

Data Sharing Review

Richard Thomas and Dr Mark Walport

Consultation paper on the use and sharing of personal information in the public and private sector

List of questions for response

We would welcome responses to the following questions set out in this consultation paper. Please follow the question order as set out in the consultation paper, leaving a blank response box for any questions not answered.

Please email your completed form to contact@datasharingreview.gsi.gov.uk

Alternatively you can send a hard copy response to:

Data Sharing Review Secretariat
5.26 Steel House
11 Tothill Street
London
SW1H 9LJ

Thank you.

Section 1: Background

Question 1.

Comments: Analytical Data Integration for Government (ADIG) is a cross government project, which has looked at the feasibility of establishing a longitudinal information base for cross government policy making, research and analysis. It has concentrated on the feasibility of a federated approach where data remain secure in individual departments and are drawn on to create anonymised datasets for analysis by linking cross-department longitudinal administrative data, survey data and research. It has not looked at the feasibility of creating a mega government database or data warehouse.

Analysis for policy making, unlike that for operational purposes, does not require the use of identifiable data. A guiding principle of ADIG has been that the process is dependent on the creation and use of anonymised datasets in order to protect the privacy of the individual.

ADIG is discrete from the Ministry of Justice Information Sharing Programme (the successor to MISC 31) to share data for operational purposes.

The ADIG Feasibility Report was delivered to the Steering Group of Permanent Secretaries from Cabinet Office, DCSF, DWP, HMRC, HMT, MoJ and ONS in January 2008. The following comments/responses are the relevant sections in the Report, a copy of which accompanies this response form.

Section 2: Scope of personal information sharing, including benefits, barriers and risks of data sharing and data protection

Question 2.

Comments:

Question 3.

Comments:

Question 4.

Comments:

Question 5.

Comments:

Question 6.

Comments:

Question 7.

Comments: **ADIG Feasibility Report Chapter 3 pages 14-18**

The process of decision making in the public sector has become an insurmountable task as civil servants have to look through disparate information repositories and systems with inconsistencies in how the information is structured and managed.

There is a requirement across government for accurate and timely information to support policy formulation and management, supplemented by a growing need for regional and local level data. There is enormous potential to use 'administrative data' that are already collected by government to meet these requirements.

As detailed in the Scoping Study, 'Administrative data' are the individual records of people and businesses already held by public authorities. The particular values of administrative data are the near total coverage of the population, the relevance to policy and administration, the frequency with which it is updated, and the potential for linking together records from separate sources using common identifiers. It is this last value which is least exploited.

Survey data are different from administrative data. Typically, coverage is weaker but quality is very good since the scope and relevance of questions asked can be much greater than those from administrative sources. The most obvious difference is the unit cost of collection. Administrative data used for research, analysis and statistics are a low-cost by-product of a departmental function; survey data collection is dependent upon costly field operation and processing procedures.

Linking these sources provides data with high levels of quality, coverage, relevance and timeliness.

The data most in demand for government policy making are those that describe interrelated factors, presented in small areas or in specific business sectors. This is by far the most difficult information to collect through surveys. It means surveys are complex, by necessity slow to react to new policy developments, and analysis of the survey outputs cannot always meet new policy needs. Outputs based on the administrative data of the relevant

department will address single issue policy requirements. The solution is to combine sources of administrative data and survey information where relevant, to produce an information source suitable for rapid, ad-hoc analysis in response to current policy needs, and to inform emerging public debates.

Access to administrative data, linking (for example) income, benefit and household composition information, would support policies to identify, monitor and redress the inequity in society. Deprived populations are often small and are clustered in ways that makes their enumeration in surveys and censuses problematic. Those with sensory disability, mental incapacity, or English as a second language will find survey and census inquiries intrusive and disproportionately difficult to complete. These populations are underreported and inadequately described in surveys, but can be the population best described in administrative data.

There are of course limitations to using administrative data. It is collected for other purposes and so may not be exactly what is required for analysis and research. Key demographic indicators such as household composition, educational attainment, etc are not usually available unless they can be matched in from other administrative data. There may also be inconsistency of treatment/data collection over a large organisation and over time. So administrative data is not necessarily a panacea for analytical needs, and may sometimes need to be supplemented with survey reporting.

Given that the data linking potential already exists for the main departments with economic data (ie BERR, HMRC, DWP), the most value is likely to be had in linking these onwards to data on educational, health, housing and criminal records. This would provide much more of the 'demographic' information otherwise missing from tax and benefits data and enable a much more refined analysis of the underlying determinants of labour market outcomes, the behavioural responses to policy changes or other interventions (eg changes in the school leaving age), and better understanding of the social and economic returns to investment in health, education, crime prevention. For example, an early years education programme has effects on later health, crime and employment. The nature of preventative policies is that it's hard to see whether they've worked till long after they've started, which means that it's crucial to form the policy using excellent evidence, and then to go on collecting evidence so that we know what to change.

Administrative data and data sharing are therefore increasingly playing a key role in providing this information. This mitigates the high cost of meeting these requirements through surveys alone and the difficulties involved in maintaining high response rates to traditional surveys and censuses. It provides a cost-effective option, placing minimal burden on respondents, as data are collected once and used many times.

The benefits of using administrative sources include more frequent and timely small area data for improved policy formulation and monitoring, improved service delivery and reduced public burden.

Policy makers are continually faced with the challenge of understanding complex and interdependent social issues. Effective policy making depends on good quality evidence to evaluate underlying causes and determine how different parts of government can combine to promote the required outcomes.

An individual's contact with public services can be complex and rarely aligns with the boundaries of a single government department. Collectively, government departments possess a rich and hugely valuable information resource that has the potential to enhance policy development and provide a better understanding of outcomes. However, today, policy makers have access to only a fraction of this potential.

The CSR07 PSA targets focus on promoting a joined up approach to addressing government priorities. These are complemented by the Service Transformation Agreement (STA) which aims to change public services so that they more often meet the needs of people and businesses. The STA identifies a number of strategic actions to achieve this, including learning from citizens and businesses through "...a real, evidence based understanding of the behaviours of people they are trying to reach ...", and making better use of the information that the public sector already holds.

The ability to develop innovative policies will depend on policy makers having a better understanding of the issues they are trying to address, and the use and effects of interventions from different areas of government. In 2004, Derek Wanless found¹ "The relationship between wider determinants and health is well established but complex. Further research in this area would be beneficial in order to fully understand the determinants of health and health inequalities." While some progress has been made in understanding this area, it highlights the problems experienced by many departments in understanding variations in susceptibility and identifying segments of the population which can be targeted to reduce later demand on public services. Whether it is health, life chances, or the environment, outcomes are affected by a range of external factors and interventions across multiple government departments.

Government spends significant amounts of money every year implementing new policies. The increasing demand on departments to deliver on efficiency targets and improve value for money places a new emphasis on optimising government interventions and evaluating new policies to ensure adequate return on investment.

Research is under way under the NCeSS Programme² (Administrative Data – Methods, Inference and Network (ADMIN) Node) to develop better tools for analysts/researchers that in turn enable them to provide more robust answers to pressing social questions. This will be achieved by exploring how analysts/researchers should best use administrative datasets and in corollary by determining how researchers can enhance longitudinal survey data by exploiting available administrative data. Producing better methods to fully exploit administrative data is important because it provides analysts/researchers with the capacity to analyse and solve social problems that currently cannot be analysed effectively. Indeed, this area of research has been identified as a high priority both by the ESRC-commissioned Strategic Review of Panel and Cohort Studies and by the National Centre for Research

¹ 'Securing good health for the whole population', Derek Wanless, December 2004

² The National Centre for e-Social Science (NCeSS) <http://www.ncess.ac.uk/> is funded by the Economic and Social Research Council (ESRC) to investigate how innovative and powerful computer-based infrastructure and tools developed over the past five years under the UK e-Science programme can benefit the social science research community. This infrastructure is commonly known as the 'Grid'

Methods (NCRM), and has been recognised as an area of weakness in social science (Smith et al., 2004).

Quantitative social research has long relied on using survey data (both cross-sectional and longitudinal) for substantive and methodological work. In recent years, policy-oriented research in particular has increasingly relied on administrative data, as researchers have begun to get access to government administrative datasets across a range of fields. For instance, in the UK, Department for Work and Pensions (DWP) administrative data have been used to evaluate some of its employment programmes such as the New Deal programmes (eg Blundell et. al., 2004) as well as some programmes for other departments such as the Department for Education and Skills Neighbourhood Nursery Initiative (see Deaden et. al., 2007). The National Pupil Database (NPD), which is held at the Department for Children, Schools and Families (DCSF), has also increasingly been used by researchers to look at what determines children's outcomes in schools (eg Wilson et. al., 2005 and Steele et. al., 2007).

The advantage of administrative datasets is that they have information on almost everybody – they are effectively a census of individuals. One of the disadvantages of administrative datasets is that they typically do not contain very rich information. Furthermore, currently, many government administrative datasets provide a partial picture on some important research issues. For instance, DWP administrative data contain good labour market information on individuals but typically do not have information on their education, although it is of course well documented that a person's labour market success is highly related to their educational qualifications. The NPD by contrast has detailed information on children's educational outcomes but does not have any information on the labour market status or education of the child's parents. Again, it is well known that a child's educational attainment is influenced by their parent's educational attainment, so this omission is serious.

Researchers using administrative data generally acknowledge these shortcomings and then use a variety of methods to try to overcome the problems of not having rich enough information on individuals.

Longitudinal surveys have long been used by UK researchers, for example, the Institute of Education (IoE) is responsible for running three of the most important longitudinal surveys in the UK – the National Child Development Survey (NCDS), the British Cohort Study (BCS) and the new Millennium Cohort Study (MCS). These, and other longitudinal surveys (such as the British Household Panel Survey (BHPS) and the Longitudinal Study of Young People in England (LSYPE)), unlike administrative data, have very rich information on individuals' personal characteristics, but by construction are only a sample of the UK population. Furthermore, because such surveys are longitudinal, they suffer from attrition problems. In addition, because many questions in these surveys are retrospective on issues such as labour market history and educational attainment, there are also potential problems of recall bias, leading to concerns about the accuracy of respondents' answers.

Research is underway to devise and test methodological approaches to overcome these weaknesses and then exploit these new approaches to provide empirical evidence on concrete policy questions and to develop methods that use administrative data to overcome the shortcomings of longitudinal survey data.

Researchers/analysts use a variety of methodological approaches to deal with these

problems, including sampling and attrition weights, imputation and other methods to tackle measurement error, but again, whether these methods actually achieve what they set out to do is not generally testable. The methods are particularly questionable when attrition or non-response is concentrated in groups that are of particular policy interest, such as ethnic minority groups. The potential to apply better solutions to the problems of attrition and recall bias in longitudinal survey data arises because these datasets are now increasingly being linked to government administrative data. These linkages mean that in the event of non-response, either on an outcome of interest or on particular covariates, one can turn to the administrative data to find alternative or proxy measures of missing variables. Furthermore, by their very nature, administrative data are a particularly good source of time-varying explanatory variables, something that is often lacking in longitudinal surveys with highly spaced interviews.

ADIG Feasibility Report Chapter 7 pages 93 – 94:

The advice of information lawyers is that existing statute and ministerial powers give little scope to progress an effective and viable cross government longitudinal information base for policy making. The barriers are not however insurmountable: a general data-sharing power purely for policy making and analysis would create an environment within which data could be integrated, subject to appropriate security and confidentiality constraints.

The legal project within the Ministry of Justice Information Sharing Programme, the successor to MISC 31, is looking at options for a general data-sharing power, not least to enable delivery of the Service Transformation Agreement. This project is concerned with data-sharing for operational purposes and is likely to provoke more controversy and scrutiny than what is being proposed for ADIG.

Existing gateways for data sharing set out the purposes for which the data can be used and the persons / entities with whom the data can be shared. To extend either of these requires additional gateways. This means that every proposal or hypothesis submitted for consideration to the Ethics Approval Process (see Chapter 8 – Governance & Ethics) would have to be individually scrutinised by each department whose data it was proposed to link, particularly if it required the matching of identifiable data. If the proposal fell outside the purposes of existing gateways, new gateways would have to be created. Further, the proposal or hypothesis would have to be for the purposes of, or for any purposes connected with, the functions of the departments' whose data was involved. Current legal interpretation would appear to make this requirement a barrier for research and analysis leading to real cross-government policy making: it is not always going to be possible to identify benefits which may accrue to contributing departments, including downstream or longer-term benefits. One way to overcome this is to revisit the definitions of Departments' functions as enacted. We know that similar concerns are being expressed about the ability of Departments to meet some of the objectives of the Service Transformation Agreement due to the same legal constraints.

The general opinion seems to be that the linking of truly anonymised data should not present problems provided there is some reference back to departments' purposes/functions. There are substantial issues around the matching of identifiable data to create anonymised datasets for analysis, even where the personal identifiers used could be said to be in the public domain.

Question 8.
Comments:

Section 3: The legal framework

Question 9.
Comments:

Question 10.
Comments:

Question 11.
Comments:

Question 12.
Comments:

Question 13.
Comments:

Question 14.
Comments:

Question 15.
Comments:

Section 4: Consent and transparency

Question 16.
Comments:

Question 17.
Comments:

Question 18.
Comments:

Question 19.
<p>Comments: ADIG Feasibility Report Chapter 8 pages 112 – 116:</p> <p>Analytical and research governance concerns the development of shared standards and mechanisms that permit the proper management and monitoring of analysis and research and, if necessary, allow sanctions to be brought in cases of misconduct. The two dimensions of ethics and governance are linked: a strong ethical culture and literacy are dependent not only on professional self-regulation but also on sound structures of formal governance within analysis and research.</p> <p>In general terms, analysis and research governance would presume that the body authorising the analysis and/or research has:</p> <ul style="list-style-type: none"> • Mechanisms in place to enable timely and expedited review of analysis and/or research

and analysis and/or research proposals. These may involve differing degrees of formality and levels of responsibility;

- Procedures that are flexible and sensitive to the differing needs of analysts and researchers and levels of risk involved in their analysis and research;
- Procedures to protect the interests of analysts and researchers;
- The capacity to deal with cases of analysis and research misconduct, complaints or appeals;
- The capacity to advise on statutory or legal considerations that might affect analysis and/or research.

Under ADIG all proposals for analysis and/or research will be required to go through an ethics process. As part of the proposal the analyst / researcher will need to have researched existing evidence; identified data sources; established the use(s) to which these may be put; explained exactly how the data will be used, methodologies to be employed etc; and how and to whom the results will be presented. They will also need to indicate whether the proposal may be subject to review as the work progresses.

Departmental data guardians or stewards will not make data available until ethical approval is given for the research to proceed.

Ethical Framework: Principle: 'Because we can, should we?'

Analysis and research should be designed, reviewed and undertaken to ensure integrity and quality in respect to the hypotheses to be tested, proposal(s) to be researched, analysis undertaken, and the use to which cross-government administrative data is to be put.

This means that analysts, researchers and the Governance and Ethics Committee (GEC) should ensure from the outset that the development and consideration of proposals is informed by a commitment to research that is accountable and of the highest quality. Accountability underlies this principle: quality is expressed through good scientific design, the anticipation of likely difficulties and how these might be addressed, and the ways in which objectives will actually be delivered during the work.

Ethical issues must always be addressed in the proposal

Legal and data requirements must be met

- Analysis and research must comply with legislative requirements and with the requirements of data providers.
- Data supplier access requirements with regard to the secondary use of datasets must be complied with at all times, including any provision relating to presumed consent and potential risk of disclosure of sensitive information.
- Data suppliers must be consulted on their particular requirements

Approval by the GEC is required before datasets are requested and before research and analysis is undertaken.

Procedures for institutional monitoring must be in place

- The GEC will establish appropriate procedures to monitor the conduct of analysis and research which has received ethical approval until it is completed, and to ensure appropriate continuing review where the analysis and/or research design anticipates possible changes over time that may need to be addressed.

- Monitoring should be proportionate to the nature and degree of risk entailed in the analysis and/or research.
- Monitoring must include consideration of best-practice procedures for the secure holding, preservation or destruction of data.

Avoiding duplication

- Where appropriate the GEC will direct that proposals be combined to avoid duplication of analysis and/or research or be undertaken in collaboration where they are complementary.

Complaints procedures should be in place

- The GEC must have mechanisms for receiving and addressing complaints or expressions of concern about the conduct of research carried out under their auspices.

Ethics Committee

The creation and function of the ADIG Ethics Committee can build on what is already happening within government and academia. In the operational area the DWP Work and Pensions Longitudinal Study (WPLS) Ethics Committee has been established for a number of years. The Committee's terms of reference can be found at:

http://www.dwp.gov.uk/asd/longitudinal_study/Terms_of_ref_WPLS_EC_05_01.doc

As the Committee's report for 2006 explains:

'The committee examines any significant new uses of the WPLS and considers for each for the following:

- *The ethical issues surrounding the proposal;*
- *Whether the proposal is accepted or not;*
- *Any modifications to the proposal, which would make it acceptable.'*

Further, the Chairman states in his foreword to the Report:

'The Committee now has well established ways of working and a clear role in respect of the use and access to WPLS. In light of the very useful role that the Ethics Committee has undertaken it has been proposed that the Committee also consider the ethical implication of all Information Directorate's use of non DWP data'.

A copy of the report can be found at:

http://www.dwp.gov.uk/asd/longitudinal_study/annual_ethic_rep_2006.pdf

ESRC's Ethical Framework -

http://www.esrc.ac.uk/ESRCInfoCentre/Images/ESRC_Re_Ethics_Frame_tcm6-11291.pdf -

is established best practice. Whilst much of the framework concentrates on research involving the participation of human research subjects, which is not directly applicable to what is envisaged under ADIG, experience of commissioning and conducting research within

the framework could be invaluable.

It is suggested that membership of the ADIG Ethics Committee should, as a minimum, comprise:

The Chairman.

Representatives of the main areas of government research – Economics, Statistics, and Social Science. These members could be nominated by the Government Heads of Analysis from within the Government Services and/or from academia.

An senior experienced independent statistician to provide professional oversight of the use of government administrative data and academic and government survey data.

A member of or representative of the ESRC

A theologian or member of the clergy

An independent lawyer specialising in human rights, data protection and freedom of information

A journalist or other communicator

Members from outside government would be appointed by advertising for applications from people with appropriate experience and following a transparent appointments process.

Section 5: Technology

Question 20.

Comments:

Question 21.

Comments:

Question 22.

Comments: **ADIG Feasibility Report Chapter 3 pages 22 – 32 and 34 – 36:**

Analytical Data Integration for Government is discrete from the operational functions of government. It is assumed that, in line with recognised good practice, individual departments will hold their data for analysis in 'information centres', distinct and separate from their operational systems. Departments will own and control the data in their 'information centres' throughout the integration for analysis process.

The proposal is therefore that an independent third party is employed to process and integrate the data to produce datasets for analysis. This will ensure that the process is, and is seen to be:

- outside the control of any one government department;

- discrete from all departments' operational systems.

Not only must ADIG separate itself in fact from the operational environment: the public perception must be that it is so.

It is envisaged that the data integration process could consist of one or more of the following, depending on the research to be done or hypotheses to be tested. All analysis will be carried out within a secure environment.

1. Anonymised datasets extracted from the information centres and analysed independently.
2. Anonymised datasets from the information centres extracted and mashed together for analysis.
3. Segmented datasets extracted and joined together. Possible segmentation is by socio-economic variables or commercially available classification systems, e.g. Experian's Public Sector Mosaic (see later).
4. Data with identifiers extracted from information centres and matched to produce an anonymised dataset for analysis. The originally extracted datasets and matched, identifiable, dataset will be destroyed when the quality assured anonymised datasets are produced.
5. An alternative to creation of an anonymised dataset is creation of a pseudonymised dataset with the identifier key held by the creator, and outside the control of the data controllers i.e. government departments.
6. Matching administrative data and survey data, both government and non-government, to create new, quality assured, datasets.

Anonymised Data and Data Mashing

All integrated datasets produced for analysis, by whatever process, will be analysed in a secure environment and will 'self-destruct'/be deleted after an agreed period of time.

For Options 1 and 2 above, the anonymised data sets may be analysed independently and then conclusions drawn or such datasets could be "mashed" together for analysis.

"The term 'mashup' is used to describe a web-based application that uses content from more than one source to create a completely new service³. 'Content' is whatever needs to be combined or aggregated: visuals, raw data, information, or working software. 'Data mashing' is therefore the combination and presentation information and data from several remote sources in a novel way. The idea of aggregating data in this way is not new. What distinguishes mashups is an approach that uses simple, open standard, freely available software tools and protocols to access and manipulate data available on the web⁴."

Innovative data mashing is neither the reserve of the research community nor of the major IT companies. Many developments and applications in data 'mashing' are likely to be products of the not-for-profit organisations, individuals and small to medium enterprises. The importance of these applications is difficult to quantify but should not be under-estimated and

³ [http://en.wikipedia.org/wiki/Mashup_\(web_application_hybrid\)](http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid))

⁴ <http://www.bio-itworld.com/newsitems/2006/january/01-17-05-news-hiphop>

could act as important drivers of innovation and entrepreneurial activity. The encouragement of data mashing of public sector data complements rather than replicates more formal data sharing across government and the commissioning of large-scale analytical projects or software from the research and private sectors. There is a need, however, to find new ways of engaging with the non-government sector and resolving issues inhibiting data access.

Although recognition of the potential value of mashups is growing there remains a number of obstacles inhibiting knowledge and development of such applications. In particular there is a lack of organisational support for exploring the potential of mashups, limited tools for creating and managing such applications, and limited opportunities or forums for demonstrating the potential of mashing techniques to non-technical decision makers.

Web 2.0 technologies have the capability of enabling the more rapid realisation of data aggregation and visualisation applications. However, these same technologies pose a number of challenges for organisations particularly in terms of the way in which IT resources, skills and security are managed.

Recommendation 6 of the Power of Information report reads:

"To promote innovative use of public sector information, the Department for Transport, with the support of the Chief Scientific Advisor's Committee, should complete the partially undertaken scoping and costing of a "data mashing laboratory" and advise the Cabinet Committee of Science and Innovation on appropriate next steps."

The Department for Transport's Data Grand Challenge sought to realise benefits of data, in particular real-time, both within and outside of Government. Work by the Department in 2006 considering a "data mashing laboratory" pointed to an "experimental environment for developing innovative information services", since development of new information services may be impeded by fragmented delivery chains, uncertain business models and the up-front costs of prototypes. The solution, termed an "information incubator", would encourage wider experimentation with data, and help persuade users and owners of data to collaborate and undertake more innovative projects. The incubator would achieve this by providing: a secure environment providing mediated access to federated datasets; experimental tools to enable development of appropriate technical and business models; and, a centre of expertise to support development of data applications from conception to prototyping and business case development.

Work to prove the concept of the incubator is currently well underway by a consortium comprising Cambridge University, Lockheed Martin, Deloitte, and Thales, working alongside Transport Direct, *Southampton University* and others. This pilot project aims to demonstrate the potential benefits of an incubator facility by developing a prototype specific information service and providing exemplar technical and business models for operating a commercial incubator facility.

The pilot's objectives are:

- Scope the concept and feasibility of a permanent incubator facility by developing and demonstrating appropriate technical and business models
- Demonstrate the benefits of the proposed incubator through developing initial business and technical models for specific information applications.

ADIG needs to stay in touch with this pilot in order to build upon the learning and to consider, at an appropriate point in time, whether to develop its own data mashing/incubator function or perhaps more realistically how to **securely** make use of the developed facility.

Segmentation

A major tenet of Service Transformation in Government is 'putting the customer at the heart of everything we do', i.e. customer-centricity. This is unachievable without step-change improvements in Government's knowledge of customers, their **current** personal details, circumstances, attitudes, behaviours and needs, in order to improve service and to materially support the Public Service Agreements imperatives. In addition, in order to deal with customers' needs in a more complete, holistic and "joined up" fashion, there needs to be a common view of the customer base, which can be used by any customer facing function of Government.

Segmentation, classifications and profiling are some of the fundamental building blocks upon which customer insight can be developed. They all provide ways of understanding people, their needs, and the ways in which services may be most efficiently and effectively delivered to them.

Segmentation is a generic term for dividing a large group of people, for example the UK population, into distinct groups, or segments. In recent years there has been full recognition that the "one size fits all" traditional model of public sector policy and service delivery is costly, inefficient, and not what today's population wants or expects. However, it is also accepted that it is not possible to design policy and service delivery to precisely and uniquely meet the needs of each and every individual.

Segmentation provides an effective half way house. Each segment contains a subset of the larger population that tend to share similar (although not identical) characteristics, and hence are likely to have similar service needs. Those characteristics and needs are very different from those of the other segments. Service may therefore be designed to meet the needs of each segment, not the needs of each specific individual.

A classification is essentially a formalised and often more complex segmentation. The segments are created not from one or two dimensions, but instead are multi-dimensional using a variety of data sources. The richness of data enables each segment to be understood in a much more holistic way, from their demographics through to their health, from their educational attainment through to their channel preferences.

Commercial organisations typically create generic classifications from several hundred pieces of data. These can be shown to be very effective means of subdividing the population based on their demands for public services.

Such generic classifications can therefore be used to cut across traditional silos and inter-departmental barriers. Each public service can design its delivery to meet the needs of each segment, but the different departments can talk to each other using a "common language".

Profiling compares a particular target group of individuals (perhaps the residents of a Local Authority, or claimants of Job Seekers Allowance) against a broader benchmark, segment by segment. It therefore enables the user to determine which segments are particularly over or

under represented in their target group.

From discussions with, and contributions from, it is apparent that Experian has a refined and focussed set of solutions designed specifically to meet the challenges faced by the public sector. Through the core functions of customer insight and spatial analysis, that enable the public sector to better understand citizens and communities and to better design services to meet their needs.

Such information is used by a variety of Central and Local Government functions for both understanding and analysis and for more operational and tactical deployment. Often this data cuts across a number of departments and information and analysis gleaned from one project can inform another project within a separate department. It is therefore important to understand the benefit of building and maintaining a cross-agency and federated dataset on people and businesses that can be used for research, analysis and informing policy decisions.

In its' contextual form, the following scenario could be possible which enables pre-defined users across government access to consistent and reliable data.

A data cleansing tool is developed that appends persistent and unique identification references to person and business level data. "Signed-up" government departments take this tool in-house, and use it to code-up subsets of their data which is then held in an anonymised data mart. Within this process, a web-based service is developed that fully verifies the user as they log on, providing a tiered-access system that presents them with their approved data options across all government. Using the unique identification references, data is subsequently merged and returned securely to the user for analysis and research purposes.

There are already organisations that offer data-sharing services for academic research purposes, such as the Census Dissemination Unit at the University of Manchester, although obviously the scale and size of operations are significantly below the potential requirement of a pan-government solution. They will have fairly strict access and user controls to ensure that commercially valuable data is not readily accessible and that confidential information cannot be easily disseminated, as such there may be some value in engaging with these organisations.

A key aspect of any data sharing operation will be to deny unauthorised access and enable a tiered approach to levels of access for authorised users. All users should be pre-approved and be provided with unique access codes that go through a series of verification-related processes.

It is worth noting that Experian are currently developing a set of persistent unique personal reference numbers that bring together the many and varied divisionally-based datasets held within the organisation. These could be appended to individuals, households or businesses and provide a unique code that identifies and distinguishes that 'unit', providing an obvious mechanism for merging diverse datasets. Within this process, at any stage the data can be anonymised, hence avoid confidentiality and data sensitivity issues.

Once developed, such an initiative could be replicated within other organisations, thus providing the opportunity to link these to other standard reference numbers such as Unique

Property Reference Number (part of National Land and Property Gazetteer).

The application of unique reference numbers and merging of datasets is considerably more effective if data is 'clean', such as names/address and postcode in recognised and consistent format. Data cleansing technology is widely used for example in address matching and geo-coding exercises.

Whilst ADIG is largely focussed on Departmental Administrative data (both structured and un-structured) data enrichment and value-added insight is also within the scope. The vast array of person, household and business oriented data, available from Experian for example could be appended to the analytical datasets (within the secure Data Hub), via the unique reference number procedure, thus providing greater insight for example into the socio-demographics, financial behaviour, channel engagement strategy and service expectations and requirements of that unique 'unit'.

Data Linking

Where data needs to be linked/integrated before analysis a key component underpinning the proposed approach is the capability within the data hub to combine variable quality data from disparate sources and provide meaningful insight to policy teams. Data sharing for transactional purposes tends to start with the question "What do we know about A Smith?" and then attempts to match records from different sources to this individual with an acceptable level of confidence. This provides a deep but relatively narrow understanding of a particular individual e.g.

A possible Data Linking process

1. The Department "data controllers" generate a random number for each case/individual in the datasets required that have already been 'cleaned' and quality assured within the Departments own Information Centres.
2. The data controllers then create data files containing only the case numbers and personal identifiers required for linking. For example the personal identifiers are taken as name, address, postcode and date of birth and any other variables deemed to be in the public domain.
3. These files are then sent to an independent third party processor. The processor knows only the generated numbers and identifiers necessary for the linking process. The processor then applies linking techniques to those personal identifiers to produce a set of matched case numbers which the data controllers can relate back to their original datasets.
4. All personal identifiers are deleted by the third party leaving the set of matched case numbers.
5. The data controllers produce datasets which retain the case numbers but which have removed all personal identifiers and any variables not required and from which personal information can be imputed.

NB. It will be necessary at this stage to ensure that variables of interest such as age, and a measure of region are maintained if required for research. This could mean departments computing extra variables at this stage such as age and the first part of an individual postcode (e.g. S10) which will be less disclosive if they are put into the dataset than retaining full date of birth and address.

6. The datasets are then given to the third party to link using the matched case

numbers. The resulting anonymised dataset, after appropriate disclosure testing to ensure no individual can be identified, is then analysed in a secure research environment, as pre-determined by the ethics process.

The above example of a data linking process presents a number of limitations for policy purposes, firstly it relies on forcing a 'same customer' or 'different customer' result for each record, this becomes less reliable where the data from different sources does not follow the same standards and data quality reduces resulting in increasing amounts of lost information. This is exactly the situation faced by government as a whole where biographic information such as surname and address can change frequently with notification to government often being patchy or non-existent. Secondly, attempting to resolve all records uniquely to an individual also fails to recognise that individuals rarely act in isolation and that social groups tend to influence behaviours and outcomes.

Entity Linking/Social Network Analysis

An approach to matching or clustering using 'entity linking' and 'social network analysis', provides a much richer and more flexible understanding of behaviours based on data from a range of sources, offers an important alternative to direct matching as outlined above. Such techniques have been used extensively by Detica in intelligence/security environments.

This approach first develops a complete understanding of the available data by building networks which capture all of the links between data items (for example names, addresses, identifiers, etc). In this way no data is discarded before analysis. The technique provides a more flexible foundation of data for analysis, for example by allowing a range of analysis dimensions such as individuals, households, locations and behaviour patterns to be performed from the same core data set. It also supports a range of analysis techniques from conventional clustering and spatial modelling through to advanced techniques such as social network analysis to identify and compare patterns of behaviour or outcomes.

Pattern matching across social networks provides one way of identifying target communities. By understanding the social networks of those known to fall within a certain area of risk, other similar networks can be identified and new service approaches can be developed to address those who are suffering through exclusion, or through earlier preventative intervention.

Example: HMRC has used network analysis in support of its tax compliance activities. By developing a better understanding of customer behaviours it has been able to assess the deterrent effect of different types of intervention. This has allowed researchers to evaluate enforcement policy and the effect of limited interventions in encouraging compliance across wider social networks.

Issues to be addressed in the Proof of Concept Pilot Study

The aim of the proposed pilot study (see Chapter 10 Way Forward) should be to undertake a proof of concept study of the benefits of creating a data hub for cross-Government data integration for policy research, analysis and decision making. It should (temporarily) draw down data from different Departments and agencies, for specific analytical purposes. The data joining options outlined previously should specifically be tested.

It is assumed, at this stage, that the pilot does not have to address the means by which

anonymised data sets are extracted from departmental information centre databases but focus upon how each dataset would need to be prepared for loading, albeit temporarily into a suitable DBMS. Metadata to describe the dataset would need to be added at this stage so as to facilitate its subsequent discovery, use, exploration of relationships between datasets and to allow straightforward and valid linkage.

While datasets may be technically linkable, in practice, the user still has choices to make regarding which datasets and what linkages are appropriate for a given analysis problem. Typically, a user might need to decide between different variable operationalisations (e.g., alternative social classifications) or, by using the metadata, to browse different datasets to discover what kinds of variables are available and what kinds of linkages are possible (i.e., what variables are in common or comparable: e.g., shared temporal or geographical dimensions, subject coverage, etc).

Section 6: International comparisons

Question 23.

Comments:

Question 24.

Comments:

Question 25.

Comments:

Question 26.

Comments:

Section 7: Additional questions

Question 27.

Comments: **ADIG Feasibility Report Chapter 5 pages 46 -49:**

As the threat landscape evolves and new ways of doing business and communicating emerge, security must evolve to become more business enabler than inhibitor. Security solutions must protect individuals and businesses at every contact giving them confidence that their information and interactions are protected however they are used. For organisations, this requires taking a more policy-driven, information-centric approach to security that is supported by a managed infrastructure.

The provision of appropriate security measures and controls will be essential in safeguarding data and ensuring that departmental data protection responsibilities are met. The Data Hub (as outlined in Chapter 3 – Proposed Approach) provides a central point of control and will be able to enforce suitable procedural and technical mechanisms for transferring data and managing its use. This will provide a greater level of consistency, transparency and control than is available through current point-to-point arrangements.

The hub will enable the introduction and enforcement of best practice controls combining technical and procedural measures, for example, the use of a single workflow system to

provide management and full auditability of all data requests and their fulfilment. It would also be able to establish preferred technical mechanisms for transferring data building on existing cross-government network infrastructure and data encryption capabilities. These measures will ensure that data is protected from unauthorised access, provide appropriate control and governance of all data movements, and remove the need for portable media (such as CDs) being physically moved between organisations.

In discussions with, and contributions from, Symantec, it was noted that where delivery of services is affected through a virtual organisation based on a “consortia” that sign up to a charter, one of the key issues will be how policy, procedure and related security measures can effectively be managed. Models for virtual organisations can be generalised into those that adopt hierarchical devolution of privileges and those that devolve based upon business function. “Supporting Decentralised Security focused Dynamic Virtual Organisations across the Grid”⁵ presents an approach that is based on a hierarchical devolution of rights and privileges, where no subordinate unit or person may have more rights than their parent authority.

Considering how large central organisations can work with and deliver services through remote agencies and organisations, over which it has no direct control, may be more closely reflected by the Viable System Model⁶ (VSM) as originated by Stafford Beer some twenty years ago. VSM presents complex organisational structures as a network rather than a hierarchy. The complexity is borne out of the relationships for the delivery of services rather than the individual organisations. The network of organisations conforms to recursivity, that is each has its own self organising and self regulatory characteristics. Within each organisation the service lines will each exhibit their own self organising and regulatory characteristics. These characteristics need to be exploited.

This model adopts the principle of granting privileges based on need to execute the organisations business. In many cases this is likely to be a devolved part of an organisational structure and possibly a part that is more susceptible to change or restructuring. Security strategy should address the detail of how delivery of the organisations services is achieved and the key roles in that delivery.

Virtual organisations may adopt a structure that the defining of roles and associated attribute assignment and access control should be the responsibility of a centralised Delegation Issuing Authority to ensure that security strategy and policy are being adhered to. The Issuing Authority should grant the rights to remote business units / agencies the responsibility of allocation of roles and ensuring that means of access and use are maintained to security policy. The Delegation Issuing Authority may also have to maintain sub-sets of role profiles to accommodate ‘local’ (business unit) restrictions and laws.

If this model can be effectively implemented, then the Authority need only monitor that policy is being used and act on breaches This approach may be considered to be more realistic as

⁵ Supporting Decentralised Security focused Dynamic Virtual Organisations across the Grid by R.O.Sinnot, D.W.Chadwick, J.Koetsier, O.Otenko, J.Watt & T.A.Nguyen of University of Glasgow, University of Edinburgh, University of Kent. <http://eprints.gla.ac.uk/3620/01/sinnott3620.pdf>

⁶ The Viable System Model as a Framework for Understanding Organizations by Raul Espejo and Antonia Gill <http://phrontis.com/vsm.htm>

it adopts a 'manage by exception' philosophy. Solutions exist that can enable security policy management of virtual organisations.

To support the development of the National Strategy for Data Resources for Research in the Social Sciences, the ESRC is seeking to develop a Secure Data Service. This service will provide controlled access to sensitive and/or disclosive personal or organisational information which cannot be released for research purposes under End User Licence or Special Licence conditions.

The ESRC currently funds (or co-funds with others) the collection of information from individuals and/or organisations and makes such data available to the research community via the Economic and Social Data Service (ESDS). Basic identifying information (names of people/organisations, addresses, detailed spatial identifiers) is always removed from such data prior to their deposit with the ESDS, in accordance with guarantees of confidentiality given to respondents and/or for legal and ethical reasons.

Many such datasets, particularly those which are longitudinal in nature, have now developed to the point where individuals or organisations could be identified by virtue of the detail available within the data that are collected for research purposes. In order to maintain guarantees of anonymity given to respondents, one approach to this problem of potential disclosure is to remove or aggregate variables in the dataset before making them available for research access via bodies such as the Economic and Social Data Service. Another approach is to make datasets available with more restrictive conditions governing their use for research purposes. These cover the security of the environment in which they are held, the status of the researcher requiring access and are backed up with severe penalties for breach of these conditions. This access regime, known as the Special Data Licence, has recently been introduced by the UK Data Archive, specifically for access to potentially disclosive datasets made available by the Office for National Statistics through the UK Data Archive.

There is a need for a third approach which, while more restrictive than the Special Data Licence, will permit researchers to carry out detailed work to link data and/or to create new analytical variables or to undertake analytical procedures which require access to detailed identifying information. Other examples include the development of linked data where the linked variables are disclosive and the provision of administrative data from government departments/agencies where the provider stipulates that such data cannot be held outside a secure environment.

The proposed ESRC Secure Data Service, will operate in a manner analogous with that developed by the Office for National Statistics Virtual Microdata Laboratory (see www.statistics.gov.uk/about/bdl/). Data placed within the service will be held on a secure server in a secure environment, together with a range of analytical tools and software suitable for data management, statistical analysis and data visualisation. Access to this environment will be managed via access and authentication procedures agreed between the service provider and the body responsible for guardianship of data to be placed within the secure environment. Remote access will be provided to researchers under terms and conditions to be agreed between data guardians and researchers, to be managed by the service provider. The secure environment must have no provision for unauthorised extraction of any data, including research outputs, by the researcher who has been granted

access. The service provider will operate methods for the release of research outputs to the researcher as agreed between the service provider and the data guardian.

Question 28.

Comments:

