Secure Data Service:

An improved access to disclosive data

by Reza Afkhami, Melanie Wright, Mus Ahmet[1]

UKDA

Abstract

The Secure Data Service is a secure environment funded by the ESRC to provide researcher access to disclosive micro data either from their offices, safe rooms in their institutions or on site at the UK Data Archive. Operation is legally framed by the 2007 statistics Act which makes possible access to the confidential data for statistical purposes. This short paper introduces this new UK Data Archive service and proposed specifications, as well as challenges facing data service providers. We envision that the proposed SDS infrastructure will meet the requirements of the data security model. The paper also aims to be an exemplar for a secure remote access practice.

Keywords: Remote access, Data security, Citrix technology, Secure Data Service

1. Introduction

Disclosure of personal information can be harmful and may result in denial of services, embarrassment and loss of reputation and trust which in turn reduces the response rate and jeopardises the future research. Research results based on disclosive data can also cause indirect harm by affecting perceptions about a group to which a person belongs.

Balancing data confidentiality and legitimate requirements of data users is a key problem of the Secure Data Service (SDS). Confidentiality of individual information can be protected by restricting the amount of information provided by adjusting the released Microdata /tables/ statistical outputs (restricted data), by imposing conditions on access to the data products (restricting access), or by some combination of these.

The UK Data Archive Secure Data Service is a new service to allow controlled restricted access procedures for making more detailed Micro data files available to some users (Approved/Accredited Researchers), subject to conditions of eligibility, purpose of use, security procedures, and other factors associated with access to the SDS data.

Building on the success of other secure data enclaves worldwide[2], and employing security technologies used by the military and banking sectors, the SDS will allow trained researchers to remotely access data held securely on central SDS servers at the UK Data Archive. The aim of the service is to provide approved academics unprecedented access to valuable data for research from their home institutions, with all of the necessary safeguards to ensure that data are held, accessed and handled securely.

The SDS follows a model which recommends that safe use of data should include safe project, safe people, safe setting and safe output (Ritchie, 2006. see Figure 1). To achieve this goal, data security depends on a matrix of technical, legal, contractual, and educational factors. The structure of this paper revolves around these factors, demonstrating how SDS has set up the necessary infrastructure to meet these requirements.

We discuss the legal and contractual responsibilities of the users and their institutions followed by issues such as user education and training prior to data access. The technical features and the system specifications are also examined. We also discuss the kind of disclosive data we aim to support in the SDS. Finally, the challenges facing the SDS operation will be examined.

2. Legal and contractual framework

Users of the SDS will be required to be either "ONS Approved Researchers"[3] or "ESRC Accredited Researchers." The first of these is defined by the Statistics and Registration Services Act 2007 as "an individual to whom the Board has granted access, for the purposes of statistical research, to personal information held by it."[4] No definition currently exists of an "ESRC Accredited Researcher," but we assume that it will have a similar status to an ONS Approved Researcher. This is a person who has been granted access for the purposes of statistical research to personal information which has been licensed to the ESDS/UK Data Archive/[5]University of Essex for dissemination on behalf of a government department or some other data provider. Neither of these two types of user will be able to use the SDS without appropriate training. Mandatory training will allow the UK Data Archive to ensure that end-users are fully aware of any penalties which they might incur if they cause a breach. We believe that if there is user approval to any penalties for breaches, and that they believe that these penalties are reasonable and necessary, we will avoid the inadvertent disclosure to which social science researchers are most likely to be prone.

The 2007 Act also allows for increased sharing of data between ONS and other Departments, subject to agreement by Parliament on a case-by-case basis. At the same

time the Act also outlines measures designed to protect the confidentiality of personal information. The Act states that a person who discloses personal information "is guilty of an offence and liable — (a) on conviction on indictment, to imprisonment for a term not exceeding two years, or to a fine, or both; (b) on summary conviction, to imprisonment for a term not exceeding twelve months, or to a fine not exceeding the statutory maximum, or both."

The SDS will immediately suspend access to the service if it believes that any user is perpetrating or attempting to perpetrate any of the breaches listed in SDS security breaches or SDS confidentiality agreement. A full investigation will follow.

Users will be required to complete an on-line form which collects personal and institutional details, information about their proposed data usage, and information which demonstrates their expertise and ability to conduct the research described in a competent and secure manner. They will also be required, if they have not already done so, to agree and sign the standard UK Data Archive End User License (EUL), and also agree/sign any Special License conditions which apply to the resource they wish to access. This application would be first checked for accuracy, sense and completeness by UK Data Archive staff, and then forwarded to the data owners for their access authorisation. Once authorised, users would be informed and requested to sign up for appropriate training (if they have not already been trained). Upon completion of training, users would be granted permission to access the secure data server, either from their own desktop if the owners of the data they wish to access permit, or from their institution's secure data access room. If their institution does not have a secure data access room, the user will have the option of negotiating access from another nearby institution's safe room (the SDS will offer 'matchmaking' introductions, but the specific arrangements must be under the control of the institution hosting the room, as audit trail security will be their responsibility) or coming to the University of Essex, to access the service onsite.

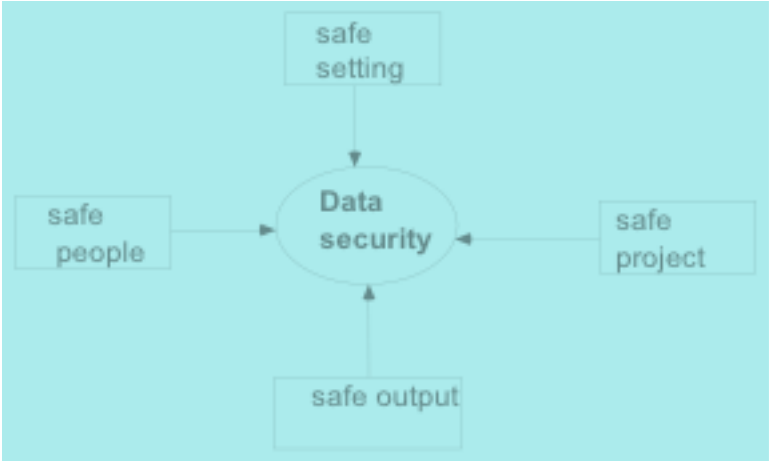

Figure 1: Elements of data security (Ritchie, 2006)

3. Education &Training

Researchers are the known weakest links in data security. Education coupled with the stricter legislative protections mentioned above, can offer another potentially efficient means of improving confidentiality, as disclosure probability can be decreased without imposing costs on rule-abiding researchers.

Before becoming an active user of the SDS, users will have to attend a mandatory training session which will focus first on the user's legal and ethical responsibilities within their SDS user license agreement, the mechanics of how to use the SDS, what they can and cannot do in a remote access setting, and the potential of the collaboratory spaces. The second part will focus on principles of the statistical disclosure control, assessment of outputs, and analysis aspects of the particular datasets in the SDS.

Access to the SDS will only be granted after users have attended an SDS training session. SDS staff will vet data analysis outputs for disclosure issues, to ensure that nothing escapes the secure data setting, which could compromise the data security (safe output). One of the purposes of the training is to give researchers the ability to recognise confidential data and distinguish it from statistical results that are safe to remove from the SDS. In effect, the training removes the 'reasonable belief' defence for a disclosure. We believe that penalties will only be an effective deterrent if they are known, and it should also be clear that we are more concerned with prevention than punishment.

4. System Specifications

The technology used by the SDS must be secure and the system adheres to the highest standards of quality. The technical model that has emerged is one which shares many similarities with both the ONS VML and the NORC Secure Data Enclave . It is based around a Citrix infrastructure which turns the end user's computer into a 'remote terminal' giving access to data, statistical software, and collaboratory spaces on a central secure server held within the UK Data Archive. The system is flexible, in that depending upon the wishes of the data custodians, access can be restricted to particular users (safe people) and/or particular locations (safe rooms/machines). It is secure because all data manipulation occurs on the server, which is maintained to very strict security protocols.

Beyond the general security policy, the secure server itself will be subject to additional security measures and controls. Approved researchers will access the proposed SDS by using VPN (Virtual Private Network/thin-client) technology, which encrypts the data transmitted between the researcher's computer and the host network. Other components of the VPN technology allow control to be established over which network resources the external researcher can access on the host network. The service will employ a Citrix XenApp server farm, which participates on two networks Mulcahy. Et al, 2008).

How the system operates?

With this technology, although all applications (SPSS, STATA, etc) and data run on a central server at the UK Data Archive/SDS, the Approved Researcher still interacts with a full Windows graphical user interface. This means that the researcher never has to install any complex applications on his/her remote computer; the only application required by the Approved Researcher is a web browser. This also means that the UK Data Archive can prevent the researcher from transferring any data from the data archive to a local computer. For example, Citrix can be configured so that data files cannot be downloaded from the remote server to the user's local PC. Similarly, the Approved Researcher cannot use the "cut and paste" feature in Windows to move data from the Citrix session into an Excel spreadsheet sitting on the local computer. Finally, the user is prevented from printing data from a local computer. The Approved Researcher logs onto the SDS system remotely via a web secure (HTTPS) browser. All data processing is carried out on a central secure server, which processes all requests centrally and returns information about the results. No data travels over the network, except the statistical results sent from the central server to the remote location by an encrypted email after the final outputs are checked against statistical disclosure controls.

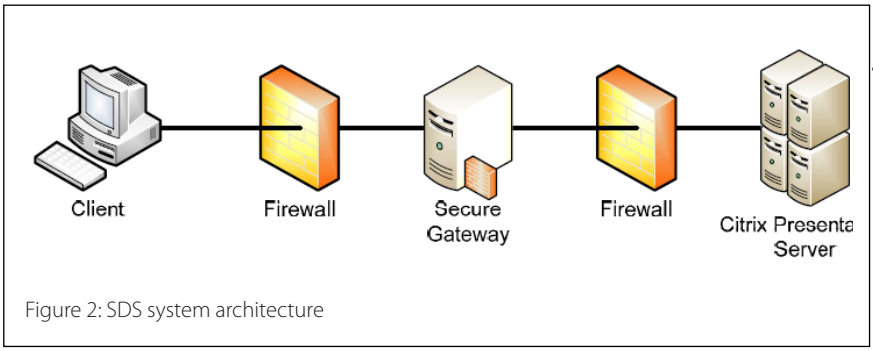

Figure 2: SDS system architecture

Key Features

- Clients cannot remove data
- Absolutely no webpage access
- Clients cannot import data
- Data transfers are logged
- All traffic is encrypted
- Smart Auditing
- Critical Security updates are applied daily

5. Benefits

This system may pose an inconvenience to the user compared to their accustomed ability to use EUL (End User License- similar to Public Use File) data on their desktop with all their favourite local software and networked resources. However, it is a price users will gladly pay for local access to data which they might otherwise have had to travel to ONS sites to access, or simply have been unable to access. The SDS will benefit users in:

- A self-contained secure 'home away from home' service with familiar analytical environments;
- Ability to work in their own private work areas or in shared areas with other approved researchers;
- Access to enhanced, highly sensitive available data storage in tandem with the related metadata through increased capacity and environmental protection;
- Possibility of data linkage exercise with using existing data in the UK Data Archive or other administrative data subject to approval of data owners/custodians; and
- Collaborative functionality including survey and document library, SPSS/ STATA code library, knowledge repository, disclosure review and technical assistance.

6. Data

A variety of data may potentially be available to users within the SDS. We are in ongoing discussions with the owners of key sensitive data resources about how SDS might assist in broadening the use and utility of these important resources, whilst assuring that legal, moral and security requirements are met. The specifications of data candidates include:

- More detailed variables from existing ESRC-funded data resources;
- More previously unavailable detailed variables from government social surveys;
- Other government data previously only available in onsite enclaves, or previously unavailable to academic researchers altogether;
- Business data which has commercial sensitivity;
- Administrative data; the SDS may be able to provide a secure environment for data linkage activities to researchers whose home institutions lack the technological wherewithal to offer it (restrictions must be negotiated and approved by the data owners/custodians upon user's application); and
- Data previously considered too sensitive or potentially disclosive due to its very nature, such as longitudinal data, medical data, etc.

In addition, the service will allow users to bring in less disclosive data from the UK Data Archive standard EUL holdings, upon request or researcher's own data subject to the standard ingest checks and approval.

7. Challenges

The two main goals of the Secure Data Service are: maximizing utility of microdata for research purposes, and protecting the confidentiality of individual respondents. Access to confidential data is an exception to the non-disclosure rule that must be justified according to the balance of the public good of the research against the risk of a breach of respondent privacy. Thus, the SDS hopes to maximize data utility while minimizing the disclosure risk, utilizing a strategy that is simple, well-communicated and acceptable to users.

However this task may be daunting as there is no consensus on the definition of what is safe data and second, even more contentious is what information loss means and how it can be measured. As any effort to implement confidentiality protection is associated with some loss of information.

Maintenance of confidentiality needs a consistent and coherent approach and we must trust the researchers as no environment is free of risk. For example, how can we prevent against a manual data copying or using photographs for researchers who remotely have access to the disclosive data or even user's memory. For the majority of researchers, data breach happens for access convenience and not out of a malicious intention and surely remote desktop access to the data would diminish that temptation. However, the possibility of disclosure is always there, the legal framework and training and education may deter users -who are approved researchers afterall - to perpetrate any confidentiality breaches.

8. Evaluation and monitoring of the outputs

Careful user vetting and the most secure analysis environment in the world cannot on its own ensure that data are not disclosed. The missing piece of the data security puzzle is not what goes into the secure data system, but what comes out of it. For the service to be able to meet the security guarantees placed upon it by the data guardians, it must offer some form of output screening. If an output has been determined to be disclosive, it will be up to the user to determine the best way to render it safe.

SDS adheres to European-wide ESSNet standards on good practice in statistical disclosure control of tabular and other statistical analytical outputs (Hundepool, et al 2009). The SDS disclosure advisor will divide outputs from SDS into two main categories:

- Safe: very low risk of disclosure – output will be released promptly
- Unsafe: high risk of disclosure – output will be blocked in its current form and won't be released; the researcher must produce safe outputs and demonstrate that they are free from the disclosure risks

There are several solutions available to protect the information of the sensitive cells:

- Combining categories of the spanning variables (table redesign). Larger cells tend to protect the information about the individual contributors better.
- Suppression of additional (secondary) cells to prevent the recalculation of the sensitive (primary) cells

9. Summary

The SDS is a secure environment funded by ESRC to provide researcher access to disclosive micro data either from their offices, safe rooms in their institutions or on site at the UKDA. It has two goals: to promote researcher access to sensitive micro data and to protect confidentiality. SDS operation is legally framed by the 2007 Statistics Act, which makes access to confidential data for statistical purposes possible.

Researcher access to microdata serves the public good both by leveraging existing public investments in data collection, and by ensuring high quality science through the replication of scientific analysis. The SDS provides Approved/Accredited researchers with remote access to microdata using the most secure methods to protect confidentiality. This is achieved by implementing technological security (Citrix gateway), applying statistical protections, enforcing legal requirements, and training researchers. The SDS also ensures that valuable data are preserved for the long term by documenting the data using DDI compliant metadata standards. In addition, the SDS aims to engage the research community in using its shared data space to share information which enables collaboration among geographically dispersed researchers.

References

Hundepool, Anco et al.: Handbook on statistical disclosure control version 1.1, ESSnet-Project. http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf. (2009)

Mulcahy, Timothy M, & John Nieszel.: Towards a Secure Data Service at the UK Data Archive. SDS Consultants' Report. (2008)

Ritchie, F .: Disclosure Control of Analytical Outputs. Mimeo: Office for National Statistics, UK. (2006)

Ritchie, F.: Disclosure control for regression outputs, Mimeo : Office for National Statistics, UK. (2007)

Ritchie, F.: Statistical Disclosure Control in a Research Environment. Mimeo: Office for National Statistics, UK. (2007)

Wright, Melanie: Case for Support – Secure Data Service. (2008)

Notes

1. This paper was presented at the IASSIST 2010 in the session "Secure remote access to restricted data". Corresponding author: rafkhami@essex.ac.uk, Dr. Reza Afkhami, Senior Data and Support Services Officer, Secure Data Service, UK Data Archive, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, Tel: +44 (0) 1206 874968, Fax: +44 (0) 1206 872003
2. Secure remote access is also developing in Denmark, Netherlands and Sweden.
3. http://www.data-archive.ac.uk/orderingData/agreements/ARFormsandNotes.doc
4. Statistics and Registration Services Act 2007 § 39 (5).
5. http://www.esds.ac.uk/aandp/access/licence.asp
6. Statistics and Registration Services Act 2007 § 39 (9).