Dadansoddi ar gyfer Polisi

YMCHWIL GYMDEITHASOL Y LLYWODRAETH

GSR

GOVERNMENT SOCIAL RESEARCH

Analysis for Policy

Llywodraeth Cymru
Welsh Government

www.cymru.gov.uk

Ymchwil gymdeithasol

Social research

Rhif/Number: 26/2014

Recommendations for Improving Research Access to Potentially Disclosive Data

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

Author

Tanvi Desai
London School of Economics

Views expressed in this report are those of the researcher and not necessarily those of the Welsh Government)

For further information please contact:

Sarah Lowe

Knowledge and Analytical Services

Welsh Government

Cathays Park
Cardiff

CF10 3NQ

Tel: 02920 826229

Email: sarah.lowe@wales.gsi.gov.uk

Welsh Government Social Research, 2014

ISBN: 978-1-4734-1083-1

Table of Contents

Glossary

ADT	Administrative Data Taskforce
ADRN	Administrative Data Research Network
ADRC	Administrative Data Research Centre
ADLS	Administrative Data Liaison Service
AFON	Access File ON-line
BIL	Business Impact Level
CESG	Communications-Electronics Security Group
DEWI	Data Exchange Wales Initiative
DMP	Data Management Policy
DPA	Data Protection Act
DwB	Data without Boundaries
DWP	Department for Work and Pensions
ESRC	Economic and Social Research Council
EUL	End User License
FOI	Freedom of Information
GSS	Government Statistical Service
HMRC	Her Majesty's Revenue and Customs
ICO	Information Commissioner's Office
IL	Impact Level (See Business Impact Level)
ISO	International Organization for Standardization
KAS	Knowledge and Analytical Services (WG)
LS	Longitudinal Study of England and Wales
MoJ	Ministry of Justice
NHS	National Health Service
NILS	Northern Ireland Longitudinal Study
NWIS	NHS Wales Informatics Service
ONS	Office for National Statistics
PC	Personal Computer
PGP	Pretty Good Privacy
RDC	Research Data Centre
SAIL	Secure Anonymised Information Linkage
SDS	Secure Data Service
SHIP	ScottisH Informatics Programme
SL	Special License
SLS	Scottish Longitudinal Study
SRSA	Statistics and Registration Services Act 2007
UK	United Kingdom
UKDS	UK Data Service
VML	Virtual Microdata Laboratory
WG	Welsh Government
WISERD	Wales Institute for Social and Economic Research Data and Methods

1. Introduction

Due to a long-standing collaboration between the academic funding councils and government, the UK has one of the oldest, best established infrastructures for providing research access to microdata in the world. UK researchers in academia and government have long taken advantage of the wealth of available data to produce internationally respected, policy relevant research. However, in recent years, new guidelines for government data handling combined with economic pressures have created a tension between the need to maximise returns to investment in data collection, and the obligation to ensure absolute confidentiality to respondents.

This tension increases the urgency for government bodies to have a comprehensive overview of data resources under their control in order to understand the level of sensitivity of each resource; for reassurance that access policies conform to legislative and ethical standards; and to enable assessment of whether these resources can be more fully exploited. To address this challenge the Welsh Government (WG) and the Economic and Social Research Council (ESRC) have jointly funded four fellowships with the aim of obtaining a set of recommendations on strategies for maximising returns to investment in microdata resources. This report is the result of the Fellowship on Improving Access to Potentially Disclosive Data.

There are many other projects underway that will impact the recommendations in this paper including Data without Boundaries (DwB)[1], Beyond 2011[2], and ESRC Big Data Network[3]. The Administrative Data Research Network (ADRN) recommended by the Administrative Data Taskforce (ADT)[4] has been referred to extensively in the text below as it is likely to have a significant impact on access to all potentially disclosive data in the UK, not just administrative data.

[1] http://www.dwbproject.org/

[2] http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes---projects/beyond-2011/index.html

[3] http://www.esrc.ac.uk/news-and-events/announcements/25683/big-data-investment-capital-funding.aspx

[4] http://www.esrc.ac.uk/funding-and-guidance/collaboration/collaborative-initiatives/Administrative-Data-Taskforce.aspx

This report will provide an outline of the methods used for the project, before giving an overview of current data access infrastructure in the UK and Europe. This will be followed by an outline of the legal and ethical issues that should be taken into consideration when developing plans to maximise access to potentially disclosive microdata and some comments on implementation and funding. The paper will then examine strategies for maximising the availability of, and access to microdata, before concluding with recommendations.

The Welsh Government's commitment to maximising returns to investment in data collection is a significant step towards efficiency and cost-effectiveness. With this in mind strategies that make use of existing infrastructure such as the UK Data Service (UKDS), the government Research Data Centres and SAIL will be prioritised, and where possible, solutions are based around these pre-existing facilities. Where the existing infrastructure and support systems are inadequate, solutions will be put forward based on international best practice.

2. Methods

The aim of the Knowledge Transfer Fellowship was to enable the Welsh Government to draw on the expertise developed by the Fellow in the area of secure data access. In order to frame any recommendations it was first necessary to identify the needs of WG Knowledge and Analytical Services (KAS) and other researchers investigating Wales, therefore a stakeholder consultation was undertaken.

The consultation consisted of both face to face interviews and requests for feedback via email and the web. Twenty-eight face to face interviews were carried out, including with twelve KAS analysts, six academic researchers, and six data sharing specialists from other government departments, two third sector researchers, and two managers of data access infrastructures. In addition a presentation was made to the Welsh Statistical Liaison Committee and an article written for the Wales Third Sector Network newsletter explaining the project and requesting feedback.

A request for feedback on data needs for research on Wales was posted on the rlab-data blog[5], and was viewed 124 times. The request for feedback was also circulated internally within WG via their intranet, and distributed widely to Welsh Universities via targeted emails to departmental administrators. It was also circulated to the MAUS[6] and SAIL[7] mailing lists. A total of 16 responses were received via the online request for feedback. All but two of these responses came from academics with the others from the third sector.[8]

The number of responses is relatively low and therefore cannot be seen as a representative sample. However the intention of the project was never to undertake a comprehensive survey of researchers working on Wales, rather to identify key themes for the fellowship to address.

---

[5] http://rlab-data.blogspot.co.uk/

[6] http://www.ons.gov.uk/ons/about-ons/who-we-are/services/vml/contact-us/index.html

[7] http://www.swansea.ac.uk/medicine/research/chiral/ehealth-and-informatics-research/healthinformationresearchunit/

[8] The consultation did not include researchers from the private sector as this was outside the remit of the fellowship.

Following the consultation the Fellow's knowledge of existing and emerging best practice was used to make recommendations of how WG could maximise the use of existing data in a way that meets the needs of their stakeholders.

It was originally planned that the Fellowship would include an audit of Welsh data resources. One of the first tasks undertaken was to send out a request for information via the KAS intranet which received not a single response. Added to this lack of feedback, it soon became clear that the information, knowledge and awareness of the data resources held by WG was so dispersed that an effective audit would require the majority of the time available to the Fellowship and it was decided in conjunction with the project manager that this would not be the best use of resources. A list of documents containing information on WG data resources is included in Annex A to assist any future audit.

3. Data Access Landscape

This section outlines the main methods that are employed internationally for providing access to sensitive microdata for research purposes, and details some of the existing infrastructure within the UK.

Countries are increasingly investigating systems to enable them to maximise the returns to their investment in data by providing researcher access to sensitive microdata. These infrastructures are habitually referred to as Research Data Centres (RDCs).

A common practice among RDCs internationally is to grant permission to access sensitive data at the project level, so that permission covers a specific researcher, project and dataset(s). This means that if a researcher wishes to carry out data analysis that does not directly relate to the project for which they are registered they must make a new application. The use of data to investigate an area for which a researcher is not registered is a breach of contract (though not usually classified as a data breach), and in some cases carries legal sanctions[9].

The main differences in RDCs tend to depend on whether remote access or remote execution is used to analyse the data and where the data can be accessed from.

A remote *access* system allows researchers to work with data as if they were working at their own desktop. Researchers remotely access a secure environment that provides them with a suite of familiar software and a link to the sensitive data they are licensed to access. All data processing and analysis takes place within the secure environment and all output (anything that is removed from the system) must be checked (almost always manually by RDC staff) to ensure no disclosive information is released. Apart from these outputs no data are ever transmitted, as all work is done remotely. Systems vary in how much data a researcher can access, in some cases e.g. UK's VML and SDS, researchers are allowed to see and manipulate full datasets, in others e.g. UK's SAIL and SHIP, researchers must request access to specific variables. While limiting access to variables does

[9] In the UK the use of ONS data for purposes other than those granted is illegal under the Statistics and Registration Services Act, and carries the possibility of a fine or a custodial sentence.

marginally increase security it significantly decreases utility. This is because unless a researcher is very familiar with a dataset it can be difficult to identify exactly which variables are required in advance (or whether the variables available will be suitable for analysis), particularly when using very complex datasets such as health records. Restricting variables to those the researcher 'needs' supports a common misconception among data providers which is that researchers work 'project to data' i.e. they develop a hypothesis and then find the data to test their hypothesis. As well as being a risky strategy (it should never be assumed a priori that the data needed exist); it is also not the way many researchers, particularly those who are more experienced data users, actually work. Many researchers who habitually use data for research work 'data to project' i.e. some characteristic or property of the data inspires their research[10]. Therefore by limiting the variables available to researchers you limit the scope of research that can be done and much of the original work that might be produced by experienced data users.

A remote *execution* system is one where the researcher never has direct access to potentially disclosive data. In this system researchers draft their analyses (programmes to run in statistical packages) which are then submitted to the RDC to be run on the data (either automatically or manually by RDC staff). Outputs are checked for disclosive information and then returned, in the same way as for a remote access system. The challenge with remote execution is how to provide the researcher with sufficient information to allow them to draft programmes that will run and produce valid output when they have not seen the data. A common solution to this is to provide researchers with access to a synthetic dataset that has a similar structure and characteristics to the original data, but does not contain real records. For example, for some German data, researchers first pay an on-site visit to the statistical office where they can have supervised remote access to familiarise themselves with the actual dataset, they are then given a synthetic dataset to use to structure all future analyses which are then submitted by email from the researcher's office for remote execution.

[10] Though as noted, if the inspiration relates to a topic area for which they were not registered, they would have to submit a new application before they could investigate in any detail.

While the risk of disclosure of sensitive data are greater for remote access systems than for remote execution systems - as, for the former researchers are allowed to see and manipulate real sensitive data - there are substantial disadvantages to a remote execution system which must be balanced against this reduced risk. Firstly, in terms of the cost of data preparation: while the main cost of both remote access and remote execution systems is in the staff time required for manual checking, the cost of data preparation for a remote access system is relatively low as, provided adequate metadata are made available, researchers can undertake their own data cleaning, create working files appropriate to their needs, and explore the data thoroughly to understand its structure and characteristics. With a remote execution system a certain amount of data cleaning and structuring must be undertaken by the RDC staff as it is far more challenging for researchers to create working files and restructure the data without being able to see it. In addition, the construction of quality synthetic datasets is a non-trivial task that has an associated cost.

When comparing remote access and execution systems across Europe, the DwB project concluded that *"...it is obvious for national data producers and Eurostat that only real RA [Remote Access] systems would meet the cumulative challenge of the high demand for confidential microdata, limited resources inside the NSIs [National Statistical Institutes] and DA [Data Archives] and researchers' needs"* (DwB 2012).

The second major difference in how potentially disclosive data are accessed relates to where they are accessed from. For example, some systems require the researcher to be physically present at the data provider's site so that they can be overseen at all times, as is the case for academic access to VML. Other data providers require access points to be restricted to dedicated safe rooms[11] as is likely to be the case when accessing much of the data available through the Administrative Data Research Network (ADRN) recommended by the ADT. Some providers, such as the SDS[12], restrict access to a place of work but allow connection from a standard

[11] As part of the ADT recommendations, the Administrative Data Liaison Service (ADLS) is leading a project to develop a national standard for safe rooms for the UK for more information see section on Safe Settings below.

[12] VML also allows remote access form standard workplace desktops but only for PCs on the Government Secure Intranet.

desktop[13]. Finally, in countries such as Denmark, Sweden and Slovenia it is possible to analyse sensitive microdata from anywhere.

Considering that researchers inevitably invest a significant amount of time in familiarisation with, and preparation of, data resources before beginning analysis, the cost of travel and accommodation when analysing data at an RDC can be significant. Therefore the location from which data can be accessed has a significant impact on the use of any RDC. RDCs that require researchers to travel in order to access data can have a significant effect on equality of access, as researchers with children, mobility challenges or caring responsibilities are likely to be restricted in their ability to spend time away from home. In addition, the high costs of travelling to RDCs are likely to restrict access for early career researchers, and those with fewer funding opportunities or large teaching commitments, for example, during the stakeholder consultation a researcher commented that the HMRC DataLab was of no use, as the cost in terms of time and money of travelling to London to access data put the DataLab out of reach of researchers in Wales.

Requiring researchers to travel in order to access sensitive data resources will restrict the use of those resources and may exclude certain sections of the population, thereby reducing the returns to investment in making data available for reuse. It is recommended that any system developed to provide access to sensitive data should be a remote access system with the minimum possible restrictions. The greater the restrictions on access points and the data available, the less the system will be used. In addition the more restricted the access system the less likely it is to produce innovative research and new methods, as the effort necessary to explore new areas and unfamiliar data sets becomes impractical.

International Landscape

European countries with established infrastructures that enable remote access to potentially disclosive data for research purposes include the UK, France, Germany, Denmark, the Netherlands, Sweden, and Slovenia. The US and Australia also have remote access systems.

[13] It may be worth noting that while there is an explicit ban in the UK (and many other countries) on academics and civil servants remotely accessing sensitive data from home. This is not the case in some commercial sectors (for example parts of the banking sector), where there is a preference for sensitive data work to be carried out at home, as there is less of a risk of disclosure to professional colleagues.

A table providing a breakdown of the characteristics of European RDCs can be found in Appendix C. The focus on Europe is because initiatives are underway to harmonise data access solutions across Europe to facilitate data sharing across borders. This means that any system developed according to European best practice may in the near future provide an access point to data from other European countries, thus maximising data access for researchers in Wales, and facilitating international comparisons.

The DwB project is working, in cooperation with Eurostat, towards establishing a set of criteria to which RDCs will have to conform in order to receive a Eurostat accreditation. Member states will then have access to an independent assessment of an RDC's security when deciding whether to allow their data to be accessed via an RDC in another member state. DwB is also carrying out an audit of data available in Europe including the level of sensitivity of data available and the access mechanisms in place.

UK Landscape

UK researchers have had access to government socio-economic microdata for research purposes for over forty years via the UK Data Archive, now part of the UK Data Service (UKDS). This long tradition of data access meant that the UK was one of the first countries to provide an infrastructure for controlled access to potentially disclosive microdata. There are now at least six RDCs providing an infrastructure for research access to various types of sensitive microdata within the UK (not including NHS facilities). An outline of the characteristics of key UK RDCs can be found below, including details of any accreditation and certification they hold. The recognised international standard for information security management systems is ISO27001, however some government departments, for example ONS, require secure data access and storage infrastructures to be specifically accredited, whether or not they hold an ISO certification.

A key component of data security for UK RDCs has been researcher training. The aim of the training is primarily to assist researchers in understanding the legal and ethical responsibilities for the data they are using. For RDCs that require outputs to be manually checked training is also a key component of resource management, as ensuring that researchers understand which outputs are more or less likely to be

acceptable reduces the number outputs rejected and therefore the burden on RDC staff. Fewer output rejections also increases user satisfaction, which helps with security (Desai, 2003). In fact, developing relationships with researchers to increase trust and cooperation is a key component of the UKDS secure branch's[14] security model (Desai and Ritchie, 2009).

ONS VML: The Virtual Microdata Laboratory based at the Office for National Statistics provides access to potentially disclosive ONS business, economic, and demographic data, as well as detailed data from the Department for Business Innovation and Skills, the UK Centre for Employment and Skills, and the Department for Energy and Climate Change. VML is a remote access system, and data can be accessed by anyone with appropriate experience in data use provided the research is 'in the public good'. Any organisation on the Government Secure Intranet can apply for a dedicated terminal with a link to VML which can be used by any civil servant who passes the application process. The VML is also available to academic researchers (including academics from outside the UK), but only from a dedicated terminal on an ONS site (London, Titchfield, Newport). Access to data via the VML is currently free.

UK Data Service: Since the consultation, the Secure Data Service (SDS) based at the University of Essex has become part of the UKDS[15]. The UKDS holds the majority of ONS data available in the VML as well as potentially disclosive data from the ESRC longitudinal studies, the Department for Education, the Department for Communities and Local Government and the Medical Research Council for Scotland. The UKDS provides a remote access system that can be accessed from any PC with a dedicated IP address on the JANET network[16] that links all UK further and higher education institutions and the academic funding councils. Permission to access the data remains with the data collectors, therefore anyone wishing to access ONS data, for example, must apply to ONS and conform to their requirements. Access to data via the UKDS is free to academic researchers.

---

[14] Previously the Secure Data Service (SDS)

[15] Funded in October 2012 by the ESRC the UK Data Service brings together the Economic and Social Data Service, the Secure Data Service, the Census Dissemination Unit, and other parts of the ESRC data support infrastructure.For more information on UKDS see Appendix D.

[16] https://www.ja.net/

SAIL: The Secure Anonymised Information Linkage databank (SAIL) is based at the University of Swansea and has long been supported by the Welsh Government. The databank holds a range of data that can be securely and anonymously linked to a number of NHS Wales data sources. The ADT report suggests that *'the establishment of the ADRC would likely build upon the experience of the Secure Anonymised Information Linkage'* (ADT 2012). While SAIL offers a rich resource there are potential barriers to the Databank being used for policy related research as there are restrictions on using some of the data for benchmarking. How this will impact the use of SAIL data for policy related research is unclear, and a more detailed report is being prepared by the Data Max Data Linking Fellow. Access to the data is controlled by an independent Information Governance Review Panel. Researchers do not gain access to the full databank, but a subset of data prepared specifically for the requirements of their project.

Others: Other secure data access infrastructures available in the UK include the SHIP, the ScottisH Informatics Programme[17]; the HMRC DataLab[18]; and the Ministry of Justice DataLab[19] which is being piloted in 2013-14.

[17] http://www.scot-ship.ac.uk/

[18] http://www.hmrc.gov.uk/datalab/about.htm

[19] http://www.justice.gov.uk/justice-data-lab

4. Legal and Ethical Considerations

When considering strategies for maximising the use of sensitive data, the legal and ethical requirements to safeguard respondents' privacy are of primary importance.

This section will outline the traditional model for data access and security before introducing a new model that is more in line with modern infrastructure. Specific legal and ethical issues that may impact WG priorities for maximising access to data will then be examined briefly.

Traditionally, access to data has been viewed as a straightforward trade-off between risk and utility, meaning that the more useful the data the greater the risk of releasing it. As minimising risk (or avoiding it altogether) has often been the primary consideration when deciding whether to release data, this simplistic model has often been used to justify non-release of data, with detailed (useful) data deemed too risky to release in the absence of any analysis of the genuine disclosure risk or the available release mechanisms. This has led to legislation being interpreted (sometimes unnecessarily) in this light (see Ritchie, 2010 & Laurie, 2011) in fact, the ADT report on Improving Access for Research and Policy states that…

'*Where a government department needs statutory powers to share data there is often a criminal sanction for unlawful sharing of data that relates to an identifiable person/legal entity, most commonly a prison sentence of up to two years. This has led to some degree of caution that may go beyond what is required by the legislation itself. This can apply to individual judgements around access to data, where possible solutions are not fully explored because of the perception of the barriers*.' (ADT, 2012).

In recent years the data access infrastructure has changed significantly. Technical and procedural advances mean that even highly detailed data can be safeguarded to a level where the risk is significantly lower than previously. In 2006 a five-part security model 'the Five Safes' was developed for the ONS VML (Ritchie 2006 and Desai et al forthcoming), which is now internationally recognised. The 2006 model was designed to address data release at all levels of sensitivity and includes five components that impact data security – safe projects, safe people, safe settings, safe data, and safe outputs – which can be balanced to safeguard data. 'Safe

projects' means that the 'projects have a valid statistical purpose and are being carried out by responsible researchers who take ultimate responsibility for inferences made'; 'safe people' are defined as having a proven research background and are 'trustworthy' with no conflict of interest; 'safe settings' relate to the setting in which the data are accessed; 'safe data' to the level of sensitivity of the data and the potential for disclosure; and 'safe outputs' to any data that is released into the public domain. Aspects of these five components can be traded off to ensure data security. For example if data are to be released for download by the general public, the projects, people, outputs and settings cannot be controlled and thus should be considered highly unsafe, therefore the fourth component – data – must be strongly controlled and therefore are likely to be highly aggregated so that they pose a minimum risk of disclosure. However if, as is the case for the Data Max project, the data to be released are highly sensitive, then strong controls must be placed on the people accessing the data, the project the data are to be used for, the setting in which it is accessed, and how the outputs are controlled.

This new 'Five Safes' framework offers the opportunity to provide access to sensitive data while maintaining data security in a way that supports the legal and ethical requirements to safeguard data, and therefore hopefully offers the opportunity for legislation to be interpreted in such a way that it can be used to promote rather than prevent data access.

While data sharing legislation remains very complex, the Statistics and Registration Services Act 2007 made research data access an express statutory function of the Statistics Board (incorporating ONS). This function is underpinned by legal penalties for anyone misusing personal data supplied by the Statistics Board. The ADT report also recommends the development of primary legislation to support the processing and linking of administrative data for research purposes. Such new legislation is likely to simplify pathways to data access while further formalising the penalties in case of misuse.

A brief outline will now be given of legal and ethical requirements that should be taken into account when developing strategies to maximise access to data.

Data sharing

As much of the data on Wales is collected and owned by government departments other than WG, data sharing is a key component in plans to maximise the use of data on Wales. However data sharing legislation is extremely complex and it can be difficult for KAS analysts (let alone academic researchers) to locate and understand appropriate legislation to support the acquisition of data from other government departments. The consultation showed that KAS analysts were unsure of where to go for advice when considering applying for access to data held by another government department.

The ADT recommendations include *'primary legislation to establish a generic gateway for research and statistical purposes that enables efficient access to, and linkage between, data held in different parts of the public realm'*. Such legislation would make a substantial change to the data sharing landscape in the UK, however such a change is unlikely to come about quickly, and therefore in the short term WG may want to consider recommendations made in other sections of this report for a central service to provide KAS and other WG staff with an identifiable source of support for their data sharing needs.

Data linking

Data linking is a key strategy for increasing the utility of existing data and minimising the expenditure on primary data collection. However there are legal constraints on data linking that must be taken into account, for example under guidelines from the Information Commissioner's Office (ICO) links can only be made for an expressed purpose and cannot be preserved once they have served that purpose. Therefore despite data linking often being a complex and time consuming process linked datasets cannot be made available for reuse. However, programmes that create links can be preserved and as noted by the SimWales Fellow *'once data has been used for research, a precedent is set which should mean that it is relatively straightforward to gain agreement for that data to be used again'* (*Examining the Feasibility of Creating a Wales longitudinal study,* Davies 2013).

Another challenge when linking data is respondent consent. The principles of fair processing of data state that a respondent must give consent for the purposes for which their data are to be used, therefore if the data are to be linked to other sources

it is assumed that explicit informed consent must be given. This poses problems for the linking of historical data where consent has not been secured. There are also issues surrounding consent in relation to administrative data as many of the systems for which such data are collected are mandatory, raising ethical questions about what the data can be used for. There are varied interpretations of the how consent to link must be applied, for example the ONS considers consent to use data for statistical purposes to be sufficient to allow in-house linking of ONS data (to other ONS sources only). Meanwhile, concerns have been raised in relation to some UK social surveys e.g. the Millennium Cohort Survey that consent to link should cover all aspects of linking for example, consent to link to employment, health, and financial data sources must all be granted separately and explicitly. These issues are key to the Programme to Maximise the Use of Existing Data and both the SimWales and Data Linking two of the other Data Max Fellowships are investigating aspects of data linking specifically.

This area is, again, one where ADT recommendations are likely to have a significant impact on the landscape. The report states that *'where linkage involves the addition of administrative data to information collected by survey methods, it is both practicable and desirable to request consent for linkage from data subjects, even though the linked data will be de-identified prior to analysis'* (ADT 2012), however the report does acknowledge that securing consent retrospectively for historical sources is likely to be impractical *'the view of the Information Commissioner's office is that…the prospect of obtaining consent for data linkage would be prohibitively expensive and, even if it could be achieved, the biases such a procedure would introduce could invalidate the research process'* (ADT 2012). As well as recommending primary legislation to develop a *'generic legal gateway'*, the ADT report also states that *'a common approach to the method of obtaining consent will be developed which will improve the efficiency of consent procedures and permit wider sharing of such linked data for research purposes'*[20].

[20] Recommendations made by ADT will take into account any requirements under the forthcoming EU regulation on data protection which will include guidelines for informed consent.

Data access

Legislation such as the Data Protection Act and the Statistics and Registration Services Act (SRSA) provide a legal framework for sharing some data. They also formalise legal penalties for misuse of data which, in the case of the SRSA can be a maximum of two years in prison or a fine.

However, the legislation under which data are collected can impact data access. A lot of government data collection is underpinned by legislation which lays out the purposes for which the data can be used. In many cases these purposes may be open to interpretation, but access is often granted on the basis of precedent and legislation seems rarely to be re-examined or reinterpreted. In this way and others the legislative framework for data collection is often used as a constraint on data access. However there is an increasing movement to use the law as an enabler for data access rather than a constraint (Ritchie, 2010; Laurie, 2011), and an aim of the ADRN is to encourage the legislation on data access to be examined in light of new technologies and procedures that are available to maintain data security.

Public Opinion

There is a legal and ethical requirement on government departments that collect data to safeguard respondent confidentiality and to ensure that there is public approval (if not direct consent) for any use to which public data are being put[21]. Balancing this, but never superseding it (respondent confidentiality is of primary importance) is the ethical requirement on public bodies to maximise the returns to any expenditure of public money. The best way to maximise returns to expenditure on data collection is to maximise reuse of the data. This means that primary data collection must be carried out in such a way as to ensure that the data can be reused for as many purposes as possible, not only to maximise the returns to existing investment, but also to minimise expenditure on duplicate data collections. Widening access to detailed microdata also has the potential to increase the efficiency of government spending by supporting innovative research that enables researchers

[21] *'the recommendations for public engagement are made by the Taskforce (see section 5) in lieu of the second [2. Research staff and participants must normally be informed fully about the purpose, methods and intended possible users of the research, what their participation in the research entails and what risks, if any, are involved] and fourth [4. Research participants must take part voluntarily, free from any coercion] principles stated above.'* (ADT 2012)

(both within and outside government) to undertake evidence based policy analysis, assisting public bodies in effectively targeting and evaluating social and economic interventions.

Another aspect of the ethical responsibility to maximise reuse of data is the requirement on researchers and statisticians to allow equal access to research resources collected with public funds. These resources should not be unjustifiably[22] restricted to a privileged group, and all research in the public domain should be verifiable and replicable, which is not possible without explicit methods for accessing data resources that have been used in publications.

There is a requirement under GSS and ICO guidelines that the public is informed of government plans for reuse of personal data. The requirement is not only due to the ethical need to ensure that the public understands what is being done with their personal data, but also to avoid any reputational damage to public sector bodies. However many of the issues that generate negative public opinion of data use are based on misunderstandings, therefore a key area of public engagement for WG could be in addressing these misconceptions. For example, the public often assume that researchers have access to identifiable data, this is an opinion that is supported by reports on government data breaches in recent years where bank details and other highly personal information has been left on laptops in the public domain. However, this misrepresents the risk of releasing data to researchers as individuals outside government rarely, if ever[23], have access to identifiable data and definitely not in a setting where it can be inadvertently released. Another example of how misconceptions arise was in reporting the 2011 census. Some newspapers sensationally interpreted the question on visitors staying in a household overnight to mean that the government could see who was having an affair. However, again, this assumes that census data available for analysis contain identifiable information. Census data are processed electronically[24] and, as for most government surveys,

[22] By necessity sensitive data are restricted to a privileged group as the expertise necessary to analyse and interpret complex microdata is not widespread. However, provided the researcher is a 'fit and proper person' as defined in SRSA there should be no barriers to them having access to any data for which there is an established gateway.

[23] In 13 years of supporting some of the heaviest academic users of microdata for socio-economic research in the country the author has never seen data containing direct identifiers such as name, address or NI number in a data source supplied by a public body.

[24] Only corrupted or defaced forms were manually coded for 2011.

data that are released for research have identifying information removed (additional measures are also applied to minimise the chance of an individual being identified through any combination of information available in the data and the public domain). Information that allows individuals to be identified in the Census is only released after 100 years in accordance with the law under which census data are collected[25]. In addition to the fact that researchers almost never see identifiable data, there are also strict laws governing how personal[26] data can be used. So even though identifiable data are not available, any use of microdata that has not been strictly sanctioned by the data owner can lead to a fine or a custodial sentence[27].

Therefore, it seems likely that were the public fully aware of the strict controls surrounding access to personal data collected by government, and the potential savings to be made from reuse of data, that there would be a high level of support for data access for research. This seems particularly likely when taking into account verbal reports on recent research undertaken to support the ONS Beyond 2011 project, that seems to indicate that the public generally believe (and accept) that government departments share and link their data. That this is not the case underpins the need not only to provide the public with more transparent information on data security and reuse, but also to actively improve education in this area. This could potentially be achieved by working with journalists, but possibly more effectively with school children, and it is recommended that WG discuss possibilities with the Getstats campaign that has been instigated by the Royal Statistical Society to improve statistical capacity in the UK[28]. The ADT report also includes an explicit remit to undertake public engagement, and to raise public awareness of the value of reusing microdata.

[25] Census Act 1920 http://www.legislation.gov.uk/ukpga/Geo5/10-11/41

[26] Note: even data classified as 'personal' for research do not tend to contain information that enables an individual to be directly identified i.e. name, address etc.

[27] In the case of the secure service at the UK Data Service misuse of data can also lead to the removal of research funding among other sanctions.

[28] http://www.legislation.gov.uk/ukpga/Geo5/10-11/41

5. Implementation and Funding

The Data Max Programme aims to maximise returns on investment in data by increasing access to and reuse of potentially disclosive microdata. This paper makes recommendations as to the best way to increase access microdata while preserving data security and minimising expenditure. Below are outlined some of the ways in which effective implementation of the recommendations might be expected to impact the current landscape.

In the current financial climate it may be a challenge to find funding to implement the recommendations. Any efforts to increase access to data resources will necessarily attract a cost, primarily in staff time to implement changes and develop and maintain systems. Additional investment in technology and, for safe settings, space will be required. In the medium to long term the cost of maintaining systems and providing data support also has to be considered.

The key to minimising costs is to make use of existing services and infrastructures as far as possible, as recommended below. Where suitable resources do not exist WG should consider partnering with funding councils, institutions, other governmental organisations and charities to develop systems, services and datasets of mutual benefit. As an example, the forthcoming ADRN is a partnership between government, funding councils and academia. KAS may also wish to investigate opportunities to become more involved with European programmes that support the development of statistical infrastructures, for example the DwB project where Wales is represented by the ONS.

A KAS member of staff expressed enthusiasm for investigating partnerships with the private sector as a way of maximising access to technology, expertise and data. While not explored as part of this fellowship, WG may want to consider opportunities in this area further, in particular partnering with technology companies to reduce hardware and software costs. When involving private sector organisations in systems that are concerned with potentially disclosive data it is important to take into account public opinion. Issues relating to conflict of interest, data security, intellectual property, copyright, and data ownership should also be carefully considered.

In addition to the financial challenges there will be logistical challenges in implementing recommendations arising from the Data Max programme. An appropriate physical space would have to be available to implement a safe setting. Setting up a central KAS data service will require some restructuring and reassignment of roles. This raises the question of whether analysts exist within KAS with the appropriate skills to deliver this service - can they be reassigned and are they willing to take on a change in role? If individuals with the right skills do not exist, is it practical or desirable to recruit to address the skill gaps? Another challenge is communicating changes so that they are effectively adopted. The implementation of a Data Management Policy would be an on-going commitment and would only be effective if it is 'owned' and a mechanism exists to enable all parts of WG to communicate relevant information with the owner(s).

However the potential for greater efficiency and saving for WG in the medium to long term is significant. For example, effective implementation of the recommendations might be expected to reduce the need for future spending on:

- Primary data collection - by easing preservation, discovery and reuse of existing resources; capturing microdata collected by consultants; increasing utility of data through linking and statistical matching.
- Data processing and analysis - by facilitating discovery and reuse of data resources; providing support for acquisition and use of data resources.

In terms of efficiency a centralised KAS data management service overseeing a formal Data Management Policy could contribute to:

- An increase in expertise available to KAS (and WG more widely) through the investment in staff specialising in issues relating to data use, analysis, and acquisition.
- An improvement in the quality of research by providing expert support for analysis of data on Wales.
- A reduction in the lead in time to research by assisting with negotiations for access to external resources; and by facilitating data discovery and implementing transparent access procedures for internal resources

- A reduction in the time needed to access data-related information and resources by facilitating WG's ability to locate and consult expert staff.

These last two are particularly important due to the requirement for KAS to provide evidence for policy makers. The disconnect between policy timescales and research timescales (where it can take years from formulating a research proposal to delivering the final analysis) means that there are times when engaging with academics and academic data access services are not practical. This further supports the argument for KAS to implement an in-house secure access infrastructure in order to minimise the time spent on navigating access protocols. A secure infrastructure would also increase access to and security of data.

Implementation of the report's recommendations could impact the public through:

- A reduction in the burden on survey respondents thus saving time and money for the public and business in Wales.
- Improved policy impact due to an increase in the range and quality of the evidence base.

6. Maximising availability

This section outlines strategies for maximising the availability of data on Wales. Section 7 below looks at how these data can be made accessible once they are available.

Microdata is usually collected for the purpose of producing a headline or official statistic, once that statistic has been published the value of the underlying data is often overlooked and hence it is not adequately preserved for reuse.

The most straightforward first step in maximising the availability of data on Wales is to ensure that data resources held within WG are exploited to their full potential. Feedback from the sample of individuals targeted during the stakeholder consultation made it clear that some KAS analysts had a low awareness of data available outside their immediate area of expertise and few were confident that they would know how to secure access to additional sources. Even analysts who were confident about data discovery admitted that securing access to data can be challenging.

Among the strategies for facilitating the discovery and reuse of WG data resources, a formal Data Management Policy for WG should be a priority. WG currently has policies in place that cover the safeguarding of data resources in the form of broad principles. A Data Management Policy (DMP) should contain explicit technical and procedural measures for safeguarding microdata at different levels of sensitivity for reuse, as outlined below. Such a policy should be reviewed on an annual basis to ensure that it is fit for purpose and that it is up to date in what is a very fast-moving field.

As well as a DMP (or as part of it), guidelines on contractor management will help maximise data availability by ensuring that the appropriate microdata is returned to WG by any person or organisation contracted to collect data on its behalf.

Finally, this section will briefly look at methods WG might use to increase the range of data that is available both to WG analysts and to external researchers through the reuse of administrative data, data linking and acquisition of resources from other public bodies.

Data Management

The most straightforward strategy for WG to maximise the availability of data for Wales would be the development of an effective DMP to enable the full exploitation of data held within WG. In common with other UK government departments WG holds a wealth of data in house. However these data can be difficult to discover, locate and use. As noted above, the consultation with a small sample of KAS staff found that analysts in KAS, while having some awareness of data held within their own area of expertise, had little knowledge of the range of data available within WG, and even less idea how they would find out. Some WG analysts indicated that if they did not have access to data resources they needed they would first search outside WG rather than examining in-house resources. This is an expensive and time consuming approach to data acquisition, indicating that a formal DMP would increase efficiency, and save staff time and resources.

To be effective, a DMP must be owned, implemented, monitored and updated. This would be a task for the KAS central data service recommended above. A DMP should consist of the following components:

A catalogue of all data held or collected by WG.

The catalogue should provide variable-level metadata via the adoption of an internationally approved scheme[29]. Catalogue records should also indicate where the data are stored and how they can be accessed, and should be made available via the WG intranet so that staff can easily locate data appropriate to their needs.

An audit was carried out for the ISO27001 certification of the WG Aberystwyth office. Alongside the KAS 'asset register' and other records, the audit could be built on for this purpose meaning that any catalogue would not have to be started from scratch[30].

[29] At least Dublin Core http://dublincore.org/, ideally ddi http://www.ddialliance.org/

[30] See Appendix A for documents that could contribute to a data audit

Access and storage procedures for data at different levels of sensitivity.

The level of sensitivity of each version of any dataset documented should be assigned a Business Impact Level (BIL)[31] (or other suitable measure of sensitivity), and a formal policy implemented to enable WG to assign appropriate security for each BIL.

A formal policy should assist WG in securing access to data from external suppliers as it will make it possible to demonstrate explicit, transparent procedures for transferring, storing, linking, matching and processing[32] data that are secure and compliant with the Data Protection Act (DPA), and any other relevant standards (e.g. CESG[33], ISO27001[34]). A formal access policy would also help WG staff to locate information about the access requirements for different datasets. Further, a visible, detailed, comprehensible set of procedures for releasing data for research purposes could help support wider data dissemination by increasing staff confidence that there is an approved framework under which they can release data.

Where the risk level is considered acceptable WG might consider building on current practices by increasing the number of datasets deposited with the UK Data Service at the University of Essex who would carry out the cataloguing, documentation, storage and where appropriate dissemination responsibilities for WG. The formation of a Wales ADRC will also provide WG with an infrastructure in which data can be deposited for linking. As the ADRCs are likely to be relatively time-consuming to access, even if their remit is not restricted to administrative and linked data, WG should consider depositing any data made available via SAIL with the UK Data Service as well, in order to maximise access.

---

[31] For guidelines on how Business Impact Levels are assigned see http://www.cesg.gov.uk/publications/Documents/business_impact_tables.pdf, for guidance on levels of protection for data at difference BIL levels see http://www.nationalarchives.gov.uk/documents/information-management/cross-govt-actions.pdf

[32] Also retention, deletion, updating and return of data (GSS n.d.)

[33] Communications Electronic Security Group.

[34] http://www.iso.org/iso/catalogue_detail?csnumber=42103

Policies for transferring and destroying data

KAS currently has two secure systems in place for data transfer: Access Files ON-line (AFON) and the Data Exchange Wales Initiative[35] (DEWI). A DMP that incorporates the use of these systems would provide guidance for WG staff and public bodies sharing data with WG, and help strengthen WG compliance with the Information Commissioner's Office guidelines on best practice for the processing of personal data. The Policy should also identify circumstances where AFON or DEWI would not prove practical mechanisms for transferring data and devise alternative systems[36].

Best practice guidelines on data destruction already exist within WG; these should also be brought into a formal policy.

An archiving policy

All data held by WG should be archived where legally possible. Data confidentiality alone should not be a sufficient reason not to archive data: if there is value in data now it is likely there will be value in it in the future, and data become less sensitive over time, for example even identifiable census data are released after 100 years.

KAS currently deposits some microdata for reuse with the UK Data Archive; however this is not a formal policy and only covers anonymised survey data for academic researchers. The author has not been able to identify any archiving policy in relation to sensitive or identifiable data collected by WG, or a policy for archiving administrative data. Therefore, a lot of data collected on Wales may be at risk of being lost as, while data may be stored, without a formal archiving policy it is unlikely that it is being adequately documented for discovery and reuse.

Administrative resources and data containing identifiable information will be of increasing value for Wales if data use is to be maximised through data linking. Therefore it is strongly recommended that KAS develop and implement a formal archiving policy for WG as part of a Data Management Policy.

---

[35] http://wales.gov.uk/topics/educationandskills/schoolshome/schooldata/ims/dataexchange/dewi/?lang=en

[36] For example WG already make use of PGP to transfer data above a certain BIL, as AFON is not considered appropriate for datasets above a certain level of sensitivity.

In order to avoid the cost of developing an in-house secure archive solution, WG might consider making use of the Data Centre at the UK Data Archive at the University of Essex[37] which is under development as a secure storage facility. The Data Centre expects to secure accreditation to hold BIL4 data allowing it to store (not disseminate) highly sensitive data (for more information on UK Data Archive services please see Appendix D).

Contractor Management

An issue that was raised in the consultation by both KAS analysts and external researchers was the management of external contractors hired to provide reports based on data collection.

The view from the consultation was that when contractors are hired to research issues that require microdata collection there is rarely any obligation for the underlying data to be supplied to WG as part of the deliverables. Some KAS Analysts were concerned that where a contractor delivers a report based on primary data without delivering the underlying data, as well as making it impossible for WG to investigate the issue further (or reuse the data in any other way), there is not even the opportunity to assess the quality of the underlying data or analysis.

An academic researcher who was contracted to provide a report to WG expressed frustration that despite repeated offers WG would not accept the underlying data. The academic believed that WG and other researchers could make further use of the data, and questioned the wisdom of paying for data collection with no intention to reuse it. However, as the data was 'owned' by WG it was not possible for the researcher to make the data available for reuse via the UKDS[38], so the data have never been fully exploited.

This implies that a formal policy on contractor management and the deposit of primary data resources could increase the range of data sources on Wales that are available, as well as minimising the chances of data collection being replicated.

An argument that is made by contractors for not passing data to government departments is that if respondents are aware that government will receive the data

[37] http://data-archive.ac.uk/

[38] Previously the Economic and Social Data Service

they are less likely to cooperate, thus reducing response rates and skewing samples. Contractors therefore tend to recommend that identifiable data are not passed on. However, even if response rates are negatively impacted by sharing data with government - and the author has found no concrete evidence in this area - investment in data collection should always be weighed against the utility of data now and in the future i.e. if it is not possible to secure data for reuse it should be carefully considered whether the investment in data collection for a one-time use can be justified, or if alternative sources might be available or other contractors used.

Therefore as a matter of policy, any contractor undertaking research for WG that requires data collection should be required to deposit the data with WG as part of their contract. Ideally, the full detailed microdata plus appropriate metadata should be deposited. In the rare cases where this is genuinely not possible, for example where data from a third party that cannot be shared has been used for sampling or linking additional information, then de-identified data should be deposited. However it should be taken into account that the less identifying information available the more difficult the data will be to link to other sources.

Administrative data

Administrative data offer a largely untapped resource for researchers and policy makers. There are a number of challenges in using administrative data sources, which are examined in more detail in the reports from the Data Max Fellowship on data linkage (Heaven 2013 and 2014). However, there is a growing interest in what they might offer and a growing consensus about the need to improve access to administrative data and to develop new methodologies to underpin their use. The ADT report states that *'Improving access to and linkage between administrative datasets for research and statistical purposes would have demonstrable effects on economic growth and would help us respond more effectively to challenges related to the health and well-being of people'*.

The UK has a very long history of using survey microdata for research purposes, and therefore methods and expertise in the processing of survey data are well advanced. However, the use of administrative data for socio-economic research is still relatively new and there are lessons that can be learnt from the health informatics field where techniques are more advanced.

A key challenge when reusing administrative data for research is that, unlike survey data, they are not collected with statistical analysis in mind. This means that the content and structure of administrative databases can reduce their utility. An example of a structural challenge is that many administrative databases are either constantly or intermittently updated, which raises serious issues about versioning: when should snapshots be taken, how often, who by? Can official snapshots be constructed and made available via a data service? Can users be allowed to tailor their own snapshots?

A further challenge is that much of the historical administrative data available is not in electronic format and the cost of digitisation can be high, particularly for sensitive data where it may not be appropriate to allow some of the more cost effective private sector companies to undertake the work.

In terms of content, administrative databases tend not to include detailed socio-economic data. Variables detailing education, labour market status, household structure, geography etc. are crucial to answering many policy-based research questions, but are rarely included in administrative datasets.

Some administrative data, in particular health records, are likely to include free text which can contain confidential information that is time consuming and complicated to identify and remove. Therefore organisations such as SAIL habitually exclude free text during the linking process for simplicity, leading to data loss. Methods for the anonymisation of qualitative data are less advanced than for quantitative data so there is scope for further research in this area.

A way of addressing the problem of versioning is to implement an infrastructure to capture, clean, document, store and disseminate administrative data for research purposes. The ADT report on Improving Access for Research and Policy (ADT 2012) made recommendations in this area and £30m of funding is being made available to develop a UK-wide network to support the use of administrative data, including the creation of an Administrative Data Research Centre (ADRC) in each country (England, Scotland, Wales, Northern Ireland), therefore the issue of versioning will be tackled by the Network.

The utility of administrative data can be greatly enhanced by the addition of socio-economic variables through data linking, for example WG have a particular interest

in being able to link DWP/HMRC data on benefit eligibility and employment in order to address the devolved responsibility for 'Tackling Poverty'[39]. In addition, data from small surveys can likewise be enhanced using techniques to impute or simulate data for a wider population using administrative data sources. These techniques are of particular interest to WG, and have been studied in more detail as part of the SIMWales Fellowship (Davies, 2013)

Increased use of administrative data for research purposes may also enable key stakeholders to influence primary data collection to include more socio-economic variables; however associated cost and privacy issues would have to be taken into account.

As noted in Section 4 above, there are issues surrounding fair processing of administrative data and consent for reuse and linking. These issues apply to all data collected by government but there are particular ethical considerations when it comes to administrative data as, unlike survey data where respondents can generally decline to take part, much administrative data is collected as part of a legal requirement, for example data required to access benefits or to drive a car. Therefore in situations where respondents are compelled to provide their personal information consent for reuse is even more important. The ADT is planning to address this issue and, as noted above, one of the areas which the proposed ADRN would tackle is:

'A common approach to the method of obtaining consent will be developed which will improve the efficiency of consent procedures and permit wider sharing of such linked data for research purposes' (ADT 2012).

A common approach to consent is only a small part of the ADT recommendations. As noted above, the recommendations will be supported by £30m funding over the next 5 years, which will lead to increased availability of administrative data and a greater understanding of techniques and methods to analyse, link, manage and protect this type of data. WG are closely involved in the planning process for the recommended Administrative Data Research Network, and will no doubt continue to

[39] http://wales.gov.uk/about/civilservice/directorates/lgc/csj/?lang=en

monitor the development of the system to ensure that Wales' research needs are met.

Data Linking

The WG Programme to Maximise the Use of existing data defines data linking as '*a technique for creating links between multiple anonymised data sources so that information that is thought to relate to the same person, family, place or event can be connected for research purposes.*' This technique has increasingly become a focus for government departments and research funding councils as a strategy for reducing expenditure on primary data collection and the burden on respondents. Without further data collection or acquisition, data linking offers the possibility of increasing the data available for analysis, particularly through the linking of administrative sources to survey data or of administrative sources to each other.

Data linking is key to WG's Programme to Maximise the Use of Existing Data and is a particular focus of the SIMWales Fellowship, which is designed to examine the feasibility of developing a Longitudinal Study for Wales (Davies 2013). A second Data Max fellowship is designed to demonstrate the added value that can be derived from linked data (Heaven 2013 and 2014).

The SIMWales Fellowship investigated the England & Wales, Scottish and Northern Irish Longitudinal Studies (LS, SLS and NILS) as a model for a proposed 100% Longitudinal Study for Wales. The SLS and NILS were created, primarily to support research in health, by using data linking techniques (Davies 2013).

Other recent data linking, that may provide useful lessons include a major collaborative project between HMRC and the MoJ (MoJ 2011); the Millennium Cohort Study link to health records[40]; and an Institute of Education, Centre for Longitudinal Studies project to investigate the impact of consent and coverage on data linking[41]. The largest and possibly the most innovative data linking project currently being undertaken in the UK is Beyond 2011[42]. The Beyond 2011 project is

[40] http://www.cls.ioe.ac.uk/page.aspx?&sitesectionid=354&sitesectiontitle=Using+health+record+linkage+in+the+UK+Millennium+Cohort+Study+to+investigate+childhood+obesity%2c+asthma+and+infections

[41] http://www.cls.ioe.ac.uk/page.aspx?&sitesectionid=1218&sitesectiontitle=Linking+cohort+study+data+to+administrative+records%3a+The+challenges+of+consent+and+coverage

[42] http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes---projects/beyond-2011/index.html

led by ONS and includes investigation of the potential for using linked administrative data as a replacement for a traditional census.

The majority of expertise in data linking in the UK, particularly when administrative data are included, is in the field of health research. WG has access to particular expertise in the linking of sensitive data for health research purposes. The WG-funded SAIL Databank at Swansea University is a leading centre for health data linking and has recently been awarded the status of a Centre for Excellence in eHealth Research by the Medical Research Council[43]. Scotland has a similar resource, the ScottisH Informatics Programme (SHIP). Both SAIL and SHIP are supported by NHS data linkage services, who act as trusted third parties able to carry out secure data linking[44] as outlined below.

A traditional barrier to data linking has been the need to use identifying data to maximise the chances of creating exact matches between records in different datasets, however there are now methods for combining information from datasets that ensure a high level of respondent confidentiality.

A common procedure for safeguarding respondent privacy during data linking involves a split file process. This separates the identifying information i.e. name, address, date of birth, NI number etc. from all analytical data, whether medical, social, financial, attitudinal etc. in each source to be linked. For each source, this creates two files, the first containing an index plus the identifiable information and the second containing an index plus the analytical data. For each dataset the identifying information is sent to a trusted third party who creates an anonymous linking field[45]. Once the linking field has been created the identifying information is destroyed leaving only the linking field and the indexes. The indexes allow the anonymous linking field to be reattached to the analytical data from each dataset creating a linked dataset without using any identifying variables. This is the procedure used by SAIL, among others, to create linked datasets.

---

[43] http://www.swansea.ac.uk/medicine/cipher/

[44] For more information on the data linking for the SAIL Databank see Ford et al, 2009

[45] The third party, while being able to see the information that identifies an individual, does not get access to any other information therefore they are not able to discover any additional information about the individuals in the dataset beyond their name, address etc.

A disadvantage of this method is the separation of the people making the link and the analysts. While the separation is necessary for data security, it means that the analyst or researcher has limited information about the quality of the link and is relying on the expertise of the trusted third party. Therefore the relationship with the trusted third party is crucial.

During the consultation a concern was raised by a KAS analyst about WG's possible over reliance on NWIS (NHS Wales Informatics Service) for data linking. It was pointed out that while WG and SAIL rely heavily on NWIS to carry out data linking there were no agreements or Memorandums of Understanding to formalise these arrangements. The development of a Wales ADRC will require the formalisation of arrangements with any trusted third parties carrying out data linking for Wales.

An alternative to 'direct linking' is statistical matching. *'The value of statistical matching is the ability to conduct analysis of new combinations of data items in situations where direct linkage is not feasible'* (Davies, 2013). A more detailed examination of statistical techniques for data linking is available in the SIMWales Fellow's report (Davies, 2013), who also recommends that *'more attention should be given to the development of a set of harmonised core questions that would be required to appear in all Welsh Government surveys so that the feasibility of pooling data across different sources is enhanced'*, and that where possible *'the utilisation of coding frames that comply with national standards should mean that the level of details used in one data set can be 'collapsed' to be consistent with that used in the other data set'* (Davies, 2013)

As noted above there is scope to develop further techniques for linking and analysing administrative data, and this is an area where the ADRN is likely to contribute in particular, as the ADT report recommended '*exploring issues relating to data linkage methods, the quality of linked data, its coverage of specific populations and its suitability for particular research purposes.'* (ADT, 2012).

Data Acquisition

An obvious method of increasing the availability of data is for WG to acquire data from outside sources. Much of the data on Wales are collected and held by other public bodies, so there is a clear incentive for WG to have a formal data acquisition policy. However, as mentioned above, the legislation governing data sharing

between public bodies is complex and challenging. This is another area where a central KAS data service could facilitate data acquisition by providing WG staff with a central store of knowledge and contacts in data sharing legislation.

In order to understand the legislative and procedural conditions under which data from external providers are held and processed KAS might consider carrying out a full audit of data WG receives and the conditions under which the shares exist. This would enable WG to identify areas where there is a potential risk of losing data due to a lack of clarity over the basis of the share. It will also give an overview of all legislation being used to share data currently, and as such may highlight legislation that can be used to negotiate data shares in the future.

To gain access to key data sources KAS may also wish to explore further the possibility of securing access directly from local authorities in Wales. As some of the data held by government departments originate with local authorities (LAs) it may be more straightforward for WG to secure data directly from LAs. Other potential sources of data include the private sector and the third sector. Data from the private sector in particular offers a large untapped resource for researchers. At present none of the UK RDCs covered in this paper make private sector data available, and the ADRN will not initially be accepting data or access requests from the private sector however *'the Governing Board will, at an early stage, investigate guidelines for access and linkage by private sector interests, as well as commissioning public engagement work on this topic'*. The forthcoming Business DataSafe aims to *'join up organisational data such as store cards, client lists held by utility companies, banking transactions, mortgage details, and records or communications held by communication providers with social scientific survey data in a safe and secure setting'*[46]. The ESRC are keen to engage with the business community so there may be scope for WG to engage with firms operating in Wales to encourage their involvement with the project.

Summary

The key recommendation for maximising the availability of potentially disclosive data for research purposes is for KAS to design, implement and provide on-going

[46] http://www.esrc.ac.uk/news-and-events/news/25683/big-data-investment-capital-funding.aspx?dm_i=XZA,1FWYX,4OC3F8,4W62O,1

monitoring for a formal WG Data Management Policy. A comprehensive policy for documenting, classifying, storing and preserving microdata will provide a foundation to support effective implementation of strategies to maximise access to potentially disclosive data. Without a Data Management Policy it is unlikely that WG will be able to fully exploit in-house data holdings in the short term let alone the long term.

Further, WG should ensure that all contracts with external organisations that undertake data collection for WG require that microdata be deposited with WG to enable quality control, reuse and where possible linking.

KAS should consider undertaking a full audit of the conditions under which data from external providers are shared and processed in order to gain a better understanding of the legislation available to WG for data sharing, as well as the status of shared data resources.

KAS may consider developing a set of core questions to be included in all surveys for Wales in order to increase their potential for statistical matching as recommended by Davies, 2013.

To maximise the possibility of forthcoming data infrastructures meeting WG needs, KAS should continue to work closely with the funding councils and stakeholders in Wales on projects such as the ADRN and Business DataSafe.

7. Maximising Access

In order to maximise access to potentially disclosive microdata a range of interrelated solutions are required to meet the needs of different users, and to address the location (and type) of data being accessed. Existing access conditions must also be taken into account e.g. are data available via an existing RDC such as SAIL or the UKDS; are the data held within WG – if so, do procedures exist for internal/external dissemination; are data known to exist, but for some reason cannot be accessed? For the purposes of this section it will be assumed that data are available, and that access conditions can and have been fulfilled. Solutions based both in government and in an academic environment will offer the opportunity to maximise the availability of the widest range of data for the widest range of stakeholders, as some data providers are more comfortable with data being held by academics and some with data being held by government. A range of access mechanisms will also maximise availability as researchers have differing needs and skill levels. Wherever possible access strategies should be based on existing solutions in order to minimise expenditure, leverage existing user support, promote national standards and streamline access.

A key component in efficient data access is the ability to locate the data therefore this section will first look at resource discovery before exploring strategies for maximising the returns to secondary analysis of potentially disclosive data. How the proposed strategies are likely to meet the needs of the various stakeholder groups will then be examined before concluding with recommendations. Where remote solutions are recommended the focus is on remote access rather than remote execution due to the increased utility and lower cost (see p11, above), however WG may choose to substitute a remote execution system at any point where remote access is recommended.

Data Discovery

Resource discovery is an important foundation for data access. However, in order to make data easy to locate there is a cost, primarily in providing adequate metadata in an internationally recognised format, but also in making the metadata accessible. Therefore to make WG data fully discoverable it will have to be documented. If a

formal Data Management Policy is implemented as recommended above most, if not all, of the documentation requirements for data discovery would be addressed as part of the Policy.

To make the metadata accessible KAS could make use of existing data catalogues such as UK Data Service Discover[47] thus avoiding the cost of developing a new resource. Other sources of information on data for Wales, such as the SAIL data holdings[48], and WISERD's DataPortal (Fry et al, 2012) could also be encouraged to link to Discover or a portal developed as part of a Wales ADRC.

This, again, is an area where ADT recommendations will have an impact. The development of a central gateway linking information for all the proposed ADRCs (now referred to as the Administrative Data Service) will assist greatly in the discovery of administrative data at no cost to WG.

Data Access Solutions

A primary motivation for increasing access to potentially disclosive data for Wales is to maximise the benefits of existing sources of survey and administrative data. There are a number of stakeholder groups who have a need for analyses of potentially disclosive data, however not all of these have the need (or desire) for direct access to the data. Therefore, this section will examine both solutions that allow researchers to access microdata and those that allow access to outputs derived from the microdata. The access solutions outlined below can be roughly divided into three categories: bespoke analyses carried out on behalf of researchers; a database of de-identified, cleaned data with front end query tool; and systems that provide secure access to detailed, potentially disclosive microdata.

Bespoke Analysis

A service to provide bespoke analyses will allow researchers who do not have the need, desire, expertise or infrastructure to analyse microdata to benefit from the information that is only available through the reuse of potentially disclosive microdata. How these analyses are provided may depend on where the data are located, and who undertakes the analysis as outlined below.

[47] http://discover.ukdataservice.ac.uk/

[48] http://www.swansea.ac.uk/media/SAIL%20Datasets%20.pdf

As a matter of principle any bespoke analyses created from sensitive data should be documented (including the reason for the analysis, any programmes used to create it and a link to any publication in which it appears). These outputs should also be assessed for their likely future value and if appropriate archived.

- Option 1. Experts within KAS

For data held within WG, KAS analysts provide bespoke analyses to support WG policy needs more widely. These analysts could also potentially work with data held within external systems that can be accessed remotely, such as the VML, UKDS, or an ADRC. Ideally these experts should form part of a central specialist data service within KAS, to enable them to be easily located by WG staff and allow effective knowledge sharing between specialists in related areas e.g. data confidentiality, legislation, analysis in different subjects etc.

Advantages of this set up could include a reduction in the number of people needing access to sensitive data. A group of specialists in particular crosscutting analytical domains e.g. data access or data linking, would enable KAS to ensure that these analysts could be trained to a high level, and could keep up to date with advances in their specialist area, which they can then communicate to analysts working in teams across KAS. These analysts will be able to develop relationships with data providers, increasing access to information and potentially data; and with data users, increasing understanding of data needs within WG (and potentially more widely). Such centralised expertise should assist in the preservation of knowledge and the location of resources, and as such should streamline operations and reduce costs.

A disadvantage to this system includes the need to recruit and maintain specialist data analysts with a high level of expertise. Also, analysts with expertise in a wide range of subject areas may be required to meet WG needs.

If WG were to implement an in-house bespoke analysis system it is likely to be of interest to other key stakeholders in particular the third sector, but also academics, journalists and the general public. However, it is likely that there would be significant issues with meeting this interest due to capacity within WG.

- Option 2. Experts outside WG

To secure access to external expertise WG regularly commissions bespoke research by academics, and contributes a substantial amount of funding to academic

departments in Wales. A system whereby academic researchers based in key departments are funded by the Welsh Government on a formal basis could improve access to high level expertise in particular subject areas. These staff should be funded on a part-time basis to enable them to maintain their wider academic interests, which should be closely related to any analyses WG are likely to require from them. Ideally these academics would be attached to a KAS central data service who, as well as providing liaison with WG staff, would also assist, monitor and archive their work.

A formal relationship with a network of academic experts would enable WG to consult specialists at the cutting edge of their fields. It would also provide the opportunity of gaining access to a wide range of experts with a depth of skills likely to be unavailable elsewhere, for a relatively small amount of funding.

Disadvantages would include the need to oversee external researchers and ensure their involvement with WG priorities and work schedules. Options for the retention of external experts are outlined below:

i. Call off-contract or framework contract

Framework contracts with academics specialising in key areas of interest to WG would provide a team of experts that could be called on at short notice without further procurement processes. This arrangement has the advantage that WG does not need to make any long term financial commitment and could pay for work as and when it was required. However, it does have the disadvantage that the academics may not respond to calls when required, particularly if the call does not relate strongly to their research interests.

ii. WG funded experts based within an ADRC

Employing academic experts based in the ADRC for Wales would enable WG to have permanent access to specialists at short notice. These posts could be either full time or part time, however unlike a framework contract this strategy would require a medium to long term financial commitment. In addition WG would have to ensure that there was adequate work to justify these posts and there would be an additional cost in on-going supervision. As staff employed by WG these posts may not have academic independence. However the experts could benefit from being based in an academic environment, particularly if encouraged to take part in academic activities

such as seminars and conferences. A setup such as this may also reduce the need for a safe setting within WG. However, to maximise the returns from investment in these staff there would have to be some mechanism for allowing them to access sensitive data that cannot leave the Welsh Government.

iii. RCUK funded experts based with an ADRC

As government analysts are not resourced to develop high level data processing and/or analysis skills for one off projects, it can be difficult for government to access staff with requisite skill levels. During the consultation the suggestion was made that as part of the funding for the ADRN, the Research Councils might fund experts based in the ADRCs to provide bespoke analyses for government. While it was explicitly stated during the bidding process for the ADRN that no funding would be made available to deliver work on behalf of government that departments could or should be doing themselves, it is likely that much of the research being undertaken in the Wales ADRC will be of use to WG. The focus of the ADRN is on research for policy and many of the data owners are government departments who are restricted by legislation in providing access to data only for projects in the 'public good'.

iv. PhD students

PhD students whose thesis topic relates to WG's areas of research interest could be employed on a part-time fixed-term basis e.g. as part of an internship scheme, to produce specific research outputs for WG. This is potentially a low cost strategy for accessing research expertise, and it offers the potential of discovering young, highly dedicated researchers. However, unless WG is very careful (and lucky) in candidate selection the cost of supervision and support is likely to be far higher than for established researchers.

v. PhD scholarships

Where WG research priorities could form the basis of a PhD thesis, WG may consider creating PhD scholarships. The existing relationship WG has with the Wales Doctoral Training Centre[49] would provide a forum to discuss possible research topics. The scholarships could be joint-funded with institution or Research Councils. In the case of PhD students it is likely to take 3-5 years to see results so

[49] http://www.esrc.ac.uk/funding-and-guidance/postgraduates/dtc/index.aspx

scholarships should only be granted for topics that do not require any short term impact.

Database with front end

Another possible solution for researchers who do not have the skills and/or are reluctant to analyse sensitive data is to develop a database with a front end query tool that supports particular statistical methods. There are broadly two approaches to this depending on the level of sensitivity of the data in the database. The first option would be to develop software that can carry out automatic checking of output to prevent disclosive information from being released (secure portal). The second is to develop a database from which no disclosive information can be released (secure database).

The main advantage of such a system would be that all KAS analysts would be able to produce their own statistics without needing a high level of expertise or access to sensitive data. Such a system could potentially be made available more widely within WG and would be likely to be of particular interest to other less specialist stakeholder groups such as the general public, the third sector (depending on the variables available), and journalists.

Irrespective of the delivery system used, the principle investment in such a service would be in preparing the underlying database so as to minimise the possibility of disclosure and maximise the likelihood of quality outputs. The expertise required to understand individual datasets, let alone linked datasets is significant. Survey microdata tend to be complicated and of variable quality, administrative microdata even more so. Expert analysts invest a substantial amount of time and effort into cleaning and preparing a data source before producing statistical outputs. For the purposes of automatic table generation additional data preparation would have to be undertaken over and above the usual measures in order to minimise the possibility of sensitive data being released.

It is unlikely that this type of access system would be suitable for the processing of linked data due to limits on the length of time for which links can be preserved. Techniques for linking data automatically on an ‘as needs’ basis may be developed in the future, but they are not well advanced at present. Add the need to secure approval for each project that analyses linked data and these issues mean that a

secure portal for accessing a database of linked data is likely to be impractical. However, were legislation to change or linking methods improve this method may be the best strategy for making the widest possible use of linked data, as users can benefit from the data without ever having to have direct access, and once set up the costs of supporting it would be lower than the other systems recommended (except a full KAS secure infrastructure as outlined below).

Other disadvantages mainly relate to cost and security. Any automatic system where potentially disclosive outputs are not checked before release increases the chances of sensitive data being published. Therefore the data preparation and software programming requirements are likely to be considerable. Even if existing software are used there is the cost of licensing and possibly the cost of customising software to WG needs. However, as mentioned above, the main cost will be that of developing the database: data cleaning, data linking, anonymisation and any other measures necessary to minimise the chances of sensitive data being released by the system. This will be costly and as such is only likely to be feasible as part of an existing project to develop WG resources such as SIMWales (Davies 2013).

Another weakness is a limitation on the statistical analyses that can be supported. Methods for maintaining security for tabular outputs are well researched, but the potential for disclosure from other more complex analysis techniques are less well understood. Therefore if methods other than tabulation were to be supported the complexity and risk would increase significantly. Also from an analysis point of view the system would represent a ‘black box’ with no ability to check results, or adjust methodological decisions made in the construction of variables in the database. Therefore a high level of trust is needed both in terms of the quality of the output but also the confidentiality of the output.

An advantage of a database with a front end would be that once it has been created it would allow researchers to query potentially disclosive data with minimum additional support and therefore could potentially be a good option for providing access for external stakeholders at minimum on-going cost. Options will be explored in more detail below, however the cost and complexity of making such a system secure and the likely level of utility offered mean that it is unlikely to be a cost effective solution.

- Option 3. A secure portal for tabular analysis

This solution would provide secure tabulations from sensitive data for a specified set of variables, with the query tool as part of the security system. The query tool would automatically assess the data requested for potential for disclosure and would only display it if it was judged to be safe. The StatsWales service could be considered as a basis for such a system, though the need to implement automatic disclosure control and restrict access to approved users will probably mean that software designed specifically for secure tabulation could be adapted more easily and cheaply to KAS needs. For example the LISSY web tabulator, used by the LIS Cross-National Data Centre[50], allows the creation of tabular outputs while providing a flexible security system. The security system can block the tabulation of variables that when analysed in combination may increase the risk of disclosure (e.g. firm size and industry for countries with companies which dominate an industry), it also enables a threshold for cell sizes to be set, which prevents the display of any cell that contains fewer observations than the threshold.

- Option 4. An anonymised database

Another possibility that was raised during the stakeholder consultation was the development of a cleaned database that could be directly accessible without a secure front end. In fact, a request that was made by two KAS staff was *'a front end on SAIL'*. In order to create a non-disclosive database with no additional controls other than the structure of the data itself it is highly unlikely that the level of detail that could be allowed would offer more than the data available via StatsWales. Therefore, any expenditure on creating an anonymised database of this kind would almost certainly be better invested in increasing the range of data available through StatsWales.

An option for making data more widely available would be to use a commercial provider. For example a respondent to the consultation pointed out that *'NHS Wales excludes data from the public domain that is available in the remainder of the UK. For example, Wales does not make data available on the Dr Fosters website'.* This method would not be appropriate for potentially disclosive data, so any data

[50] http://www.lisdatacenter.org/

deposited with commercial providers would have to be de-identified and highly anonymised.

Secure access to microdata

The options above offer strategies for maximising the benefits of reusing potentially disclosive microdata that remove the requirement for researchers to access the data directly. The benefits offered by research access to microdata can only be fully realised if a method for enabling 'fit and proper persons' to analyse the data for their own research exists, therefore this section will examine strategies for providing secure access to potentially disclosive microdata.

Any plans to maximise access to potentially disclosive microdata should take into account the existing services. However, if WG have data that cannot be deposited elsewhere that they wish to make available for reuse, either to researchers accessing from a safe setting or via an in-house service, it is likely that full secure infrastructure would need to be developed and overseen.

- Option 5. Existing UK RDCs

One of the most straightforward and least expensive ways to improve access to microdata, for academic researchers in particular, is to use one of the established secure remote access systems, such as VML, the UK Data Service or SAIL. WG could increase access to their own data holdings by depositing them in one (or more) of the services outlined below.

The UK Data Service is the most widely used of the existing RDCs and WG could build on their existing relationship to increase the range of data on Wales deposited with the Service. For example, WG can deposit data at different levels of confidentiality[51]. For potentially disclosive data deposited in the secure access section of the UK Data Service WG has the option of controlling access, as all applications to access WG data can be passed to WG for approval as is the case for other providers who have deposited sensitive data e.g. ONS, DfE, DWP, ESRC. This secure access system has the added advantages that WG can not only restrict access to projects for which there is a legal justification under data sharing legislation, but could also gain more detailed knowledge of how researchers wish to

[51] For an outline of the different data licenses available at the UK Data Service see Appendix D

use WG data and a record of research outputs relating to Wales. There would be no charge for WG to deposit data in the UK Data Service for access by academic researchers; however there would be a cost associated with the preparation of data for reuse. While the UK Data Service would produce the majority of documentation and value added materials associated with a dataset, WG would still have the responsibility of providing a dataset with adequate metadata for it to be reusable. It should be noted, though, that the majority of data preparation needed for deposit at the UK Data Service would be covered by the recommendations above for a formal Data Management Policy. If that policy were implemented the extra resource requirement to prepare data for release would be low. The main cost that cannot be absorbed by other projects would be that of answering data-related queries. While the UKDS would act as first point of contact for all data queries it is inevitable that there will be queries that their staff cannot answer and that will therefore have to be referred back to the data providers. However, any efforts to increase access to data resources will necessarily include an increase in data support requirements. Providing access to data via established data services will reduce the support burden to the minimum level possible.

The UK Data Service is accredited to hold data up to BIL4 (in aggregate), and to remotely disseminate data at BIL3. Improvements in data storage may increase the BIL level the UK Data Service can hold, and the development of a UK standard for Safe Settings could raise the level of data that can be disseminated to institutions with approved safe settings. These new initiatives are likely to come into place in 2014.

Another option is SAIL, which has undergone independent security audits[52], but is not currently formally certified under ISO27001 or to hold data at a particular Business Impact Level. However, if granted funding for an ADRC, SAIL would be accredited as part of the wider network.

Where possible, data should be sent to SAIL in order to secure the gains available from data linking. For this solution to add maximum value SAIL would need to

[52] *'RSM Bentley-Jennison…qualified to provide Information Systems Assurance undertook the work'* (Ford et al, 2009)

provide the wider research community with better information about its holdings, and support services and pathways to access for other researchers.

A further option is the ONS VML. The obvious restriction would be that only data that is held at ONS can be accessed via a VML link (and not all ONS holdings are available through VML). However, as ONS is the National Statistical Institute for the UK and holds a large amount of data on Wales, it is recommended that WG investigate the legal and logistical possibility of having a secured area within VML that could be dedicated to data on Wales.

To install a VML terminal, there must be a connection to the Government Secure Intranet. In addition the PC must be positioned so that the screen cannot be overlooked. There is currently no requirement for a VML terminal to be housed in a safe room or other secure setting, however with the advent of the Administrative Data Research Centres and other opportunities to access increasingly sensitive data, KAS might consider installing a safe setting on site as will be discussed in more detail below. To access data held in the VML researchers must submit applications to the ONS Microdata Release Panel for approval. As access has to be from within a government department this option is less practical for stakeholders outside WG as any solution that requires a researcher to travel in order to access data will reduce equality of access to data as noted in Section 3 above. On-site access can also be very costly considering the time it takes a researcher to understand and prepare a dataset before they can produce quality outputs. Where a researcher does not have on-site access locally the cost of accommodation can make using that data source impractical even for highly funded senior academics, let alone PhD students and 3rd sector or early career researchers.

VML does not currently possess a formal government accreditation certifying the BIL of data it can hold, though data are accessible at BIL3. VML is currently going through the ONS accreditation process which is expected to be completed by early 2014.

Finally, the proposed ADRN will significantly increase access to administrative data for research. KAS should continue to work closely with the ADRN and the ADRC for Wales in particular to ensure that any developments meet the needs of researchers

investigating Wales, and that the system is developed in such a way as to enable as much WG administrative data to be deposited as possible.

- Option 6. KAS secure infrastructure

Four possible ways of providing on-site access, in ascending order of security (and cost) are: a PC on the WG network with access to data on a server; a standalone PC onto which data are loaded for a specific research project, with output being manually checked before removal from the PC; access via a dedicated terminal to a KAS secure infrastructure; access to a KAS secure infrastructure from within an on-site safe room.

To implement a fully operational in-house secure data access solution based on international best practice, KAS would have to develop an infrastructure comprising a strategy for secure storage and analysis of WG data resources, and for the release of outputs. The system would need to include a technical solution for storing, accessing and analysing data to a level that is secure enough to allow for data linking; procedural measures to ensure ethical practice and data security; an team to oversee and control access and to vet outputs from the system to ensure that no disclosive data is ever released.

Such a system would not only require a high level of initial investment, but also on-going supervision, maintenance and development, although the secure data storage solution recommended as part of the Data Management Policy above would form the basis of a secure infrastructure. KAS might also have to develop and manage an appropriate training programme for staff (and other individuals if KAS extends the service to external researchers) wishing to use the system, though it is likely that the UK training and accreditation programme that the ADRN is required to setup would meet KAS needs without further investment.

A KAS secure infrastructure has the advantage of providing an environment in which WG could undertake major data linking projects in house thus maximising the potential of their data holdings, particularly as WG have data resources that could be linked and exploited in house that cannot legally be shared with others.

As discussed above there are limitations to on-site access therefore to fully maximise the use of potentially disclosive microdata KAS may consider

implementing a remote access component to any secure infrastructure that is developed.

If WG wish to maximise access to own data by making it available via an academic safe setting it is likely that full secure infrastructure would need to be developed and overseen in-house, as it would not be possible to use the ADRC to process microdata that cannot legally leave government departments.

- Option 7. In-house safe setting

KAS could consider establishing a simpler and cheaper secure data access infrastructure based around an internal safe setting[53] (safe room), with access to a secure server (setup as part of a formal Data Management Policy recommended above). With this infrastructure WG analysts could process any data that can be held in house, as well as ADRC resources if necessary. An in-house safe setting may also assist with access to data from other government departments, in particular direct access to government RDCs e.g. HMRC or MoJ Datalabs. While for some internal data holdings KAS may choose to provide staff to undertake statistical disclosure control of outputs to ensure that no sensitive data is released, for data held by external RDCs disclosure control would be carried out by the staff of the RDC.

A safe setting is a controlled environment in which sensitive data can be processed. These occasionally rely on standalone machines, however as that would not be a practical solution for accessing external RDCs, for the purposes of this paper safe settings that offer additional physical and procedural controls for remote access will be considered. The basic principles of a safe setting include: a physical space with access controls linked to a secured network; staff to oversee access and security; no printing, recording or copying from any machine (or peripherals) located in a safe setting; users must give up phones, laptops, cameras, and any other electronic storage devices before entering the setting (in some cases paper and pens/pencils are also banned). A project to determine the national standards for safe settings is being funded as part of the ADRN and is being led by Dr. Chris Dibben of the University of St. Andrews[54].

[53] Set up according to national standards established by the ADRN

[54] The project is expected to report in early 2014

At least one safe setting will be funded in Wales as part of the ADRN, but will almost certainly be based in an academic institution. While it is likely that WG will be able to make use of any safe setting in Wales, it is important to note that, at least initially, ADRN safe settings will prioritise access to data held in ADRCs (and possibly government Datalabs such as those at HMRC and MoJ). Firstly, this means that the safe setting would be designed to access highly sensitive linked data, potentially leaving WG with no infrastructure to access sensitive data that is not linked. Secondly, WG staff would be subject to the same access procedures as anyone else, which are likely to be very time consuming at least in the first years of the ADRN.

WG may choose to maximise access to sensitive data on Wales by allowing external researchers to access an in-house safe setting, however this is likely to increase staff and other resource costs due to the need to accommodate and oversee external users.

Stakeholder Access

This section will examine the suitability of the access solutions outlined above for different stakeholder groups that provided feedback to the consultation.

Welsh Government Staff

Most KAS analysts interviewed reported having no need to access potentially disclosive data; those that did fell into one of two groups when questioned. One group said that while they needed statistics derived from sensitive data they had no interest in accessing the data themselves. Reasons given for this included a lack of expertise, but more often unease at the idea of handling sensitive data. For example, one respondent was interested in analysing health outcomes for people with varying levels of educational attainment; however they admitted that they would be *'nervous about holding and accessing data at that level'.* The second group wanted and needed access to sensitive data and in many cases had it, but few were entirely satisfied with the range of data available to them. In many cases the data required are not owned by WG, so the issues are not only of access but also availability and therefore it is not only infrastructure that is lacking. However an improvement in the secure infrastructure available to KAS analysts for data access and analysis can only

contribute positively to negotiations with data owners outside WG for data resources on Wales.

While StatsWales[55] received good feedback during the consultation it was clear that the service did not cover all the needs of KAS analysts. Therefore for KAS analysts who do not wish to process potentially disclosive microdata themselves but nevertheless cannot currently access the information they need through publically or internally available aggregate statistics, a bespoke analysis service (options 1 and/or 2 above) would be the ideal solution. These solutions would offer flexibility in the type, topic and detail of analyses they could access while removing the need for them to see and process potentially disclosive data.

For KAS analysts requiring access to sensitive data, a Data Management Policy including a secure access strategy supported by some level of internal secure infrastructure (as described in 6 and 7) would increase the range of data they can analyse. A shorter term, lower cost solution would be to make use of existing RDCs as outlined above. For example, SAIL allows virtual remote access to their database therefore an access terminal could potentially be installed at WG. It should be noted that any WG research projects that make use of data not wholly owned by WG may have to undertake the same application and licensing process as any other user. Also, as mentioned above there are restrictions on what the existing Databank can be used for which may limit the uses to which WG can put the data. However, SAIL holds a rich and valuable collection of Wales-specific data that can be linked for research purposes, and that are highly unlikely to become available elsewhere. Therefore access to the SAIL databank would be a key component of any strategy to maximise WG's evidence based policy research.

The establishment of an ONS VML terminal within KAS would not only give analysts access to a range of detailed microdata that they have not previously had, but it could also potentially provide WG with a suitable infrastructure for accessing and analysing sensitive data at very little cost. KAS reached an agreement with ONS for a remote access VML terminal over 4 years ago; however the system was never implemented. As none of the staff currently working within VML were in post at the

[55] https://statswales.wales.gov.uk

time of the negotiations with KAS it is likely that any application for a VML terminal would have to be restarted.

Finally, while the Secure Data Service was not accessible to researchers working outside academic institutions for reasons of security and cost[56], since it has been re-funded as part of the UK Data Service there is an explicit remit to explore the possibility of access beyond academia. Therefore there is an opportunity for KAS to negotiate use of the system for their analysts. As well as gaining access to the existing collection KAS could also investigate whether a legal basis exists to transfer data to the UK Data Service, and whether the Service would be willing to restrict access to some data to KAS analysts only, thereby providing a secure access system for data on Wales that does not have to be managed by WG. It is likely that there would be a cost associated with such use, but it would almost certainly be lower than developing and maintaining a bespoke in-house solution.

Academics

The cheapest and most straightforward strategy for increasing access to data for academics is to deposit data on Wales in one or more of the existing RDCs with which academics are familiar (see option 5 above). From a researcher's point of view a familiar access system reduces the lead-in time necessary to produce meaningful analysis, thus increasing research output. The level of sensitivity and therefore detail of data that can be deposited in each RDC will depend on the service's infrastructure and the level of risk WG is willing to accept[57].

For data that cannot leave WG, on-site (or ideally remote) access within WG offices means that researchers can be monitored while analysing sensitive data. The cost of an on-site solution to WG is in the need to develop infrastructure and training, in staff to oversee researchers, and in the dedicated space required. However on-site access gives researchers the possibility to access all WG data holdings and, where legally possible, to link their own primary data to WG data holdings, thus maximising

[56] Part of the SDS security depended on the use of the JANET network. Software licenses are all academic licenses.

[57] WG may wish to make use of BIL levels to accredit services to hold their data at different levels of sensitivity.

any investment in data collected by researchers by increasing the utility of their data and therefore quality and scope of research into Wales.

On-site access for academics should only be considered as a last resort when the research is valuable enough to justify access to data that is deemed too sensitive to access via other means.

Another option is for KAS to provide a service that produces bespoke analyses to meet researcher needs, though as outlined above this is likely to fall outside KAS's capacity. While bespoke analyses may suit some academics, particularly those who are not expert in data analysis, many academics will want to have input into methodological decisions relating to data structure and analysis which may not be practical or secure in the case of bespoke outputs.

Third Sector

While the coverage of third sector stakeholders in the consultation was low the issues raised were such that the author believes it is safe to assume that they are relevant to other similar organisations in Wales.

The key messages arising from the consultation were that third sector organisations did not want access to highly detailed potentially disclosive microdata, as they had neither the expertise to analyse it, nor the resources to develop a secure setting in which to hold such data. However, there was a high level of frustration at the lack of data they might use for decision making and funding allocation. The lack of access to potentially disclosive microdata seemed to have a particular impact on third sector organisations as they tend to operate in areas that may require information on subjects deemed particularly sensitive, such as mental health, disability, poverty and criminal activity; or have a need to understand small populations.

A bespoke analysis service as described in 1 and 2 above is likely to be the preferred solution for third sector stakeholders as it removes the need for them to access potentially disclosive data or to employ staff with sufficient skills to undertake complex data analysis. It is unlikely that staff in the third sector would have the same concerns as academics about any methodological decisions reached in producing the outputs the data as the feedback received is that they would like 'official statistics' i.e. reports containing estimates at the aggregate level that have been approved by WG.

On-site or remote access to a WG secure infrastructure is an option for providing access for researchers from third sector organisations, but as it would require third sector organisations to have in-house expertise in data analysis, or to employ external experts, it is far from an ideal solution. In addition, the costs to WG that would be associated with such a solution, as outlined above, make this solution unlikely to provide value for money, though as the ADT report states that *'charities and third/voluntary sector organisations may themselves undertake research in the public good… [they] will be afforded the same status as publically financed research institutions'*, the ADRN may support many third sector research needs at no additional cost to WG.

It is recommended that a further study is undertaken to engage with the third sector in Wales to better understand their data needs and to develop data access arrangements that support these organisations' efforts to improve the quality of life of people in Wales. A joined up approach also offers the opportunity for the third sector and WG to pool resources and expertise in assessing needs and implementing a solution.

Members of the Public

Members of the public with the interest and expertise to analyse sensitive data are likely to be relatively few and far between, therefore it is unlikely that any infrastructure set up specifically to serve public needs in this area would receive a good return on investment. In addition, when developing access procedures it should be taken into account that allowing members of the public direct access to sensitive data will increase the risk and the cost of any access system. To mitigate risk there would need to be strong procedures to establish a public person's motive for accessing the data. In the case of an academic, government analyst or charity worker it is far more straightforward to understand the reason for which they wish to access sensitive data. In addition civil servants and academics run a strong risk of losing their job and or reputation for data misuse. While members of the public are subject to the same legal sanctions for data misuse as all other researchers, the aim of secure access systems is to prevent misuse not to punish it. Therefore the cost of oversight would be likely to be higher than for other stakeholder groups. It is also likely that members of the public have less expertise in data management and

statistical analysis than academic or government researchers and are therefore going to be more costly to support.

Taking these issues into account, there are increasing efforts from the current UK government to leverage entrepreneurship and innovation through public access to government data, through initiatives such as Open Data[58] and hackathons[59]. Though these initiatives do not involve disclosive data, the possibility that members of the public have a genuine need to access sensitive data should not be overlooked. A member of the public may be a 'fit and proper person'[60] with a research proposal that serves the public good, and therefore should be considered when licensing conditions, legislation or data access infrastructure are being developed.

For example, the consultation captured feedback from a community activist who wanted access to potentially disclosive microdata to better understand and support community cohesion in their local area. Due to a lack of information on suitable data resources to research this area the respondent's only option, as far as they understood, was a Freedom of Information (FOI) request to WG. However, the responses indicate that staff answering FOI requests may have a low level of awareness of WG data holdings. The respondent had undertaken lengthy correspondence which he did not feel had provided a satisfactory answer, and it transpired that if he had been informed of the existence of the 'Statistics for Wales: Catalogue of Outputs', this would have gone some way to satisfying his request. This highlights another area where a central KAS data service would assist in the delivery of WG priorities - by providing support for FOI staff in answering data-related queries.

An in-house data access infrastructure, as described above, would be a significant investment for WG, therefore there is no suggestion that such a solution should be implemented specifically to provide data access for members of the public, as demand is unlikely to be high enough to justify the expenditure. However, it is

---

[58] http://data.gov.uk/

[59] http://www.theodi.org/news/healthy-start-hackathons-odi ; http://www.manchester.gov.uk/news/article/6484/inaugural_manchester_hackathon_held ; http://www.bis.gov.uk/ukspaceagency/news-and-events/2012/Oct/space-solutions-hackathon

[60] As defined in Statistics and Registration Services Act.

recommended that should an on-site system be developed for other stakeholders KAS consider the possibility of allowing members of the public access.

Therefore, as for third sector researchers, a bespoke service allowing members of the public to commission analyses derived from sensitive data would probably offer the most secure and practical solution to widening access to statistical information for Wales for this stakeholder group.

It may be appropriate to implement a system to monitor data-related requests from members of the public, whether via FOI or other mechanisms, to gain a more accurate understanding of the level of public need.

Summary

This section has outlined a range of possible solutions for maximising access to data on Wales taking into account the needs of different stakeholder groups. It is not expected that KAS will implement all the solutions outlined above as they overlap and not all would be necessary let alone affordable. The solutions likely to offer the most impact are summarised below.

The most efficient and effective strategy for maximising access to data on Wales is to make as much as possible available via one of the existing data services and RDCs, in particular SAIL and the UK Data Service. Implementation of a data management policy as recommended in section 6 above would mean that the tasks necessary to prepare data for release, such as assessment of the level of sensitivity and the generation of metadata to enable reuse, would have already been undertaken suggesting that the marginal cost of releasing data via an existing service would be low.

To support sensitive data use by WG analysts KAS should investigate whether best use is being made of existing services such as VML, SAIL, UKDS and ADRN. To maximise access to these services and to facilitate analysis and linking of sensitive data in-house a safe room should be implemented within KAS.

For data that cannot leave WG, or for cases where the licensing process for external services are incompatible with government timetables, KAS will require additional infrastructure to fully exploit existing data sources. Apart from the implementation of a safe room, a staged strategy is recommended where the components of a secure

data access infrastructure, from a central ‘server’ for storage and management of sensitive data to a full secure remote access system (as outlined below), are established incrementally with each stage building on the last. Such a strategy would enable KAS to assess the suitability of the existing infrastructure after the implementation of each stage, and only progress if a need is demonstrated and the funds are available.

For a full secure access infrastructure that conforms to international best practice KAS would have to develop an environment in which data can be analysed but not removed, and a system for checking any output a researcher wishes to remove from the system for disclosure before it can be released. If WG decide to invest in a service of this type a remote access component should be included to maximise access and therefore the returns on the investment. As any system that requires manual checking of outputs is expensive and difficult to maintain due to the need to find capable staff, KAS may wish to monitor the need for such a solution over the next 2 years to assess whether other internal initiatives and external services meet the requirements of researchers using potentially disclosive data on Wales.

To streamline the provision of data analysis for policy WG may wish to develop a network of experts in particular subject areas that can be called on at short notice to provide expert input. These experts may be in-house, or they may be academics or other specialists working outside WG. These researchers would still require access to secure infrastructures for any research based on potentially disclosive microdata, therefore where the data are held and what the access mechanism is would have to be taken into account when selecting experts (and assessing the type of in-house infrastructure to develop).

Any solution developed by KAS to support the analysis of potentially disclosive microdata is likely to be of interest to external researchers, in particular from the third sector. Therefore when developing solutions consideration should be given as to whether WG can provide adequate capacity to support these external stakeholder groups.

8. Conclusion and Recommendations

This paper has presented the findings of the WG/ESRC Fellowship on Improving Access to Potentially Disclosive Data. The paper has outlined the data access landscape nationally and within Europe, and the legal and ethical considerations that must be taken into account when making potentially disclosive data available for research. It then went onto discuss implementation and funding before looking in more detail into strategies for maximising the availability of data from a range of sources and access to data for different stakeholder groups. While recommendations have been made throughout the paper the key recommendations likely to have the most impact on access to potentially disclosive data for Wales are listed below.

Recommendation 1: Implement a Data Management Policy to enable all WG data resources to be recorded, archived and documented, and where possible reused.

To be effective, a Data Management Policy must be owned, implemented, monitored and updated. This would be a task for the KAS central data service as outlined in recommendation 3 below. A DMP should consist of the following components:

A catalogue of all data held or collected by WG.

The catalogue should include variable-level metadata to an approved international standard, and should indicate where the data are stored and how they can be accessed.

Policies on access and storage.

Each dataset held by WG should be classified according to the potential for disclosure, and a formal policy implemented to enable WG to assign appropriate access and storage solutions at each level of sensitivity.

Policies for transferring data

A formal policy for transferring data between organisations should incorporate WG's existing solutions, AFON and DEWI. It should also address circumstances where these solutions are not appropriate and offer suitable alternatives.

Policies for destroying data

Best practice guidelines on data destruction already exist within WG; these should also be brought into a formal Data Management Policy.

An archiving policy

All data held by WG should be archived where legally possible. A formal data archiving policy should provide suitable solutions for preserving all possible WG collections and outputs at all levels of sensitivity.

Recommendation 2: Establish a KAS Data Service to increase the efficiency of data users within WG.

A service that provides support for data users could oversee and implement the Data Management Policy, handle negotiations for data sharing with other government departments and manage the supply of data to external researchers. It could also maintain contacts with key external stakeholders and other government departments. Any bespoke data analysis requirements could also be handled by this service. There is potential for such a central service to significantly increase WG returns to investment in data, raise research and data quality, improve data security, reduce duplication of data collection, reduce the time taken to negotiate access to data from other government departments, and generally act as a data advisory service for KAS, the WG and (potentially) Wales.

Recommendation 3: Develop a network of experts to provide bespoke analyses of potentially disclosive data.

To streamline the provision of data analysis for policy WG may wish to develop a network of experts in particular subject areas that can be called on at short notice to provide expert input. These experts may be in-house, or they may be academics or other specialists working outside WG. These researchers would still require access to secure infrastructures for any research based on potentially disclosive microdata, therefore where the data are held and what the access mechanism is would have to be taken into account when selecting experts (and assessing the type of in-house infrastructure to develop).

Recommendation 4: Require all contracts with external organisations undertaking data collection for WG to include clauses which specify that

microdata must be deposited with WG to facilitate quality control, reuse and where possible linking.

As a minimum, subcontractors should be required to deliver de-identified data to WG, along with appropriate metadata sufficient to enable deposit at the UK Data Archive or other facility. Ideally identified microdata should be delivered to maximise reuse and facilitate linking.

Recommendation 5: Make full use of existing infrastructure in order to minimise expenditure, leverage existing user support, promote national standards and streamline access.

Where ever possible, KAS should investigate the suitability of existing services such as the UK Data Service, SAIL, and VML, before investing in in-house solutions. KAS should also make use of standards being developed at the National and European level to ensure that any systems implemented conform to international best practice and offer the best possible chance of integration into a wider data access network.

Recommendation 6: Implement a Safe Setting to maximise access to external RDCs, and the potential for secure in-house analysis of WG data.

A ‘safe setting’ is an environment with physical, procedural and technical controls, which can be used to support remote access to sensitive data stored elsewhere. A ‘safe setting’ should be a priority component for an in-house infrastructure, as it will support KAS use of the forthcoming ADRN, increase chances of gaining access to existing government RDCs, and provide a secure environment for in-house data linking and other sensitive analyses. A ‘safe setting’ should be based on standards being developed for the ADRN.

Recommendation 7: Investigate the costs and benefits of developing a secure data access infrastructure in-house.

For data that cannot leave WG, or for cases where the licensing process for external services are incompatible with government timetables, KAS will require an in-house solution.

For a full secure access infrastructure that conforms to international best practice KAS would have to develop a secure environment in which data can be analysed but not removed, and a system for checking any output a researcher wishes to remove

from the system for disclosure before it can be released. If KAS decides to invest in a service of this type a remote access component should be included to maximise appropriate access to data.

KAS may wish to monitor the need for such a solution over a set time period to assess whether other internal initiatives and external services meet the requirements of researchers without requiring further investment.

Recommendation 8: Continue to monitor and engage with national and international projects to maximise access to funding opportunities and to increase the chances of any forthcoming data projects meeting WG needs.

Appendix A: Sources of information for Data Audit

KAS Asset Register

Data Audit carried out for ISO27001 certification of WG Aberystwyth office

Statistics for Wales Catalogue of Outputs

Appendix B: Data requested by stakeholders

The list below shows the datasets requested during the stakeholder consultation, sorted by topic area, in some cases requests are listed under more than one topic.

Key Entries which are duplicated under another topic are in red

Where linked data were specifically mentioned the entry is in purple

Where the request specified detailed geography the entry is in green (including section for the level of geography requested)

Where the respondent stated that the data should be in public domain the entry is in blue

Agriculture

Farming facts and figures survey

Child protection

Child protection registration

Referral for parental support

Community cohesion

Community 'connectedness'

Looked after people's experience of volunteering

Crime

Hate crime

Offending / re-offending rates

Disability

Recruitment, retention, attainment of students with a disability, by type of disability, in further education

Trajectories of disabled students at 16

Education

Recruitment, retention, attainment of students with a disability, by type of disability

Trajectories of disabled students at 16

Attainment at KS3 by eligibility for free school meals, gender and LA

Education attainment at various levels

School truancy

School sickness

School counselling referral

Environment

Environment perception (crime & safety, 'pleasantness', etc…)

Geography

Super output area

Postcode

SPARQL endpoint

Ward level

North wales/south wales

Why no EUL data for Welsh boost to APS?

Health

NPSA reporting data - interorganisational comparisons

Hsmr (hospital standardised mortality ratio)

Dr fosters

1000 lives dataset

Accelerometer (e.g. Actigraph) data from many individuals with coincident GPS data

Yearly BMI data for everyone

Georeferenced health data

Mental health

Count me in Census

Number of children in care due to mental health issues

Employment sickness absence

Prescription data

Repeat admissions

Waiting times for counselling etc.

Range of talking therapies

Amount and range of health services available through Welsh or community languages

Substance misuse data

Public health issues - e.g. % smokers / obese / alcohol intake

Housing

Garden size data for individual properties

Detailed house info (incl. no. Rooms, age, etc.)

Homelessness data

Data on temporary accommodation

Labour market

Employment figures Cardiff city region - by LA, employment sector, total distance travelled to work approximate wage.

Income data linked to LS

Household level income

Employment sickness/absence

% In work / training

% of employers who have mental health policies for staff

Earnings indicators

Financial indicators

Population

Pop/hhlds within 5 miles of a railway station in Cardiff city region - by LA

Pop/hhlds within 1 mile of a railway station in Cardiff city region - by LA.

Ward level census data for Wales

More convenient access to cams

Life expectancy

All data broken down by race, sex, age

Poverty

Poverty indicators

Household level deprivation

Social services

Social care - children (requested by two respondents)

Social care - adults

Benefits data linked to LS

Looked after people's experience of volunteering

No of children in care due to mental health issues

Transport

Road freight data

TTWA by LA, mode of transport

Welsh language

Amount and range of health services available through Welsh or community languages

Appendix C: An outline of the characteristics of European RDCs

The Data without Boundaries project undertook a survey of Research Data Centres in Europe. The survey concluded that there are a number of common characteristics among all RDCs in Europe, these are:

1. Confidential data does not leave the research data centres
2. Outputs are checked[61] (either all outputs or a sample)
3. There is a process of accreditation before researchers can get access
4. All RDCs use Microsoft Active Directory to authenticate and authorise users
5. The internet provides the remote connection between researchers and data with relatively standardised technologies for encryption of communication
6. Surveillance is carried out during research activities[62]
7. RDC staff upload data for researchers after checking it, often based on the project related “need-to-know” principal

8. Standard statistical software is made available free of charge (SAS, Stata, SPSS). Additional software is possible, but users may have to cover costs (DwB 2012)

[61] for disclosive data before being released to researchers

[62] Half of the RDCs monitor complete sessions; some log sessions; others record only log-on and log-off information

RDC characteristics by Country

	Denmark	France	Germany	Netherlands	Sweden	Slovenia	UK[63] VML	UK Data Service
Access via	PC	thin client	Thin client	PC	PC	PC	Thin client	PC
Connection via	citrix	bespoke	citrix	citrix	microsoft	citrix	citrix	citrix
Encryption	SSL VPN	SSTP port 443	SSL-VPN port 443	SSTP port 443	SSL VPN and RDP	Encrypted VPN CICO tunnel and ICA basic encryption	?	SecureICA (2048-bit encryption)
RDC accessed from	anywhere	academic institutions	safe centres only	academic institutions	anywhere	anywhere	UK government intranet	academic institutions
Authentication								
userid and password	x		x			x	x	X
IP range	x	x	x			x		X
Smart card biometrics (fingerprint) - no password		x		x				

[63]For the UK, two RDCs are used as examples the VML based at ONS, and the UK Data Service (previously SDS)

Smart card creates one time password					x			X
Certificate bound to the hardware (thin client)		x	x					
RSA SecureID token[64]	x							

[64] http://en.wikipedia.org/wiki/SecurID#March_2011_system_compromise

Appendix D: UK Data Archive

The UK Data Archive has been preserving data and providing access for research purposes for over 40 years. Funded by the ESRC, the UK Data Archive now forms part of the UK Data Service (UKDS).

The UK Data Service provides access to data produced by government departments, researchers, funding councils, Intergovernmental Organisations, and any other data that might be useful for research. The data primarily cover socio-economic topics and include census, longitudinal, international, and qualitative resources. Data are made available across most of the access spectrum, from anonymised to sensitive. An outline of the type of data and licenses available is provided below.

End User License

Data available under an End User License (EUL) is anonymised microdata with a low possibility of disclosure of personal information. All identifying information such as name, address etc. is removed, and additional measures are taken to mask any potentially disclosive responses.

To access EUL data researchers must register with the UKDS and agree to the 'terms and conditions of use of data', once registered they can download data to their desktop.

For some EUL data there are additional 'special conditions'. The exact conditions depend on the data provider, but they are usually agreed by the researcher online during download.

Special License

Data available under Special License (SL) are anonymised microdata, but contain more detail than is available under the EUL, and in particular are likely to contain variables that are considered too sensitive to release under EUL, for example age, year and month of birth, more detailed geographical variables, or detailed information on occupation.

SL data are available to download, however there are conditions on where and how the data can be stored and processed and researchers must make additional commitments on data handling as outlined in *Microdata Handling and Security:*

Guide to Good Practice[65]*.* For example, cells containing one or two cases cannot be reported and nor can geographies below Government Office Region.

Unlike EUL data, SL data cannot be downloaded automatically on application as requests for SL data are forwarded to the data owners for approval. Researchers must prove that other data available are not fit for their purpose before being granted access to SL data.

Secure Data

Data available through the UKDS's secure service are highly detailed, de-identified data. There is a risk of disclosure with these data which contain sensitive information such as postcode, date of birth, firm level data, or health information.

Sensitive data cannot be downloaded to a researcher's desktop; all analysis takes place in the controlled environment of the UKDS's RDC. All output is checked manually by RDC staff before removal. The RDC is a remote access system that allows access from the researcher's institution.

Applications for access via the secure service must always be approved by the data provider. Researchers wishing to use the secure service must undertake a mandatory training course.

For more information on the UKDS's terms and conditions of access see http://data-archive.ac.uk/conditions/data-access

Data Centre

The UK Data Archive is also developing a data centre that will provide highly secure data storage facilities. The data centre has been developed to a level suitable for holding data at BIL4.

The UK Data Archive has been designated as an official Place of Deposit by The National Archive. They are ISO27001 certified, and accredited by BIS to hold data up to and including BIL4 and disseminate data up to BIL3 (under certain circumstances).

[65] http://www.data-archive.ac.uk/media/132701/UKDA171-SS-MicrodataHandling.pdf

References

ADT *see* Administrative Data Taskforce

Administrative Data Taskforce (2011), *Terms of Reference*, [online]. Available at http://www.esrc.ac.uk/funding-and-guidance/collaboration/collaborative-initiatives/Administrative-Data-Taskforce.aspx [Accessed 10/07/2012]

Administrative Data Taskforce (2012), *Improving Access for Research and Policy*, UK Administrative Data Research Network. http://www.esrc.ac.uk/_images/ADT-Improving-Access-for-Research-and-Policy_tcm8-24462.pdf

Cabinet Office (2008). *Data Handling Procedures in Government: Final Report*, Cabinet Office Report http://www.cesg.gov.uk/products_services/iatp/documents/data_handling_review.pdf

CESG (n.d). *Cross Government Actions: Mandatory Minimum Measures,* HMG http://www.nationalarchives.gov.uk/documents/information-management/cross-govt-actions.pdf

CESG (2009), HMG IA Standard no. 1, Technical Risk Assessment, HMG http://www.cesg.gov.uk/publications/Documents/is1_risk_assessment.pdf

CO *see* Cabinet Office

Commissioners for Revenue and Customs Act 2005 (c.11) http://www.legislation.gov.uk/ukpga/2005/11/contents

Davies, R (2013). Examining the Feasibility of Establishing a Wales Longitudinal Study.

Desai, T (2003). *Providing Remote Access to Data: The Academic Perspective.* Work Session on Statistical Data Confidentiality, UNECE/Eurostat, Luxembourg http://www.unece.org/fileadmin/DAM/stats/documents/2003/04/confidentiality/wp.9.s.e.pdf

Desai, T and Ritchie, F (2010). *Effective researcher management*.,Work session on statistical data confidentiality 2009, UNECE/Eurostat, Bilbao http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2009/wp.15.e.pdf

Desai, T, Ritchie, F and Welpton, R (forthcoming). *'Five safes': designing data access for research.*

DwB *see* Data without Boundaries

DwB (2012). Report on the state of the art of current SC in Europe. DwB http://www.dwbproject.org/export/sites/default/about/public_deliveraples/d4_1_current_sc_in_europe_report_full.pdf

Fletcher-Cooke, G. (2003). *Effect of the Statistical Legislation Framework in the UK on the Work of the Government Statistical Service.* Statistics Commission Report no. 13. London http://www.statisticsauthority.gov.uk/reports---correspondence/archive/statistics-commission-archive/research/report-13--effect-of-the-statistical-legislation-framework-in-the-uk-on-the-work-of-the-gss--december-2003-.pdf

Ford et al (2009). *The SAIL Databank: building a national architecture for e-health research and evaluation*. BMC Health Service Research 2009, 9:157. http://www.biomedcentral.com/1472-6963/9/1617

Fry, R., Berry, R., Higgs, G., Orford, S. and Jones, S. (2012), *The WISERD Geoportal: A Tool for the Discovery, Analysis and Visualization of Socio-economic (Meta-) Data for Wales*. Transactions in GIS, 16: 105–124. doi: 10.1111/j.1467-9671.2012.01308.x http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9671.2012.01308.x/full

Government of Wales Act 2006 (c.32) http://www.legislation.gov.uk/ukpga/2006/32/contents

Government Statistical Service (n.d.). *Building a Framework for Data Sharing for Statistical and Analytical Purposes*, Government Statistical Service Subgroup on Data Sharing for Statistical and Analytical Purposes, Newport.

Government Statistical Service (2009). *Stepping Stones to Data Sharing for Statistical Purposes*, Government Statistical Service, Newport.

GSS *see* Government Statistical Service

Heaven, M (2013) Data Linking Demonstration Project - Journey mapping for patients with multiple chronic health conditions

Heaven, M (2013) Data Linking Demonstration Project - Examining fuel poverty using Home Energy Efficiency Data and routinely collected health data

Heaven, M (2014) Data Linking Demonstration Project – Flying Start

ICO *see* Information Commissioner's Office

Information Commissioner's Office (2007). *Framework Code of Practice for Sharing Personal Information (consultation draft)*. Information Commissioner's Office, Cheshire.

http://www.ico.gov.uk/upload/documents/library/data_protection/practical_application/ico_information_sharing_framework_draft_1008.pdf

Jackson, P.J. *Limitations of Current Legislation for Data Sharing.* (GSS(SPSC-DS) 33). Government Statistical Service

Jackson, P.J. (17th January 2008). *The Statistics and Registration Act and Implications for Data Access.* [Powerpoint slides]. Presented at ESDS Government meeting on changes in access to the government surveys and Labour Force Survey/Annual Population Survey user meeting. http://www.ccsr.ac.uk/esds/events/2008-01-17/

Kruten, T. (2008)

Laurie, G. (2011). Reflexive governance in biobanking: on the value of policy led approaches and the need to recognise the limits of law. Journal of Human Genetics, 130(3), pp347-356. doi: http://dx.doi.org/10.1007/s00439-011-1066-x

Ministry of Justice (2011). *Offending, employment and benefits – emerging findings from the data linkage project,* Ministry of Justice, London. http://statistics.dwp.gov.uk/asd/asd1/adhoc_analysis/2011/offending_employment_and_benefits.pdf

MOJ *see* Ministry of Justice

Morgan (2012). Feedback to the first draft

Public Audit (Wales) Act 2004 (c.23) http://www.legislation.gov.uk/ukpga/2004/23/contents

Ritchie F. (2006) *Access to business microdata in the UK: dealing with the irreducible risks*. Work session on statistical data confidentiality 2005, UNECE/Eurostat, pp239-244

Ritchie F. (2010). *Access to sensitive data: satisfying objectives, not constraints*. WISERD working paper no. 7. http://www.wiserd.ac.uk/wp-content/uploads/2012/02/WISERD_WDR_007.pdf

Statistics and Registration Service Act 2007 (c.18) http://www.legislation.gov.uk/ukpga/2007/18/contents

Welsh Government (n.d.). *Wales Accord on Sharing Personal Information*. http://www.waspi.org/home.cfm?orgid=702

Welsh Government (2011). *Programme for Government*. (WG 13124). Crown Copyright 2011. http://wales.gov.uk/about/programmeforgov/?lang=en

WG *see* Welsh Government