

Lost in Reality: The Case for Virtual Safe Settings

Mustafizur Rahman¹, Marina Jirotko¹, William Dutton¹

¹University of Oxford, UK

Email address of corresponding author: mustafizur.rahman@begbroke.ox.ac.uk

Abstract. Despite significant advances in secure data transmission, even over public networks, the common data sharing practices for many organisations remain dated, such as sharing data with remote locations by transporting data via physical storage media. Likewise, sensitive data often reside on laptops that are carried out of the confines of secure office buildings. Major government data breaches, which are tied to such practices, have made headlines in the UK (BBC, 2007a, 2007b, 2007c). The tension between respecting confidentiality and enabling easy access by researchers to sensitive data has been an on-going challenge for data custodians. This paper highlights some of the most relevant initiatives aimed at creating safe settings for researchers to access such data without jeopardizing their security. These are examined in the context of collaborative e-research, where it is envisioned that data sets will be analyzed remotely from geographically dispersed locations. This review leads us to identify many of the common misconceptions surrounding barriers to data sharing, sensitivity of research data, and the basic legal obligations. The paper concludes by outlining the most realistic approaches to virtual safe settings that balance security risks with the facilitation of advanced research in the e-social science community.

Introduction

Researchers can benefit greatly from a multitude of datasets generated by other researchers, institutions and government agencies. Governments around the world publish data to fulfil legislative mandates, such as those compiled by the Office of National Statistics (ONS) in the UK. Similarly, the US Census Bureau publishes anonymized microdata files that consist of individual records containing values of variables for a single person, business establishment or other economic unit (CDAS and FCSM, 2001). Highly sensitive personal information, such as the respondent's name, address, gender, and ethnicity as well as other data that can identify individuals, is protected from unauthorized disclosure under numerous ethical guidelines and legal regimes. Various privacy and data protection acts govern the release and use of personal information. Public-use of microdata files, which we will refer to as microdata, have their personal identifiers removed when they are released for research purposes. Even these processed data are subjected to procedures that limit the risk of potential disclosures, such as through profiling.

Traditionally, data collection on individuals and organisations has been conducted via surveys, fieldwork and the routine collection of administrative data. Data are used to understand contributing factors and trends in economics, demographics, health and other aspects of everyday life, and increasingly, national security (Lane, 2007). Recent progress in Grid computing and Cyberinfrastructure has increased the potential for the collection and

dissemination of personal information at an unprecedented level, fundamentally changing the way scientists are collecting information and modelling human behaviour. For example, greater computing capacity can encourage the integration and use of entire population datasets, rather than sample surveys.

Safe Settings

There are data that are so sensitive that they simply cannot be made available without close supervision and under severe restrictions. Many agencies have therefore opted for supporting Restricted Data Centres (RDCs) or Reading Rooms. RDCs are secure sites that are located within the statistical agency responsible for the dataset, or at approved locations. Limited access over a specified period to these physical sites is granted under very restricted conditions following a lengthy research proposal assessment and security clearance. Even when a proposal is approved and an individual researcher is cleared and has taken mandatory training, access is restricted to pre-defined hours with an agency officer physically stationed at the site. Researchers may not take anything in or out of the RDCs. External data may be merged by RDC staff upon request and the resulting file is sanitised. All systems at the RDC are completely disconnected from any external networks, limiting access to a centralised system at the agency's data centre. All printers are located in a separate locked room with all print-outs examined by the host agency. Furthermore, types of analysis may still be limited with all materials removed from the site being subject to further disclosure analysis. These procedures create major costs for the operation of RDCs, both for the host agency to maintain and for the researcher to access.

Virtual Safe Settings (VSS)

A VSS provides an attractive alternative to the RDC. To access a VSS, researchers use a variety of resources, typically accessible at their home institution. They may require use of specialised computer systems and software packages, access to their own data, data available from their centralised services or other institutions, access to libraries and other such resources. Equally important, researchers may not have access to their colleagues who they may routinely collaborate with, especially in the case of those working on quantitative social science research. Lane (Lane, 2007) reinforced the view of a panel on the National Commission on Vital and Health Statistics (USHHS, 2006) that high-quality research is best promoted by allowing access to data directly from a researcher's offices. In this respect, a VSS not only promises to make access more economical, but also more effective in shaping the quality of research.

Many have advocated similar developments for sharing qualitative data. Secondary analysis of qualitative data, through the availability of original data, such as interview transcripts or recordings, could enable not only greater access to qualitative research, but also potential advances in theory and method. However, qualitative researchers have traditionally been highly resistant to the sharing and reuse of qualitative data. This is tied to the inherent personal aspects of qualitative research in which the analytical orientation cannot be separated from the context in which data were originally produced (Carusi and Jirotko, 2007). Moreover, similar to quantitative secondary analysis, qualitative data contains sensitive personal data (Fielding, 2006). Qualitative research characteristically emphasises personal experience, often addressing sensitive issues, eliciting intense emotional responses and therefore requires careful judgement on the part of the original researchers on what to disclose in publications and any release of data, such as transcripts.

Researchers conducting secondary analysis of qualitative data may also pursue interests quite different from the original study, applying a different perspective, conceptual focus, or purpose to the primary research. Furthermore, researchers with limited social science background, who are not socialized in the norms governing the protection of qualitative data, are increasingly working on interdisciplinary projects that could enable access and demand working with qualitative data.

A number of initiatives on VSS that are underway were investigated. There are a growing number of initiatives that may be considered and appropriately combined and/or modified for the UK e-Social Science Infrastructure in order to facilitate remote access to sensitive research data. These include innovations such as:

- New Licensing Agreements, in which limited data is released for a specified period and which dictate data security requirements, including unannounced inspections and researcher training.
- ONS's Virtual Microdata Laboratory (VML), which provides direct, on-site access to microdata at ONS following proposal assessment and researcher training.
- ABS's Remote Access Data Laboratory (RADL), enabling differential access methods developed by the Australia Bureau of Statistics (ABS) ranging from CD-ROM, batch-style analysis to limited access to microdata over the Internet. Unrestricted access to microdata is only available on-site.
- Buffered Remote Access, which allows limited statistical programmes to be executed against microdata held remotely. Results of the analysis are sent back to the researcher following automated disclosure analysis.
- Data Enclave, an enhanced version of Buffered Remote Access in the USA, which utilises remote desktop technology to execute statistical programmes against microdata from registered systems. Direct access to microdata is only available on-site (at the National Opinion Research Center, or NORC).
- Remote Analysis of Microdata Files, which is similar to the Data Enclave but allows access from anywhere in the world, some incorporating biometric authentication (UNECE, 2007).
- Online Query Systems, which are Web-based interfaces to analytical back-ends with restricted functionality.

Audio/Video: The Qualitative Challenge

Audio, and increasingly video data, pose major challenges in complying with legal and ethical obligations. Unlike microdata, audio-visual data can be exponentially more complex to anonymize without losing its value. For example, it may be possible to identify subjects based on a broad range of familiarity such as background noise, voice, language, accent, hair colour, skintone, clothes, and the presence of background objects, such as a painting on the wall.

However, it is possible to create a closed user group. MiMeG (Shaukat and Fraser, 2007) demonstrates promising advances in establishing a trust framework based on contract systems, underpinned by Digital Rights Management (DRM), to share video through controlled access. Additionally, control can be exerted in real-time such that access is available only in the (virtual) presence of the data custodian. Other groups are experimenting with systems for automatically replacing subjects with their virtual duplicates (digital objects) to protect the privacy of individuals (Fan et al, 2005). These emerging systems can help facilitate sharing of video data whilst preserving confidentiality obligations under many

circumstances. Nevertheless, they obviously lose data that many qualitative researchers judge to be essential.

Images are also increasing in popularity with the widespread availability of digital cameras and digitised medical equipment, enabling digital video to slip into work that is not done within a rigorous qualitative tradition. Individual images can be blurred appropriately to preserve confidentiality, but again by losing data of possible importance. Furthermore, access can be controlled via DRM systems ranging from encryption and watermarking (fingerprinting) to the use of virtual appliances (Rahman, 2006).

To Share or Not to Share

Interviews were conducted with a range of social scientists and data providers to elicit information about their current practices. Other researchers who are exploring implementing various models of electronic safe settings were contacted as well to discuss the practical limitations of alternative approaches, despite many advances in technology. Site visits to national archives, including participation in a mandatory on-site training session in handling sensitive microdata, were employed as well.

The Lone Researcher

Although social scientists do have a long history of collaborating with their colleagues, the bulk of the actual research is typically carried out individually. Specifically, social scientists have traditionally analysed, if not collected the data, largely on their own and subsequently shared the findings with their peers. This has been particularly true for qualitative data, partially due to many of their concerns over quality and the protection of their subjects (Carusi and Jirotko, 2007). While this differs by discipline, most qualitative researchers are taught early in their career to err on the side of caution in order to protect the data subject, regardless of their actual sensitivity. However, statistics quoted by a data provider stated that only about 35% of the data deposited in the leading archive cannot be shared, largely due to lack of appropriate consent.

Many more researchers are gradually realising the merits of data sharing as collaborative research projects are becoming the norm. However, they are limited in part by the lack of collaborative infrastructures, combined with traditional ideas of data custodianship that are very much ingrained into current work practices.

Besides ethical issues, concerns over intellectual property rights (IPRs) and appropriation of due credit were also cited as disincentives. Other legal issues such as copyright protection can impose additional obstacles to the use of data. Some of these obstacles, especially public data such as ordinance surveys held by government agencies and public/private partnerships, may be attributed to potential loss of income, control, or influence.

Even in the case where researchers may want to share their data, there is currently a notable lack of tools and software systems to facilitate this process. For instance, systems such as QDA and NVIVO have been slow to extend their functionalities for collaborative work such that various researchers may code different parts of the same data sets. Emerging Virtual Research Environments tend to provide communication tools, as noted by one respondent.

Institutional Research Organisations

Government departments require personal and organisational information, especially in light of increasing popularity with evidence-based policies, for legislative purposes and for the common good. The ONS has developed a trust model allowing researchers to access

microdata via the VML. They rely on the following principles to minimise any misuse:

- Safe projects – projects must show that they provide direct or indirect benefits to the ONS and contribute towards greater knowledge of the data set.
- Safe people – access is only granted to researchers from approved organisations.
- Safe settings – technological solutions ensure that no data leave the laboratory without ONS permission and all input/output are archived.
- Safe output – all researchers must undergo training sessions highlighting disclosure control mechanisms with all outputs and publications vetted by ONS staff.

Additionally, VML is actively policed with sanctions in place for any researcher attempting to circumvent disclosure controls. Many of the other European countries have similar facilities in place whilst more restricted access is generally in place in the United States. Australia, on the other hand, has moved towards providing differential methods of access through CD-ROM, RADL and on-site. Similar to the ONS, ABS requires researchers to sign legal undertakings, attend training and comply with data security guidelines.

Data archive and data service organisations such as UKDA and ESDS do not generally place restrictions on users as data sets are generally sanitised or consent has been acquired. The Medical Research Council makes it explicit that it expects all data to be made available to the scientific community with as few legitimate restrictions as possible (MRC, 2008). It does allow for a limited, defined period of exclusive use to encourage innovations.

The Ties that Bind

Data sharing, especially through linking of datasets, can potentially transform the landscape in the social sciences through powerful analytical capabilities facilitated by emerging e-Infrastructures. They, however, also increase the risks of disclosure and profiling. These risks are neither readily understood nor can they be fully anticipated in advance. Data providers therefore provide strict controls on data linkage. However, one of the interviewees noted that complex data linking, especially a link that may lead to effective disclosure, is not an easy task.

A Socio-technical Approach

Recent news headlines have been awash with stories of detailed personal data ending up in the wrong hands, including equipment disappearing from security conscious personnel at the Ministry of Defence (BBC, 2008a). Thus data providers, especially those acting as the data custodians on behalf of data collected by governments, must strictly adhere to legislation protecting the confidentiality of any data collected on individuals and organisations. In large part, this requires the establishment of safe data handling procedures. Rather than transporting copies of data using insecure physical medium, or at the other extreme allowing costly access only through RDCs, electronic safe settings could facilitate managed access to confidential data under tightly controlled environments. It must be emphasised that no system can provide absolute guarantees against misuse. For instance, there is no protection against someone memorising particular microdata at a highly monitored RDC and replicating that information once they leave the centre. Or, trained personnel at the US State Department might misuse their privileges to improperly access records outside their regular duties (BBC, 2008b). However, privacy laws and ethical guidelines are in place to pursue legal actions for any breaches in confidentiality. Virtual safe settings could provide workable solutions by allowing electronically monitored differential access to substantial portions of research data.

Critical Success Factors for VSS

The primary goal for a VSS is to allow electronic access to research data without breaching legal, ethical and institutional obligations on behalf of both the data provider and the researcher. Risks of disclosure cannot be reduced to zero under any scheme. Data should also be searchable and curated such that they remain accessible for generations to come. For widespread adoption, mechanisms for crediting the primary data collector with due acknowledgement should be an integral part of the overall architecture. This study found that the following should be taken into account for virtual safe settings to succeed:

- Research training - ensure that the stakeholders are updated on handling of research data. The message should be delivered emphasising the merits of data sharing for the advancement of knowledge, especially in the social sciences. Researchers should be further trained on approaching their subjects for appropriate consents. Community trust model(s) should be developed with researcher made aware of their obligations, both legal and ethical, and potential sanctions upon breach of confidentiality.
- Controlled Access – access to highly sensitive data should be controlled based on per user request. A trust model similar to ONS combined with the RADL functionalities could be utilised. Access may be controlled by the primary researcher, perhaps for a designated time period, or delegated to institutional data providers.
- Credit Report - VSS should ensure that primary researcher receive due credits for data used for secondary analysis and all relevant researchers are duly cited in publications. Sanctions should be in place such that appropriate actions, legal or otherwise, can be taken if anyone is found not providing appropriate citations.
- Monitor Access - every access to data should be logged. This will not only provide usage data but will also provide evidence in case of improper/unauthorised access and disputes with credit reporting.
- Technologically Secure - appropriate technical systems (such as those highlighted earlier) should be selected and appropriately combined ensuring that data are only accessible in the manner specified for any particular VSS. All data should be transmitted using proven secured channels upon appropriate authentication. Depending on the sensitivity of the data, different systems and multiple levels of access privileges may coexist to serve the specific needs.
- Minimise Logistical Barriers - too often, security systems are in place that protect the data but hinder seamless access. This may be due to localised firewalls at departmental level, acquiring and deploying digital certificates or other hurdles such as obscure username and/or password requirements. Such hurdles should be minimised.
- Technical Support/Deployment - systems should be designed to facilitate the deployment of appropriate applications on their systems with minimum technical skills. Help desk support should be provided where possible in case of difficulties.
- Eternal Data Curation - systems, software, as well as data formats will evolve over time. VSS should set guidelines and provide assistance to researchers to translate their data into supported formats such that they can be curated appropriately into the future.
- Describe and Share - the potential of VSS can be realised if powerful search tools exist to locate the relevant data. Data curation procedures should incorporate appropriate meta-data technologies to tag new data. Data should be well-documented, describing how they were collected, conditions under which they were collected,

meaning of variables, terms of usage/consent, etc. This will enhance the search facilities and make relevant data more easily accessible.

- Sustainability/Maintenance - VSS must be backed up, maintained, upgraded and transitioned to newer systems over time. Financial models should be considered to ensure personnel and systems are supported for continuous access to data.

Towards an e-Infrastructure for the Social Sciences

Data collection will continue to soar, especially as innovative (electronic) sensors are deployed across the globe, which will capture an ever increasing number of activities around us. Much of the data collected will be extremely useful for research, particularly in the social sciences. However, the tension between respecting confidentiality and providing access to sensitive data, especially from unmanned remote locations, remains a formidable challenge.

It is evident that it is not possible to provide absolute guarantees against breaches in confidentiality. This remains true even despite the highly regimented access at the RDCs. Thus, the possible risks need to be carefully considered, taking into account current legal and ethical obligations, to come to a reasonable compromise on how remote data access may be facilitated securely. Further research needs to be conducted in this area through a complete review of data access requirements involving research institutions and government bodies, in conjunction with the public, to come to an agreement on the acceptable level of risk.

Good science needs to be repeatable and further knowledge can be extracted from secondary analysis of existing data. Without greater access to such valuable data, the future of science, healthcare and prosperity are at risk of being compromised, as one interviewee highlighted. Virtual Safe Settings, if designed properly, can be adapted under certain conditions to provide access to data remotely and thus contribute to the building blocks of future e-Infrastructures for social science research. However, the challenges are significant, as discussed in this review.

References

- Australian Bureau of Statistics (ABS) (2006). ABS Remote Access Data Laboratory (RADL): User Guide Version 4. Australia.
- British Broadcasting Corporation (BBC) (2007a). Discs 'worth £1.5bn' to criminals, http://news.bbc.co.uk/1/hi/uk_politics/7117291.stm. [Accessed 10 February, 2008]
- British Broadcasting Corporation (BBC) (2007b). Data of 60,000 on stolen computer, http://news.bbc.co.uk/1/hi/northern_ireland/7133194.stm. [Accessed 10 February, 2008]
- British Broadcasting Corporation (BBC) (2007c). Up to 3,000 patients' data stolen, <http://news.bbc.co.uk/1/hi/wales/7143358.stm>. [Accessed 10 February, 2008]
- British Broadcasting Corporation (BBC) (2008a). Details of Scots on stolen laptop. <http://news.bbc.co.uk/1/hi/scotland/7214464.stm>. [Accessed 28 February, 2008]
- British Broadcasting Corporation (BBC) (2008b). More US passport 'file breaches'. <http://news.bbc.co.uk/1/hi/world/americas/7315813.stm>. [Accessed 27 March, 2008]
- Carusi, A and Jirotko, M. (2007). From data archive to ethical labyrinth. In *Proceedings of the 3rd International e-Social Science Conference*, Ann Arbor, Michigan, USA,
- Cohen, S. (2002). Access to Confidential Statistical Agency Data. Bureau of Labor Statistics, USA.

- Confidentiality and Data Access Committee (CDAC) and Federal Committee on Statistical Methodology (FCSM) (2001). Brochure - Confidentiality and Data Access Issues Among Federal Agencies, Federal Committee on Statistical Methodology, USA.
- Fan, J., Luo, H., Hacid, M., Bertino, E. (2005). A novel approach for privacy-preserving video sharing, Proceedings of the 14th ACM international conference on Information and Knowledge Management, pp 609-616, Bremen, Germany.
- Fielding, N (2006). The Shared Date of Two Innovations in Qualitative Methodology: The Relationship of Qualitative Software and Secondary Analysis of Archived Qualitative Data. In *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, [Online Journal], 1(3). Available at <http://qualitative-research.net/fqs/fqs-eng.htm>. [Accessed 17 October, 2007].
- Lane, J. (2007). Optimizing the Use of Micro-Data: An Overview of the Issues. *Journal of Official Statistics*, Vol.23, No.3, pp. 299—317.
- Medical Research Council (MRC) (2008). MRC Policy on Data Sharing and Preservation. Available at <http://www.mrc.ac.uk/PolicyGuidance/EthicsAndGovernance/DataSharing/PolicyonDataSharingandPreservation/index.htm>. [Accessed 19 February 2008]
- Rahman, M. (2006). Securing IPR in e-Health, MSc thesis, Software Engineering, Exeter College, University of Oxford, UK.
- Schouten B., Cigrang M. (2003) Remote Access Systems for Statistical Analysis of Microdata. *Statistics and Computing*, Volume 13, Number 4, pp. 381-389(9).
- Shaukat, M. and Fraser, M. (2007). A Security Framework for Data Distribution in Qualitative Analysis Tools: Digital Rights Management in MiMeG, in Proc. ICeSS 2007, Ann Arbor, Michigan, USA.
- United Nations Economic Commission for Europe (UNECE) (2007). Managing Statistical Confidentiality & Microdata Access – Principles and Guidelines of Good Practice, Geneva, Switzerland.
- U.S. Department of Health and Human Services, National Commission of Vital Health Statistics (USHHS) (2006). Workshop on Data Linkage to Improve Health Outcomes. Available at <http://ncvhs.hhs.gov/060918tr.htm> [Accessed 8 October 2007]