data providers.

On the other hand, we believe that if access is managed sensibly and securely, the issue about how sensitive the data is perceived becomes less important, particularly in a situation where analytical outputs are screened to ensure that individual observations from the data cannot be identified.

Other characteristics shown in the table below clearly demonstrate the HMRC Datalab sits parallel to services such as the VML and Secure Data Service (see Figure 1). Specifically:

- Proposals to access data are assessed by a review body (the MRP)
- Interactive training is provided to new users of the Datalab
- Civil and criminal procedures against researchers who break their conditions of access can be taken
- Only direct identifiers have been removed from the data
- Access is only permitted in a supervised setting
- Outputs are screened to ensure they cannot be used to reveal the identity of individual observations

HMRC Datalab does differ with both the VML and Secure Data Service in terms of IT security. Possible solutions to this are identified in the next section.

2.2.6 Alternative Mechanisms for Accessing Data

In this section, we consider whether access to HMRC micro-data could be distributed more widely.

UK Data Archive License Agreements

The Economic and Social Data Service (Secure Data Service) (managed by the UK Data Archive and operated in partnership with the University of Manchester) has established a range of licensing arrangements for data to which it provides access.

The addition of the Secure Data Service means that UK Data Archive can now offer a full spectrum of data access by the level of detail, and the risk of disclosure when users access the data.

As the diagram in Figure 1 illustrates, data are available under a variety of access modalities: some data is available for download from the website following registration and agreement of an End User License (EUL). These data are more anonymised and contain little detailed data that could identify a survey respondent.

Others require researchers to provide more information and sign up to more detailed service agreements which condition how the data can be used. These data are more detailed, and therefore present a small risk of disclosure, albeit very small.

Could HMRC business data be released as an (EUL) or Special License file?

While access to these data would be tremendous in terms of research outcomes, we think this would be unlikely for two reasons:

First, it would be necessary for the data owner (i.e. HMRC) to produce different (more anonymised) versions of the data that they have already produced for researchers to analyse in the Datalab. An assessment about the variables which could be included in the licensed version of the dataset would have to be undertaken.

The resources of HMRC would be required to produce these additional versions. However, if such "stripped down" data were to vastly increase useful policy-relevant research, such an investment might well be viewed as reasonable and proportionate.

However, (and most importantly), the nature of business data is that they cannot be usefully anonymised. Suppressing direct identifiers such as names and addresses does not eliminate the possibility of deducing a company from the observations. To do this, additional characteristics would also have to be removed (such as turnover and employment), but then the data will not be useful to analyse. For example, removing employment or turnover variables from the dataset will undoubtedly hinder any analysis with these data.

Nevertheless, HMRC can benefit from producing licensed versions of data, where this is appropriate (depending upon the confidential nature of the data). For example, if in the future, HMRC decide to make available data on individuals (such as Tax Credit data), then we advise HMRC to consider investing in the production of licensed versions for wider dissemination.

The benefits for the research community of making data available through EUL and Special Licenses are enormous. More research can be conducted, and these outcomes can feed into HMRC/HMTs' policy-making process.

Another benefit to HMRC is promotion. Researchers will realise the potential of applying to access the more detailed version of the data from HMRC Datalab. HMRC would only have to make a small one-off investment in preparing data to be distributed via these licenses. The Department would then benefit in the long-term from this investment, via applications to use more sensitive data at the Datalab.

HMRC can reduce the burden of distribution on its own resources by working with institutions such as the UK Data Archive (which has over forty years experience of disseminating and documenting data).

Access arrangements such EUL are unlikely to be suitable for HMRC business data at present. If data about individuals are made available, then we believe there may be future scope and justification for creating a licensed version, and the UK Data Archive could work with HMRC to make this potential a reality.

Remote Execution

In a remote execution system, researchers are provided with dummy or 'synthetic' data to work with at their office. The data files are structured and appear the same as real data files, but the data are 'fake'. Researchers can use these files to write syntax to for generating their results. They email this syntax to the RDC, where a member of staff runs the syntax for the researcher. SDC is undertaken on the outputs, which are then returned to the researcher if they are satisfactory.

This may be a solution for the most risk-averse data owner, but it is highly resource intensive. Staff with expertise of statistical software are required to produce synthetic data (in itself this is an extremely resource intensive exercise, and there is ongoing debate about the validity and usefulness of such data) and run syntax, in addition to performing statistical disclosure control of outputs. Researchers also have to wait longer to receive their outputs.

We feel that such a system would not be beneficial for HMRC. Not only is it resource-intensive, but the research benefits that will flow to HMRC and HMT from making access to these data available will not be realised due to the inevitable delays of returning results.

Remote Access

Remote access allows researchers to access the data remotely (using a network connection). Virtual Private Network (VPN) technology allows such connection to take place securely. Examples of remote access services include the VML, Secure Data Service and NORC (in the US) where researchers access data and statistical software for analysis. Technological restrictions prevent data from leaving the secure area.

These systems have the advantage that they are relatively inexpensive to establish and manage. Researchers undertake all the (research) work themselves. Resources may be required to train researchers and undertake SDC of outputs (although quality training of researchers in SDC techniques will minimise the SDC work of RDC staff). But it remains a cost-effective solution because the marginal cost to the RDC from each visit by a researcher is negligible.

This solution is ideal in meeting the requirements of researchers who wish to access sensitive and confidential data that cannot be made available to them through alternative arrangements,

and data owners who are concerned about data security. However, location and the cost of accessing the RDC providing remote access remains an understandable concern to the research community.

2.3 Realising the Research Potential

The researchers we have spoken to have all expressed the desire to link datasets together. For example, the research potential of the data which HMRC are to provide access to will be enormously enhanced if the data are able to be linked to business data from other sources, in particular, the business data made available by the VML.

For example, linking data can increase the number of research hypotheses that can be tested. A range of surveys are conducted by the ONS, examining:

- Innovation
- R&D
- Workplace relations
- Investment
- Capital Stocks
- Business Survival
- Earnings

Linking the Corporation Tax data to these surveys could produce an interesting array of proposals from researchers, which would benefit HMRC and HMT. For example, the behaviour of firms and individuals could be modelled more accurately by combining sources of data together. Indeed, the research agenda of HMRC/HMT would more comprehensively be addressed by the availability of linked data.

Other Government Departments (see section 3.2 below) have also expressed their desire to utilise HMRC data for inter-departmental policy objectives. This would also involve linking HMRC data to ONS data.

There is no more risk to the security of the data by combining, since these data will only be accessed within the secure environment of an RDC.

For this reason, we recommend that access to these data is made possible alongside business data available in the VML and Secure Data Service. There are different options for achieving this, and we present three different scenarios for HMRC to consider:

Scenario 1: HMRC deposits data with ONS VML to distribute access to researchers

HMRC would send data to the ONS VML. The VML would then provide access to these data, together with other ONS business data, to researchers via its secure on-site facilities (currently available in London, Newport, Glasgow and Belfast). In addition, the VML could be accessed at HMRC Bush House in London.

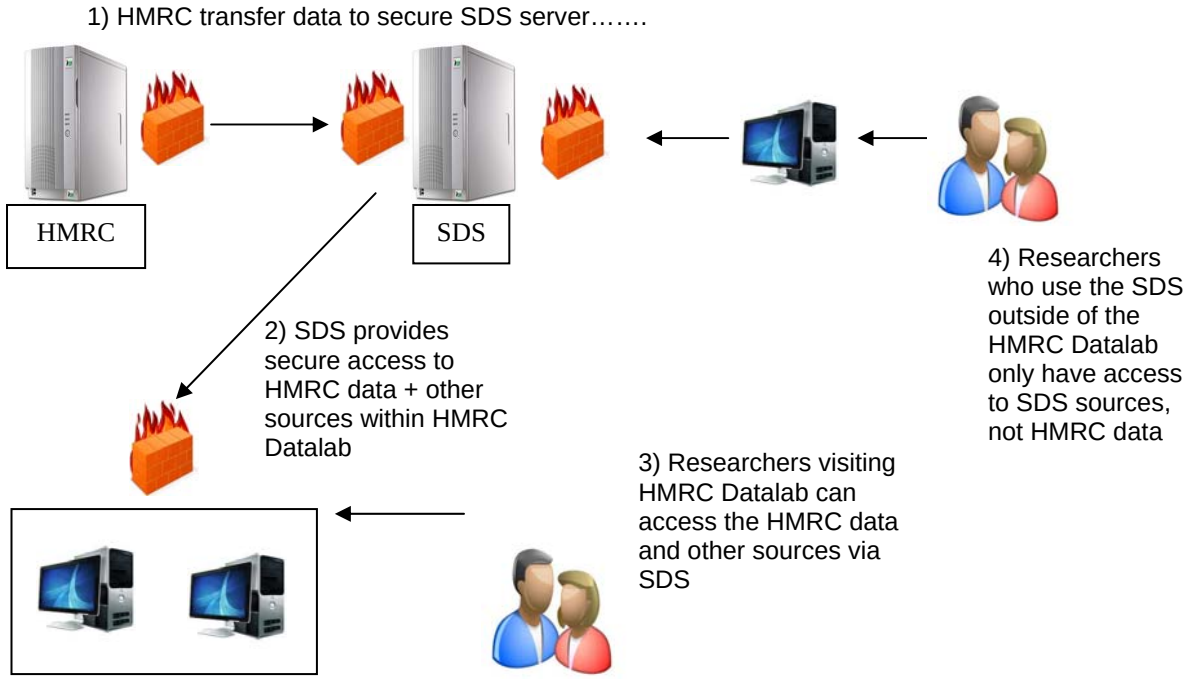
The advantage of this method is that HMRC would have the reassurance that its data was distributed via another Government Department. It would allow researchers to potentially combine HMRC and ONS sources of data together. HMRC would also benefit because the

infrastructure can already be provided by the VML. It should be noted that VML will charge researchers a daily fee of approximately £100 per day beginning April 1st 2011.

Scenario 2: HMRC takes advantage of existing Secure Data Service infrastructure

Secure Data Service technology enables the UK Data Archive to provide access to the service at a variety of locations. It would be possible for Secure Data Service to provide access to its service within HMRC Datalab. HMRC could provide its data to the Secure Data Service, and Secure Data Service could provide access to a combination of ONS data and the HMRC data solely within HMRC premises. This is illustrated in the diagram below:

Figure 2: Solution to data linking



A researcher who is approved to use HMRC data and other business data could access the Datalab in HMRC, and have access to these data via the Secure Data Service. If the same researcher logs in to the Secure Data Service outside HMRC premises, they would not have access to HMRC data.

This solution is ideal because HMRC can take advantage of:

- Secure Data Service secure data infrastructure (which meets international security standards – ISO 27001 and CHECK, see above)
- UK Data Archive experience of managing data (e.g. standards in data documentation, meta-data and user support)

This would be ideal for researchers because neither Secure Data Service nor HMRC will charge a daily fee for using the service.

Scenario 3: HMRC provides services in-house

HMRC would continue to provide access to its data using its existing resources. If the Department decided to allow researchers to link its data to other sources, it would negotiate for access with each of these sources (e.g. ONS, BIS etc) to be used inside the Datalab.

Assessment

The third scenario is the least desirable, both for the research community and for the HMRC. Negotiating for access to a variety of data sources is time and resource intensive. This seems particularly unnecessary at a time when resources in the public sector are constrained, and given that the Secure Data Service has already negotiated terms of access to sensitive and confidential ONS data.

The second solution is the least expensive for HMRC. The Secure Data Service has already invested technology that meets international infrastructure and security standards. HMRC would simply have to provide inexpensive dumb terminals to connect to the service. The remaining marginal costs of supporting the researchers would be borne by the Secure Data Service.

Also, the UK Data Archive can fulfil many of the data documentation and user support functions on behalf of HMRC. In particular, the organisation can advise and assist HMRC on researcher training, documentation and meta-data standards.

In addition, the Secure Data Service can, in conjunction with the ESRC, impose additional penalties on researchers who abuse their access conditions. In the worst-case scenario, the ESRC can withdraw funding for entire institution for five years. The Secure Data Service can therefore provide the incentive mechanism that encourages researchers to work responsibly.

An assessment should ideally be made about whether the VML or Secure Data Service should support HMRC Datalab in the provision of data services to the research community. HMRC would make a decision based on either the reassurance of another Government Department providing a service and receiving HMRC data, or an organisation, supported by a major Research Council, with vast experience of data management and compliance with international IT security standards undertaking services on behalf of HMRC.

Before such a decision can be made, we recognise that it is important to allow the HMRC a period of time to develop, and to reassure the MRP that the access mechanism that has been established is trustworthy and reliable. We therefore conclude that working with a service such as the Secure Data Service to provide access to more data can be considered for implantation in the medium-term.

2.4 Assessing the Impact of Research

Research Councils UK (RCUK) is a body representing the seven Research Councils, including the Economic and Social Research Council. RCUK's 'Pathways to Impact' programme embodies RCUK's belief that investment in research should have academic impact (i.e. research should contribute towards academic advances across disciples), and should also have an impact on society (for example, by "increasing the effectiveness of public services and policy"[2]).

We believe that HMRC would benefit from assessing the impact of the research undertaken in the Datalab. It would allow the Datalab to feed relevant research into HMRC's policy-making areas.

However, to ensure that assessing the impact of research outputs does not commit HMRC resources, we believe researchers should be responsible for assessing the likely impact of their research. Secure Data Service has already considered this. Upon requesting a research output from the secure environment of the Secure Data Service, researchers must provide information about the impact of their research. For example, researchers are asked to state the key findings of their research, and whether their findings are relevant for policy-makers. In the future, the Secure Data Service can use this information to feed research outcomes generated by researchers to interested Departments and other stakeholders. We encourage HMRC Datalab to adopt a similar approach.

[2]See 'Pathways to Impact' at http://impacts.rcuk.ac.uk/content/impactmeans.htm

3.0 Liaising with other services

This section considers how HMRC Datalab can benefit from liaising with other Government Departments and similar data services. Developing partnerships and working relationships with these other organisations can enhance the operations of the Datalab. The first part of this section will consider how HMRC Datalab can benefit from immersing itself into the data infrastructure community. We follow by presenting findings of our discussions with other Departments. Finally, we consider the benefits to HMRC that can be accrued from working with researchers.

3.1 Working with Other Data Services

There already exists an established infrastructure community with expertise in managing data, which HMRC Datalab can participate in. A large pool of expertise can be tapped into. Many of the issues which HMRC are working through with regard to setting up a data enclave, have already been thought about and solved.

What are the relevant forums for sharing experience?

There are now at least two formally organised forums where RDCs meet to share their experiences:

Workshop on Data Access

This yearly workshop brings together directors of RDCs across the world. So far, meetings have taken place in Nuremberg (2007), Cardiff (2009) and Ann Arbor (2010). With each successive meeting, more and more representatives from RDCs have attended.

IASSIST

This conference brings together staff from RDCS as well as Data Archives around the world. The programme focuses on technology and data (including access).

In addition to these, the following two organisations regularly undertake projects to introduce or formalise standards concerning the management of data (this encompasses many activities, including data collection, documentation, meta-data, management of data, security standards).

CESSDA

The Council of European Social Survey Data Archives (CESSDA) is an organisation of Data Archives throughout Europe. Regular meetings and projects are established to share experiences and develop standards in data dissemination and access. The Council regularly works with National Statistics Institutes (NSIs). This may be an appropriate platform for HMRC to join, not only to provide advice but to learn from the experience of Data Archives.

ESSnet

The European Statistics System Network (ESSnet) is an organisation of representatives from NSIs in Europe. Projects are established with the intention of producing common standards in statistics, but also include data management, dissemination and access. This may be a forum that HMRC could usefully participate in to share and gain experience of managing HMRC Datalab.

Why work with other data enclaves (RDCs)?

The above forums and organisations regularly meet to discuss the following issues (for which a large amount of expertise is available:

Infrastructure	For example technological solutions (hardware and network), and other computing solutions (such as thin-client computing, cloud computing, website development and interactive researcher tools).
Administration	Practices and procedures for managing data, researchers (including applications to access data), communication, publications.
Statistical Disclosure Control	Recent developments in detection, discussion of new research methods and implications for SDC, managing SDC efficiently
Operational Issues	Bottlenecks and solutions. For example, services can share problems they have encountered and develop methods to solve them.

How does the wider research community benefit from this collaboration?

The benefits of collaboration between RDCs are summarised below:

Data providers	RDCs can distribute secure access to confidential data to the research community. RDCs specialise in this function and have considerable expertise. This knowledge-base expands when RDCs work together to share lessons, and will benefit data providers because RDCs provide a service on their behalf.
Policy-makers and Researchers	Using their expertise to provide access to data benefits policy-makers because researchers can access the data they require to undertake analysis.

This research feeds into the policy-making process. New technologies allow researchers to access data and exploit it, enabling better research to be conducted.

When RDCs share lessons about this technology it benefits the researchers who analyse the data, and the policy-makers who receive the evidence they require to make informed decisions.

Case Study: Establishing the Secure Data Service

The Secure Data Service has followed the ONS VML model of providing access to data. In order to establish the Secure Data Service, the UK Data Archive has worked closely with the VML on a number of issues. These include: managing researchers (including training and applications), managing datasets, SDC, and technology issues (operating systems etc). In order for the ONS to be satisfied that the Secure Data Service can manage the service effectively, the Secure Data Service and VML have worked together to ensure that a number of areas (such as researcher training) are satisfactory. The establishment of the Secure Data Service is a good example of co-operation between RDCs.

Case Study: Canada and US Bureau of Census

The Canadian RDC Network and the US Bureau of Census RDC Network meet twice a year to discuss operational issues. In addition, board members sit on each others' committee (i.e. a member of the Canadian RDC Network will join the US Bureau of Census Board.).

Case Study: Canada and NORC

The Canadian RDC Network joined forces with NORC in 2009 to produce a proposal known as 'Digging into Data'. The proposal examined developing meta-data elements that related to SDC management.

3.2 Working with Other Government Departments

There are many reasons for working closely with other Government Departments because research can benefit the policy decision-making process. In this section, we highlight the benefits to HMRC from working collaboratively with other Departments.

Why should HMRC collaborate with other Departments?

Firstly, a large amount of resources are invested into collecting data. HMRC can increase the return on this investment by providing access to researchers, who regularly undertake research either on behalf of other Departments, or who produce research that is still relevant for policy-making across Government.

The research agenda can be influenced by Government. In particular, each Department will often have its own programme of research that it wishes to pursue. For example, policy-makers may wish to analyse the impact of a particular programme that they have introduced, so research is driven by Government.

On the other hand, researchers can influence the direction of policy by making public the findings of their research. For example, researchers can make Government aware of how policies impact society. Departments may wish to consider undertaking further analysis, and data from a number of Departments may be required to provide a better understanding of a particular issue.

Another reason for collaboration is because not all Departments will have the infrastructure to allow researchers to analyse their data securely. Therefore, researchers are prevented from undertaking analysis which would benefit those Departments.

We have spoken to two other Departments which have expressed an interest in using the data already made available by HMRC. This is discussed in more detail below.

Example 1: Department for Business, Innovation and Skills (BIS)

One of the key policy areas of this Department is innovation, and understanding the role of innovation in the economy. The data which HMRC is making available from the Corporation Tax return form, is relevant to the research undertaken by BIS (in particular, the Department is interested in the R&D Tax Credit data which companies supply).

BIS regularly outsources its research work to the academic community. Academic researchers play an important role in generating research and analysis that feed into the BIS policy-making process. Access to new sources of data is clearly of interest to the Department in order to provide the best evidence and advice to ministers who are required to make policy decisions.

Example 2: Intellectual Property Office (IPO)

The IPO has embarked on a large programme of work, with the intention of quantifying the number of intellectual property rights (IPRs) in the economy (for example patents, trade marks and design copyrights). The Department wishes to investigate the behaviour of those firms that hold IPRs, compared to those firms that do not. Another programme of work includes the creation of an R&D 'Index' by researchers at Imperial College to assess the size of R&D activity as a proportion of GDP is strongly supported by the IPO.

The IPO is working with a number of researchers and institutions throughout the UK to improve their decision-making process. . As well as data on IPR, they have a strong interest in evaluating the characteristics of firms that hold IPRs, since the Department does not collect this information. Data from other sources, including the R&D component of the Corporation Tax dataset from HMRC dataset, would therefore be very welcome.

Research Potential

We have provided two examples of other Government Departments that HMRC could liaise with. Not only would the research community benefit from such collaboration, but so too would these Departments. The analysis would also benefit HMRC because the research interests of these Departments relate to the research agenda that HMRC has established (see section 2.1.3 What Setting?).

The feedback from both Departments, as well as from researchers we have spoken to (including at the IFS, NIESR, University of Oxford, LSE and Imperial College), is the usefulness of combining sources of data together.

But innovation analysis (and the R&D tax credit data) will surely be the tip of the iceberg. There are undoubtedly many other avenues of research that could be explored by combining these Corporate Tax data with other sources. As HMRC increases its catalogue of data, then opportunities for further research will grow.

4.0 Promoting the Service

The final part of this report reviews different options for promoting HMRC Datalab to ensure effective take-up by the research community. Our intention is to assist HMRC in developing a communications strategy, through which different methods of promotion will work harmoniously to create a service 'identity'. Services such as the Esecure Data Service, for example, are instantly recognisable by the research community.

This section is organised into four parts. First, we provide guidance on how HMRC can develop its communication strategy. We then provide options for promoting the service for the consideration of HMRC: they are based on the experience of data enclaves around the world. We also identify research networks that HMRC can tap into, and promote the service to help ensure there is take-up. Finally, we briefly consider how HMRC could work directly with researchers to promote the service.

4.1 Developing the Communications Strategy

The purpose of a communications strategy is to ensure that all methods of communication are working harmoniously towards achieving the same goal: ensuring take-up of HMRC Datalab by the research community.

The following aspects contribute towards this strategy:

Communications 'culture'	Creating a culture of communications is very important, especially to establish the principle that HMRC Datalab is trying to reach out to the external research community. It should not be the responsibility of one member of staff to deal with communications – everybody with an interest in making the service a success also has a communications role to play. Actions that may develop this culture can include: • Establishing a short-life working group to plan communications activities • Electing a communications 'Champion' to co-ordinate activities • Make 'communications' a default item on the agenda of internal meetings
External perceptions	HMRC should evaluate the general external perception of its existing data services. Our experience of talking to the research community is that HMRC data would be very valuable for use in research, but overly-restrictive policies prevent access to data.

Statement of objectives	This could be written to state what it is HMRC are trying to achieve in terms of promoting the Datalab. For example, is the objective to install in researchers' minds that the Datalab is the first place to go to access tax data? In terms of the Datalab, what are the strategic targets for HMRC in, for example, 5 or 10 years time? The statement should also state the activities that HMRC are going to undertake to achieve these objectives.
Target audiences	We understand that HMRC are trying to engage with the research community. But which research community? Researchers often move in close circles. And which circles to penetrate will depend on the data which HMRC have to offer, as well as the research interests of HMRC and HMT (since applications will be tied to this agenda). An evaluation should be made by HMRC about such groups of researchers before undertaking promotional activities.
Budget	The ESRC's Communications Toolkit recommends that approximately 5 per cent of a project's budget should be devoted to communications.
Evaluation	Before and after evaluation to determine whether the communications strategy is effective is important. External awareness of the service should be measured. Other aspects of evaluation include: • Does research translate into analysis or policy? • Feedback and participation from events • Monitoring website usage

In addition, HMRC should at all times be transparent, and provide as much information as possible to researchers. This enable researchers to plan their work in advance (and apply and secure the necessary funding from Research Councils and other sources). Research proposals of a good standard are likely to be delivered to HMRC if researchers have good information about the data available, and how to use the data.

4.2 Options for promoting an RDC

Drawing upon the experience of the UK and international data enclaves, including the VML and Secure Data Service, we now identify activities to establish the reputation of the service.

The table below presents some options for promoting the service. The list is not exhaustive, but they are examples of activities undertaken by data enclaves around the world.

Table 2: Promotion Activities

Activity	Objective	Target Audience	Impact (success measure)	Example
Launch Event	Raise awareness of the Datalab, stimulate new data users and applications, inform policy advisers.	Researchers, other data enclaves, policy-advisers (from other Departments)	Increasing number of queries and applications to use the service.	December 2009: Launch of the Secure Data Service at the Royal Statistics Society.
Conference Presentation	Promote the service and available data (inform potential users about current use, stimulate new growth in demand, receive feedback for service improvements from current users).	Researchers and policy-makers.	Increasing number of queries and applications to use the service.	VML and Secure Data Service have made presentations at a number of conferences, attended by researchers of varying disciplines. The CAED (Comparative Analysis of Enterprise Data) would be an ideal opportunity to raise awareness.
Workshops (May be specific to the service, themed around a research topic or around a data source)	Promoting the service and available data. Bringing researchers, data suppliers and policy-makers together.	Researchers, data suppliers and policy-makers.	Increase in applications to use data. Improvements to data collection methods. Improvements to data dissemination/access. More/better evidence-based policy.	VML regularly organises workshops. Researchers using the service are invited to present their findings to peers, policy-makers and suppliers. Positive feedback from other Government Departments about this service (e.g. BIS).
E-Bulletins (including email distribution lists e.g. JISCmail).	Maintain relevance and raise awareness of new data and services. Encourage new/continued use of data, and new deposits of data.	Researchers (although policy-makers also find these updates informative).	Responses to survey requests. Requests to be added to mailing list.	VML produce a quarterly bulletin to inform users of service developments.
Word-of-mouth (including support for researchers, networking activities and conferences and workshops etc).	Spreading news of service developments	Researchers.	Interest and applications to access data.	Researchers move in close circles and regularly present work to peers. Message of Datalab can spread quickly.
Data Briefs (a short research document analysing a particular aspect of using the data).	Promote specific data by highlighting features. Encourage applications to access the data.	Researchers (potentially policy-makers).	Interest and applications to access data.	VML periodically publishes data briefs about particular data sources.

Activity	Objective	Target Audience	Impact (success measure)	Example
Website	Provide information about the service (such as applications, data, documentation, meta-data). Potential for providing frontline support services.	Researchers, policy-makers, data enclaves.	Interest and applications to access data. Minimal support role by Datalab staff.	The new Administrative Data Liaison Service (ADLS) has an excellent website with information about data sources and how to access them. Publications are also posted on the site.
Liaising with other services	Promote the Datalab.	Researchers.	Interest and applications to access data.	Secure Data Service, VML and ADLS recognise the benefits to researchers when promoting each others' services and strengths.
User group meetings	Promote use and understanding of particular data.	Researchers, policy-makers and data providers.	Increase in applications to use data. Improvements to data collection methods. Improvements to data dissemination/access. More/better evidence-based policy.	BIS organises regular user groups meetings for its innovation data. Each session is attended by researchers, data suppliers, data enclaves and policy-makers. This regular exchange of ideas feeds into policy-making, improvements to data collection and access dissemination.

In addition to supporting research, RDCs may engage in activities such as developing standards. For example, the VML has developed standards for Statistical Disclosure Control, which have been adopted by Eurostat.. The Secure Data Service provides an advisory function to other data enclaves (for example, advising the Scottish Health Informatics Programme). These activities will often include making presentations at workshops for data professionals, thereby raising its profile.

4.3 Working with other Services

In this section we discuss how HMRC Datalab could benefit by working with similar services to promote data.

4.3.1 Evidence from Canada

Evidence from overseas data enclaves suggests that joint collaboration between data services is an effective method of promoting a service.

For example, the Canadian RDC Network works with research institutions, Research Councils and federal agencies to promote the service. Institutions provide information about the Network to its researchers and students. Research Councils provide funding to researchers who often access the Network for data.

The Councils support the Network by providing promotional literature. In both of these examples, the objective is to remove knowledge barriers to accessing data (i.e. ensuring researchers are aware of the Network).

Federal agencies often contract researchers to undertake analysis on their behalf, and so they benefit from promoting the Network, because researchers are aware of the resources that are at their disposal to produce analysis of an excellent quality.

4.3.2 The Data Coalition

The Administrative Data Liaison Service (ADLS), Secure Data Service, VML and HMRC Datalab, all share a common goal: to promote access to confidential data. Each service is in regular contact with a number of researchers and other Government Departments. By pooling resources to promote each others' services, researchers will benefit from improved knowledge of data sources and how to access them.

The VML, Secure Data Service and ADLS are already working closely to promote each others services at training seminars and conference presentations. Since the services complement each other, this partnership will benefit researchers and their analyses.

HMRC Datalab could join this 'Data Coalition' – it would be a cost-effective way of promoting the service and bringing it to the attention of researchers. The Secure Data Service have already confirmed that they will promote the new service at its training seminars.

4.3.3 Economic and Social Research Council (ESRC)

As in Canada, Research Councils provide a large proportion of funding for researchers wishing to analyse data. The impact of this research will be more effective if researchers have access to better data resources. We therefore suggest that HMRC work closely with the ESRC to promote the Datalab. This is already happening as funding has been made available for a small number of projects using the data.

An important event in the academic calendar is the ESRC Research Methods Festival. This is an opportunity for HMRC Datalab to promote its service to academics of many disciplines, particular to young researchers. This is especially important because in recent years, the ESRC has prioritised research funding for young researchers undertaking applied research.

4.3.4 Administrative Data Liaison Service (ADLS)

The ADLS was launched in December 2009, alongside the Secure Data Service. Its purpose is to act as an intermediary between researchers who wish to access administrative data, and Government Departments who wish to provide these data.

We recommend that HMRC and ADLS work jointly to promote HMRC Datalab. The ADLS is an excellent source of information for researchers, and HMRC would benefit from this publicity.

4.4 Identifying Research Networks

This section summarises the research we have conducted, investigating the research networks which HMRC may wish to understand in order to promote its Datalab.

As explained in Section 2.1.3 (What Setting?), proposals to use HMRC Datalab will be accepted by the MRP if they relate to the research interests of HMRC/HMT. Given this agenda, potential topics of research that the Departments would be interested in may include:

- Behavioural economics
- Labour economics
- Industrial organisation
- Tax
- Finance
- Forecasting (and effects of tax structures)
- Research & Development/Innovation

The following is a small selection of research institutes that HMRC could liaise with to promote the Datalab. The list is not exhaustive. However, it does provide a starting point for HMRC.

- Institute for Fiscal Studies (IFS)
- National Institute for Economic and Social Research (NIESR)
- Policy Studies Institute (PSI)
- Institute for Public Policy Research (IPPR)
- Institute for Employment Research at University of Warwick
- Euromod (University of Essex)
- Centre for Market and Public Organisation (CMPO) at the University of Bristol
- Centre for Economic Performance (LSE)
- London Business School
- Oxford University Centre for Business Taxation
- Imperial College, London

Behavioural economics is a relatively new discipline in economics. Two centres that undertake a large amount of research in this area include CEDEX (Centre for Decision Research and Experimental Economics, University of Nottingham) and EXEC (Experimental Economics, University of York).

After speaking to researchers at the above institutions, they find the following aspects of data critical to their research:

- Data as a panel (ability to track individuals or companies throughout time)
- Ability to analyse the structure of companies throughout time
- Ability to link sources of data

4.4 Working with the Academic Community to Generate Research

A more discreet form of promoting the Datalab can include collaboration with researchers. Word-of-mouth is an important element of promoting these data services. Researchers often move in very close circles, and message of the Datalab's existence will soon spread.

Two options for working with researchers include:

- Tenders to conduct research on behalf of HMRC
- ESRC Knowledge Exchange Small Grants Scheme

HMRC has already explored the first option. Working with the ESRC, grants have been made available for researchers who can propose to use the data made subject to the research agenda of HMRC. This could be extended in another way. Tenders could be advertised for researchers willing to organise the data (undertaking data cleaning routines, documentation, meta-data etc.), for the benefit of HMRC and other users of HMRC Datalab.

The second option is also very effective. Two examples of the ESRC Knowledge Exchange include the contracting of an academic researcher working for the Department for Business, Innovation and Skills (BIS), and the contracting of the LSE Data Manager to the ONS VML. Both examples are characterised by:

- The benefits to the organisations (both personnel provided new ideas and techniques to the Departments)
- The benefits to the personnel involved (both gained experience of working in Government)

When the personnel return to their original workplaces, they are likely to pass on knowledge of the Department. Acquiring an academic for a period of time would be useful for HMRC because of the network of contacts that they would possess. To employ somebody working full-time in the Datalab to improve data and documentation is likely to result in the universal dissemination of knowledge of the Datalab's existence.

5.0 Recommendations

We now summarise our findings and make recommendations to HMRC about how the Department can successfully engage with the academic research community, other Government Departments, and other data services. We make recommendations for:

- Informing academic researchers and other Departments about HMRC Datalab
- Liaising with similar data enclaves
- Distributing access to HMRC data

With the austerity measures facing the public sector, it is more important than ever before to put efficient data access infrastructure in place. The consequences of denying access to data on the grounds of cost would be a serious blow for the research effort in the UK. Evidence collected from micro-data research continues to play an important part of the policy-making process, and efforts that strengthen this evidence-gather should be encouraged.

HMRC has made an important achievement in providing access to its micro-data. We understand that data suppliers have justifiable concerns about allowing researchers to examine data. However, this report has suggested some of the mechanisms that can enhance the experience of researchers using the Datalab.

This report also highlights the benefits that can be accrued by data suppliers and policy-makers when researchers are provided with access to micro-data. Given these benefits, we strongly believe that emphasis should not be placed on whether data is too sensitive to be accessed. Instead, a discussion about sensibly managing the risk of providing access to data should ensue. The experience of data enclaves around the world concludes that, if managed effectively, the risks from providing access to data is negligible.

5.1 Informing Research Stakeholders

Informing these stakeholders will be essential if HMRC wish to maximise the take-up of its Datalab.

We recommend that HMRC establish a coherent communications strategy, comprising of various promotional techniques that work together to effectively promote the Datalab. One aspect underlying all of the promotional activities that HMRC can undertake is the need for transparency in decisions by the MRP to grant access to projects.

Researchers should be aware of what HMRC expects from them in terms of their application. This will enable them to plan their research accordingly. Similar information is needed about using the service, in particular, disclosure control of outputs. Researchers should know what is expected of them, and what they can expect of HMRC, when they are using the Datalab.

Specifically, we recommend that a ‘launch workshop’ is hosted by HMRC to promote its existence. We strongly suggest that the individuals from the four stakeholder groups are invited. There is much benefit to be derived from gathering data owners, researchers, policy-makers and service providers in one location.

An effective website will also play an important role within the communications strategy. This will act as a key centre for disseminating information about the Datalab (for example, the data that is made available and how to apply for access). In the future, the website can also provide some of the support roles that would otherwise require the resources of HMRC.

We believe that attendance at conferences will be essential for HMRC Datalab staff. This will undoubtedly spread the message of the Datalab's existence. The recent CAED (Comparative Analysis of Enterprise Data) held at Imperial College in London would be an example of such an event.

Also, other data services, such as the VML, Secure Data Service and ADLS, are also effective forums for promoting HMRC Datalab. Working with these services to jointly promote access to each others' data and services, will benefit HMRC Datalab enormously, since these services have access and are in regular communication with a large number of researchers that would have an interest in applying to analyse HMRC data.

5.2 Liaising with the 'Access to Secure Micro-Data' Community

A number of forums exist which HMRC Datalab can participate in. We are in no doubt that the Datalab will benefit from the experience of other data enclaves that provide secure access to confidential micro-data.

In particular, we recommend attendance at conferences such as the Workshop on Data Access (WDA) and IASSIST, which facilitate an excellent exchange of ideas and problem-solving, and how to effectively manage the risk of providing access to data.

The VML, Secure Data Service and HMRC Datalab are three services that share the same purpose: providing secure access to confidential business micro-data to the research community. All services will be in constant communication with researchers and data providers. We expect that a researcher using one data service (e.g. the Secure Data Service) for a particular project might also benefit from learning about the data provided by another service (e.g. HMRC Datalab).

We therefore recommend that the four services work in collaboration to promote access to each others' datasets, in a 'Data Coalition'.

We also recommend that HMRC use promotional activities to target other Departments. Better use of resources, and consequently, better evident-based policy-making, are the fruits of collaboration. Two Departments have already expressed an interest in working with HMRC Datalab. Expressions of interest may arrive from other Departments if the existence of HMRC Datalab is communicated effectively.

5.3 Wider Dissemination of Access

As stated previously, HMRC should be congratulated for its achievement in establishing a Datalab. We recognise that a period of safe use is desirable to ensure that all stakeholders are satisfied with current access arrangements.

Dissemination of access through services such as the Secure Data Service is unlikely to occur in the foreseeable future. However, we propose a solution which would allow researchers using the Datalab to also access data made available in the Secure Data Service. We recommend a consultation to investigate this further.

The solution is highly desirable for researchers and other Government Departments, and inexpensive to achieve. We believe that the infrastructure that the Secure Data Service can provide is flexible enough to satisfy HMRC data providers (in terms of security, management of data and management of researchers), and researchers who wish to combine HMRC data with other sources. We therefore recommend further dialogue between the services to consider whether this solution can be realised.

If data about individuals (such as Tax Credits) is made available in the future, we recommend that investment is made in licensed versions (such as the EUL). There are many advantages that both the research community and HMRC would benefit from.

We recommend that developing working partnerships with organisations such as the UK Data Archive would help HMRC to achieve dissemination and promotion of these data, especially given the experience of the UK Data Archive. The UK Data Archive could also play a role in documenting data, and producing meta-data, on behalf of HMRC.

References

Desai, T. and Ritchie, F. (2009), "*Effective Researcher Management*", paper presented to the Joint UNECE/Eurostat work session on statistical data confidentiality, available to download from http://www.unece.org/stats/documents/2009.12.confidentiality.htm.

Lane, J., Heus, P., and Mulcahy, T. (2008), *"Data Access in a Cyber World: Making Use of Cyberinfrastructure"*, Transactions on Data Privacy 1 (2008), 2-16.

Ritchie, F. (2009), "*Designing a national model of data access*", available to download from http://felixritchie.co.uk/publications/felix_pubs.html.

Ritchie, F. (2010), "*International Access to Restricted Data - A Principles-Based Standards Approach"*, Presented at WDA 2010 (U. Mich., May) and IASSIST 2010 (Cornell, June), available to download from http://felixritchie.co.uk/publications/felix_pubs.html.

Brandt, M. et al (2010), *"Guidelines on the checking of output based on microdata research"*, recommendation prepared by ESSnet for Eurostat, available upon request.

More information about the RDCs mentioned in this report can be found by visiting their websites:

ONS Virtual Micro-data Laboratory:
http://www.ons.gov.uk/about/who-we-are/our-services/vml/about-the-vml/index.html

ESRC Secure Data Service:
http://securedata.ukda.ac.uk/

NORC:
http://www.norc.uchicago.edu/DataEnclave/

Canadian RDC Network:
http://www.rdc-cdr.ca/

The director of the Canadian RDC Network is:

Charles (Chuck) Humphrey
Academic Director, Alberta Research Data Centre
1-01 Rutherford South
University of Alberta
Edmonton, Alberta T6G 2J4
CANADA