Data Access Project
Department of Social Services

Final Report

This report has been prepared for the Australian Department of Social Services by Elizabeth Green and Felix Ritchie of Bristol Economic Analysis at the University of the West of England, Bristol.

This report has benefited from discussions with various interested parties in the Australian government and the research community, as well as respondents to an online survey; in particular we are grateful for the engagement of the Department itself for identifying a number of areas of confusion, ambiguity or factual error. All remaining errors and omissions are the responsibility of the authors.

The views expressed in this document are those of the authors and should not be taken to imply a policy position by the Department.

3rd July 2016

Executive summary

This report reviews the context and options available to the Department of Social Service (DSS) in developing its data access strategy, and proposes future developments based upon international practice. The structure of the report is separated into 4 parts.

Part I: context for the study

Part I provides the context for the report, considering current practices, law and the wider strategic framework. It notes that DSS already has a number of data access mechanisms in place. In addition, there are a number of off-the-shelf solutions currently operating in Australia, which may be relevant to DSS. Three typical DSS user groups are identified and defined based on their statistical skills and understanding of DSS; 'non-expert' (general public, journalists, non-PhD students), 'professional researchers' (Australia-based academics and PhDs) and 'other researchers'.

Part II: international experience

Considering first the operational perspective, the report uses the Five Safes structure to break down data access mechanisms into five separate components: the project purpose, the people, the settings, the level of detail in the data, and checks on any statistics produced. In respect of those elements:

Project purpose

- Data should be made available for research purposes if the expected benefit to society outweighs the potential loss of privacy for the individual; however, it is not clear that the claimed benefits are effectively realised. There is widespread agreement that data should only be used for statistical research and must be sanctioned by an appropriate authority. The design of the approval process can make a substantial difference to the effectiveness of data access, particularly if precedents or classes of project are used. Data access agreements should encompass physical, time, environmental and behavioural elements, instructing the user to best practice; they often include much more but it is not clear that this is justified or wise.

People

- Traditionally users have been treated as customers of the service that the data owner delivers. More recent models have raised awareness of the importance of understanding the psychology of users and, if possible, building a 'community of interest' between data owners and users. This has been demonstrated, albeit in a small number of situations, to deliver improved security and greater usefulness as lower costs. It does require a greater commitment to training both researchers and data owners, but this is increasingly seen as best practice.

Settings for data access solutions

- Distributing data is well established and relatively uncontroversial. Concerns do arise over users' ability to maintain their research environment appropriately, but there is little reliable evidence that this is a significant risk. For the most sensitive data the virtual RDC (vRDC) is almost universally seen as best practice; most vRDCs operate in very similar ways and most

hold fully detailed data with minimal de-identification carried out. Remote tabulation is a well-established technology, but Australia is currently leading the field in new developments. A number of organisations are now considering creating synthetic data for educational/training purposes.

Level of detail in the data

- This is a very mature field where off-the-shelf tools and a vast academic literature means that data can be created to any level of 'safety' desired. However, some recent authors have argued that the applications of the conceptual models is seriously flawed, and have demonstrated that a more evidential approach can generate both more secure and more user-friendly outcomes.

Checks on the output

- In the last ten years a new field of 'output statistical disclosure control' (OSDC) has developed, specifically to ensure that the results of research use of sensitive data do not breach confidentiality. vRDCs are more likely to use 'principles-based' OSDC, but this does require a commitment to training researchers. Distributed data solutions still tend to rely on sending users a set of instructions with the dataset. However, some modern training tools are being developed.

The report also considers the wider institutional and ethical concepts. This has an important influence on the effectiveness of any operational solution, and it can lead to very different outcomes. The report contrasts two approaches to data access planning: the 'traditional' model and the 'EDRU' model.

The traditional model is fundamentally defensive in nature; the focus is on the costs and risk to the data owner, and it assumes that the primary aim of any data access strategy is to prevent malicious misuse. This model therefore makes extensive use of worst-case scenarios and protection against hypothetical possibilities. The traditional model is default-closed; that is, it assumes that no access will be granted unless it can be proven to be safe.

The evidence-based, default-open, risk-managed, user-centred (EDRU) model reverses almost all of these precepts. It is evidence-based: hypothetical possibilities have little or no place in decision-making. It is default-open: data is assumed to be released, and the question for the data owner is how to manage confidentiality risks. It accepts that their data access choices are subjective and made in conditions of uncertainty, and that the relevant metric for costs and benefits is society, not the data owner; therefore, decision-making is a balance of subjective probabilities. Finally, it is user-centred, focusing on the usefulness of outputs rather than the risk to the data owner.

Breaches of confidentiality by the research community are extremely rare. It has been argued that the lack of breaches of confidentiality is evidence that the traditional model works well. As the traditional model has, until recently been the universal ethos for all data access solutions, this statement cannot be challenged. However, an increasing number of data access solutions adopting the EDRU ethos provides evidence that the traditional model has been over-cautious and missed opportunities to create synergy with the research community; as such it has not served society well.

This reports concludes that the EDRU ethos provides a more sustainable world-view and, on the limited evidence available, is more likely to provide a secure and useful data access solution; it also seems better suited to exploit the gains from increased data access by engaging with researchers more. The report acknowledges that this is very much a minority view, but a growing one which seems likely to become much more significant.

Part III: mapping users and solutions for DSS

Following the broad identification of user groups in Part I, and international understanding of good management of the user community in Part II, this section describes user groups in detail and the proposed data access solutions mapped to each user group:

- Remote tabulation would serve the non-expert user group well; in the longer term, synthetic data may have value in educational contexts.
- Expert researchers need access to micro level data; the full data with minimal de-identification should be delivered through vRDCs, but Scientific Use Files (SUFs) with a lower level of detail in the data should also be part of DSS' strategy.
- For researchers working abroad or in private sector organisations, our recommendation is that the vRDC/SUF combination also suits this market.

Institutional factors can have a strong influence on the effectiveness of operational decisions, and this report covers them separately. The recommendations made are based on the EDRU ethos and so many of them are at the edge of current knowledge about how to manage data access effectively; some of them may be seen as very radical. However, based on the experience of the last ten years, we expect that this approach will become more important, not less, and we believe this is a rare opportunity for DSS to lead such developments.

Overarching issues which need to be addressed at the institutional level are:

- Common operational issues: standardisation of both institutional and personal access agreements; classification of data types for mapping onto technical solutions; training institutional signatories about the data community and their role within it; developing training programs for researchers.
- Attitudes: formal statement of intent for 'default-open' institutions, clarification of staff roles and duties within data accessibility, and a modern approach to risk management and the likelihood of human error.
- Stakeholder management: developing a strong relationship with data providers and researchers to share findings, particularly in terms of developing whole-of-government solutions to data access in Australia.
- Perceptions: early discussion with privacy campaigners and with users; the development of an advisory board with representatives from different institutions, the general public, journalists and privacy campaigners to discuss data sharing practice.

The report notes that DSS has already gone a considerable way down this route, with high-profile statements about attitude and a draft Data Policy which conveys much of the EDRU ethos. These developments can be usefully augmented in a number of ways.

Part IV Implementation

Part IV provides a roadmap for future development, based on the three user groups (plus overarching issues), considered over the short, medium and long term. The timing is based on both the natural order of changes, and the feasibility of targets. The list of proposed actions is long, but the elements should be complementary. The programme is also ambitious in its scope, but we believe this is feasible and will provide DSS with a solid base for an enviably coherent long-term strategic direction.

The strategic roadmap provides short-term and long-term guidance on the following areas:

- The creation of synthetic data files and associated metadata for Public Use Files
- Development of TableBuilder for the dissemination of SUF and metadata
- Simplifying the provision of SUFs and clarifying the access arrangements
- Scoping a potential pilot for a virtual RDC for hosting detailed datasets with minimal de-identification and
- Institutional change promoting a positive 'default open' attitude amongst users, institutions, stakeholders and the public.

Definitions

Definitions used in this report are as follows (note, these may differ slightly from similar definitions given elsewhere):

Types of data

- **Microdata**: individual unit records about a person or organisation, such as information collected from surveys or administrative data
- **Raw data**: the source data collected by the Department
- **Identified data**: data which includes information allowing the recipient to be directly known just from one or two fields in the data (such as name, or social security number)
- **De-identified data**: data which includes sufficient detail to allow the data subject to be identified, but only with effort and with less certainty (for example, a combination of gender, age, type of employer, salary range and disability status)
- **Anonymous data**: data which does not include sufficient detail to allow the data subject to be identified, under any reasonable conditions
- **Confidentialisation**: the act of reducing the likelihood or harm of re-identification by reducing detail or perturbing the dataset
- **Confidential data**: data which, for legal and/or ethical reasons, must not be made available to anyone who does not have appropriate authorisation
- **Sensitive data**: data where release to an unauthorised person is likely to cause non-negligible harm or distress to the data subject; for the purposes of this report, we assume that all sensitive data is also confidential
- **Protected information**: a specific term used in various ways in Australian legislation, that includes referring to information 'about a person' but not necessarily identifying a person; anonymous data can still be 'protected information' under Australian law

Disclosure and breaches

- **Unauthorised disclosure**: the unauthorised release of information about an identified data subject
- **Statistical disclosure control (SDC)**: applying statistical measures to (a) determine if there is a substantive risk of unauthorised disclosure in a dataset or publication, and (b) make changes to the data or publication to reduce that risk
- **Input SDC**: the application of SDC methods to raw data to reduce data risk before it is released to the users
- **Output SDC (OSDC)**: the application of SDC methods to potential publications after the analysis has been carried out, to guard against residual risk
- **Breach of confidentiality**: the release of identified or de-identified data to an unauthorised system, environment or person; a breach of confidentiality may not mean a disclosure as it will depend on the circumstances
- **Breach of procedure:** failure to follow appropriate operating procedures, irrespective of whether a breach of confidentiality occurs

Forms of access:

- **Distributed data:** sending microdata to users under licence, to analyse on their own machines
- **Distributed access:** restricting the physical location of the data, but allowing users in others locations to carry out analysis and take retain statistical results (but not microdata)
- **Table server:** a system which allows users to generate their own tables from the data flexibly, but without seeing the source data; a form of distributed access
- **Remote job server:** a system allowing a range of complex analyses to be carried out, not just tabulations, without seeing the source data; a table server is a remote job server which has only one function
- **Remote access:** a system which allows users to 'see' and manipulate the source data
- **Research data centre (RDC):** a restricted access facility where users can manipulate the source data without restriction as if on their own computers; but the environment is made secure so that users cannot bring information into or take data out of the facility without approval, and additional services (such as internet access) are normally very restricted; typically provided by **on-site access**, where the facility is hosted on the organisation's premises
- **Virtual RDC or Remote RDC (vRDC):** an RDC where technology is used to provide equivalent security to a physical site and to separate the RDC from the actual location of the data, such as SURE
- **Public use file (PUF):** data file without restrictions on use or onward access
- **Scientific use file (SUF):** data file which retains some non-negligible confidentiality risk and so therefore has circulation restricted to authorised users for specific research purposes
- **Secure use file (SecUF):** data file which contains non-negligible confidential information therefore circulation and use is restricted to authorised users for specific purposes

Part I Project context

1. Context, aims and objectives for the Data Access Project

1.1 Department of Social Services data access strategy

The Department of Social Services (DSS) "aspire[s] to be Australia's pre-eminent social policy agency. [The Department's] mission is to improve the lifetime wellbeing of people and families in Australia"[1]. One part of that mission is providing external users with access to statistics and raw data collected by DSS, so that third parties can potentially enhance DSS activities with their own research, and generally boost service provision from Federal, state and local council programs and non-government services.

The public benefits expected to arise from this project are:

- greater ability of the general public to identify useful information specific to their circumstances, such as small area data
- improved capability for professional researchers using detailed information to produce policy-relevant analyses
- streamlining of DSS' data access procedures, including integration with the wider Australian Government strategy
- better understanding of user needs by DSS
- methodological commentary and quality control arising from an expanded pool of expert users and
- greater understanding and acceptance of data collection and use by DSS.

This strategy is consistent with *Australian Government Public Data Policy* (AGPDP)[2] and associated report *Public Sector Data Management* (PSDM)[3], which provide high-level guidance on how public sector data resources are to be managed for the benefit of the community. DSS is investigating how to best implement the AGPDP, particularly in respect of access to confidential or sensitive microdata. Some practical elements of this investigation are already in place or are being piloted, by the Department on its own and in collaboration with other agencies (such as the Multi-Agency Data Integration Project).

A key element of this strategy is developing a Trusted User Model to manage access to confidential data. This has proven difficult for governments in the past, not just in Australia but around the world. Solutions are typically over-cautious and fail to encourage users to engage positively with the service. The AGPDP and PSDM propose a conceptual framework which challenges traditional perspectives on data access, and encourages data access to develop along 'modern' lines: the evidence-based, default-open, risk-managed, user-centred (EDRU) model which is increasingly seen as a more effective and cost-effective strategy for delivering data access. The Australian Bureau of

[1] DSS website "About the Department". https://www.dss.gov.au/about-the-department.
[2] Australian Government (2015) *Australian Government Public Data Policy*. Office of the Prime Minister and Cabinet. https://www.dpmc.gov.au/sites/default/files/publications/aust_govt_public_data_policy_statement_1.pdf
[3] Australian Government (2015) *Public Sector Data Management*. Office of the Prime Minister and Cabinet. https://www.dpmc.gov.au/sites/default/files/publications/public_sector_data_mgt_project.pdf

Statistics has already redesigned its data access strategy along EDRU lines, and DSS has already made moves towards adopting a similar approach.

The Trusted User Model is being developed to provide a platform for providing access to relevant administrative data to "trusted" users. DSS is committed to developing a Trusted User Model in alignment with the AGPDP alongside coordinating similar activities across organisations; to help develop a sustainable system.

1.2 Objectives of the project

The Department has commissioned the University of the West of England to:

- Review international good practice in data access
- Consider how the needs of different user groups accessing DSS data might be met by alternative (technical and organisational) solutions; particular attention should be paid to institutional relationships
- Identify where tools and practices already exist which DSS uses or may be able to exploit; and where gaps in policy or practice exist
- Recommend a development path, focusing particularly on access to confidential microdata and the development of the Trusted User Model and
- Advise the Department with practical proposals on how such a development path might be implemented, including timing and priorities.

While the focus is on meeting the needs of DSS and its data users, consideration should also be given to integration with other Australian initiatives, to identify potential areas for collaboration and avoid creating data solution silos in different Departments.

This report considers how to make best use of the microdata that DSS holds; that is, record level data. Access to this data allows users to make their own analyses of the data. At one end of the spectrum, effective use of the microdata this might be achieved by allowing journalists to create their own tabulations of the data, but without ever seeing the source data. At the other end, expert statistical analysts may have unrestricted access to identifiable microdata (subject to legislation such as the *Privacy Act 1988* and security requirements imposed by DSS when receiving data), with only their publications being subject to confidentiality checks. Standard statistical analyses produced by DSS or bespoke statistical services provided by DSS are not considered in this report.

1.3 Structure of the report

The remainder of Part I summarises the current position in respect of DSS's interests and activities.

Part II reviews international practice, both on operational issues and in terms of wider institutional issues. It then introduces the tripartite structure (operational issues, institutional matters, public relations) which will be used for the proposals and recommendations.

Part III reviews the specific combinations of options and draft provisional recommendations. It also considers the wider institutional context, and where DSS may need attitudinal statements.

Finally, Part IV proposes specific actions and a timetable to achieve the recommendations in Part III.

2. Relevant considerations

2.1 Current facilities and solutions

2.1.1 DSS paths to access

General access to data

DSS data is provided widely to universities and other institutions on request subject to meeting legislative requirements under social security law, family assistance law etc. and the *Privacy Act 1988*. The information is typically provided after a Public Interest Certificate (PIC) or Protected Information Disclosure (PID) has been approved under the legislation which required the collection of the data for Government purposes. The data is supplied to the researcher with restrictions on what the researcher can do with it. We will refer to these datasets as Scientific Use Files (SUFs), to contrast them with Public Use Files (PUFs) which are unrestricted. DSS does not at present produce PUFs. A third type of data is a Secure Use File (SecUF), a detailed data file made available to researchers remotely through facilities controlled by the data owner. Currently DSS does not produce SecUFs.

To the casual observer, DSS does not appear to make its data available for third-party analysis. There is no statement on the website about access to data, a general data strategy, or to indicate that data might be available.

National Centre for Longitudinal Data

An exception to the general invisibility of data is DSS longitudinal data resource. DSS created and maintains several longitudinal studies, the earliest from 2001 and the most recent from 2013. These are managed by DSS' National Centre for Longitudinal Data (NCLD[4]). The microdata is distributed to researchers through a transparent and apparently long-established process, described in detail on DSS website.

The key features of this distributed data approach relevant to this project are:

- Data are released as SUFs with two levels of confidentialisation: 'General Release' (lower risk) and 'Unconfidentialised' (higher risk).
- Both datasets have to be kept on systems with defined security standards; the Unconfidentialised datasets require additional electronic and physical locks; both are subject to DSS spot-checks with three hours' notice.
- Users can apply for an individual access agreement.
- Alternatively, they can apply through their organisation's licensing agreement with DSS (for some categories such as Honours students this is the only route).
- When an organisation has an agreement with DSS, a data manager needs to be appointed at the organisation to take responsibility for the implementation of procedures.
- Eligible users are Australian academics and students, research organisations, government users; all research must be for public benefit.

[4] https://www.dss.gov.au/about-the-department/national-centre-for-longitudinal-data

- The same eligibility criteria apply to overseas researchers, but they are only able to access the lower risk General Release files, not the Unconfidentialised files.
- The finest detail in the Unconfidentialised files is date of birth, which is very detailed by international standards.
- There have been a small (single-digit) number of breaches of procedure from using these data since 2003; mistakes and poor handling practices are counted as breaches; it is possible that some of these might have led to breaches of confidentiality, but the impact of any breaches appears to have been contained.

Development projects

DSS is participating in two relevant development projects.

The Multi-Agency Data Integration Project (MADIP) is a collaboration between DSS, the Australian Taxation Office, the Department of Health, the Department of Human Services and the Australian Bureau of Statistics (ABS), with the latter leading the project. This project shares data between the agencies to allow personal and health data to be combined for policy analysis. The Trusted User Model being developed for that project has a significant potential overlap with this project; hence developments in MADIP will be important for recommendations on this project. However, ABS is already committed to the EDRU model of data access, and so differences across projects should be a matter of degree rather than principles.

DSS is developing its own Trusted User Model project. This project will provide research access to the income support dataset via a virtual Research Data Centre (vRDC), the Secure Unified Research Environment (SURE) of the Sax Institute. Proof of concept in terms of physical access to systems and agreement on approval of projects is underway, with user agreements based on personal contracts. The ambition is that the final Trusted User Model will reflect the EDRU model and, where possible, be compatible with developments across government. The findings of the Data Access Project will directly feed into the post-pilot implementation of the Trusted User Model project, and so this report will pay particular attention to the management of vRDCs.

2.1.2 Data access resources within or accessible to the Australian Government

At this stage, there are several data access systems in Australia which could either be exploited or used as precedents for DSS solutions.

Distributed data solutions are:

- **NCLD**, already in operation at DSS for research access to DSS longitudinal studies
- **Confidentialised Unit Record Files or CURFs**, the ABS' microdata files for distribution and
- **data.gov.au**, the Australian Government website which facilitates the distribution of data.

On-site solution is:

- **In-postings,** used by the ABS to provide access to an on-site RDC (the Australian Data Lab or ADL) for government researchers.

Distributed access solutions are:

- **TableBuilder**, a tabulation tool designed by SpaceTime Research for the ABS, providing real-time statistical disclosure control before generating the tables; it has generated considerable interest amongst other statistical institutes
- **Secure Unified Research Environment (SURE),** a vRDC run by the Sax Institute which appears to run along similar lines to the dominant model in the rest of the world
- **Virtual Microdata Laboratory**, a remote access version of the Australian Data Lab under development and piloting by the ABS and
- **Remote Access Data Laboratory (RADL),** a remote-job server for more detailed CURFs, run by the ABS.

In terms of the institutional framework, there are examples of both institutional and personal contracts in these systems. The RADL provides some good examples of where problems can arise, particularly when the relationship between the researchers and the data provider is not strongly positive.

2.2 User groups

DSS has identified a number of user groups who need to be considered for this project. These are categorised in Table 1, along with provisional solutions to help in considering alternative solutions. 'Knowledge' is classified by whether users have a good understanding of DSS activities, and whether they have strong statistical skills. The possibility of contracting is based on the practicality, credibility, meaningfulness to the user, and difficulty of enforcement.

Table 1 Summary of DSS user groups

Group	Interest	Assumed knowledge and statistical skill	Contract possible?		Potential solutions
			Personal	Institutional	
General public	Simple tabulations	Very little	No	No	Remote tabulation
Journalists	Complex tabulations	Good knowledge of DSS, limited statistics	Possibly	Possibly	Remote tabulation
Honours/Masters students	Microdata for analysis	Little knowledge of DSS, basic statistics	Possibly	No	Synthetic data, remote tabulation
PhDs	Microdata for analysis	Good knowledge of DSS, good statistics	Yes	Possibly	Distributed microdata, vRDC
Academics (university and research institutes)	Microdata for analysis	Good knowledge of DSS, good statistics	Yes	Yes	Distributed microdata, vRDC
Government researchers	Tabulations, microdata	Good knowledge of DSS, good statistics	Yes	Yes	Distributed microdata, vRDC
Private sector researchers	Tabulations, microdata	Good knowledge of DSS, good statistics	Possibly	Yes	Remote tabulation, distributed microdata
Foreign researchers	Tabulations, microdata	Good knowledge of DSS, good statistics	Possibly	Possibly	Remote tabulation, distributed microdata

On this basis, the report classifies these further as 'non-expert' (general public, journalists, non-PhD students), 'professional researchers' (Australia-based academics and PhDs) and 'other researchers'.

2.3 Legal framework

2.3.1 Laws covering access

DSS data is covered by several laws and determinations reflecting the range of activities carried out by the department. This part discusses laws which discuss the governance of data collection. Although the premise of the laws are similar, there are differences within the acts. The laws which predominately affect data governance are briefly outlined and discussed below.

Social Security (Administration) Act 1999 (SS(A) Act)

This Act addresses 'protected information'- information which is about a person but which does not necessarily identify a person; hence even anonymous microdata is still covered. The gateway for research access is provided in s202(2C), which allows research use by individuals in "matters of relevance to a Department". S208(1)(a) offers an alternative approach as the Secretary of the Department may provide the information "…to such persons and for such purposes as the Secretary determines", by providing a 'Public Interest Certificate'. Moreover, the Secretary can develop guidelines to settle matters of access in both cases and classes of cases. S2081)(a) appears to be the clause under which research data access is given. However, the guidelines specified in relation to the access clause (defined in the Social Security (Public Interest Certificate Guidelines) (DSS) Determination 2015) make clear that, for the Certificate to be granted, research still has to be relevant to the department.

The Act also allows DSS to have an institutional arrangement with another government department, where data is shared for the purposes of that department.

Note that, in case of breaches of confidentiality, the intention of the miscreant has to be taken into consideration (s204). On the other hand, asking for protected information to be provided unlawfully is an offence. In other words, the law appears to provide sensible protection against mistakes whilst explicitly criminalising the "can I borrow your data without authority" attitude. Both of these are useful messages to send to researchers.

A New Tax System (Family Assistance) (Administration) Act 1999

This provides research access to data under very similar wording (s167) to the SS(A) Act s207. Again, the Secretary is allowed to approve guidelines on access and classes of access. The *Family Assistance (Public Interest Certificate Guidelines) Determination 2015* provides more detail on what will count as sufficiently important to be 'in the public interest'.

This Act does not directly provide for research access to data (unlike s202(2C) of the above SS(A)Act), but it does have the same references to error, strict liability, and encouraging others to break the law.

Privacy Act 1988

The privacy act extends to all data collected by the Australian government and affects much data collected by the private sector including credit, health and insurance information.

2.3.2 Ethical approval

Ethical approval is not required by DSS where administrative data already exists or is part of routine data collecting related to an application for government support (e.g. when data is collected when someone applies for a pension). Ethical approval is however required when conducting a new survey or linking new data which obtains further data related to existing data.

2.3.3 Organisations working in data privacy

As well as legal constraints on data access, and ethical constraints raised by the Department, the success of data access initiatives may be affected by the support or opposition of organisations with an interest in data privacy.

On the anti-access front, one organisation is the Australian Privacy Foundation, who in the past has made some extreme statements about data access (for example over the ABS' plans to retain name and addresses on the Census data)[5].

On the more pro-access side, the Menzies Foundation has been commissioning work looking at public concerns and preferences over linking government and health data.[6]

Finally, the Office of the Australian Information Commissioner (OAIC) may be a powerful ally or a major limit on activities. The EDRU approach is likely to be welcomed by the OAIC because it tends to produce demonstrably more secure solutions at lower cost; this has been the experience in other countries.

In all of these cases the lesson from international experience is that early and meaningful engagement with interested parties matters as much as compliance with legislative requirements.

2.4 Wider strategic framework

In terms of data access, it is important to build on previous developments. First, this enables one to learn from others. Second, and more importantly in government, following precedents enables difficult decisions to be taken more easily. For example, the fact that the PIC/PID requests and the NCLD appear to have been securely providing very detailed data for several years without confidentiality problems suggests that (a) potential problems are manageable in practice, and (b) DSS has the experience manage risk effectively. A third reason for looking to other Departments is that several bodies are working through the same issues, all perhaps driven by the AGPDP. This provides a unique opportunity to consider whole-of-government approaches. While exactly the same solution may not suit all parties, there should be enough commonality to avoid re-inventing wheels, or building data sharing systems on principles or organisational attitudes that are incompatible.

Therefore in section we consider systems in existence as well as potentially useful precedents.

[5] 'ABS slammed for breach of trust over 'intrusive' 2016 Census data matching plan'. *Financial Review* (2016), 9th March
[6] Menzies Foundation. (2013). *Public Support of Data Linkage for Better Health*. http://www.menziesfoundation.org.au/pdf/Data%20Linkage_16aug13/Menzies%20Foundation_Public%20support%20for%20data-based%20research.pdf

Since the beginning of 2015, the ABS has developed strategic data management plans which bear substantial similarities to the needs of DSS. Traditionally the ABS was seen by statistical organisations in other countries as one of the more conservative bodies, but the strategic decision to adopt the EDRU approach and use the Five Safes (see below) as an organising framework means that the ABS is currently, in our opinion, a world leader in modern approaches to data access. This makes the ABS approach a useful precedent for DSS' data access strategy, as well as a potential partner in developing a whole-of-government approach. The change in perspective has not necessarily required wholesale changes in the specific solutions (for example, distributed data solutions are essentially unchanged), but has radically changed the decision-making process and the way different elements of the data access strategy knit together[7].

In a process linked to the Policy Statement, the Australian Productivity Commission in undertaking an inquiry into the collection and dissemination of data, including research strategy, across all levels of government. Preliminary responses are due by the end of July 2016, after this project has completed. Although the Productivity Commission (PC) primarily deals with data relating to businesses rather than individuals, and so the stakeholder group is different, information on the PC's perception of data access may usefully enable DSS to start building a wider coalition of knowledge.

In November 2015 the Department of the Prime Minister and Cabinet, sponsors of the AGPDP, established a 'Public Data Branch' which 'aims to maximise the economic, social and environmental benefits from the use of data whilst maintaining a strong focus on and commitment to privacy and security'[8]. At present this branch appears to be mainly concerned with making better use of open (unrestricted) data, but this branch also has an interest in cross-government developments and coordination of confidential data access.

Finally, in 2014 the National Health and Medical Research Council produced a set of 'Principles' of data access[9]. Although pre-dating the AGPDP, the broad ideas are very similar to the subsequent AGPDP, and have already been suggested as a possible model for adoption by other bodies[10].

[7] A similar process took place in the UK Office for National Statistics in 2011: a Data Access Policy was produced which changed systems very little in practice, but put all of ONS' procedures into one conceptual framework and showed how this could be integrated with legal obligations to develop specific guidelines for future changes. See ONS (2011) *Data Access Policy*.

[8] https://www.dpmc.gov.au/pmc/about-pmc/core-priorities/public-data-branch-within-dpmc

[9] National Health and Medical Research Council (2014) *Draft Principles for accessing publicly funded data* https://consultations.nhmrc.gov.au/files/consultations/drafts/draftprinciplesaccessingpubliclyfundeddata141209.pdf

[10] The Wellcome Trust (2015*) Enabling data linkage for better public health research*. https://wellcome.ac.uk/sites/default/files/enabling-data-linkage-to-maximise-value-of-public-health-research-data-phrdf-mar15.pdf

Part II International practice

3. Introduction

This part reviews international practice with the aim of identifying relevant precedents, experiences, and solutions for DSS.

Note that the aim is to 'review' rather than 'identify best practice'. This is an evolving field, and while there is some agreement on current best practice, there is more agreement on what works and what doesn't work in different environments. Hence, for example, Eurostat uses the phrase 'Expert advice is that…' rather than 'Best practice is…' This seems to us more helpful then trying to identify a gold standard, particularly when the preferred outcome is likely to be sensitive to the specific environment.

3.1 Sources of information

Practice and theory in access to government data is dominated by National Statistical Institutes (NSIs) generating personal (non-health) and business data, such as the Office for National Statistics in the UK, the Census Bureau in the US, and the Australian Bureau of Statistics. Whilst other government bodies also release data, NSIs are often influential in setting the standards under which government bodies release data. Almost all of the literature on statistical aspects of data access is targeted at and/or sponsored by NSIs; and development of data access solutions is mostly driven by NSIs. In some countries data archives or research institutes have played an important role in improving access to data, but these work closely with NSIs when dealing with government data.

NSIs also generally produce multiple types of data for analysis, and so may try to develop coherent data access strategies across a range of customer types. This is the relevant model for DSS to follow. Hence, although the discussion below does make use of examples from other government bodies, much of the discussion references NSIs.

Other than NSIs, most discussion about data access arises from epidemiology/public health. Research in this field largely focuses on three topics: data linking; consent; and public perceptions of data use. The first two topics are not directly relevant; they were reviewed in the 2015 Public Health Research Data Forum report *Enabling Data Linkage to Maximise the Value of Public Health Research Data*[11], and will not be covered here in detail. That report noted that public health was lagging significantly behind social sciences in practical matters; that is, most of the activity in public health is focused on ethical issues, and the practicalities of data access (alternative technical solutions, researcher management, stakeholder management, risk evaluation) do not, generally, reflect the advances made in access to personal and business data this century.

Public health is much more advanced then social science in the understanding of public perceptions of data access. Important studies have shown that public opinion is typically hostile by default to government data sharing but very amenable to persuasion; and the successful management of public opinion is, in some ways, more important than any technical solution for providing access.

[11] Wellcome Trust (2015), ibid.

In summary, the following review draws mainly on the experience and perspectives of NSIs who drive much of the work in this area, supported by examples from other government departments providing personal microdata, data archives and research institutes, and public health bodies.

It should also be noted that the evidence base is sparse. Breaches of confidentiality amongst research users of government data are exceedingly rare; and almost every government release strategy has the same, defensive, conceptual framework. As a result, identifying cause and effect in public data protection is very difficult, and relies heavily on expert opinion.

Much of the recent thinking about effective data management has come from the vRDC world, where different data strategies have been tried; there are also a handful of examples from the world of distributed data. Evidence does not show that breaches of confidentiality have been prevented – this is almost impossible to show – but that the perceived likelihood of a breach has been reduced, while user value has not been reduced and costs have not increased. As a result, when the report comments 'evidence shows…' this is a shorthand for 'what little empirical evidence there is, is supportive of…'

It is slightly easier to determine what is perceived to be 'good practice', and where there are differences between schools of thought. Note however that these perceptions are also largely unevidenced.

3.2 Structure

The second section in this part looks at the practical elements of data access solution, following the 'Five Safes' structure widely used for analysing data access solutions. This breaks down complex solutions into five independent but linked questions around the managerial and statistical components, enabling one part of the data puzzle to be addressed at a time. These are covered in Section 4.

Such practical information needs to be put in the institutional context. One of the great changes in recent years has been the realisation that the effectiveness of any practical implementation is highly dependent upon the institutional attitudes that pervade the organisation: whether the perspective is default-open or default-closed, whether worst-case scenario planning should be used, what role do users play in the decision-making process, and so on.

Allied to this is the wider topic of managing public expectations; this may seem like an operational issues – and in many cases is so – but increasingly it is recognised that pro-active PR can bring significant benefits, and to be effectively pro-active requires an accommodating organisational stance.

Hence, Section 5 considers all of these overarching issues. It differs from the second section where there is broad agreement on what works. There is much less agreement on institutional perspectives, largely because it has not been addressed, and so this section has less of an evidence base to work with.

Finally in this Part II, Section 6 draws together the operational aspects, internal perceptions, and external perceptions to provide an overview the strategic planning process. This provides the framework for the next two Parts looking at specific solutions.

4. Components of a data access solution

4.1 The five safes[12]

The Five Safes is a framework for organising thinking about data access. The basic premise of the framework is that data access can be seen as a set of five 'risk dimensions': safe projects, safe people, safe data, safe settings, safe outputs. Each dimension provokes a question about access:

Safe projects	Is this use of the data appropriate?
Safe people	Can the researchers be trusted to use it in an appropriate manner?
Safe data	Is there a disclosure risk in the data itself?
Safe settings	Does the access facility limit unauthorised use?
Safe outputs	Are the statistical results non-disclosive?

These dimensions embody a range of values: 'safety' is a measure, not a state. For example, 'safe data' is the dimension under which the safety of the data is being assessed; it does not mean that the data is non-disclosive. Nor does it necessarily specify how the dimensions should be calibrated. 'Safe data' could be classified using a statistical model of re-identification risk, or a much more subjective scale, from 'very low' to 'very high'. The point is that the user has some idea of 'more safe data' and 'less safe data'.

Any data access solution needs to consider all five dimensions (even if simply to note that a particular dimension is not relevant), which is done in Part III of this report. However, each element can be reviewed independently for risk characteristics and evidence of appropriate practice; this is how we use the framework in this section, evaluating each of the individual elements.

4.2 Settings: options for access

There are four options for providing access to data:

1. Secure on-site research facilities (research data centres or RDCs)
2. Distributed data (licensing)
3. Distributed analysis, broken down into
    a. Remote tabulation
    b. Remote job servers
    c. Remote or virtual RDCs
    d. Analysis services and
4. Synthetic data.

[12] For a detailed discussion of the Five Safes, see Desai T., Ritchie F., and Welpton R. (2016) *The Five Safes: designing data access for research*. Working papers in Economics no. 1601, University of the West of England, Bristol. January

The first option relates to access which requires the user to be in the same physical place as the data, in a facility controlled by the data owner. As these are usually implemented nowadays by virtualised technology (and so the limitation to be in a physical research facility is the choice of the data owner) we will consider these to be a subset of option 3, distributed analysis.

4.2.1 Distributed data

Sending data to users has been the dominant solution in social science and epidemiology for providing access to research data for most of the last fifty years. As such, current good practice is well-established and uncontroversial:

- Determine whether the data to be released is a 'public use file' (PUF) or 'scientific use file' (SUF); the former can be disseminated without significant disclosure risk, whereas the latter are assumed to retain some non-negligible risk and so have access limited to approved users
- For SUFs, validate the identity of the user
- For SUFs, require a commitment to a level of IT security where the data will be held and used
- Distribute the data to the user (via a physical medium or download); for SUFs an acknowledgement of receipt should be required
- For SUFs, the data may also be encrypted; if so, the encryption key should be sent by an alternative medium (such as SMS or email) once receipt is acknowledged and
- For SUFs, provide advice on management of the data and the publication of outputs.

However in practice, not all of these steps are followed. For example it is rare but not unheard of for both data and encryption keys to be sent by email to the same address. Thankfully incidence of this has become increasingly rare practice with no major data-holding organisations presently making this mistake.

The advantage of distributing the data is that the data owner can concentrate on a specific statistical problem (making the data safe enough for this user group) and not on the need to think about the research process. If a fixed set of data files are produced and made available to multiple users, then the cost of that production can be spread over many units; this is the case for PUFs and for some, but not all, SUFs.

The concern that arises from distributed data is the lack of control of the distributing organisations in respect of SUFs. These arise from concerns about storage, appropriate use, and unauthorised sharing. For example, users of data downloaded from the UK Data Service under the End User Licence (EUL) are merely required [edited T&Cs][13]:

> *to use the data in accordance with the EUL and to notify the UK Data Service of any breach you are aware of [summary text, condition 1]*
>
> *to give access to the data collections only to registered users with a registered use (who have accepted the terms and conditions, including any relevant further conditions) [condition 6]*

[13] Edited extract. Full text: http://ukdataservice.ac.uk/media/455131/cd137-enduserlicence.pdf

to ensure that the means of access to the data (such as passwords) are kept secure and not disclosed to anyone else [condition 7]

Most distributed data providers have similar terms and conditions.

There seems to be very little evidence about how well such T&Cs are followed, but there is a general impression amongst data managers with experience of universities that:

- academic researchers do mean stick to the purposes for which they applied for the data, but projects may 'drift' over time
- there is a strong suspicion that academics do not delete the data after project completion, as 'research' is never complete
- disclosure control of outputs is not applied effectively in academia and
- academics are over-confident in their own abilities to manage data securely.

In contrast, government researchers appear to place their trust in government IT systems and seem more inclined to follow rules than academics. Finally, for both government and academic researchers, mistakes are thought to be more common than taking deliberate actions to get round restrictions on use.

These perceptions apply to SUFs where acquiring them is relatively easy, for example by download after registration as a verified user. There is some support for the idea that, where the effort of acquiring the data is larger, adherence to the rules is likely to be higher as the researchers have a stronger incentive to protect their interests. One team of researchers argued that the slowness of the Eurostat application process provides protection against frivolous or casual access to SUFs[14]. In France prior to the introduction of SecUFs in 2011, researchers wishing to get access to business data SUFs were required to travel to Paris and make a formal presentation to an ethical committee. In the UK, Census microdata distributed to the universities of Southampton and Manchester required certification of IT systems and processes to government secure standards with impromptu inspections possible.

Unfortunately, these perceptions are based upon evidence that is ad hoc and anecdotal. We are not aware of any systematic attempt to quantify how well rules are followed for distributed data. For some secondary organisations hosting very sensitive data, or significant amounts of data for onward redistribution (such as the UK Data Archive), periodic inspections are feasible, but this is not generally the case; for example, the UK Data Archive receives roughly 1,000 new registrations and 400 re-registrations for its services each month. Most evidence comes from informal visits or discussions with university IT colleagues, which is the source for the idea that mistakes are much more common than deliberate infractions. The general belief is that researchers try to do the right thing; but that they do not, in general, try very hard.

An important qualification is that almost no evidence, anecdotal or otherwise, has come to light which suggested deliberate attempts to re-identify research data. We explore the very small number of exceptions in the section on 'Safe People'.

[14] Hafner H.-P., Lenz R. and Ritchie F. (2016) "User-centred threat identification for anonymized microdata". Forthcoming.

Table 2 Distributed data: summary of advantages and disadvantages

	User	Data owner
Advantages	• Access in local environment • Ability to use own analytical tools	• Responsibilities limited to creating an appropriate dataset for the environment, and possibly providing guidelines to researchers • Fixed or semi-fixed costs of producing files can be spread out over many researchers
Disadvantages	• Less detail	• Cost of producing PUFs • Concerns over appropriate usage of SUFs

4.2.2 Distributed analysis 1: online or remote tabulation

Remote tabulation tools are designed to allow users to create their own tabulations of the data, rather than relying on the data owner's choice of tables or bespoke tabulations. As well as data tables, homologous tools can produce geographical images or time series. What distinguishes these tools is that the output statistic is a simple linear value (sum, mean, index) broken down by categories under the control of the user.

The value to the user of remote tabulation is twofold. First, it allows the user to have data presented in a useful form to his or her demands without needing to manipulate microdata. Second, it allows the data owner to present results from data which may be confidential and not suitable for release as microdata, but which nevertheless can produce secure tabulations.

There are many tools for displaying data using non-confidential data, but relatively few tools which can produce secure tabulations. Confidentiality is ensured in one of two basic ways; restricting the input or restricting the output.

To restrict the input, one option is to apply standard anonymisation techniques to the underlying microdata before analysis. This is the approach taken in the US Census Bureau's Table Creator[15], and appears to be the mechanism underlying the OECD's online tabulation tool[16]. The assumption is that the data is near enough non-confidential. A slightly more flexible approach is to identify all acceptable combinations of key variables (allowing for differencing between possible tables) and then only allow any analysis on the confidential data which uses an acceptable combination; this is partly the approach taken by the ABS's CData system (now supplanted by TableBuilder).

Restricting output rather than input allows for more flexibility in table creation; the web tabulation tool at LISSY[17] (see below) allows tabulations on the same restricted microdata access through its remote job service, and only redacts small cells. This does raise concerns about confidentiality breaches caused by differencing (different numbers of observations in tables with slightly different selection characteristics, allowing observations with distinguishing characteristics to be separated out). Standard rules-based output SDC (see below) cannot address this problem.

[15] http://www.census.gov/cps/data/cpstablecreator.html
[16] http://stats.oecd.org/.
[17] http://www.lisdatacenter.org/data-access/lissy/

The solution adopted by the ABS' TableBuilder[18] is to add a small amount of random noise to each table cell, so that the true value of the cell is uncertain (and the true difference between two cells is much more uncertain). To get round the objection of users that adding random noise makes results difficult to compare across tables, TableBuilder permanently attaches the noise to the cell using the combination of key variables to index the data. Thus the same table requested twice, or the same cell appearing in different tables, should appear the same. Whilst these concepts are not new, at present TableBuilder appears to be an outlier in implementing them in a production environment (rather than as an academic exercise)[19].

Table 3 Remote tabulation: summary of advantages and disadvantages

	User	Data owner
Advantages	• Easy to manipulate data and get tables, maps, graphs etc. quickly • User friendly interfaces	• Large number of off-the-shelf tools • Fixed cost of producing base dataset spread out over many researchers
Disadvantages	• Limited to univariate statistics on complex categories	• Ongoing cost of the system • Need to create base dataset and apply some form of SDC

4.2.3 Distributed analysis 2: remote job servers

Remote job servers are a generalisation of remote tabulation tools to allow a range of statistical analysis to be carried out on a confidential dataset. Users submit code to the server (nowadays mostly through a web interface, but in some cases still through email), which runs the code on the data and generates results. These results may then be checked by someone to make sure there is no residual disclosure risk; if the outputs are acceptable, they are returned to the user. The RADL system developed by the ABS and adapted for New Zealand is an example.

The best example of a successful remote job server is the LISSY system supporting analysis of the Luxembourg Income Study (LIS)[20]. This has been providing a remote job interface for over a decade, and claims to have serviced over 50,000 job requests; recently a web tabulation tool was added to interrogate the same data.

In theory, the advantage to both parties is that the user can carry out unrestricted analysis without ever seeing the microdata, therefore meeting the goals of both sides. In practice this is not the case:

- Users do in practice need to see the microdata (for example, when creating new variables); hence the ABS RADL allowed users to display a small number of records for checking purposes, but this meant the goal of not letting users see the microdata was not achieved.
- Users need to see the results of analysis which might be confidential (for example to study outliers); again, if these are not blocked then microdata are visible to researchers.

[18] http://www.abs.gov.au/websitedbs/censushome.nsf/home/tablebuilder

[19] TableBuilder is currently being evaluated by Statistics Sweden as a potential service; see Andersson K., Jansson I., and Kraft K. (2015) *Protection of frequency tables – current work at Statistics Sweden* in UNECE/Eurostat Worksession on Statistical Data Confidentiality 2015, Helsinki http://www1.unece.org/stat/platform/download/attachments/109248612/Session%202%20-%206%20-%20Sweden%20%28Jansson%29.pdf?version=1&modificationDate=1440606581076&api=v2

[20] http://www.lisdatacenter.org/data-access/lissy/

- Developing and testing code is difficult (for example, when Stata fails it just gives an error code and does not return the line on which the code failed); hence researchers are restricted in their activity in practice, even if not in theory.
- If code takes a long time to run, researchers may want to store intermediate versions of the dataset (particularly during development where mistakes are most likely to be made); this may not be possible.
- The code-input system needs to be relatively tolerant of coding faults, and researchers need to be instructed on how to use reference the input data.

There is also an unresolved theoretical debate over whether code and outputs should be checked for inappropriate content and, if so, whether this should be 'default-allowed' or 'default-blocked'. In the former, codes and outputs are allowed unless there is a reason to block them; this is likely to miss many cases of code or outputs which should not have been allowed. In contrast, the default-blocked position (only pre-defined commands and outputs are allowed) is likely to severely restrict user flexibility.

Finally, as noted below in the 'People' section, the remote job model relies upon researchers being trustworthy. In Australia this is particular problems, as two breaches of the RADL have made this contention hard to sustain[21]. In the worst case, a team of researchers systematically misused the facility over a period of time to download data to their own machines. Part of this has been attributed to a breakdown in the relationship between the user community and the data owner, and two opposing lessons could be taken away from this. The first is that the environment needs to strengthened to prevent misuse; the opposite view is that the restrictions imposed by the environment created the behaviour they were supposed to deter.

For these reasons, remote job servers have largely dropped out of the frame for analysis; they satisfy neither user not data owner. The development path of Australia (and New Zealand) has not had a wider impact (although Statistics Canada has acquired and is adapting the ABS' DataAnalyzer). The US Census Bureau and German NSI DeStatis have supported concept systems, but remote job servers have mostly been pushed out of the frame by virtual RDCs.

In this light it is worth considering why LISSY stands out as a counter-example. The consensus is that LISSY is successful because:

- The data is relatively simple, well understood, and does not change dramatically; this helps with the development of documentation too.
- There is a large and well-established user community.
- The system takes a light-touch approach to user and confidentiality management (and yet still has an excellent record on maintaining security).

Finally, the LIS began providing remote access to analysis via email long before remote job or remote RDCs became available. It therefore had a large reservoir of goodwill when it developed its remote job solution, which has become 'the way' to access the LIS. This goodwill seems to have been difficult to replicate in other systems.

[21] The more serious incident was widely reported in the newspapers; one other was discussed in NSI circles but did not have the same impact or publicity. If there were other breaches, the project team is not aware of them.

Table 4 Remote job server: summary of advantages and disadvantages

	User	Data owner
Advantages	• Ability to carry out complex analysis on microdata	• Little/no direct user access to data
Disadvantages	• Analysis limited by not having direct access • Need to be able to code and debug effectively	• Cost of developing system • Ongoing cost of the system • Not possible to guarantee no direct access to data without severely restricting output

4.2.4 Distributed analysis 3: RDCs and Remote (virtual) RDCs

Prior to remote access technologies, researchers were required to visit 'Research Data Centres', RDCs: sites located in the data owners' offices (or controlled and monitored by them). In return for physical travel and isolation in a secure environment, researchers were given access to the most detailed microdata. Many statistical organisations set up such facilities; in some places (such as Canada) the academics took the initiative and secured funding to set up the secure facilities. In the US and Canada, geographical constraints were limited by having duplicate facilities set up in multiple locations under the direct supervision of Census Bureau or Statistics Canada staff.

In 2002 the Danish NSI set up the first virtual RDC (vRDC), offering the same facilities as an on-site system but accessible from the desktop of researchers across Denmark. This was quickly taken up across much of Europe with the UK, the Netherlands, and Sweden being early adopters, followed by Italy, Finland, Slovenia, NORC (US), and Mexico in the third wave. Most European NSIs, as well as the US, Canada, Mexico, South Africa and Japan, now have or are planning some form of vRDC system. Some countries have more: in Germany both DeStatis and the Employment Department research unit (IAB) run vRDCs, as did the Bank of Italy and IStat, the Italian NSI. In the UK vRDCs are run by ONS, the UK Data Archive, the Tax Department and a network of four universities; the latter are exploring doubling up for both social science and epidemiological research.

Almost all facilities use Citrix 'thin client' software (or the Microsoft clone, Terminal Services); the US Census Bureau uses Unix to achieve the same end. Most European government-run vRDCs cite the Danish/Dutch system as the mode for their technology. However, NORC is more often cited as the technical model for systems set up by academics.

The 'virtual' in vRDC relates to the way it can be used, not the way it is used. The Census Bureau and UK tax department only allow access from restricted sites; vRDC technology simplifies management but does not necessarily change the user experience. The ONS vRDC, the 'Virtual Microdata Laboratory' (VML) allows access to users across the UK government network but not beyond. The UK non-ONS facilities allow access across the university network, as does the Danish system. The Dutch and French systems are unrestricted, although eligibility requirements for international access, for example, are stringent.

Despite this, there are various common features[22] :

[22] This data comes from a small survey of vRDCs carried out as part of the project. See the Appendix for specific results.

- Most have institutional as well as personal agreements, and both parties are liable in the event of a breach.
- Some RDCs make the same data available to all, whilst others restrict detail based upon organisation and user, separating the users into 'classes' based on the organisation they work for.
- Most allow the most detailed data to be available to researchers (such as postcode); some even allow direct but uninformative identifiers, such as tax reference numbers.
- Almost all provide training on security awareness, mostly through face to face or online courses rather than passive use of documents.
- Almost all provide training on output checking and output SDC.

The growth of vRDCs has led to several changes in perspective. First, the level of control has allowed the detail of data to be increased. Second, the efficiency gains from principles-based output SDC (see below) have encouraged training which emphasise engagement. Third, engagement is also encouraged because of the high cost of having untrustworthy users in the vRDC. Fourth, more control has increased confidence in allowing non-typical users (for example, private sector organisations) to get access to data. Finally, the unrestricted nature of the research carried out in a vRDC has led to the new field of output-based SDC targeted specifically at research.

Two recent innovations have raised new possibilities for accessing secure data through vRDCs.

The French research organisation Reseau Quetelet developed 'secure thin clients' to provide access to their vRDC. These patent-pending boxes were designed specifically so that they could be plugged in anywhere that was connected to the internet, and they would provide a secure end-to-end connection, using both password and biometric markers to authenticate. They have enabled France to become the leading provider of international access to a vRDC, and have also been adopted for the Eurostat "Decentralised and Remote Access to Confidential Data in the ESS project" (DARA), replacing the original plan to build a dedicated secure network[23].

Secure Pod is a complete module, roughly six square metres, which is also designed to provide a secure facility for access to a vRDC[24]. Users enter the pod using a biometric key only at an approved time; only one user can be in the module at a time. The module contains a screen, keyboard, mouse and secure link, but nothing else, and is monitored by a remote camera connected the vRDC team. It was envisaged as the sort of thing that could be installed in a university library where access from a researcher's desktop was not feasible or not appropriate for extra-risky data. At present these have not been widely adopted, partly because the underlying ethos of the module is that 'the researcher can't be trusted'; this is in direct conflict with the EDRU ethos adopted by the Administrative Data Research Network team behind the pods. Nevertheless, they remain an interesting option where exceptionally secure remote solutions might be needed.

[23] http://ec.europa.eu/eurostat/cros/system/files/final_report_ESSnet_DARA_20140321_publishable.pdf_en

[24] Dibben C.(2015) *Micro, remote, safe settings (safePODS) – extending a safe setting network across a country* in UNECE/Eurostat Worksession on Statistical Data Confidentiality 2015, Helsinki http://www1.unece.org/stat/platform/download/attachments/109248612/Session%204%20-%20Univ.%20Edinburgh%20%28Dibben%29.pdf?version=2&modificationDate=1442327968194&api=v2

Table 5 vRDCs: summary of advantages and disadvantages

	User	Data owner
Advantages	• Full access to data	• Very high level of control and monitoring possible
Disadvantages	• Location may not be convenient • Some researchers resent restrictions on uploading and downloading	• Cost of developing system • Ongoing cost of the system

4.2.5 Analysis services

Prior to the development of vRDCs, some organisations offered serviced research requests: researchers would write code and send it to the data owner, who ran the code, checked the outputs, and sent the results back. This was the 'remote data access' offered by Statistics Canada, for example, or by the ONS Longitudinal Study.

We note this for completion. Although some services such as the ONS-LS continue to operate, the analysis services suffer from many of the drawbacks of remote job servers, with the additional cost and delays of everything being done manually.

Table 6 Analysis services: summary of advantages and disadvantages

	User	Data owner
Advantages	• Ability to carry out complex analysis on microdata • Access to expert support team	• No direct access to data
Disadvantages	• Analysis limited by not having direct access • Need to be able to code and debug effectively • Delays in transmissions	• Cost of specialist support staff

4.2.6 Synthetic data

Synthetic data developed with the simultaneous publication of articles by Rubin and Little in 1993. Both proposed that confidentiality problems could be solved by replacing risky data with imputed data that was sufficiently 'close' to the original data to allow for valid analysis, but without the confidentiality risk.

A synthetic dataset has the original data replaced by random draws from a distribution reflecting the original data. For analytical work, multi-dimensional distributions are specified so that some of the correlations between data points are maintained. Multivariate relationships are modelled, and then used to generate predicted values. Additional rules can be specified; for example, to ensure that gender and gender-specific illnesses are consistent.

There are two general-purpose pieces of R code around to generate synthetic datasets. SimPop was developed in Germany and is being used, amongst other things, to produce PUFs for Eurostat. Perhaps more relevantly for DSS, SynthPop was developed in Scotland to produce synthetic PUFs from linked health and social care data.

Synthetic dataset can be either fully or partially synthetic. The latter will only synthesise key identifying variables. The advantage is that most of the data is genuine, and only the variables of concern have been falsified. On the other hand, a fully synthetic dataset is easier to sell as 'risk-free'.

In fact, synthetic datasets are not automatically risk-free. The rules which specify their creation determine how much risk there is. If the aim is to provide a set of variables of the right size and shape, then the dataset can be made safe by synthesising each variable independently, subject to some logical checks. However, if the aim is to reproduce the characteristics of the original data faithfully, the synthesis may not protect data adequately; for example, if the aim is a good representation of household structures, then household structure outliers might re-appear in the synthesised data.

This broadly typifies the two directions that synthetic data production has gone. In the US, the view is that synthetic data can effectively reproduce original data for research, and some papers have been published using this data. Hence US researchers have spent much effort on providing support for synthetic data analysis, including servers which will provide confidence intervals around synthetic estimates based on comparisons with the original data. These synthetic files are relatively high-risk as they are designed to be close to the original data and so, in some cases, even the synthetic data are only available through secure access facilities.

In contrast, European research seems to be heading down the route of making synthetic files a good approximation of structure and distribution for the original data, but without any commitment to analytical validity. These files are suitable for users wanting to get a feel for the data (such as developing code before using a restricted access facility), or possibly even produce simple broad totals, but not for analytical purposes. This has driven the interest in synthetic data to produce PUFs for Eurostat and the Scottish Health Informatics Project. This is also the reason it was considered by ONS in the late 2000s, to produce 'test files' with a realistic structure to make sure researchers used their time in the restricted access facility more effectively.

In short, in the US synthetic data is seen as a way of widening the analytical options; in Europe, it is seen as helping researchers to prepare for more detailed analysis but not in itself of analytical value.[25]

4.2.7 Settings: summary

Distributing data is well established and relatively uncontroversial. There are concerns over whether users maintain the local environment appropriately, and some informal evidence to suggest that the expectations of data owners are not fulfilled; but there is very little evidence that this has led to an unacceptable level of risk.

For more sensitive data, the virtual RDC is the recommended option. These have been proven to be secure, popular with researchers, and relatively easy to manage. They are more costly than distributing data but, as described below, much of the cost can be offset or additional gains can be realised. There is now a great deal of experience in setting up and running such facilities.

[25] This is of course a simplification, and there is a lot of cross-fertilisation between the two camps. For example, Jerry Reiter in the US and Jorg Dreschler in Europe are leading researchers in and epitomise the relevant continental stance, but the two also work together.

There is no agreement on where the boundary lies between SUFs and vRDCs; how much risk should there be in the data before access needs to be controlled? However, most vRDCs now are expected to hold, by default, detailed data with little or no confidentiality adjustment apart from basic de-identification; there is strong evidence to support the contention that these are very low risk facilities, even when training is not up to international standards. Where desktop access to a vRDC is allowed, detailed SUFs are being dropped in favour of vRDC access to the original data; SUFs are being pushed towards the PUF end of the spectrum.

Remote tabulation is widespread with many off the shelf tools available. However, most systems have a low-risk underlying dataset. The ABS' TableBuilder is an exception, not yet widespread but attracting interest from international bodies.

Creating synthetic data seems to be an option for PUFs, but this is at an early stage of development, and general acceptance by researchers is currently unknown.

4.3 Project approval

4.3.1 Lawfulness

All project accesses, except for some non-personal data, require a demonstration of the legality of the release of data. Whilst informed consent from the data subject is seen as the gold standard for research use[26], most social science and epidemiological research use relies upon statutory gateways allowing access for research purposes. While the wording varies substantially between countries (and even within countries with respect to different organisations), broadly the research gateways specify that:

- The re-use of data for research purposes is lawful if the public benefit expected from the re-use outweighs the private cost to the respondent of having their privacy breached.
- Steps should be taken to mitigate the private cost to respondents as far as possible given the intended use.
- An appropriate authority which will make the judgment is identified.
- Only statistical research is allowed.

Most legislation does not specify much more than that; this is left as an operational decision by the authority, which may be a person (such as the Departmental Secretary in the case of Australian Public Interest Certificates or government minister in the case of UK Chancellor of Exchequer notices) or an organisation (in the case of Canada, the UK - the NSI is the decision-making body – or the European Commission).

Despite this commonalty of principle, actual practices vary wildly. This is true even in the European Union, where there are cross-national regulations, directives and statutes on data protection. There is very little agreement even on definitions such as 'personal data' or 'anonymous', even often within countries.

[26] See Wellcome Trust (2015, ibid.) for a detailed discussion of the issues surrounding the use of informed consent as a gateway.

As has been pointed out[27], 'the law' is seldom unambiguous and closed to interpretation, particularly in the case of data access. This makes decisions on the implementation of the legal framework subjective, influenced by the objectives of the data owner. The traditional approach to data access begins by asking "**are** we allowed to do this?"; this makes the law a shield, part of a defensive, default-closed strategy. An alternative is to consider the law as one of the tools to be used in designing data strategies; the appropriate question is "**how** do I lawfully achieve what I want?" This alternative approach, of deciding objectives and studying the legal framework to see how an objective can be achieved, is a key part of the EDRU ethos (see below) but not yet widespread..

In summary, the statutory framework across countries is very similar in principle, and often in details too. However, actual practice varies widely because organisations tend to interpret the statutory framework in respect of their own internal perspective. While 'the law' is often discussed as if it were unambiguous and inviolable, in practice legal interpretation is at the service of the organisation, sometimes as a shield and sometimes as tool. Very few developments in data access have come about as a result of the change in the law (France provides a good exception); most have arisen from new interpretations, with legal changes following (the UK is an example).

4.3.2 Contract details

In practice, almost all data access agreements specify four constraints:

- which data is to be used
- by whom
- for what purpose and
- for how long

and two instructions:

- researchers are expected to act in accordance with standards specified by the data owner and
- the research environment should be set up and maintained in accordance with standards.

This specification is consistent with a wide range of statutes and operating principles. However, there are almost no examples of organisations sharing common agreements. For example, in the UK, ONS, the UK Data Service and the four vRDCs of the Administrative Data Research Network share common training, common technology, and a common conceptual framework; but access agreements for individuals are specific to the services.

This seems to be because access agreements like to refer to specific legislation, but it is not clear why. Very few laws require the user to be aware of them to have force, and so adding specific detail does not increase guilt. Indeed, authors have argued that excessive detail on agreements reduces the responsibility of researchers, by making the contract essentially unreadable[28]. An analogy is with

[27] For a detailed discussion of these topics, see Ritchie F. (2014) "Access to sensitive data: satisfying objectives, not constraints", J. Official Statistics v30:3 pp533-545, September. DOI: 10.2478/jos-2014-0033

[28] For a critique, see Hafner H-P., Lenz R., Ritchie F., and Welpton R. (2015) "Evidence-based, context-sensitive, user-centred, risk-managed SDC planning: designing data access solutions for scientific use", in UNECE/Eurostat Worksession on Statistical Data Confidentiality 2015, Helsinki
http://www1.unece.org/stat/platform/download/attachments/109248612/Session%204%20-%20Various%20%28Hafner%20et%20al.%29.pdf?version=1&modificationDate=1442327222025&api=v2

the tick boxes used to accept terms and conditions when installing software, for example; courts have begun to take the perspective that these do not count as informed consent.

In short, it is not clear that any further information is needed; but in practice common contracts are limited by a desire for each organisation to have its own wording, presumably to cover it for not making users aware of their specific responsibilities.

4.3.3 Institutional versus personal agreements

For anything other than PUFs, access arrangements almost always require the individual who will be working with the data to sign an access agreement. This may be directly with the data owner, or it may be with the third-party distributor (for example in the case of the UK Data Service).

Many agreements also require an institutional signatory; for example, an organisation must be a recognised Research Entity before any researcher can access SUFs from Eurostat. This is common for vRDCs, but less common with SUFs.

In theory the institutional agreement provides an additional safeguard, but it is not clear in practice how well this works. Some organisations such as IPUMS[29], which distributes confidential Census data internationally, make strenuous efforts to ensure that institutional signatories are aware of their obligations. This is because IPUMS, being an international distributor, has relatively few levers it can pull to control usage; in Five Safes terms, most of the control is invested in the 'Safe Projects' and so it expends considerable effort in getting the institutional arrangement right.

Other arrangements are less clearly beneficial. Informal opinion amongst data access professionals suggests that generally, institutional signatories are not fully aware of their responsibilities[30]. For example, the penalties for misusing the UK Data Service vRDC include, in the worst case, a five-year ban on the institution receiving funding from the Economic and Social Research Council. This is sufficient to close down any social science department, and potentially leaves the organisation open to being sued by its own staff. However, it seems unlikely that this risk appears in any corporate risk registers.

As was the case with opinions about whether users follow IT guidelines, there is almost no hard evidence as to whether suspicions about institutional engagement are well-founded. On the other hand, there is also no evidence against, and psychology would suggest that, for some organisations at least, institutional agreements are seen as merely another form to be signed. Whilst most data owners or distributors offer some form of personal training, no organisation appears to offer training for institutional contacts on a systematic basis.

In summary, institutional agreements should in theory improve confidence in security arrangements. However, it is not clear how well this follows through in practice.

[29] www.ipums.org

[30] In a case personally known to the authors, a departmental administrator refused to accept responsibility for the misconduct of a member of staff, arguing that academics were all free agents; and was surprised when his department was temporarily blacklisted for failing to take appropriate internal action.

4.3.4 Ethical approval, due diligence and the use of precedents

All access arrangements for confidential data have some form of check on the purpose of the project. These can range widely. For example, access to the IPUMS SUFs requires institutional and personal registration and electronic acquiescence to an agreement to use the data for statistical purpose; access to French business microdata, prior to 2010, required researchers to travel to Paris for an extended presentation and examination in front of an ethics panel.

Less clear is the use of project classification, internal precedence and external due diligence.

Project classification means identifying 'types' of project so that approval processes can be streamlined around the characteristics of the class. For example the UK Data Service has four classes of project:

1. 'open data' where registration is required for auditing purposes but no further checks are made
2. automatic approval, subject to registration and electronic commitment to standards of behaviour and data security, for very low risk (but not public) datasets
3. mostly automatic approval with data owner confirmation, for higher-risk datasets, or where the data owner has requested final oversight of projects and
4. specific project scrutiny for access to datasets in the secure environment.

Internal precedence means identifying similar classes of projects so that decisions can be made more quickly following an earlier precedent. For example, the UK ONS Microdata Release Panel does not formally have different classes of output, but it has established a wide range of precedents for data access, ranging from Freedom of Information requests to the supply of microdata to other government departments. Once a precedent is established, projects can be approved by the Secretariat without waiting for the periodic meeting of the full panel.

External due diligence means accepting the capacity of other organisations to carry out due diligence checks on the value of projects or the trustworthiness of researchers. For example, in the UK the existence of funding from one of the academic Research Councils is taken by ONS as sufficient evidence of significant public benefit. In general however it has proved hard for organisations to give up their taste for scrutiny. As the Wellcome Trust noted[31], everyone agrees that allowing one body to check the project credentials and all others to accept the decision is an efficient solution; but everyone also thinks that they are the only body that can do the checking. This seems to be a particular problem in countries where data is spread amongst federal bodies: Australia, Canada, or Germany, for example.

These three processes are designed to speed up the ethical process whilst maintaining proper oversight. For example, the first application was received for the ONS Virtual Microdata Laboratory in 2003; it established a precedent for the type of projects which were acceptable for the VML. When the law changed in 2008, a second precedent-setting application was scrutinised by the panel. Those two applications took some weeks to gain approval; the other five-hundred-plus applications were approved in days (sometimes hours) by the secretariat.

[31] Wellcome Trust (2015, ibid.)

However, all three options are controversial. Defining types or classes of projects too broadly might allow inappropriate projects so slip under the radar. The criteria for precedents might creep over time. Third party due diligence might not cover all the criteria of the data owner.

It is difficult to find statements about the use of these process accelerators. However, anecdotal evidence suggests that precedent-setting is the most effective way to make the application process more efficient. The reason may be because it allows the streamlining to be hidden. Defining classes of users or delegated due diligence highlights the fact that a decision is being automated. If the data owner is adopting a defensive position, automation immediately leads one to start considering where the automation would fail. In contrast, precedents are not defined in advance; they are responses to circumstances and only apply in similar circumstances, and so it is harder to argue that a precedent is committing the data owner to inappropriate actions. Setting precedents also allows the organisation to learn, it can be argued. Finally, precedents can be a good way of changing attitudes; as data owners tend to be risk-averse, the continual accretion of successful instances can demonstrate the safety of a process it a way that theoretical arguments cannot.

Note that the defensive perspective is driving this way of thinking. From an EDRU perspective, all three options are valuable and defensible; precedent is a sensible way of defining types and areas for delegation, and the acceptance of mistakes provides a mechanism for avoiding creep in classifications or precedents.

So, there is no internationally agreed perspective on what is the best practice for project approval, and few organisations state explicit views on streamlining. There is, however, informal evidence that this can make a significant difference to the resources needed to run a data operation.

4.3.5 Composition and perception of ethical approval authorities

It is increasingly common practice now for ethics boards to represent four archetypes:

1. Data owners
2. Independent legal or ethical advisors
3. Researchers
4. The general public.

The last is a relatively recent innovation, and more common amongst data operations run by specialist academic groups. The intention is that the external voice allows the ethical committee to identify potential problems (in terms of public acceptance of their activities) early on. This seems a sensible approach but there is little hard evidence for this.

The relationship between the ethical approval committee and the other elements does seem to be crucial. The Wellcome Trust[32] noted that a major block or spur to successful data access in epidemiology was the success of the ethics board. A ‘successful’ ethics board understood the research environment, the needs of data providers, and the wider public perception; was default-open (see below); and was viewed by the researchers as providing a positive stage in project development. In contrast, an ethics board, which was driven by the demands of the data owner, was

[32] Wellcome Trust (2015,ibid.)

defensive in approach and guarded its absolute right to make decisions. This instilled antipathy and unwillingness to co-operate in researchers.

4.3.6 Identifying benefits

As noted above, laws tend to specify that projects can go ahead if the benefit to society outweighs the costs of ensuring that the privacy of the individual is adequately protected. Given that the marginal cost of providing the data (especially for SUFs) is close to zero, the expected public benefit from a researcher producing some statistical analysis does not need to be substantial. In many countries, the existence of a valid research project itself is sufficient public benefit. While any one project may not lead to substantial public benefit, overall the belief that research contributes to the public good means that there is a positive expected benefit from each project.

Some organisations specify the 'public benefit' effect in detail. For example, both the US Census Bureau and the UK tax department require that the research support the operations of the relevant tax office – supporting general principles is not sufficient. Such requirements can be interpreted more flexibly if the data owner is willing. One organisation required researchers to show how their research contributed to one of the data owner's strategic objectives, but the nature of research meant that almost anything could be said to 'contribute'. In practice, the organisation was more concerned that access was justifiable to the general public.

Use of data by individuals following a specific agenda (for example, a journalist or member of the public looking for evidence for a particular stance on migration) cannot be stopped on public use files, as by definition they are available without restriction. When the project requires approval by the data owner, this becomes more difficult as the act of approving the access may be seen as condoning the purpose of the research. Moreover, it has been argued that all research is biased in some way; trying to ban some types of analysis is implicitly condoning others.

To avoid this dilemma over whether and how to limit some kinds of research, most organisations publish the criteria for projects to be accepted. The criteria focus on the statistical nature of the research, the need for research to be lawful (for example, not breaching discrimination laws or the consent under which the data were gathered), the credibility of the user to make appropriate use of the data, and any external evidence that the project has been deemed worthy of support (such as public funding). The criteria also note that the data owners are under an obligation to treat all valid research projects equally. However, most data owners leave themselves some leeway by allowing that access can be removed if the user makes 'inappropriate' use of the data in some way.

There are exceptions to the 'all research is valuable' rule when private gains seem likely to take precedence over public ones.

First, the use of data by students below PhD level is often taken to have negligible public benefit. Master and Honour students produce research for their own benefit which is not normally circulated.

Second, private sector companies might be expected to acquire a commercial advantage; it is less clear whether this works to the general 'public' benefit. Some organisations argue that, because commercial organisations only use the data if it leads to a commercial gain, there should be no access. Others have argued that, as long as the same data is available to all parties on the same

terms, parties who see an opportunity to make use of that data should not be disadvantaged. A third group argues that the public gain can be maximised by ensuring that all outputs produced by the commercial firm are made publicly available for others to use. This third option seems to seriously disadvantage the private company, which expends effort on research only to see results made available to its rivals; but in practice this is less important, as the research process is often as valuable as the final results. In summary, there is no clear consensus on whether commercial organisations should be allowed access to confidential data, or how to treat the results of their research.

Aside from a general public benefits, more specific benefits to the organisation have been identified, particularly around methodology and data collection. At its simplest level, the data owner has limited resources to exploit its data assets; more users of the data mean that more should be known about the data. This was an explicit reason for the UK Department of Energy and Climate Change to create public-use versions of its energy-use data[33], for example.

The more detailed the data, and the more expert the users, the more likely the chance of useful methodological input. For example, the UK Low Pay Commission has been making extensive use of ONS data, as well as commissioning research reports using those data, since 1998. Amongst the consequences of this intense scrutiny have been the complete redesign of two major surveys and changes to the questions in a third, as well as papers on data accuracy and measurement error.

Actually realising these potential gains seems hard. Unless the data owner has a large internal research group (for example, as at the US Census Bureau) then the data owner needs to take active steps to engage with researchers to get meaningful feedback (as opposed to trawling through academic papers to find critiques of the data). For example, both the Administrative Data research Network in the UK and the German Labour Ministry Research Department (the IAB) run 'user conferences' attended by both researchers and data producers.

Methodological input seems to be more valuable to NSIs and other organisations collecting data for statistical purposes. For organisations that mainly accumulate data through administrative processes, data acquisition is driven by operational need rather than statistical value, and so the commentary of users may be of less relevance.

In general, while the specific benefits to the data owner exist in theory, empirical evidence of these gains being realised or substantial is scarce.

Finally, it is not clear what benefits, if any, accrue from international users. Such users would be able to contribute to methodological issues, but may have less incentive to do so than locally-based researchers. However, international users are more likely to be doing research comparing countries, and it could be argued that this is even more important than local research to local policy-makers. Overall, those few organisations who do allow international access to confidential data seem to use the same benefit criteria for local researchers.

[33] Gregory M. (2014) "DECC's National Energy Efficiency Data-Framework – Anonymised dataset". http://www.turing-gateway.cam.ac.uk/documents/Gregory.pptx

4.3.7 Need-to-know versus standard datasets

Should researchers be given access to a standard dataset containing all the data – which may include variables of observations they do not need – or should a dataset be defined for a particular project, with the minimum set of information?

The argument for the need-to-know dataset is a legal one. If data are not needed for the research to be carried out, then providing those data is unlawful. There is also a confidentiality argument - providing more data than is strictly necessary increases the likelihood of an unauthorised breach.

The counter-argument is that 'need-to-know' cannot be precisely defined. In practice, research by its nature is uncertain and so what data is 'needed' cannot be determined ex ante. It could even be argued, to take an extreme position, that 'need-to-know' followed to its logical conclusion means specific observations should be excluded if they have missing values in key variables, for example.

In practice, those operating on a 'need-to-know' basis (such as the Scandinavian and Dutch NSIs, or the Administrative Data Research Network in the UK) interpreting this as "provide data corresponding to the selection criteria and set of variables identified by the user as being likely to be useful". In contrast, organisations providing standard datasets (such as the IAB in Germany, NORC in the US, and NSIs in the UK, Mexico, and South Africa) argue that a dataset is an atomic structure, and therefore the relevant question is "do you need to know the data in *this* dataset?"

In short, it is clear that this is a question of degree: given ex ante uncertainty over what is strictly necessary, how closely do you draw the lines around what is likely to be needed?

There are however, practical implications. Extraction of a unique dataset is likely to be more costly than copying an existing dataset. Where the datasets are de-identified but otherwise original, then creating a file is simply a question of extracting the subset of observations and variables. However, if disclosure control is to be applied to the data (to create an SUF, for example), this needs to be done for every dataset. An alternative is to apply SDC once to the original data but this is likely to overprotect the data as not all of the original data is being released.

In summary, both models are usually justifiable in law; conceptually the need-to-know model seems preferable to the standard-dataset model, but in practice the latter prevails for operational simplicity.

4.3.8 International access

Very few countries allow international access to confidential data. One author lists as example of successful outcomes:[34]

- The IPUMS (International Public Use Microdata Series) project www.ipums.org harmonises and shares confidential but anonymised Census microdata from sixty-plus countries.
- 'Mesodata' models create semi-aggregated microdata from country-specific disclosive microdata with a view to a particular type of analysis.

[34] Ritchie F. (2013) "International access to restricted data: A principles-based standards approach". *Statistical Journal of the IAOS* v29:4 pp289-300. DOI 10.3233/SJI-130780

- The 'RDC-in-RDC' model allows researchers in the US to access a vRDC at the German ministry of labour, the IAB; the Eurostat-supported project 'Data Without Boundaries' (DwB) expanded this model to one site each in the UK and France.
- The Dutch statistical office used contractual arrangements within the umbrella of European law to allow Italian researchers to have live access to the Dutch vRDC.
- The 'Lissy' remote job submission system allows researchers around the world to run queries on confidential earnings data.

However, the author notes that these are exceptional outcomes, as is the current DSS practice of allowing international researchers access to SUFs. In general, microdata have to be fully anonymised to be made available internationally for research[35].

An unresolved question is whether remote access to a dataset counts as international sharing. To take an analogy, consider someone standing on the shore at Dover holding up a card with numbers which could be read by someone on the French coast with a pair of binoculars. Has the data crossed the border? Or is it merely an image of the data?

In summary, there is very little common agreement on whether international access is feasible. One NSI told the authors that, although the law allowed international data sharing, the organisation didn't do it because it was concerned about begin unable to prosecute if something went wrong. Again, attitudes appear to be more important than any specific legal framework.

4.3.9 Projects: summary

Good practice is difficult to identify, as most organisations claim that their freedom to approve or deny projects is closely limited by the legal environment. Nevertheless it is clear that it is possible to draw up some simple principles which cover all access:

- Data may be made available for research purpose if the expected benefit to society outweighs the potential loss of privacy for the individual, taking into account the cost of reasonable actions to protect the individual's privacy.
- Data may only be used for statistical research, and the purpose must be sanctioned by an appropriate authority.
- Data access agreements should state the data to be used, the time period, the users, and the reason for use, and also that users are expected to follow instructions on behaviour and managing their local environment; anything beyond that may be counter-productive and open to legal challenge.
- Identification of 'public benefit' is difficult; most organisations make claims for the potential benefits but it is not clear how well these are realised; the justification for making data available may be more that the marginal cost of providing that data is effectively zero.
- A positive relationship between the ethical approval committee and the research managers provides substantial benefits, as does representation from the general public.
- Streamlining ethical approval can be done in several ways; using precedents, grouping applications, accepting due diligence of third parties; the appropriateness of these techniques seems to depend on the institution's attitude.

[35] Sharing internationally for operational purposes/processing, is, ironically, much easier legally than sharing for research.

- All data is released on a need-to-know basis, but there are differences of opinion as to whether the dataset or the individual record is the appropriate level for selection; however, the practical benefits mean that most organisations take the dataset as the atomic entity.
- International access is rare, and the examples are largely exceptions.

4.4 People

4.4.1 How trustworthy are users?

The confidentiality literature assumes users are 'intruders', malicious individuals aiming to breach confidentiality protection for reasons of their own. Several types of intruder have been suggested:

- 'nosy neighbours' who want to find information about family, friends or associates
- campaigners who seek to embarrass the data owner by proving that security is poor and
- individuals with a commercial interest, such as blackmail.

This can be a sensible assumption when developing statistical theory: all models are being assessed against a common, if infeasible, baseline[36].

This intruder model is also a common assumption of those taking a defensive attitude to data access: if deliberate intruders can be kept out, then surely non-malicious intrusion is guarded against. This model has particular power amongst those who do not have familiarity with data or the research process, such as IT managers or legal experts.

However, there is almost no evidence to support this for the research community, despite over fifty years of access to sensitive data around the world. The authors have no hard evidence of any researcher behaving like the posited 'intruder', and only hearsay about potential cases.

Two responses could be made to this. First, that successful malicious intrusion is unlikely to be noticed, and even less likely to be discussed. Second, that past behaviour does not rule out future misbehaviour. Whilst acknowledging the conceptual truth in these arguments, EDRU proponents point out that placing possibilities on an equal footing with overwhelming empirical evidence is unhelpful and obstructive; it is the equivalent of saying that, because quantum theory allows a chair to spontaneously mutate into a penguin, all offices should keep bucket of fish handy just in case.

The EDRU school also argue that this focus on intruders is damaging because it distracts attention from the genuine risk. There is extensive evidence of mistakes being made in all systems, and there is also evidence that researchers who find procedures obstructive will find ways round them. Examples of breaches of procedure include researchers:

- taking notes from the screen in vRDC sessions
- using mobile phones or laptops in restricted environments
- taking confidential data out of the office to work on at home
- downloading data without authorisation from a restricted facility
- ignoring guidelines on producing safe output
- using data for projects other than the one approved

[36] See Hafner et al (2015, ibid.) for a discussion of this point and the other issues raised in this section

- asking those who have access to data to carry out analysis, rather than applying for access themselves and
- retaining data after the time allowed.

These breaches are caused by:

- ignorance of procedures or methods to be followed
- ignorance of the consequences of their actions and
- preference for taking simplest path to outcomes.

This has led some authors[37] to argue that the 'intruder' model should be replaced by the 'lazy idiot' model as being far more relevant; Eurostat prefers the term 'human' model, as being less provocative, and uses it in its training[38].

It has been argued that researchers based in private sector companies have more incentives to breach confidentiality than academics. Consider this hypothetical example: a researcher working for a consulting firm who manages to extract commercially sensitive information from confidentiality data may be promoted rather than fired.

Again, this is a theoretical possibility but there is little evidence to support it. While allowing access to commercial firms is not common, organisations that do allow it do not seem to find any significant difference between commercial and academic researchers.

Part of this may be confusion over 'commercial' firms. In the popular mind, these are organisations such as insurance companies or supermarkets wanting to target individuals. In practice, many 'commercial' firms are private sector consultancies or charities with a research arm, carrying out very similar research to academics, and with the same need to protect their access to data.

Hence, to date there does not seem to be strong evidence that researchers from private sector companies act differently to those in academic research organisations.

Finally, some organisations provide more or less access based on the characteristics of the researcher: either the type of organisation the researcher works for, or the researcher's seniority or qualifications.

There is some evidence that government researchers given access to confidential material are lower risk; there is speculation that this is to do with government being a naturally more regulated environment than academia or the private sector. However, there are fewer government researchers, and some have been involved in breaches of procedure or confidentiality. A fair assessment may be that government researchers are 'safer', not 'safe'.

In terms of personal characteristics, there is little evidence to support the contention that seniority equates with trustworthiness. Indeed, many data professionals dealing with academic would argue that the more senior the individual, the less likely he or she is to follow instructions and take training

[37] Hafner et al (2015, ibid.); see also Desai T. and Ritchie F. (2010) "Effective researcher management", in Work session on statistical data confidentiality 2009; Eurostat http://www.unece.org/stats/documents/ece/ces/ge.46/2009/wp.15.e.pdf
[38] http://ec.europa.eu/eurostat/web/microdata/overview/self-study-material-for-microdata-users

well; a PhD student is a much safer bet than her supervisor. Similarly, qualifications have been used by some organisations but are not generally seen as good evidence of ability to act appropriately.

In short, the consensus seems to be that all researchers should be treated the same in terms of their personal characteristics; but that government employees might be more used to working under regulations. There is no consensus on researcher attitudes, with the EDRU school arguing for the 'human' model, but most data owners sticking with the 'intruder' model.

4.4.2 Can users develop and/or be trained?

The previous section noted that confidentiality breaches are the result of mistakes or researchers avoiding onerous procedures. One solution is to design an incentive-compatible release mechanism – that is, one that encourages good behaviour by default. The other is to train researchers so that mistakes are not made and procedures are not skipped. All vRDCs now have some form of researcher training; there is much less training available for users of SUFs.

There is very strong evidence that users generally want to do the right thing and appreciate understanding the context in which they work. They also react badly to being preached at, instructed, policed – essentially being treated as potentially naughty children.

There is a wide literature on this in criminology and psychology: the two key concepts are 'procedural justice' (if people believe what you are doing is fair and right, they are more likely to support you) and social conformity (people will want to work with you if they like and trust you).This has only recently caught on in the data world, which is still dominated by the 'police' model of training researchers: warn them, frighten them, and make clear that responsibility lies with them for any wrongdoing.

'Active researcher management' (ARM) is now the dominant training model for vRDCs[39]. The 'active' is about actively trying to engage the researcher, help them to understand the concept of security, and try to develop a shared sense of purpose. The aim is that the researcher sees him- or herself as part of a research community where the security of one depends on the collective action of all.

The evidence to date suggests that ARM is not only more secure than the 'police' model, it is cost-effective: researchers become self-policing, and willing to work with the data owner. Examples of self-policing include:

- researchers informing the data owners of minor misuse by others
- a researcher threatening a miscreant with exposure if the latter did not correct hi actions and
- a PhD student setting up additional training in good behaviour for fellow PhDs.

Note however, that the number of breaches is so small that it is difficult to make strongly evidenced statement. Nevertheless, the experience of the vRDCs applying ARM is that the number of breaches of procedure occurring on over 10,000 visits is in single figures, with no breaches of confidentiality amongst trained researchers; there were however a very small number of breaches of

[39] See Desai and Ritchie (2010, ibid.) for the initial description of the problem; for an excellent practical example of an organisation implementing this strategy, see Wolters A.(2015) Researcher management and breaches policy https://www.ukdataservice.ac.uk/media/604140/14_5safes_safepeople_wolters.pdf

confidentiality caused by individuals who had not been required to undertake the 'active researcher management' training.

ARM training is usually delivered face-to-face, because of the importance of developing a bond between the researcher and the research manager/data owner. However, there are some online courses taking the ARM approach: the Eurostat Researcher and Administrative Data Research Network Refresher courses are explicitly designed with ARM in mind; the SURE training is not, but covers many of the same ideas. To date, the Eurostat course appears to be the only ARM-consistent course targeted specifically at SUF users. The effectiveness of these online courses has not been tested yet as these are mostly recent developments.

Strong advocates of ARM have also argued that this increased engagement can be turned to the data-owners advantage. By encouraging a sense of partnership, engagement with the data owner over methodological issues or over the publication of results is also encouraged. It is not possible to ascribe this solely to the training, as most ARM training schemes are part of an EDRU-school approach and so the data owner is also prepared to engage with the user. Nevertheless, ARM advocates argue that the training is an important opportunity for the data owner to sell its user-friendliness.

4.4.3 When researchers go bad

There have been some events to provide a counter-argument to the 'no researchers deliberately try to breach security' assertion:

- Netflix released user information without testing whether the de-identification was strong enough.
- Researchers at Harvard failed to anonymise effectively recipient data.
- A lecturer in a North American University asked his students to re-identify respondents in a dataset as an exercise [related second-hand - unverified].
- A researcher in a north-eastern US town identified a senior politician's medical records as part of a campaign against the release of the data.
- A researcher given identified data on patients with a sensitive disease used the information to try to blackmail patients [related second-hand - unverified].

Note that in all bar the last case, the aim of the researchers was to prove that the anonymisation was badly done, not for personal gain. Rather than countering the argument that researchers cannot be trusted, this supports the idea that researchers are generally keen to make sure that data is used safely – they do recognise the value of it, and the risks of misuse. They tend to do things wrong because they do not understand either the right thing to do or the consequences of doing the wrong thing.

4.4.4 People: summary

There is more agreement in people management than on projects. However it is noticeable that most of the developments in people management have come from the vRDC world. Distributed data solutions tend to rely still on sending out admonitory guidelines on good behaviour with data.

Common understandings are that:

- People remain the biggest risk to any data release proposal , apart from PUFs and remote tabulation.
- Current best practice suggests face-to-face training, but this is only likely to be cost-effective and tolerated by users for the most secure data solutions; for distributed data solutions such as SUFs, online training seems to be the only feasible option.
- Most training still follows the 'police' or 'intruder' model; applying Active Researcher Management to the 'human' model seems to provide better results and is the preferred solutions for vRDCs.
- There are a small but growing number of ARM-based online modules, although their effectiveness is still under review as they are new developments.

4.5 Data management and input SDC

4.5.1 Confidentialisation and anonymity

A 'fully-anonymised dataset' is generally understood to be suitable for unrestricted public release; a 'partially-anonymised' retains some confidentiality risk and so is released under conditions. Producing datasets to a given level of anonymisation is a well-understood practice, having been studied for some forty years. There are off-the-shelf tools such as sdcMicro and muArgus, as well as an extensive literature.

Trying to create fully anonymous datasets for public use runs into problems of definition. True anonymity cannot be proved for a dataset (unless every record is an exact copy of at least one other record), so what is the standard for 'anonymous'? For example, the law governing the ABS states that no statistics should be produced which have the 'potential' to breach confidentiality; as aggregate statistics have in the past led to breaches of confidentiality, a strict interpretation of the law could mean that the ABS is not allowed to produce any statistics.

German law comes closest to providing a practical definition of anonymisation with its concept of 'de facto anonymisation': a dataset is assumed to be anonymous if the likely benefits from re-identifying a respondent are outweighed by the likely costs of the re-identification. In general however the effectiveness of the anonymisation is considered independently of the evidential likelihood of it being breached.

4.5.2 Anonymisation as a residual

Anonymisation is often considered as the key protection technique for distributed data; once it has been decided to create an SUF, for example, there is much advice on how to produce a dataset appropriate for release. This approach is consistent with the defensive approach to data access: anonymisation is there to ensure that even in the worst case of malicious intruders, risk of re-identification is acceptably low.

However, some authors have begun to argue that anonymisation should be treated as a residual: what to do when you have no better ways to preserve confidentiality. The rationale for this is twofold. First, the aim of distributing data is to support research and analysis; anything that damages the data is likely to restrict that analysis, and therefore should be avoided by using non-statistical protection measures. The second rationale is that evidence should be used to identify an

appropriate level of anonymisation, not theoretical issues. This model is recognisably part of the EDRU approach to confidentiality.

Work by the UK Department of Energy and Climate Change[40] showed that, even in a PUF where data protection is the only option, an evidence-based assessment of the anonymisation can produce more useful data with increased confidence in the safety of the data. For an SUF, a Eurostat project[41] showed that addressing non-statistical issues first (the five safes excluding 'data') and then considering the evidence base for residual threats leads to a substantially different anonymisation technique than the 'defensive' technique applied before: targeting of very specific threats mean much less perturbation, but much better protection for problematic observations. Eurostat, which commissioned this re-thinking of anonymisation, has subsequently adopted this approach in a second dataset.

This approach is still unusual; if nothing else, it requires a substantial degree of buy-in from the data owner, whereas the traditional defensive approach is uncontroversial. Nevertheless, the existence of successful EDRU-style anonymisation may set a precedent for users to demand greater care paid to usefulness and evidence.

4.5.3 Spontaneous recognition

A frequent concern is in 'spontaneous recognition': a researcher looking at the data recognises (accurately or not) one or more of the respondents. This is most likely to occur when there are well-known outliers; for example, in business data, unusual health events, large families in small geographies, and so on. Because it is not possible to prove that spontaneous recognition does not exist, this can encourage an overly defensive anonymisation strategy. In particular, this is sometimes used to justify anonymising data in a vRDC, despite the range of non-statistical controls that are already in place.

Some recent work in the UK, Germany and Australia has focused on this, concentrating on business data. Noting that it is equally hard to prove that spontaneous recognition can occur or has occurred, and that it does not appear to be unlawful in any case, the researchers have suggested that a more useful concept is 'identity confirmation': a breach occurs when a researcher takes action to confirm or relate to others the possible identification, which is clearly unlawful. This is qualitatively different problem, as it is amenable to managerial solutions, not just statistical ones. This was one of the reasons cited for the much-reduced anonymisation of the Eurostat SUF[42].

Again, this is a developing field. The consensus among data owners is that spontaneous recognition needs to be considered as a risk; 'identity confirmation' may replace it, but that is likely to be some time away.

4.5.4 Digital object identifiers

The practice of giving datasets unique permanent identifiers is relatively recent; it is not yet widespread, but is increasingly used by those supplying data to researchers. The most popular

[40] Gregory (2014, ibid.)
[41] Hafner et al (2015, ibid.)
[42] Hafner et al (2016)

scheme is the Digital Object Identifier collection[43]. The advantages of DOI registration to the data owner are that (1) uses of the data set can be tracked more easily and accurately (assuming the researcher includes the DOI in the publication), and (2) the management and naming of the dataset can be separated from the discoverability of the data.

For the reader or paper referee, the DOI is actionable; that is, clicking on a DOI reference takes the reader straight to the dataset, allowing easier verification or replication of results.

Giving a dataset a DOI is now established good practice. However, there is still uncertainty about how a dataset should be identified when it changes. Should multiple bespoke copies taken from the same source data have their own DOI, or should they use the source data DOI? When a dataset gets updated, should we treat this as a new dataset, or retain the current DOI even though previous uses of the DOI now refer to an outdated data source? These issues are still unresolved.

4.5.5 Data: summary

SDC of datasets is a mature field; the only question is the context in which it should be implemented, which is outside the scope of this dimension:

- Anonymisation is well established, and there are multiple tools for creating datasets at different levels of protection; as a result, data should be treated as a residual, something that can always be created to suit the situation.
- Most literature and advice follows the 'intruder' model, leading to defensive SDC and over-protection of the data; however, some new developments show that applying the EDRU model can make substantial changes to the value of the data without cost to confidentiality.
- Despite the range of security features for vRDCs, some facilities still insist upon applying SDC to the data; the biggest justification for this is spontaneous identification, but there is little or no evidence to support this as a realistic concern, and it may be the wrong concept anyway.
- Use of Digital Object Identifiers seems likely to be the norm for datasets in future, but this does pose problems if datasets are to be created on the fly for specific projects.

4.6 Output SDC[44]

4.6.1 The origins of modern OSDC

Generalised output SDC is a relatively new field. Until recently, most SDC research focused on tabulations. However, researchers produce a much wider set of outputs than NSIs, and so the need arose for a more generalised solution.

Safe output is most easily managed in restricted access facilities, where facility managers can check all statistics being released and, if necessary, block their release. In other situations, the data owners must rely upon the training and knowledge of the researchers. SUFs are often distributed with some guidelines on good practice when producing outputs; the data owner then hopes that the user will read the guidance and act appropriately, but cannot guarantee it. For example, SUFs distributed by

[43] www.doi.org
[44] This section is largely taken from Desai et al (2015, ibid), 'The Five Safes'

Eurostat come with extensive guidance on what is expected in terms of minimum cell size and dominance in tables.

4.6.2 'Safe statistics'

Central to OSDC is the concept of 'safe statistics'. This is a system for classifying types of output (such as tables, regressions, or odds ratios), acknowledging that many research outputs pose no disclosure risk because of their functional form. For example, regression coefficients count as 'safe'; they pose no meaningful disclosure risk. In contrast, frequency tables are 'unsafe'; their structure provides a clear theoretical risk, and so specific releases must all be checked before publication. This allows the data owner to concentrate resources on the most risky outputs. This approach also provides a justification for the historical focus on tables to the exclusion of most other outputs.

4.6.3 Principles- versus rules-based OSDC[45]

OSDC can be applied using either a 'rules-based' (RB) or 'principles-based' (PB) approach. RBOSDC means that hard-and-fast rules as to what is acceptable are applied; for example, every table must have at least ten units in every cell, and a dominance rule must be applied to every magnitude table. This is well-suited to automatic checking systems (such as remote job servers), or where a large number of similar outputs are being repeatedly produced (as in NSI outputs). Output checkers require little statistical skill or knowledge: RBOSDC is a 'box-ticking' exercise.

However, the rules-based approach suffers from the lack of context. This typically means that non-disclosive outputs are blocked unnecessarily, but a more subtle concern is that inadequacies in the rules might go unnoticed. This can only be countered in the rules-based approach by making the rules very strict, which increases the likelihood of unnecessary blocking of non-disclosive outputs.

In contrast, the principles-based alternative places context at the forefront of decision-making. All outputs are considered potential candidates for release, and 'rules-of-thumb' are used to provide a first approximation as to whether to approve the output or not. An output checker may decide that the output can be cleared even if it breaches the rules-of-thumb; or it may be blocked even if it meets the rules-of-thumb. A researcher can always make a case that a blocked output should be released; the output checker should consider such a case, but is under no obligation to accept it.

PBOSDC only works if both output checker and researcher understand the rules of output clearance, and therefore this requires training of researchers. To operate efficiently, thought also needs to be given to the incentives for researchers to produce good output, as this can dramatically affect the resources needed. However, if effective training can be implemented, then PBOSDC, when combined with the 'safe statistics' approach, is demonstrably more efficient and safer than other methods of output checking. It is also easily scalable, as the bulk of the work is done by the researcher who is incentivised to produce good output.

4.6.4 Current practices

Currently most vRDCs employ PBOSDC (although a smaller number explicitly acknowledge it), and have training programs reflecting that. In contrast, most guidelines devised for circulation with SUFs

[45] For a detailed discussion of the two alternatives see Ritchie F. and Elliot M. (2015) "Principles- versus rules-based output statistical disclosure control in remote access environments", *IASSIST Quarterly* v39 pp5-13

are implicitly rules-based (in that the guidelines rarely explain how the context is important), and focus on simple tabular outputs. An exception is Eurostat's new online training for researchers, which takes its reader through the basics of PBOSDC and 'safe statistics'.

4.6.5 Outputs: summary

Until recently output SDC meant simply protecting tables. Modern output SDC was developed for vRDCs, and then percolated outwards, so that there is a great deal of consistency amongst the relatively small group of organisations that actively consider output SDC:

- Principles-based output SDC (PBOSDC) works best for research environments, and although costly to train, can bring auxiliary benefits in terms of researcher engagement.
- PBOSDC is also more scalable than rules-based models, if researcher incentives are set appropriately.
- The 'safe statistics' classification can significantly reduce workload and improve security.

5. Managing institutional expectations

5.1 The role of defaults

Institution attitudes and behaviours are often overlooked when considering the development and progression of data sharing. However, these are often a significant barrier to data access.

Many variables can influence an institution's attitudes, such as the institution's reputation, social responsibility and influence. When a new or unknown challenge arises institutions will often use their current institutional blueprint to frame the situation. This often leads to a conservative closed (defensive) mentality, particularly if the institution members have been conditioned to react this way towards potentially risky situations[46].

The defensive mentality puts the onus on those advocating change to justify their position, and, as noted above, a literature focused on theoretical negatives makes winning such an argument very difficult. The default perspective of the organisation therefore has a significant effect on the implementation of strategy. In particular the default-open/default-closed debate directly affects how decisions are taken.

Almost all organisations releasing data today are default-closed in respect of data access. This makes sense in the context of defensive decision-making in government, but there seems to be overwhelming agreement among public servants that this is inappropriate and default-open should be the organisational strategy[47]. However, very few organisations have actively tried to follow such a strategy, ABS being a recent and rare exception. There is therefore a massive disconnect between what individuals believe is the right way to act and the way they do act.

Part of the reason for this may be that asking questions about attitudes forces individual to focus on their public service ideals, whereas day-to-day decision-making is likely to be strongly influenced by institutional factors. Research suggests that hypothetical questions are not an ideal way to model individual behaviour, particularly in the complex hierarchies which predominate in government. Nevertheless, the few data access groups that have adapted the default-open model do seem to find that staff welcome a positive approach.

5.2 Costs and charging

Attitudes to whether data provision should be charged for vary considerably. There are several conflicting imperatives; arguments put forward for charging, or not, include:

- The cost of providing elective services should be recovered from those using those services
- The full cost of providing services should be covered
- The marginal cost of providing a service should be covered

[46] For a review of the literature on risky decision-making in government, see Ritchie F. (2014b) "Resistance to change in government: risk, inertia and incentives". Working papers in Economics no. 1412, University of the West of England, Bristol. December.

[47] Hafner et al (2015, ibid.) report the result of show-of-hands polls across a wide range of data professionals over two years. Almost invariably, 90% of respondents questioned about their personal attitude say "I am default-open"; but in a follow-up question to the same group, again 90% of respondents say "…but my organisation is default-closed".

- Data should be considered a public asset, and only the costs associated with provision of the data should be covered
- Costs should make a contribution to the collection and organisation of data
- Public benefit is maximised by not charging at the point of access
- A charge helps to discriminate between frivolous and valuable projects
- A large charge might discourage worthwhile projects which have not been able to attract external funding and
- Small charges might not bring in enough to cover the administrative costs of charging.

As a result, the only consistency in international access is that PUFs are not charged for, as these are freely available. Elsewhere, practice varies enormously, for example:

- In the UK most data access is provided free to users, with central funding covering the costs; but the ONS vRDC did charge users a daily rate for some years, then stopped, and is now reconsidering charging again; and some data linkage is chargeable.
- In the US, institutions hosting a Census Bureau RDC had to cover the whole cost of the facility, including on-site Bureau staff, which was then recovered from users; a similar facility in Canada was centrally funded.
- The SURE vRDC charges for the level of computing power required.
- The Dutch NSI's vRDC charged on the basis of number of outputs, with the first two outputs being free.

There are some consistencies. Where data is together for a specific project (for example, in Scandinavia and the Netherlands), this is usually chargeable; where the data is created once and then access to it granted, this is less likely to be charged. Were vRDCs have a daily charge for use of the facility (and many do), the going rate for this in many countries seems to be about AU$100/day; there is no obvious reason for the consistency.

The question of charging is also complicated because each data access system has its own particular costs structure; see Table 7 for some very simple indications of the relative cost of solutions.

Table 7 Indicative cost structures for access options

	Set-up cost	**Ongoing maintenance cost**	**Ongoing user cost**	**Marginal cost of an additional user/project**
PUF	High	None	None	None
Remote tabulation	High	Low	None	None
SUF (standard file)	High	None	None	Low
SUF (bespoke file)	Low	None	None	High
vRDC	High	Low	Low-medium	Low
RDC	Medium	Medium	Medium	Medium

In summary, every charging model seems to be used somewhere; and because a country tends to choose a charging regime and stick with it, there is very little evidence about the impact of charging on demand for or supply of data.

5.3 Stakeholder/customer relationships

5.3.1 Identifying stakeholders

As noted above, ethics boards seem to be more effective when there is a high level of engagement. This holds true for other stakeholders. For example, the support of the UK Treasury was crucial in nurturing ONS' vRDC through its initial stages, particularly when funding was under threat.

It has been argued that having user groups on board reduces the risk burden placed upon data owners[48]. Typically, data owners carry the risk if a data release strategy goes wrong; if however the strategy works perfectly, the users benefit, not the data owner. This imbalance is cited as one of the reasons data owners prefer defensive decision-making. Getting the users to actively acknowledge and support the risk taken by the data owners may allow the data owners to get some cover for the risk, and also be rewarded for taking the risk. While this is an appealing theoretical construct, there is very little evidence to back this up. Very few organisations formally make joint statements about accepting the risk, and very few data breaches occur, and so the idea of the user as white knight has not been tested to date.

One stakeholder group increasingly seen as key is the privacy lobby. A number of bodies (such as the UK Administrative Data Research Network) have ethics committees or advisory boards which contain people who, in theory, are fundamentally opposed to the operations being considered. For these groups, the rationale for involvement is quite different. Instead of "would you like to help make this happen?", the question is "something is going to happen in some way; would you like to help make sure it's done as well as it should be?" Feedback from such arrangements seems to be positive: privacy campaigners have their concerns listened to and sometimes addressed, whereas data managers challenged to justify their actions find that the rationale for their work is strengthened.

5.3.2 Identifying benefit for the data owner

The data owner is often under an obligation to provide access to the data. However, some organisations have tried to reverse the data owner's perspective, seeing data release as a positive benefit. As noted above in the discussion on 'safe projects', two justifications are typically made:

1. Methodological input, with expert data users feeding back information to the data owners, and hopefully engaging in dialogue on use and technical matters and
2. Outputs which are relevant and interesting to the data owner.

As with many topics discussed in this section, these seem like good ideas but there is very little evidence one way or the other for the difference that they make. There are numerous specific example of researchers having input into the work of the data owner, but very few organisations systematically integrate external users into their internal review mechanisms. Similarly, there are examples of data owners requiring non-technical summaries of their work to be written on project

[48] Ritchie F. and Welpton R. (2012) *Data access as a public good* in Work session on statistical data confidentiality 2011, UNECE/Eurostat.

completion (for example, at the IAB in Germany), but again it is not clear how well this actually feeds back into the organisation.

5.4 Public buy-in[49]

Most data owners are aware of the potential public relations impact of approving access to confidential data. Irrespective of the legality of data sharing and the ethical considerations discussed by expert committees, public expectations can have a profound effect on the prospects for data access.

For example, in the UK in 2014 a plan called "care.data" was unveiled to improve the use of GP data for research. This project became a source of much media interest; whilst scientific journalists made careful critiques of the plans, in the popular press it was characterised as "government selling your health data to insurance companies". The programme was abandoned with little serious public discussion over the programme's pros and cons, or whether the shortcomings could be addressed; public health professionals reported that the extremely negative reaction has made data owners more wary of data sharing generally.

However, this media storm is not representative of public perceptions. A number of studies show that in the context of medical, by a significant majority, the public is comfortable with:

- their data being used in research, particularly by academics
- their data being made available to 'trusted' organisations
- broad consent being used to carry out studies, and no need to obtain consent for specific projects and
- multiple datasets being linked to carry out research.

Very few data owners in the social sciences actively devise strategies for public awareness (although health organisations seem to be slightly better at this). Some have strategies for engaging the academic community – for example the NORC Data Enclave runs an annual programme of events and conference stands for the academic and government community. However, engagement with the non-academic public is much rarer. This may be affected by the experiences in public health, which show that, while the general public is generally very well disposed to medical research, the answers are very sensitive to the way the questions are asked and the wrong answer can have major repercussions, as in the care.data case.

A potential model is given by the Administrative Data Research Network, set up in 2013 to facilitate research access to UK government administrative data. The PR team was appointed early, privacy campaigners were invited to contribute substantively to the design of processes, and the ESRC produced a cartoon to publicise the role of the ADRN[50]. The cartoon is thoughtfully designed: most of it describes the benefits for policymaking and society in general from using this data for research. The security of the process is treated in a very low-key manner, as if it were not worthy of attention; one would expect the data to be managed to best international standards, so let's move on. Given that the whole ADRN project has much potential for adverse comment, the lack of any such comment to date suggests that an early, active public relations programme can draw dividends.

[49] This section is based on the findings in the Wellcome Trust (2015, ibid.) report
[50] https://www.youtube.com/watch?v=E3e4D2bHxa8

5.5 The Nordic example

An extraordinary system is currently under development in the Nordic countries[51]. Norway, Sweden, Denmark, Finland, Iceland and Greenland have agreed to allow de-identified register data from the six countries to be combined for cross-Nordic research projects. The dataset created for the project will then be made available to the researcher at his or her workplace through one of the vRDCs operated by Denmark, Sweden or Finland. So, an Icelandic researcher could carry out research on Norwegian and Finnish data hosted in the Swedish vRDC.

To get this working, the six countries have agreed[52]:

- a common application form containing the information needed from all Nordic countries
- a common Nordic security agreement (to be signed by all researchers as well as each relevant NSI)
- an agreement between the relevant NSIs on the data transfer to ensure a common understanding of the regulatory environment, including elements on data security and
- a structure for communication between the NSIs about data security breaches.

It could be argued (and the management team acknowledges) that this sort of project could only be done in these countries. All of them have similar register-based government data systems, have a similar legal framework, and are culturally similar. They are exceptionally open to the re-use of government data for research; there is widespread popular support for the concept and trust in the management of data. Finally, all have similar principles for their existing approaches to data access, even if practices differ.

Nevertheless, this is still a remarkable achievement. As noted above, even agreeing on a common access contract has eluded many organisations within countries, let alone across countries. The Nordic agreement is up and running barely a year after the idea was considered in a review. There are potential lessons for DSS in the way that attitude can bring common goals forward.

[51] Nielsen and Thaulow (2015) provide detail, but a clear presentation is available at http://www1.unece.org/stat/platform/download/attachments/109248612/Session%204%20-%20Denmark.pptx?version=1&modificationDate=1444313553135&api=v2

[52] Adapted from Nielsen and Thaulow (2015)

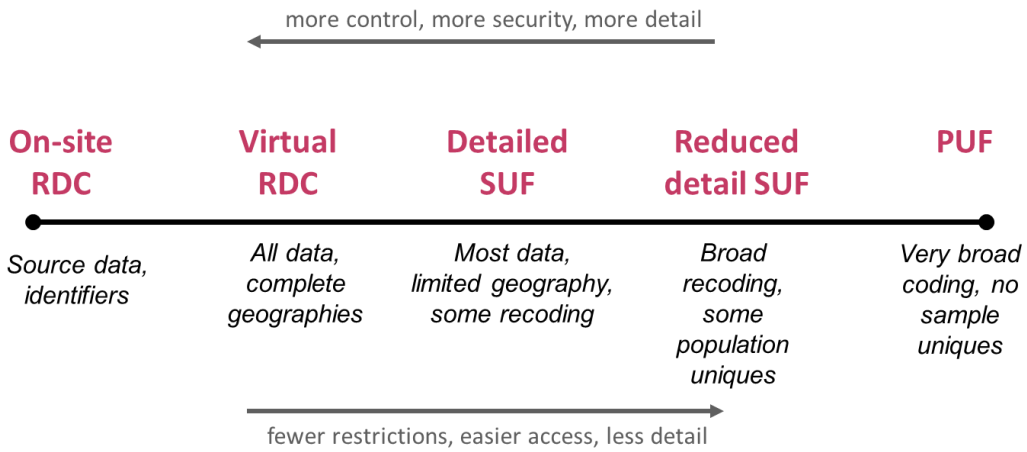
6. Developing a strategic direction

6.1 The data access spectrum

Developing a strategy for the organisation requires combining the operational and the institutional elements. To do that, a number of organisations simplify the operational aspects into a linear relationship, often called the 'data access spectrum' (DAS) or 'continuum of access', the latter term being coined by Statistics Canada and the Canadian RDC Network.

This takes as its starting point the idea that the non-statistical controls in the Five Safes model tend to move jointly, with 'safe data' left over as the residual, determined by the amount of non-statistical control being exercised. Users care about first and foremost is whether they have an appropriate amount of detail available to them (and they may not need very much), but are also able to balance this with accessibility. We can then derive a linear relationship with varying levels of data detail; the non-statistical controls are embodied in the type of access:

Figure 1: The Data Access Spectrum



This two-dimensional representation is appealing as it shows how different access systems relate to each other. Ideally it should cover (almost) all of any potential user's needs, helping them to decide how much detail they are willing to trade off for other restrictions on their research. It also has value as a planning tool, by identifying gaps in current provision for certain types of data or facility; and it was used as such to provide a rationale for two new facilities in ONS' portfolio[53].

6.2 Devising a strategic overview

The data access spectrum provides a simple representation of the operational options available. However, a fully strategic approach needs to consider the context within which the operational element sits: what is the internal attitude to designing data access and solving problems? And what does the wider public think about what we are doing?

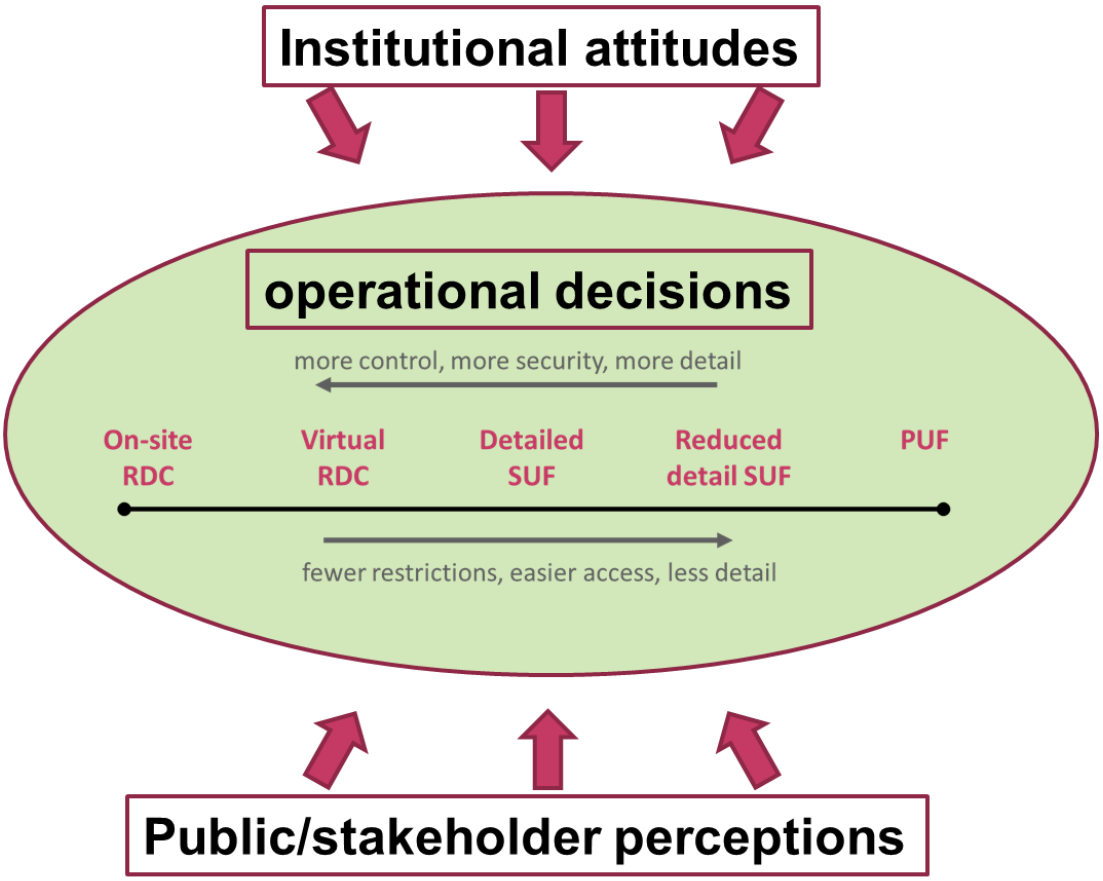
The Five Safes is a useful way of structuring discussions, and has helped data owners to exploit developments in technology such as vRDCs. But substantial gains in efficiency and access have also

[53] ONS (2011) *Data Access Policy.*

come from changes in institutional attitudes, which strongly affect the way that operational decisions can be made. Finally, an area which is relatively unexplored, but which seems likely to become more important as access to administrative data increases, is winning the battle of public perceptions.

Figure 2 illustrates the relationship between the elements. This tripartite framework will be used in the remainder of the report to discuss proposals, options, and activities.

Figure 2 Attitudes, operations and public relations



Part III Mapping users and solutions

Part II compared alternative data access solutions, and studied institutional barriers to effective data access. This was all done in the broad context of international experience. In this part we consider the specific case of DSS and its data access solutions. The next section maps the users identified in Part I to the solutions identified in Part II. It considers where DSS is currently, and what specifically needs to be developed. The structure for part III is as follows:

- Section 7 develops a template for analysing the characteristics of the users, and uses this to analyse the users from Part I and put them into three groups: 'non-expert', 'professional researcher', 'other'
- Section 8 maps these groups to particular solutions and
- Section 9 covers overarching issues, such as attitudes and public perceptions.

7. User descriptions

We begin by considering the groups discussed in Part I in more detail, broken down into three sections. The overall characteristics which need to be considered are as follows:

Table 8 Descriptions of characteristics to be reviewed

Topic	Sub-topic	Question to be asked
Statistical needs	Likely use	What sort of analyses will this group run, in general?
	Need for detailed categories	How interested will they be in the more complex categories in the data?
Limits on usage	Willing to travel to approved site	Is the user willing to physically travel to a site to get access?
	Availability of funding	Is the user able and willing to pay for access, including the costs of travel to sites?
	Can wait for access to be approved?	Is a slow application process (days or weeks, rather than minutes) likely to turn off users?
	Will wait for results to be released	If requested analyses are not instantly available, does this discourage users?
Skills	Knowledge of DSS data	Will the user know what data they are looking at?
	Statistical skills	Will the user know how to interpret or manipulate the data correctly?
	Support needed on…	On what topics are they likely to ask for support?
Benefit to DSS	Research likely to be of value to DSS?	Is the analysis likely to be of direct interest to DSS?
	Can contribute to DSS methodology?	Is the user's perspective likely to usefully inform DSS' methodology (ie what benefits can DSS get)?
Contracting	Willingness to register	Can we expect the user to register themselves?
	Identity confirmable	Can we set up systems to reliably confirm the user's identity?
	Personal contract	Is a contract with the user feasible?
	Personal and institutional contract	Is a contract with both the user's home organisation and the user feasible?
	Accreditation from other bodies	Can we rely upon/remove accreditation from bodies other than DSS?
Incentives	Identifiable repeat user	Is this user likely to be a regular user?
	Relevance of legal sanction	Will legal sanctions be seen as important by this user?
	Relevance of employment sanction	Can actions be taken to reduce the user's employment prospects?
	Relevance of funding sanctions	Can action be taken to restrict the user's access to research funds?
	Relevance of access sanction	Can action be taken to restrict the user's access to DSS or other data?
	Relevance of reputation	Is the user's professional reputation important?
Options	Distributed PUF	Fully anonymised data, unrestricted circulation
	Distributed SUF	De-identified, partially but not fully anonymised, restricted circulation
	Remote access SecUF	De-identified data available only through secure link to virtual RDC
	Remote tabulation	Live tabulation server with confidentiality protection
	Remote job submission	Access to run code on de-identified, possibly partially anonymised data
	Secure pods	Access to SecUF via virtual RDC from dedicated secure physical space
	Physical RDC	Access to SecUF from DSS RDC
	In-posting	Access to DSS facilities on-site
Precedents		Are there useful precedents from Australia or other international practice for the proposed solution?
Training	Compulsory?	Should compulsory training be required?
	Purpose of training	What is the aim of the training?
	Delivery	What delivery methods are preferred?
	Content	What is the content of the training?

7.1 The 'non-expert' group

The first group is the 'non-expert' group: the general public, journalists and non-PhD students. While there may be considerable expertise in the group about DSS activities we assume that this group will not be interested in statistical detail. The main focus will be on relatively simple statistics, delivered quickly on an ad-hoc basis. Undergraduate and Master's students are likely to be interested in microdata but do not require access to detailed data.

Table 9 Characteristics of the 'non-expert' group

		General public	Journalists	Honours & Masters students
Statistical needs	Likely use	Simple tabulations	Complex tabulations	Microdata for analysis
	Need for detailed categories	No	No	No
Limits on usage	Willing to travel to approved site	No	No	No
	Availability of funding	No	Yes	No
	Can wait for access to be approved?	No	No	No
	Will wait for results to be released	No	No	Yes
Skills	Knowledge of DSS data	Little	Good	Little
	Statistical skills	None	None	Basic
	Support required for metadata	Yes	Yes	Yes
Benefit to DSS	Likelihood of value to DSS?	No	Probably not	No
	Contribution to methodology?	No	Yes	No
Contracting	Willingness to register	Yes	Yes	Yes
	Identity confirmable	No	Yes	Yes
	Personal contract	No	Possibly	Possibly
	Personal and institutional contract	No	Possibly	No
	Accreditation from other bodies	No	No	No
Incentives	Identifiable repeat user	No	Possibly	No
	Relevance of legal sanction	No	Yes	Yes
	Relevance of employment sanction	No	No	No
	Relevance of funding sanctions	No	No	No
	Relevance of access sanction	No	No	No
	Relevance of reputation	No	No	No
Options	Distributed PUF	No	Yes	Main
	Distributed SUF	No	No	No
	Remote access SecUF	No	No	No
	Remote tabulation	Main	Main	Yes
	Remote job submission	No	No	No
	Secure pods	No	No	No
	Physical RDC	No	No	No
	In-posting	No	No	No
Precedents		UKDA, TableBuilder	TableBuilder	UKDA, TableBuilder
Training	Compulsory?	Optional	Optional	optional
	Purpose of training	Understand DSS activities	Understand DSS activities	Understand DSS activities
	Delivery	Online	Online	Online
	Content	DSS data	DSS data	DSS data

For the non-expert group, speed of service is essential. The lack of reliable identification of users and the inability to have contracts or effective sanctions means that DSS has very few options to control

use of the data. PUFs and remote tabulation appears to meet the user needs. Optional training might secure some additional benefit to DSS by encouraging knowledge and understanding of DSS activities – and could play a role in engaging the wider public; see below - but there is limited chance for DSS to engage with and gain from this user group.

7.2 The 'professional research' group

This group comprises PhD students, academics at higher education and research institutes, and researchers in government departments (we do not distinguish between federal and state government, although we understand there are legal implications).

This group is assumed to be statistically expert and interested enough in DSS activities to become expert in those as well; willing to invest time and money in gaining access if the data quality merits it; amenable to good-practice security training; and willing to engage with DSS in methodological discussion or dissemination strategies. Not all academics are keen to engage with data providers, but as the international review showed, academic disengagement seems to be more a sin of omission rather than commission: communication channels not suited to academic interests tend to be ignored, even if there is a compulsory element to reporting.

Two key features of this group are:

1. the likelihood of repeated and long-term engagement and
2. a wide range of sanctions (law, employment, reputation, funding, access) available in the event of any breach of procedures or confidentiality.

In theory, this makes these users ideal candidates for the distribution of very detailed SUFs, as the users have a strong incentive to look after the data and communicate with DSS. Unfortunately, the certainty of mistakes in following procedures makes this a high-risk strategy. Hence most recent work has gone into developing vRDCs which give local-like access but without the data being local. Less detailed SUFs then provide a complementary service to those who do not need access to the complete data.

As noted in the review, training is increasingly seen as the key battleground to win 'hearts and minds' of researchers; in particular, selling the idea that the restrictions are a positive thing because they help to protect the researcher from committing crimes accidentally. This is where the strategy of engagement with DSS starts, and hence for this group face-to-face training is universally adopted.

Table 10 Characteristics of the 'professional researcher' group

Group		PhDs	Academics (university and research institutes)	Government researchers
Statistical needs	Likely use	Microdata for analysis	Microdata for analysis	Tabulations, microdata
	Need for detailed categories	Yes	Yes	Yes
Limits on usage	Willing to travel to approved site	Yes	Yes	Yes
	Availability of funding	No	Yes	Yes
	Can wait for access to be approved?	Yes	Yes	Yes
	Will wait for results to be released	Yes	Yes	Yes
Skills	Knowledge of DSS data	Very good	Very good	Very good
	Statistical skills	Fully competent	Fully competent	Fully competent
	Support needed on…	data detail	data detail	data detail
Benefit to DSS	Research likely to be of value to DSS?	Yes	Yes	Yes
	Can contribute to DSS methodology?	No	Yes	Yes
Contracting	Willingness to register	Yes	Yes	Yes
	Identity confirmable	Yes	Yes	Yes
	Personal contract	Yes	Yes	Yes
	Personal and institutional contract	Possibly	Yes	Yes
	Accreditation from other bodies	Yes	Yes	Yes
Incentives	Identifiable repeat user	Yes	Yes	Yes
	Relevance of legal sanction	Yes	No	Yes
	Relevance of employment sanction	No	Yes	Yes
	Relevance of funding sanctions	No	Yes	Yes
	Relevance of access sanction	No	Yes	Yes
	Relevance of reputation	Yes	Yes	Yes
Options	Distributed PUF	Yes	Yes	Yes
	Distributed SUF	Yes	Yes	Yes
	Remote access SecUF	Main	Main	Main
	Remote tabulation	Yes	Yes	Yes
	Remote job submission	No	No	No
	Secure pods	Yes	Yes	No
	Physical RDC	No	No	No
	Inposting	No	No	No
Precedents		Many places, not all	Everywhere	MADIP, VML, ABS
Training	Compulsory?	Yes	Yes	Yes
	Purpose of training	Engagement and security	Engagement and security	Engagement and security
	Delivery	Face-to-face	Face-to-face	Face-to-face
	Content	DSS data, policy, security	DSS data, policy, security	DSS data, policy, security

7.3 The 'other researchers'

The third group comprises private sector researchers and foreign researchers. We assume these two groups have the same skills as the Australian public-sector researchers, but these two groups are treated differently as there are concerns about the incentives for this group and the effectiveness of sanctions; in the case of overseas researchers, there is also the question about the legality of cross-border data access and the jurisdiction of contracts.

Table 11 Characteristics of the 'other researcher' group

Group		Private sector researchers	Foreign researchers
Statistical needs	Likely use	Tabulations, microdata	Tabulations, microdata
	Need for detailed categories	Yes	Yes
Limits on usage	Willing to travel to approved site	Yes	No
	Availability of funding	Yes	Yes
	Can wait for access to be approved?	Yes	Yes
	Will wait for results to be released	Yes	Yes
Skills	Knowledge of DSS data	Very good	Good
	Statistical skills	Fully competent	Fully competent
	Support needed on…	data detail	data detail
Benefit to DSS	Research likely to be of value to DSS?	Yes	Yes
	Can contribute to DSS methodology?	Yes	No
Contracting	Willingness to register	Yes	Yes
	Identity confirmable	Yes	Yes
	Personal contract	Possibly	No
	Personal and institutional contract	Yes	Yes
	Accreditation from other bodies	Yes	Yes
Incentives	Identifiable repeat user	Yes	Possibly
	Relevance of legal sanction	Yes	No
	Relevance of employment sanction	Possibly	No
	Relevance of funding sanctions	No	No
	Relevance of access sanction	Yes	Yes
	Relevance of reputation	No	Yes
Options	Distributed PUF	Yes	Yes
	Distributed SUF	Yes	Main
	Remote access SecUF	Main	possibly
	Remote tabulation	Yes	Yes
	Remote job submission	No	No
	Secure pods	Yes	Yes
	Physical RDC	No	No
	Inposting	No	No
Precedents		NCLD, UK, NORC	IPUMS
Training	Compulsory?	Yes	Yes
	Purpose of training	Engagement and security	Security
	Delivery	Face-to-face	Online
	Content	DSS data, policy, security	Security

For the private sector researchers, there is enough evidence to support the contention that this group can be trusted to use restricted access facilities. There might be concerns raised about access

to SUFs where use is unsupervised, but again there is little evidence to suggest this is a significant risk. The main reason for preferring SecUFs in an RDC is that private sector researchers are likely to be more 'collegiate', ironically, than academics and are used to working with data in an administrative sense; they are therefore more likely to share access to the data and to focus on 'interesting' rather than statistically valid findings. However, this is largely speculation; there have been no formal studies on this topic, and this may reflect the prejudices of the public sector (including us) rather than any evidenced risk. The public relations aspect of this also needs to be managed; providing access to private sector researchers often alarms the general public. This is where the secure pods might be appropriate, but the UK evidence suggests that the pods increase concern over security rather than diminishes it, and so these are not recommended in the first instance.

For foreign researchers, as well as extra-territorial legal difficulties, there are geographical issues which affect the outcome. First, users are unable to attend face-to-face training, and may also be unwilling to engage in conversation with DSS if there are time zone problems. Foreign users are also likely to be less interested in DSS policy interests then locally-based ones, and hence the gain from engagement is smaller.

8. Mapping users to solutions

8.1 The 'non-expert' group

As previously defined the 'non-expert' group comprises of the general public, journalists/ non-specialist business users. For the general public and journalist/ non-specialist business users a Remote tabulation tool should be sufficient in meeting this group's needs. Users can create their own tabulations of the data through the tabulation tool, rather than relying on the data owner's choice of tables or bespoke tabulations. Tabulation tools do not require user training, allowing for the data to be accessed remotely.

It is not viable for untrained groups to access 'detailed' datasets as this group may not possess sufficient research training to analyse the data sufficiently. When considering synthetic PUFs (1) there would be a risk of the group treating the data as non-synthetic, and (2) the group may not be able to analyse the data appropriately. With a detailed SUF disclosure control also becomes an issue. Finally, the group is likely to be very large, and trying to process these users through a vRDC model would be too resource intensive from a disclosure and training perspective, even if they had the technical skills.

Pre doctoral students may also use remote tabulation tools, but they may also want synthetic PUFs which are expected to have the same characteristics as the real data but are imputed from statistical models. There is no disclosure risk from invented data, but the analytical value in purely invented data is limited; hence it is useful as a teaching aid but not for research. In public health the value of synthetic data seems low given the importance of accurately assessing recording health events; however, synthetic data have been used in data fusion models to generate simulation models for policy analysis which can be relevant for pre doctoral students. Both tabulation tools and synthetic PUFs do not require user training, allowing for the data to be accessed remotely without limit.

8.2 The 'professional research' group

This group comprises PhD students, academics at higher education and research institutes, and researchers in government departments. In practice this group will require access to the full spectrum of data, from basic tabulated data to SUFs to full detail in the vRDC.

For this group it is important that the whole range of options is available. Not all academics need all levels of data, and academics can discriminate. The basic 'rule of thumb' used by the UK Data Service is something like:

1 vRDC project = 10 SUF projects = 100 open data downloads.

For this group we would expect the full detail to be available in the vRDC, given other controls and the experience of academics working in vRDCs in other countries. SUFs should also be available, created once rather than for each project; these are likely to be lower cost for DSS, and so it would be desirable to direct users to these if they do not need to full SecUF detail. It might also be worth considering whether multiple different 'levels' of SUFS are necessary, perhaps with slightly more or less stringent approval requirements.

The group of researchers would be expected to undergo and appreciate the importance of training in using confidential data. For the vRDC users, face-to-face training should be required; this does require additional time and cost given Australia's geography, but should be seen (and sold) as an investment by both parties. For SUFs, a modern on-line training program such as the Eurostat or SURE modules should be sufficient.

8.3 The 'other researchers'

Once the infrastructure for the other groups is set up, this group can be thought of as the 'professional researchers' group but with extra restrictions, for example reflecting public concerns about access for private companies. Hence we do not devise any new solutions for this group, but suggest that DSS might consider whether it is comfortable with this group having access to data at the same level of detail as the other group.

8.4 Summary: user-solution map

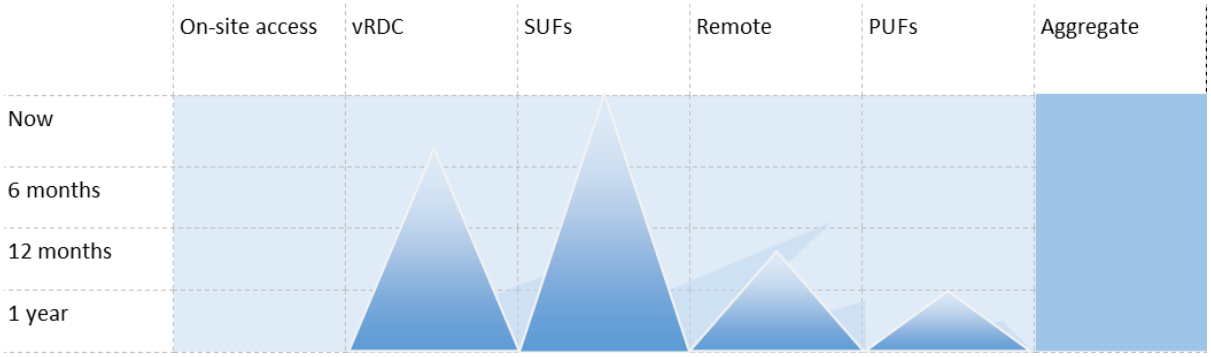
Table 12 below summarises our mapping of the user requirements and feasible solutions. It also suggests a preferred 'order' if one data access mechanism were to be tackled at once. In the next section, it is assumed that teams could be working on multiple solutions at a point in time.

Table 12 User-solution map

Proposed solutions		**Synthetic PUFs**	**SUFs**	**vRDC**	**TableBuilder**
Population	general public				Y
	journalists/non-specialist business				Y
	pre-doctoral students	Y			Y
	doctoral students	Y	y	Y	Y
	academic/gov researchers	Y	y	y	Y
	international researchers	Y	y	y	Y
	specialist business researchers	Y	y	y	Y
Timing	Priority	Last	Third	First	Second
	Rationale	lowest value user group	Something in place already works	The most controversial; get it right first	Existing solution works outside DSS; need to integrate re-use
Information required to set up solution		Demand for synthetic data	Set of (minimal) appropriate SUFs to create Willingness to pay for data	Set of (minimal) appropriate SecUFs to create Willingness to pay for access	Demand for live tabulation

Given our proposed priorities and the existence of SUFs at the moment, we would see the DAS evolving over time as follows, so that within a couple of years DSS has a complete suite of data access options available to the different groups:

Figure 3 Completing the data access spectrum for DSS



9. Overarching issues

The previous section covered practicalities of specific solutions. This section addresses the overarching issues which relate to DSS' general strategy.

9.1 Common operational issues

There are a number of operational issues which are better tackled as a corporate issue rather than being related to specific solutions.

9.1.1 Mechanics of access and personal agreements

Currently DSS has different applications arrangements and access agreements depending on the type of access. It should be feasible to streamline these significantly, for example by having a single landing page on the website for those requiring access to microdata or bespoke tabulations. This would include information on the criteria for a successful application. A model might be the Eurostat microdata website.

It was noted earlier that there is a tendency for access agreements to include excessive amounts of information, and that a simpler version would be no less binding in law but should be more defensible. DSS has the opportunity to draft a simple access agreement which could be used for all its data access operations. For example, consider the sample access agreement text in Figure 4 below. This text has been designed to reproduce the messages from training, in language that is relevant to the users. Points to note are:

- The aim is to encourage good behaviour, rather than threatening users: "I must…" instead of "you must not…"
- The references to DSS could be replaced by 'the data owner' or some other wording that would make this generic across Australian government departments.
- The references to mistake reflect specific provisions in the laws covering DSS data; however, they are really there to demonstrate understanding of human frailty and the adult line taken by DSS.

Figure 4 Simplified access agreement text

I, [researcher name], have been granted access to the following dataset(s) for the purpose of project [project reference] for the period [period].

[list of datasets]

I understand that:

- I must only use the data for the purposes and period specified in the project application
- I must only discuss details of the data with my co-researchers listed on the project application
- I must only share the data with those listed on the project application as having approved access to unit record data
- I must not link other unit record data to these data without prior approval
- I must apply appropriate statistical disclosure control techniques to my outputs, as specified in my training
- I must follow all instructions about data handling as given in my training
- If I make a mistake in any areas of data handling, this will not be penalised as long as (1) I notify DSS as soon as is practical after discovering the mistake and take remedial action as requested (2) I do not repeat the mistake.
- If I observe others handling data inappropriately, I must notify DSS; if I observe others making mistakes in data handling, I may encourage them to report themselves to DSS, and I need only inform DSS if the other parties do not do so
- Any changes to data handling (that is, different people on the project, a different purpose, a different period of access, or a different dataset) must be agreed with DSS team in advance of acting on those changes and
- Where any uncertainty exists, or if I need specific guidance, I should contact DSS team for clarification.

I also understand that:

- Access to data for research is a public good which must be protected
- As a researcher, I am in a position of trust for which I have received training
- Any breach of access conditions may lead to a combination of legal, financial and operational sanctions against myself and my institution and
- My institution will be involved in any breach of access conditions where sanctions are applied.

9.1.2 Ethical and operational approval

Ethical approval arrangements are not clear within DSS and there is no overarching mechanism for ensuring that data access requests are directed to the right access arrangement and do not cause conflict (for example, having a valid public benefit in one case but not another).

A potential model for microdata approval might be the ONS' Microdata Release Panel, which has oversight of all requests for microdata from ONS. It keeps this manageable by the (informal) use of classes of data access and the (formal) use of precedent.

An approval process wanting to follow best practice could also:

- take the position that the default assumption is the project will be approved; the role of the approvals panel is to check that there are no reasons why the project should not go ahead, and suggest solutions if necessary
- allow delegation of approval in certain circumstances, such as minor changes to project specification, extensions to the timetable or researchers involved, or a project very similar to a precedent
- identify where external approval (for example, from health service ethical boards) will be accepted and
- put information about all of this on the DSS website.

9.1.3 Institutional access agreements

DSS has expressed an interest in having institutional access agreements as well as individual ones. This would be supported by international experience although, as Part II noted, there is very little evidence to suggest that researchers or institutional administrators are aware of the implications of such an agreement.

DSS, and the wider Australian Government, could help to define good practice by trying to develop in institutional administrators the same sort of active engagement that is desired of researchers. For example, it should be feasible to develop a training course when an institutional agreement is first signed; the institution and DSS could share information on activity; and the administrator could provide perhaps a short annual statement of accession to various standards.

A system such as this implies a trained cadre of institutional staff, whom we will refer to as "Institutional Responsible Officers" (IROs). Identifying individuals might be problematic, particularly in universities. Academics generally tend to avoid posts with an administrative element, and there would be a natural tendency for universities to appoint a data protection officer to this post, whereas the recent experience shows that having research-active staff involved in administration is one of the most effective ways of keeping the project user-centred and developing new ideas. One solution would to use the training for IROs to develop sense of the community responsibility of the role.

As there is no precedent for this, DSS would be pathfinding for others, and there are likely to be numerous problems. However, it would demonstrate DSS' commitment to actively managing risks, rather than relying on the ticking of boxes.

9.1.4 Charging policy

As noted above, there is no international consistency on charging. Our suspicion is that unless the full cost of the service is to be recovered, the arguments for central funding (that is, free at the point of use) are stronger than the cost recovery argument.

SURE will be charging DSS on a user/project basis for the initial pilot. In the short term, the SURE costs can be absorbed into the overall costs of the pilot. In the longer term, this may not be feasible.

9.1.5 Data and trusted user maps

Some organisations have found the Data Access Spectrum helpful in defining types of data. For example, in 2010 ONS defined two data types for its own vRDC, the VML, and the vRDC run by the UK Data Service, the SDS:

1. VML data: source data at postcode level with only direct meaningful identifiers (names, house number, street) removed and
2. SDS data: as VML data but with unique identifiers (company or tax registration numbers) removed, and no data less than a year old.

This had two effects. It simplified discussion over what data should go in which facility; and it also set a default standard, against which data managers unwilling to release data had to argue. This was an important change. Prior to that date, the VML team used to spend time each year arguing with divisional data owners whether a certain level of detail was really necessary; after the change, no divisional data owner challenged the definition. We would suggest DSS take a similar approach to clarify how data should be allocated to the different options. This also helps to plot gaps in the data provision: is there any data which is not available in some way?

9.1.6 Training programme

Best practice clearly indicates the use of a training programme, for both SUF and SecUF users. In the short term, there are several options:

- The SURE online training provides a good overview of the principles of good data management, and is suitable for both SUF and SUF users; on its own, however, it is not sufficient for SecUF users as it does not address the issue of the community of users[54], and the OSDC module is limited and does not cover modern OSDC principles; this is not surprising, as SURE is a facility where clients have much flexibility to set up the gateways in the way that they want with the controls that they want.
- The ABS Secure Researcher training does cover all the necessary elements of safe use, but not the SURE environment.
- The Eurostat online training was designed specifically to reflect modern thinking in data security (including the EDRU ethos and modern SDC), and was targeted mostly at SUF users.

It therefore seems that the immediate training needs of DSS could be met as follows:

- SUF users: SURE or Eurostat training and/or
- SecUF users: SURE training as preparation, followed by ABS training.

In time, it is likely that DSS would want to develop training that reflects some of its interests, but this does not necessarily mean that separate training needs to be developed. In the UK, for example, integration of training across ONS, the tax department, education and health services has been achieved, to some extent, by ensuring that a wide enough range of examples is included.

[54] That is, while the SURE training is good on personal responsibility and avoiding mistakes, it does not actively promote the message that all users and the data owners have a shared interest (and hence responsibility) in effective data security.

Finally, it should be noted that output checkers will require training as well. We suggest joint training with their equivalents in ABS would be cost-effective.

9.1.7 Demand identification

All the above proposals assume that demand for DSS data products is known. This is unlikely to be the case, and international experience suggests that data demand and supply will evolve over time. We therefore propose that, in the medium term, DSS begins some market research to identify the scope and scale of the market for its data.

9.2 Attitudes

DSS has already made strong commitments to the 'default-open' and EDRU strategies, and has a Data Access Policy which makes many positive statements. However, at present this statement does not cover attitudes, and in our view, is not yet sufficient to support those countering defensive strategies and resistance to innovation.

Specific comments are:[55]

- Principle 1 needs extra clauses to emphasise uncertainty and judgment, and there is also a need to acknowledge that risks are acceptable, and mistakes are expected; the important thing is contingency planning. For example, additional clauses could include:

  "Any decisions about data access have costs and benefits. Even the decision not to release data has a cost, in that society cannot use that information to make judgements. DSS will aim to look at the whole costs and benefits from society's perspective."

  "Costs and benefits are uncertain: we do not know how the world will turn out, and therefore any decision is risky. DSS accepts that decisions must be made on the balance of probabilities, using the evidence available at the time of the decision, which implies that some DSS decisions may turn out to be wrong."

  "Decisions about data access are made by humans, who will make mistakes. It is more important that we quickly identify mistakes and learn from them, rather than apportioning blame."

- We would suggest re-ordering Principle 3 before Principle 2, to emphasise positive aspects.
- The current principle could establish the default-open status more clearly:

  "Line areas should assume that their data will be made available, and their role is to identify how to do this cost-effectively (maintenance of confidentiality is a cost)"

  "All decisions about data access should be based on evidence; in particular, decisions to restrict detail or access should demonstrate the threat being targeted and the credibility of that threat."

[55] These comments relate to v1.0, dated February 2015

- Principle 5 needs to bring the data user in, so the community of interest is being built early on (that is, the "we're all in it together" ethos). DSS might be legally responsible, but the whole EDRU aim is to get everyone to accept some share of responsibility.
- A start could be to change the title to include "but will seek to engage third parties", and add a clause:

  "DSS will seek to engage third parties (including researchers and policy users of research) in sharing responsibility for data releases."

- The detail of Principle 6 works well but perhaps the title could be changed to something like "Principle 6: DSS favours controlled access rather than distributing data for the most sensitive data".
- A clause could then be added that distributing data is an acceptable alternative where the release mechanism (including but not limited to restrictions on data detail) provides confidence that the release is low risk.
- Principle 7 seems problematic, and liable to legal challenge; it also limits detail by default, rather than keeping it open by default. We would suggest rewording as:

  "Researchers will only be given access to the datasets that they need to carry out their project. In exceptional circumstances, researchers may be restricted to a subset of the data."

In summary, existing policy documents, and statements made by senior DSS staff, provide a strong starting point; but some small changes in wording would emphasise the attitudinal change.

9.3 Stakeholder management

Given the concurrent developments in other Commonwealth bodies, there seems a strong benefit in early and specific negotiations across government to avoid slightly different competing solutions being developed. The Nordic model may be a useful guide: agree on the things that are possible, allow some variation, and accept that some things will have to be worked out over time.

The ultimate goal, which may be unachievable in the medium term, would be to have a single government solution on all aspects of data access: attitudes, forms, accreditation, training, and PR.

There is an opportunity to identify gains to data providers, including feedback from researchers, and build a system which can exploit those gains. This might include gains from TableBuilder and the PUFs, as well as from the hardcore research community. DSS may want to consider how it can create occasions for data providers and researchers to meet and share findings.

9.4 Perceptions

An early positive (but low-key) campaign, perhaps taking lessons from the Administrative Research Data Network, would be useful. Activities of the ADRN included:

- Production of high-quality information films on the public benefit value of data linking
- Production of guides aimed at the public, particularly journalists
- An active public engagement strategy, often in collaboration with the Royal Statistical Society and the Statistics Authority

- Inclusion of privacy campaigners on the ADRN Advisory Board
- Regular meetings with the Information Commissioner
- Regular meetings with heads of analysis at government departments and
- A dedicated YouTube channel containing presentations from meetings and events, to demonstrate transparency in activities.

We would recommend an early discussion with privacy campaigners and with users, and the development of an Advisory Board to allow a forum for different opinions to be expressed. We would recommend that the Advisory Board be composed of individuals who have authority in their own organisation to initiate change; this will be a key mechanism for helping to disseminate good (or at least common) practice across government.

Both of these activities could be developed by DSS independently, but there are substantial potential benefits from developing a whole-of-government approach. We would suggest DSS begins early talks with other departments engaged in developing data strategies (currently, ABS, Productivity Commission, DPMC) to explore the options for collaboration.

Part IV Items for action and roadmap

This Part details the recommendations relating to the solutions identified. The next section identifies, for each of the recommended solutions and for the institutional position, elements which are in place, the gaps, and potential ways to fill the gap. Section 3 then provides a timeline for the changes.

10. Items for action

The four options in this section are ordered by our assessment of priority to DSS: vRDC, TableBuilder, SUFs, and PUFs. We then consider the institutional changes needed to implement and fully benefit from the practical developments.

10.1 Virtual RDC

This project has been identified as the highest priority as a pilot and is imminent. The vRDC is also likely to be the most controversial, and so while this is high risk, this also affords the most opportunities

	Project		People		Settings		Data		Output	
Progression	**Action needed**	**Comments**	**Action needed**	**Comments**	**Action needed**	**Comments**	**Action needed**	**Comments**	**Action needed**	**Comments**
Things in place at DSS not needing significant change				Training is essential for a vRDC, and is the best chance to engage researchers; SURE offers training to its users	SURE environmental controls	The SURE system works and, in the absence of comparable systems, should be considered for use. There may need to be some configuration of the SURE environment				

	Project		People		Settings		Data		Output	
Progression	**Action needed**	**Comments**	**Action needed**	**Comments**	**Action needed**	**Comments**	**Action needed**	**Comments**	**Action needed**	**Comments**
Things done by others (not DSS) to be adapted			SURE training; ABS training Eurostat guidelines on good researcher behaviour	Guidelines for researcher behaviour need to be transparent and visible; in the short term the Eurostat guidelines are a good proxy, but in the longer term these could be replaced with guidelines suitable to the Australian context	Creation of SURE project environments Implementation of secure user environment: locked-down screen and session recording Secure thin clients and secure pods to be considered as long term options for extra-sensitive datasets	SURE produces guidelines for users; these need to be discoverable from DSS website			ABS training for researchers and output checkers Eurostat guidelines on PBOSDC and 'safe statistics' as background UK practical manual *Guidelines for output checkers* to be adopted and adapted	In the short run, the ABS training seems closest to international good practice
	Transparent and discoverable guidelines on process and requirements for approval	There is a need to demonstrate to stakeholders that the projects generate a benefit to DSS and the wider public							Transparent and discoverable guidelines on what is expected of researchers	
	Training for institutional contract-holders									
	Institutional agreements									
	Personal agreements									
	Mechanism for research findings to be disseminated meaningfully amongst data providers									

10.2 TableBuilder

Using the ABS' TableBuilder allows relatively quick delivery of an additional access solution for the general public, of whom researchers and lecturers are a part. Using TableBuilder also enhances the impression of joined-up government, and means that users do not need to learn a new tool. Finally, if both ABS and DSS datasets are hosted at the ABS, then users do not need to worry about which site to access.

	Project		People		Settings		Data		Output	
Progression	**Action needed**	**Comments**	**Action needed**	**Comments**	**Action needed**	**Comments**	**Action needed**	**Comments**	**Action needed**	**Comments**
Things in place at DSS not needing significant change		There is no requirement for project approval		Information on why TableBuilder produces some results inconsistent with other published tables should be transparent and discoverable, but not actively highlighted; this may just be referenced on the ABS website		The local setting is irrelevant for TableBuilder outputs		TableBuilder should run on the SUFs (see below), producing a balance between detail and confidentiality risk		There is no need for further output control beyond that done in TableBuilder
Things done by others (not DSS) to be adapted			ABS documentation/user guides referenced				Current ad hoc data requests to be formalised and developed as SUFs	Metadata may need to be created in a specific form for the ABS Mechanisms for delivering the data and metadata to ABS need to be specified		
Things to be developed anew			Guidance on understanding limits in TableBuilder				Delivery of data to ABS Metadata			

10.3 Scientific Use Files

DSS already produces SUFs but these seem to be generally on an ad hoc basis. Producing a finite set would allow DSS to exploit economies of scale, and they could also be used to supply TableBuilder. We identified the development of SUFs as a lower priority as DSS already has a process that works. However, producing SUFs would enable the development of TableBuilder as an additional tool.

	Project		People		Settings		Data		Output	
Progression	**Action needed**	**Comments**	**Action needed**	**Comments**	**Action needed**	**Comments**	**Action needed**	**Comments**	**Action needed**	**Comments**
Things in place at DSS not needing significant change	Authority for approval and ethical criteria	There is a mechanism for project approval in place at DSS but it is not transparent, visible or discoverable					Ad hoc dataset creation replaced with fixed set of SUFs	The aim is to develop a fixed set of files rather than ad hoc extractions for efficiency's sake; this contradicts a narrow interpretation the 'need to know principle', and so this needs to be addressed		Whilst the risk is low, good practice recommends some basic training in output SDC
	Institutional agreements	The ad hoc nature of current arrangements may not work on a larger scale; decision-making by precedent would allow the process to be streamlined					Metadata associated with the fixed files			
	Personal agreements						Create justification for off-the-shelf files			
	Metadata									
							Review anonymisation procedures in line with EDRU model	If current SDC processes are not based on EDRU principles		

	Project	People	Settings	Data	Output		Project	People	Settings	Data
Progression	**Action needed**	**Comments**	**Action needed**	**Comments**	**Action needed**	**Comments**	**Action needed**	**Comments**	**Action needed**	**Comments**
Things done by others (not DSS) to be adapted	Approval process => approval by precedent Decision-making process/criteria to be transparent	Agreements with institutions as well as individuals aligns SUF and vRDC practices Training for institutional agreement-holders raises awareness of data responsibilities and can be used to encourage engagement in the data community	Current guidelines to be reviewed in the EDRU framework Eurostat guidelines/test to be required reading for researchers in the short term	In the short term, the guidelines at Eurostat can be used as a short cut; in the medium term, DSS/Australian specific guidelines should be developed	Description of local environment to be reviewed for practicality of implementation and compatibility with behaviour	Guidelines on how to set up environments should be discoverable and transparent			Eurostat guidelines/test to be referenced	The Eurostat guidelines provide a temporary solution, and are compliant with the EDRU framework; in the longer term, specific guidelines using examples relevant to DSS may be better
					Transparent and discoverable guidelines on local environment to be drawn up	Guidelines should be reviewed in the light of the modern approach being taken by DSS				
Things to be developed anew	Transparent and discoverable guidelines on process and requirements for approval Training for institutional agreement-holders Mechanism for research findings to be disseminated meaningfully amongst data providers	Findings need to be published to demonstrate the value of research to DSS and the wider public	Transparent and discoverable guidelines specific to DSS context, building on Eurostat guidelines, in the medium term				Transparent (except with reference to specific parameters) and discoverable guidance on how the SUFs were created		Transparent and discoverable guidelines specific to DSS context, building on Eurostat guidelines	

10.4 Public use files

This project has been identified as the lowest priority. PUFs provide an additional way for users to get access to data. Synthetic data can be a very cost-effective way of getting minimal-risk datasets which behave somewhat similarly to the source data; it also means the data can have the same structure as the underlying data, so users of the vRDC, for example, can develop their code off-site on the synthetic data. The alternative, anonymisation of source data is more likely to be complex and expensive if any consideration is taken of the user. However, the creation of synthetic PUFs is a relatively new field and there is not much evidence yet on how the user community reacts to them. Therefore this element is a long-term option for DSS.

	Project		People		Settings		Data		Output	
Progression	**Action needed**	**Comments**	**Action needed**	**Comments**	**Action needed**	**Comments**	**Action needed**	**Comments**	**Action needed**	**Comments**
Things in place at DSS not needing significant change		Not relevant		It should be emphasised to users that the data is all imputed, and hence should not be used as exact data		Not relevant		This would be a new departure for DSS, but seems to offer good cost/benefit ratios, compared to anonymisation to create 'traditional' PUFs		Not relevant
Things done by others (not DSS) to be adapted										
Things to be developed anew			Guidance on using/interpreting synthetic data				Synthetic dataset			

10.5 Institutional components

In this subsection we consider the activities which need to occur at the institutional level (or indeed across the whole of government). The rationale for specific items are not given as it should mostly be self-explanatory; however the overarching reasons are:

- A commitment by DSS to a “pro-access” attitude, which will influence: design, implementation and delivery of systems; the more the new system deviates from what is already in place, the more clear leadership helps
- Principles can provide common ground and a basis for discussion where ideas of implementation differ strongly and
- Engaging stakeholders (including the public) early and meaningfully improves the prospect of success.

This is a unique time in the development of Australia’s infrastructure, and co-operation and the identification of common problems may allow much more cost-effective and sustainable gains than separate solutions for each Department.

Attitudes		Stakeholder Management		Perceptions		Common Operational Issues	
Aim	**Action**	**Aim**	**Action**	**Aim**	**Action**	**Aim**	**Action**
Formal commitment of DSS to EDRU approach	Redraft and circulate policy statement from DSS management with additional attitudinal elements Develop brief guide for staff on intranet on what this means in practice	Integration with other Commonwealth developments	Negotiate to adopt EDRU perspective in ABS, OPMC, PC, AIC and other relevant bodies; potentially develop cross-government statement on perspectives Develop common cross-government documents (eg personal or individual agreements for data access), with references to, for example, specific statutes separated and located in another discoverable place Agree a single (cross-government) accreditation process for individuals and organisations Develop joint training (active and passive) and information materials	Develop positive public attitude for data sharing	Develop web pages and literature focusing on the societal benefits, with security as an operational issue In the short-term, steal the ESRC cartoon; in the longer term, develop some similar publicity for Australia	Develop a streamlined transparent application and data access process	Streamline applications so it is dependent on the amount of detailed data the institute/ researcher requires. Provide information on the criteria for a successful application on DSS website. Develop an overarching model for ethical/ operational approval for microdata access/ release. Utilise current training programmes (such as SURE and Eurostat) for researchers.
Avoiding defensive decision-making	Develop formal corporate advice that ○ objectives and attitudes must be agreed <u>in advance of</u> specific discussions about legal/technical feasibility ○ (potential) users must have a substantive input when setting objectives ○ the risks/costs of not going ahead with projects (including to the wider public, not just DSS) must be considered user representatives must acknowledge risks, and contribute substantively to the risk management plan	Obtain positive engagement from data providers	Hold early talks with data providers inside and outside DSS, deploying evidence to counter opposition Identify gains to data providers, including feedback from researchers Create occasions for data providers and researchers to meet and share findings (DSS attendance at academic conferences; webinars by DSS aimed at academics)	Address concerned groups head-on	Invite privacy groups to early meetings, focusing on the 'if you don't do it this way, then how?' argument Set up an advisory board including data providers, researchers, policymakers and privacy activists	Formalise cross institutional arrangements.	Introduce the role of Institutional Responsible Officers within the institution. These individuals will help set precedence and streamline communications between the institution's researchers and DSS. Develop training courses for IROs Scope the future demand for data supply, and also the feasibility of charging for data access.

11. Roadmap

Area of development	Short-term (1-3 months)	Medium term (3-6 months)	Longer-term (6-12 months)	Strategic Developments
Public Use Files	None	None	None	Creation of synthetic data files and associated metadata
TableBuilder	Terms of use agreed with ABS	Delivery of metadata and sample datasets to ABS	Delivery of SUF files and metadata for dissemination	All
Scientific Use Files	Application process visible using existing documents Users required to undergo online training and take test regarding Scientific Use Files (investigate the Eurostat online training as a potential option).	Fixed set of SUFs identified by user demand Approval mechanism redesigned to establish and use precedents	Fixed set of SUFs created on EDRU basis and delivered to TableBuilder SUF metadata created Online training appropriate to DSS created	Review of dataset use – are more or less justified?
Virtual RDC	Pilot users agreed Hosting with SURE agreed. Agreement to require ABS Safe User training and/or SURE training. Adoption of Principles-based output SDC and training of staff. Personal contracts written for pilot phase only. Detail in vRDC defined but recognised as pilot only Delivery of pilot data and metadata to SURE	Approval mechanism redesigned to establish and use precedents General application process agreed and made visible ABS training redesigned to incorporate DSS-specific references Redraft of *Guide for Output Checkers* in association with ABS, possibly specific to Australian use Review of feasibility of 'production line' for vRDC data and metadata	Review of hosting with SURE, including data manager, stakeholder and user views Review of training effectiveness, including DSS needs	Review of SURE system, training and metadata Review feasibility of Secure Thin Clients/Secure Pods for extraordinary datasets (that is, where an evidenced risk much higher than normal research use exists)
Institutional changes	Statement of DSS policy (including attitude) to data access including formal adherence to EDRU approaches Top-level data strategy using the Data Access Spectrum Advisory Board set up and members enrolled Review of potential areas for cross-government sharing of tools and documents Form designed for users to produce non-technical summaries of research and tested with past users	Default definitions of data suitable for each element, with exceptions needing to be proven Dedicated area of the website for microdata, including publications Users required to produce non-technical summaries of research 1st meeting of Advisory Board Practical cross-government workshop on designing commonality into Commonwealth data access Pilot training for Institutional Responsible Officers	Publicity material targeted at general public about value of research data access Single register of DSS users Joint accreditation of DSS/ABS users Use of common documents for all access and across government if possible • Institutional access agreements • Personal agreements • Default local environment specifications for SUFs • SDC guides Statements of data access intention and strategy	Annual report by DSS and commentary by Advisory Board

Appendix: virtual RDC survey results

As part of this project, the team asked respondents at 12 RDCs around the world about the management of their RDCs. The results are discussed in section 4.2.4 of the main report. The detailed results are presented below.

Question	Response
Who is the data access agreement with?	27% The researcher 73% Both the institution and researcher
Who may sanctions be applied to in the case of a breach?	9% Just the researcher 9% Just the organisation 82% Both the researcher and the organisation
Is there more or less detail available depending on the type of users?	55% No difference in detail based on user 45% Different level of detail available based on both organisation and user
Are there different classes of users? (more/less trusted?)	55% No different classes of users 45% Different classes of users based on organisation and individual
In theory, from where can the data be accessed? (tick all that apply)	18% No response to the question 18% The researcher's organisation 54% Onsite facilities 45% Anywhere 18% Approved facilities
In practice, where is access normally granted? (tick all that apply)	54% Own site 36% Approved facility 27% Researcher's organisation 36% Anywhere
What is the most detailed level of geography made available to researchers for business data?	36% Exact 18% Local 36% None 10% No response to question
What is the most detailed level of geography made available to researchers for personal data?	72% Postcode 18% Local 10% No response to question
Are unique identifiers available if the project requires them? (tax numbers, health service numbers)	45% No 10% Yes with restriction 45% Yes with no restriction
Do you check all outputs before it is released from the centre?	82% Yes 18% No
Type of output checking, if used	45% Rules based 45% Principles based 10% None
Are users allowed to archive their workspace once the project has finished?	82% Yes 18% No
Can users retain the code they develop after the project finishes?	100% No
If the data is created specifically for the project, is it retained after the project is completed?	82% Yes 18% No
Is training provided on security awareness?	90% Yes 10% Optional
How is the security training delivered?	56% Face-to-face 35% Online course 9% Online guide
Is training provided on using the system?	64% Yes 36% No

Question	Response
Is training provided on statistics eg aspects of data linkage?	10% Yes 45% Optional 45% No
Is refresher training required?	10% Yes 27% Not recent 63% No
Are researchers trained in checking output for disclosure risk?	73% Yes 27% No
Is explaining your role to the wider public part of your RDC's or your organisation's usual objectives	45% Yes 37% Not directly 18% No
How do you engage with the public	18% Active and passive engagement 26% Active engagement 56% No response to question