Comparison of Two Remote Access Systems Recently Developed and Implemented in Australia

Christine M. O'Keefe[1], Phillip Gould[2], and Tim Churches[3]

[1] CSIRO Computational Informatics
GPO Box 664, Canberra ACT 2601, Australia
Christine.OKeefe@csiro.au
[2] Australian Bureau of Statistics
Locked Bag 10, Belconnen ACT 2616, Australia
phillip.gould@abs.gov.au
[3] Sax Institute
PO Box K617, Haymarket NSW 1240, Australia
Tim.Churches@saxinstitute.org.au

Abstract. National Statistical Agencies and other data custodians are vital sources of data for research and policy analysis. However, external researchers must be provided with access to data in such a way that privacy and confidentiality are protected. We discuss two recently-implemented research data access systems. The first was developed by the Australian Bureau of Statistics for use with certain of its data collections. The second was developed by the Sax Institute, a non-profit health research non-government organisation, for use by population health and health services researchers to analyse complex, linked administrative health and related data sets provided by a range of data custodians. Although these organisations both chose remote access systems, it is interesting that there are significant differences between the two systems. We discuss the drivers for and consequences of the different choices made.

Keywords: Privacy, Confidentiality, Remote Access, Remote Analysis.

1 Introduction

The use of population-level data in research has come to underpin the generation of information for government policy and operations, health services and population health research, as well as advances in many other areas. National statistical agencies and other data custodians make data available to both internal and external researchers under strong confidentiality protections. External researchers are typically located in universities or government agencies, and undertake data analyses ranging from simple descriptive tabulations to the fitting of complex statistical models. In this paper we will discuss approaches taken by two organisations, the Australian Bureau of Statistics (ABS) and the Sax Institute, for making data available for research while protecting confidentiality.

J. Domingo-Ferrer (Ed.): PSD 2014, LNCS 8744, pp. 299–311, 2014.

ABS is Australia's national statistical agency and a major provider of population-level data for research. The Census and Statistics Act 1905, states:

1. The Statistician shall compile and analyse the statistical information collected under this Act and shall publish and disseminate the results of any such compilation and analysis, or abstracts of those results.
2. The results or abstracts referred to in subsection (1) shall not be published or disseminated in a manner that is likely to enable the identification of a particular person or organisation.

The Sax Institute is a non-governmental research institute, providing data access infrastructure to health services and population health researchers. It is a partner in the Population Health Research Network (PHRN), a consortium of research service providers co-funded since 2008 by Australian national, state and territory governments. The PHRN has enabled establishment of record linkage services for population-based administrative health and health-related data across Australia. These linkage services, which use internationally accepted privacy preserving data management and linkage protocols, enable the provision of linked de-identified data for approved research projects. The services which comprise the PHRN operate under Australian national and jurisdictional privacy legislation and regulation, with example provisions including:

– Health information reasonably expected to identify individuals should not be included in a generally available publication.
– The confidentiality of participants and their data should be protected in the dissemination of research results.

Most relevant legislative statements about confidentiality focus on preventing *identity disclosure*, that is, the identification of an individual or organisation represented in the data. Only some include the additional objective of preventing *attribute disclosure*, that is, the disclosure of attributes of an individual or organisation, though this is not always made explicit. In personal data, attribute disclosure is usually only of concern if identity disclosure is a possibility.

1.1 The Changing Research Data Environment in Australia

With regard to the external researcher environment, data custodians are experiencing changing user expectations and differing levels of user sophistication and analytical requirements [15]. In particular, users are increasingly expecting access to richer microdata from an expanded range of collections in a flexible range of access modes or mechanisms. The types of richer microdata include: more detailed, hierarchical, linked, administrative, longitudinal, and business, as well as combinations of some or all of these. Users are also becoming more sophisticated in their adoption of the latest technologies, including online access and sophisticated data analysis and data-mining tools. In addition, researchers are increasingly forming large, multidisciplinary teams and using collaboration platforms and tools for sharing data and results in conducting their research.

These trends are expected to continue, for example, the recent Australian National Commission of Audit [1] recommended that *...the Government, recognising the need to safeguard privacy concerns, rapidly improve the use of data in policy development, service delivery and fraud reduction by: ... extending and accelerating the publication of anonymised administrative data ...*

At the same time, according to a recent survey of the Australian community's attitudes to privacy [9], the Australian environment has become one of enhanced community understanding of privacy, concern for privacy, knowledge of privacy rights, and willingness to take responsibility and change behaviour because of concerns about the handling of personal information. For government agencies, nearly all Australians (96% of respondents) believe that they should be told how their personal information is stored and protected.

Recently there have been a number of high-profile data breaches in Australia, and the Australian Office of the Information Commissioner handled 61 data breach notifications in 2012-13, a 33% increase since 2011-12 [8]. Although there is little or no evidence of privacy complaints or breaches in research on Australian data [10], the growth in number of data archives, custodian organisations, and researchers, together with the changing external researcher environment, may lead to a growing risk of data breach unless appropriately strong protections are put in place.

1.2 The Evolution of Data Access Mechanisms in the ABS

The ABS has traditionally made ABS census and survey data available via Confidentialised Unit Record Files (CURFs), as follows. CURFs are produced from the original unit-level data by the application of a (manual) confidentialisation process involving removal of name and address information, controlling the amount of detail and changing a small number of values through the application of statistical disclosure control techniques. CURFs are produced in increasing levels of detail, from Basic, through Expanded, to Specialist. Access to a CURF is granted to an organisation at the discretion of the Australian Statistician - then a researcher affiliated with the organisation can apply for registration and access. ABS can also grant access to CURFs to individuals. A precondition of organisational or individual access is the establishment of a legally binding Undertaking setting out the Terms and Conditions under which the access is approved. Basic CURFs are available on CD-ROM for the researcher to analyse on their own computer. Alternatively, a researcher can attend a Data Laboratory on-site at an ABS office in the nearest Australian capital city, in order to access Basic, Expanded or Specialist CURFs. In this case any statistical output derived from Expanded or Specialist CURFs is manually cleared before the researcher can remove it from the on-site Laboratory.

Around ten years ago or so the ABS implemented the Remote Access Data Laboratory (RADL) for access to Basic and Expanded CURFs. The RADL is a secure online data query service that clients can access via the ABS website. Users submit queries written in the SAS, Stata or SPSS statistical programming languages through a web interface, although some commands, functions and pro-

cedures are disabled to protect confidentiality, and there are restrictions on the size and nature of allowable outputs. The queries are run against the requested CURF that is kept within the Australian Bureau of Statistics environment. The results of the queries are checked for confidentiality by ABS staff and then made available for download to the users via their web browser.

More recently, the ABS has developed the *TableBuilder* and *DataAnalyser* systems to allow registered users to build their own custom tables and undertake regression analyses on secured ABS microdata, respectively [15].

TableBuilder is an online tool with a menu-driven interface allowing registered users to create confidentialised user-specified tables of count or continuous variables. Requested tables are produced and confidentialised on-the-fly as standalone outputs or as inputs to more sophisticated analyses such as regressions. Under the confidentialisation process, all cell values, subtotals and totals are randomly slightly adjusted to prevent any identifiable data being exposed. The adjustments are done in such a way that consistency of cell values across different tables constructed from the same data set is maintained. TableBuilder has been operating successfully for several years on Australian Census data and is being expanded to include survey data.

DataAnalyser is an online system that allows users to undertake analyses of detailed ABS microdata in real time. It allows users to remotely conduct certain data transformations and manipulations, basic exploratory data analysis, create summary tables and run regression analyses including linear (robust), logistic, probit and multinomial. For the first version of DataAnalyser, a low level of manual confidentialisation is applied to the microdata before loading into the system. The microdata are kept within the ABS secure environment behind a series of firewalls, requests are submitted through a menu-driven interface, and confidentialised outputs can be either viewed on screen or downloaded to the user's own computer. The confidentialisation processes are:

- a menu-driven interface is used to restrict the allowed variables, as well as the range and nature of data manipulations and analyses available
- counts are perturbed
- for regression, a small number of randomly-selected records is removed
- for regression, a model is rejected if it pertains to fewer than a minimum number of records, it has greater than a maximum number of parameters, there are fewer than a minimum number of records for each parameter, any record has a leverage above a given threshold, the sum of the leverages of two records exceed the threshold, or if the summary table constructed with the response variable against any of the categorical explanatory variables contains a zero
- for regression, the score function is perturbed prior to the estimation of the regression parameters
- scatter plots are replaced with hex plots on data where each hexagon with fewer than a minimum number of observations is suppressed

DataAnalyser is planned to be released as a beta product at the end of June 2014. Initially, invited users will be able to access the Australian Census Longi-

tudinal Dataset and Australian Census-Migrants Integrated Dataset. The ABS plans to add additional survey datasets in the future and may consider requests for access to the beta trial from interested users.

The confidentialisation routines applied in TableBuilder and DataAnalyser are applied not at the unit record level, as is the case with CURFs, but at a level of aggregation relevant to the analysis. The level of confidentialisation required is therefore lower, leading to substantially reduced total variances [2].

1.3 The Evolution of Data Access Mechanisms in the PHRN

The PHRN is facilitating the creation of a nationwide data linkage infrastructure in Australia, with nodes servicing States and territories, as well as national linkage capabilities. It includes amongst its nodes the successful Western Australian Data Linkage Branch (established in the mid-1990's) and (NSW and ACT) Centre for Health Record Linkage, established in 2006. The PHRN linkage nodes interface with numerous routinely-collected Australian and State or Territory population-based databases, including Registrations of Births, Deaths and Marriages, Cancer Registries, and Emergency Department and Hospital Admitted Patient Data Collections. Specially-collected data from research studies such as from the 45 and Up study [14] can also be incorporated into the PHRN data linkage infrastructure and operations. All PHRN nodes enable linked, de-identified data to be provided to researchers, using a privacy-enhancing separation protocol involving linkage keys [5]. Under the protocol, the PHRN data linkage units receive *only* demographic information (name, address, sex and date-of-birth) and researchers receive *only* the health or other content data items. Researchers are able to assemble all records for each individual using project-specific de-identified linkage keys provided by the data linkage unit.

Thus, commencing with the Western Australian Data Linkage Branch operations in the mid-1990's, approved researchers in Australia have been provided with de-identified data files for approved population-based studies, after an appropriate user agreement has been signed and compulsory training has been completed. The provisioning of these data has recently been improved using on-line encrypted data transfer technologies.

In the last couple of years, one of the PHRN nodes, the Sax Institute, has developed the Secure Unified Research Environment (SURE) [13] as an alternative to providing linked, de-identified data files directly to researchers. SURE is a remote-access computing environment that allows researchers to access and analyse linked health-related data files for approved studies in Australia. The remote environment is accessible over encrypted internet connections, and effectively replaces a user's local computing environment. For each research study hosted by SURE, a project workspace is established to host virtual computing desktops for the researcher or team of researchers conducting the study. The research datasets are stored on virtual servers also located within the confines of each project or study workspace - thus, an entire virtual network is provided for each study, remotely accessed by researchers who use a facsimile of the screen of their remote virtual computing desktop on their local computer screen to

manipulate and analyse the data. A range of standard and optional software is available on each SURE virtual workstation, including statistical packages such as R, SAS, SPSS and Stata, together with add-ons and libraries for each. Users can request other, more specialised software to be installed, if required, subject to cost and licensing conditions.

Although researchers using SURE can directly view microdata, and conduct unrestricted data manipulations and statistical analyses within the SURE remote-access environment, the only way that a file, such as a supplementary data file or a file of analysis outputs, can enter or leave SURE is via a single audited portal called the *Curated Gateway.* It is possible that there are issues of confidentiality associated with analysis outputs which researchers may wish to remove from SURE, for example for publication in the academic literature. Because such outputs cannot be assumed to be free from disclosure risk, outbound files uploaded to the Curated Gateway for use outside of SURE need to be assessed for confidentiality risk and treated with confidentialisation measures if necessary. This is currently the responsibility of the study's chief investigator, though it could also be done by an independent senior investigator or custodian representative as appropriate. Note that the compulsory training provided to all SURE users includes training in privacy and confidentiality regulatory regimes, and in the principles of statistical disclosure control for protecting confidentiality.

Traditionally, it has been the responsibility of the individual researcher and the Curated Gateway reviewer to ensure that analysis outputs removed from the SURE environment do not represent a disclosure risk. A recent project of CSIRO and the Sax Institute has reviewed confidentiality issues associated with public health and health policy research analysis outputs generated in a secure analysis laboratory such as SURE [11]. The outcome of the project has been endorsement of the current two-stage confidentiality protection process for SURE, comprising the existing data preparation and output confidentialisation stages. In the data preparation stage, data custodians and/or SURE administrators apply some basic confidentialisation measures to the dataset before making it available to researchers within each study or project workspace, but this confidentialisation is as lightweight as possible, and typically involves removal of all direct identifiers such as names, street addresses and medical record numbers, as well as removal of data items which substantially increase the risk of re-identification, such as exact date-of-birth, or high resolution spatial attributes of place of residence. These measures are designed to reduce, but not entirely eliminate, the risk of both spontaneous recognition by researchers and disclosure in analysis outputs. The residual risks are managed in the output confidentialisation stage, where the Curated Gateway reviewer ensures that published outputs generated in SURE comply with confidentiality protection requirements. In the CSIRO-Sax Institute project, a checklist was developed to assist reviewer and researchers using SURE to assess confidentiality risks in their analysis outputs, and apply confidentialisation treatments to reduce the risks to acceptable levels. In the future, this step should be able to be at least partially automated, or

tools could be provided to enable researchers and reviewers to efficiently carry out such steps as part of a routine workflow.

2 Comparison of ABS DataAnalyser and PHRN SURE

The ABS and the Sax Institute/PHRN are organisations seeking to facilitate the use of routinely-collected data by researchers external to the organisations which collected the data. Both are currently responding to the changing research data environment with the implementation of new data access mechanisms designed to augment their traditional data dissemination channels.

Interestingly, both the ABS and the PHRN, through the Sax Institute, have very recently chosen to develop and implement remote access systems, with several features in common:

- detailed de-identified datasets are held in a secure environment,
- users require registration and/or approval and sign user agreements,
- users access the datasets via a secure channel on the internet, and
- users submit analysis requests and receive analysis outputs

However, the details of the two systems are quite different. Perhaps the major difference is the degree of user access to the dataset. In DataAnalyser, the user has no direct access to the data, in fact, the user cannot even view individual dataset records. This type of remote access system is sometimes called a *remote analysis* system. In contrast, in SURE, the user has unrestricted access to the data and can view every dataset record. This type of remote access system is sometimes called a *virtual data laboratory* or *data enclave*.

Internationally, examples of remote analysis systems include Table Servers developed by the National Institute of Statistical Science (NISS) to disseminate marginal sub-tables of a large contingency table [3,4], and the Microdata Analysis System under development by the U.S. Census Bureau to allow users to receive certain statistical analyses of Census Bureau data, including regression analyses, without ever having access to the data themselves [6]. Examples of virtual data laboratories include the UK Secure Data Service, providing secure remote access to data operated by the Economic and Social Data Service [16] and the US NORC Data Enclave, providing a confidential, protected environment within which authorised social science researchers can access sensitive microdata remotely [17].

In this section we compare the ABS and Sax Institute/PHRN systems and examine the drivers for and consequences of the different choices. In this comparison, we have assumed correct implementation and operation of the the information security functions necessary for the trust characteristics of each solution, including appropriate architecture, firewalls, authentication, monitoring and audit. In practice, this assumption must be carefully verified through independent design reviews and implementation audits.

2.1 Drivers for ABS DataAnalyser and PHRN SURE

In this section we focus on the drivers in the research data environments of the ABS and Sax Institute/PHRN, see Figure 1.

	ABS	PHRN
Mission includes	enable broad use of ABS data and data products	make health and related data available for research
Legislative Requirements	identification should not be likely	identity should not be reasonable to ascertain
Range of data	broad range of census and survey data types	health and social administrative data
Range of users	broad range of users with diverse requirements and statistical sophistication	academic population health and health services research community
Research governance and ethical oversight	data access for statistical purposes	project approved by data providers and Human Research Ethics Committee(s)

Fig. 1. Drivers in the research data environments of ABS and PHRN

The ABS is seeking to deliver on its mission and strategic objective of supporting the informed and increased use of statistics [15]. In response to this driver, the ABS is seeking new data dissemination technologies that minimise actual and perceived barriers to accessing ABS holdings. New data dissemination technologies must therefore deliver infrastructure for real time dissemination of ABS data, increase the detail and the range of collections available, reduce the resources required, and improve timeliness. A broad range of users with a range of levels of sophistication and analytical requirements is contemplated, including: government agency and large corporation employees, individual university researchers, and consultants. The range of data to be made available includes: census, social and business surveys, economic, demographic and land-use data. Since the obligations of the Census and Statistics Act 1905 must be upheld regardless of the type of user, the type of data, or the kind of analysis being undertaken, the ABS needs to implement a one-size-fits-all approach to provide confidentiality protection across a multitude of users and purposes.

The Sax Institute and the PHRN are seeking to deliver on their mission of supporting public health and health services research of national relevance in Australia [12]. In response to this driver, the PHRN is seeking new data dissemination technologies that enable researchers to more efficiently conduct the sort of studies that have been traditional in public health and health services research, although with richer and greatly expanded data collections. Researchers using the PHRN are generally from universities and government health agencies, and the PHRN seeks to grow its user base in these communities. The datasets made available through SURE are predominantly administrative health and social datasets, though research study data can also be included. The PHRN

currently enables the provision of linkable, de-identified datasets directly to researchers for use in their own computing environment. The SURE system is designed to be functionally not more restrictive than the current arrangements.

2.2 Summary of Features of ABS DataAnalyser and PHRN SURE

In Figure 2 we summarise the main features of the technological systems implemented by ABS and PHRN, focussing on confidentiality protection.

	ABS DataAnalyser	PHRN SURE
Dataset Preparation	light manual confidentialisation	light manual confidentialisation
User can browse metadata	yes	yes
User can request any data set	within the scope of data sets provided in DataAnalyser	user can only access project datasets with provider and ethics committee approval
User direct access to data including viewing de-identified records	no access	full access
Data manipulations	restricted	unrestricted
Range of queries	restricted	unrestricted
Queries	modified/restricted	unmodified, unrestricted
Software available	only DataAnalyser software	broad range of standard software and some custom
Range of outputs	restricted	unrestricted
Output	confidentialised	reviewed at Curated Gateway

Fig. 2. Features of ABS and PHRN remote access systems

First, and as mentioned in Section 2 above, DataAnalyser prevents the user from viewing any data records, while SURE gives the researcher full access including viewing all (de-identified) data records. In order to provide adequate confidentiality protection in each case, the different levels of direct access to data are balanced by different levels of other measures. In DataAnalyser, researchers can browse and request analysis of any of the data sets which ABS has approved for access via DataAnalyser, while in SURE, researchers must have their project approved by the relevant data providers and by a Human Research Ethics Committee, and can only access the data set and data items approved for that project.

The second major difference is that DataAnalyser applies strong restrictions on the range of data manipulations, range of queries, and range of outputs available to the researcher. DataAnalyser applies modifications to some analyses, for example, it perturbs the score function for a regression, and applies further automatic confidentialisation routines to outputs before returning them to the researcher. In contrast, the researcher using SURE is unrestricted in the data manipulations and analyses they can apply, and there are no restrictions on the

types of output they can obtain. Outputs are not modified by SURE, however are subject to review for confidentiality protection at the Curated Gateway. SURE relies on the researchers and/or the Curated Gateway reviewers to confidentialise analysis outputs before publication.

2.3 Comparison of ABS DataAnalyser and PHRN SURE

Types of Users and Data. First, the ABS cannot assume a uniform or even a minimum level of sophistication of its users. Therefore, DataAnalyser is initially targetted to a core group of users with a medium level of sophistication, including: policy analysts and social and economic researchers. The menu-driven system is well suited to these users and makes fully automated confidentiality protection achievable for realistic cost. Future versions of DataAnalyser may have extended capability in order to address the needs of more sophisticated users. The ABS also cannot assume uniformity across its datasets, which are extremely diverse.

The main drawback of the DataAnalyser is that there is significantly reduced flexibility offered to users, for example, DataAnalyser offers users only prescribed data manipulations, methods and outputs. The ABS may never be able to anticipate and provide functionality for the full range of analyses that its very broad user base may wish to perform. If a researcher requires more flexibility or a different analysis, they must use a different ABS data dissemination channel.

SURE can assume a reasonable level of sophistication amongst its researchers, since each project hosted by SURE has been approved by an ethics committee convinced that the outcomes will be of sufficient merit to outweigh any confidentiality risk, and which has thus considered the qualifications and experience of the researchers involved in the project. In addition, normally researchers seeking to use SURE embark in what can be a lengthy negotiation phase to establish whether their proposed study is feasible using available linked data sets. SURE has been set up to enable collaborative team-based storage and workspaces for project teams. SURE is designed for administrative health and social data.

Both ABS and PHRN make use of a user registration process, normally also involving the user's employing organisation. SURE makes use of strong three-factor authentication of users at the web interface.

Scope of Trust. The difference in trust of researchers is also an important drivers for the choices. The level of trust extended depends on the dataset, the custodians, the researcher(s) and the research questions being asked.

DataAnalyser contemplates a broad range of external users of varying levels of sophistication. The appropriate choice has been made to extend a lower level of trust to the users and instead to rely on the automated confidentiality protections built into DataAnalyser technology itself for preventing disclosures.

In contrast, SURE extends a higher level of trust to approved researchers and their computing environments, providing access via a virtual data laboratory mechanism with much lighter automated confidentiality protections. The SURE approach of trusting researchers and/or reviewers to assess confidentiality risk

and confidentialise outputs is underpinned by a tighter and more formal research governance practice involving: custodian and Human Research Ethics Committee approvals, targetted training in confidentiality protection, strong user agreements, post-study reporting, and strong sanctions for breaches.

Consistency of Analysis Results. In the case of the ABS, a researcher could analyse the same data via several different data access channels, for example, CURFs, TableBuilder and DataAnalyser. In order to avoid inconsistencies in the application of confidentialisation processes across the range of ABS data dissemination modes, possibly leading to either inconsistent results or unexpected confidentiality risks, the ABS has developed general perturbation algorithms that can be incorporated into a broad range of analysis methods including summary tables, summary statistics and statistical regressions.

In contrast, the nature of the projects hosted by SURE means that it is unlikely that exactly the same data subset is used in more than a handful of studies, so the problem is not so pressing. In cases where the same dataset is used for a number of studies, often it is the same group of researchers and they can ensure consistency as they are applying the confidentialisation methods themselves. More broadly, SURE users are required to actively seek to publish or otherwise disseminate their results, increasing the likelihood that researchers are aware of research outputs published by other groups using the same datasets.

Summary. Marsh et al. [7] noted that a successful disclosure involves first an attempt at disclosure, then success of that attempt. In probabilistic terms, this is: $\Pr(\text{disclosure}) = \Pr(\text{attempt}) \cdot \Pr(\text{disclosure} \mid \text{attempt})$. The ABS environment requires it to assume that $\Pr(\text{attempt})$ is close to 1, and therefore to seek to minimise $\Pr(\text{disclosure} \mid \text{attempt})$. The PHRN works to ensure that $\Pr(\text{attempt})$ is negligible, and therefore does not need to minimise $\Pr(\text{disclosure} \mid \text{attempt})$.

3 Discussion and Conclusions

We have described the evolution of data access mechanisms in two important Australian organisations providing or enabling data access to researchers, namely, the Australian Bureau of Statistics (ABS) and the Sax Institute node of the Population Health Research Network (PHRN). In the last couple of years, both of these organisations have implemented a new remote access system, however it is interesting that they have chosen different types of remote access. We have analysed the reasons for these differences through a comparison of the context and environment for each system, and the technological responses to them.

In the current international environment of open government and data sharing, organisations are seeking to make more and more data available for research and policy analysis. Both the ABS and the Sax Institute/PHRN are responding to the evolving Australian community environment of increasing concern for privacy and knowledge of privacy rights, by increasing transparency about their data holdings and data access arrangements. Both organisations are responding to growing researcher interest in richer detail across an expanded range of collections by implementing increasingly automated data access technologies. Both

organisations make use of appropriately targetted researcher registration and agreements, and have sanctions in place for breaches of the agreements. A stand out observation is that both organisations have chosen types of remote access.

The ABS, in focussing on a broad range of users with varying levels of sophistication, has chosen remote analysis. Under the DataAnalyser approach, the lower trust level implied by providing access to a wide range of users requires a less flexible system and restricted outputs. The Sax Institute, in focussing on a community of more sophisticated users, has chosen a virtual data centre. Under the SURE approach, the higher trust level implied by strongly restricting access allows a more flexible system. A useful way to compare the systems is to note that a disclosure requires first a disclosure attempt, then success of that attempt. The ABS focusses on reducing the likelihood of success of any disclosure attempt, while PHRN focusses on reducing the likelihood of an attempt.

We remark that the U.S. Census Bureau has adopted an automated output confidentialisation approach for its Microdata Analysis System, similar to the ABS DataAnalyser, noting that both are examples of remote analysis systems. In the NORC Enclave, a virtual data centre, any export request from a researcher is scrutinised by a NORC statistician to ensure that it does not contain disclosive data. This is similar but more restrictive than the SURE approach.

Our two detailed examples show that there is no single solution for protecting confidentiality while making data available for research, since differences in context and focus will lead to different requirements and different approaches making use of different combinations of protections. Both of the systems we have discussed have advantages and disadvantages in terms of scope of access and flexibility. In each of our examples there is a combination of individual protections, none of which is sufficient alone but the aggregation of all of which provide strong confidentiality protection for data during research.

One challenge associated with remote access is the need to go through the sometimes lengthly funding application, registration and approval processes before any analysis of the data can be conducted. In some cases, this can be a real problem if it is subsequently found that the data are not suitable for addressing the proposed research question. Both the ABS and Sax Institute/PHRN are seeking to address this question by seeking to make available low risk datasets for initial data exploration and methods development under a lightweight approvals process. The ABS is investigating the use of model-based synthetic datasets, and the Sax Institute is investigating the use of massively perturbed datasets, such as are generated by data swapping with extremely high swapping probabilities.

We conclude with the observation that: *... recent events in the development of remote analysis servers herald the dawn of a new era in automated confidentiality protection for analysis and we look forward to invigorated research collaborations among NST's and academic institutions to further this research ...* [15].

Acknowledgements. This paper draws on work done while Christine O'Keefe was on secondment to the Australian Bureau of Statistics, and during a collaborative project between CSIRO and the Sax Institute. We thank the Australian

Government Education Investment Fund Super Science Initiative for part funding of the latter project through the Population Health Research Network.

Disclaimer. Views expressed in this paper are those of the authors and do not necessarily represent those of the Australian Bureau of Statistics. Where quoted or used, they should be attributed clearly to the authors.

References

1. Australian Government: National commission of audit. Report, Phase I, http://www.ncoa.gov.au
2. Chipperfield, J., Lucie, S.: Analysis of micro-data: Controlling the risk of disclosure. Research Paper - Methodology Advisory Committee 1352.0.55.110, Australian Bureau of Statistics (2010)
3. Karr, A., Lee, J., Sanil, A., Hernandez, J., Karimi, S., Litwin, K.: Web-based systems that disseminate information but protect confidentiality. In: McIver, W., Elmagarmid, A. (eds.) Advances in Digital Government: Technology, Human Factors and Public Policy, pp. 181–196. Kluwer, Amsterdam (2002)
4. Karr, A.F., Dobra, A., Sanil, A.P.: Table servers protect confidentiality in tabular data releases. Commun. ACM 46(1), 57–58 (2003), http://doi.acm.org/10.1145/602421.602451
5. Kelman, C.W., Bass, A.J., Holman, C.: Research use of linked health data best practice protocol. Australian and New Zealand Journal of Public Health 26(3), 251–255 (2002)
6. Lucero, J., Zayatz, L., Singh, L., You, J., DePersio, M., Freiman, M.: The Current Stage of the Microdata Analysis System at the U.S. Census Bureau. In: Proc. 58th Congress of the International Statistical Institute, ISI 2011 (2011)
7. Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D., Walford, N.: The case for samples of anonymized records from the 1991 census. J. Roy. Stat. Soc. Ser. A 154, 305–340 (1991)
8. Office of the Australian Information Commissioner: Annual report 2012-13
9. Office of the Australian Information Commissioner: Community attitudes to privacy survey (2013)
10. O'Keefe, C., Connolly, C.: Privacy and the use of health data for research. Med. J. Aust. 193, 537–541 (2010)
11. O'Keefe, C., Westcott, M., Ickowicz, A., O'Sullivan, M., Churches, T.: Protecting confidentiality in statistical analysis outputs from a virtual data centre. Working Paper, Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Ottawa, Canada, October 29-30, 10 p. (2013), http://www.unece.org/stats/documents/2013.10.confidentiality.html
12. Population Health Research Network: website
13. Sax Institute: Secure Unified Research Environment (SURE). Website, www.sure.org.au
14. Sax Institute: 45 and up (website), https://www.saxinstitute.org.au/our-work/45-up-study/
15. Thompson, G., Broadfoot, S., Elazar, D.: Methodology for automatic confidentialisation of statistical outputs from remote servers at the Australian bureau of statistics. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Ottawa, Canada, October 28-30, 37 p. (2013)
16. UK Data Archive: Secure data service (website), securedata.data-archive.ac.uk
17. University of Chicago: NORC (website), www.norc.org