Annex 1

Improving the security of qualitative data in a digital age: a protocol for researchers

Judith Aldridge & Juanjo Medina

Introduction

This protocol for improving the security of digital qualitative research data was developed by Judith Aldridge and Juanjo Medina in the course of conducting the ESRC funded research ‘Youth Gangs in an English City’. This document describes a set of guidelines that were developed after we revised our security procedures after the theft of a laptop containing sensitive data. These take into account the varying forms that digital data are held, the numerous physical locations and storage devices that must be considered, the individuals with a range of roles on a research team, and the phases of research.

Background to the protocol

Our research involved collecting highly sensitive data in the form of interviews and fieldwork notes from gang members and former members, gang associates, and others in the community. After considerable time and effort, we were finally successful in gaining the trust of key actors; we were rewarded in hearing them talk candidly about their own lives and those of others, and about serious criminal events. Should this data be made public, this could result in danger to interviewees themselves or people they discussed, from others in the community or from the police.

The original ethical protocol that we developed in proposing the research addressed issues around securing our digital data. We knew that it was no longer sufficient in a digital age that the ‘locked filing cabinet’ (to store paper copies of data with identifying information) could provide sufficient data security. However, we understood that using digital approaches to holding qualitative data had advantages that could be important for us, particularly facilitating team-working. Our protocol sought to:

1) minimise paper versions of interview data (transcriptions) and fieldwork notes in favour of holding these digitally from the start or as soon as possible;
2) minimise the time period during which data was held outside of our university offices (with fieldworkers, with transcribers);
3) anonymise data as early as possible in the period of holding the data;
4) destroy non-anonymised versions of data (i.e. voice recordings of interviews) as soon transcription had occurred, and again, as early as possible in the process;

5) back-up all digitally held data onto a transportable medium (CDs or mini-disks) to be stored in a locked filing cabinet; and
6) use passwords on all word-processed documents (e.g. interview transcripts, fieldwork notes) and to protect all data held on desktops and laptops through ‘Windows’ passwords.

We felt that these procedures together were fairly rigorous, and that whilst they did not guarantee the security of our data, they went a considerable way towards it.

Only a few months into our fieldwork, an unexpected event put our procedures to the test. A laptop containing a handful of both digital voice files of recorded interviews – and their transcriptions – was stolen from the home of a fieldworker after a break-in. Our first response was to report the theft to the police and alert the University of Manchester Ethics Committee (to whom we were obliged to report unexpected events affecting ethical aspects of our research). We then set about informing each of the people with whom we had conducted the stolen interviews of the loss. We were fortunate that not one of these interviews (all had occurred early in our research) involved interviewees discussing other people, or highly sensitive events such as those we succeeded in getting data about in later stages of our research. We were also fortunate that our interviewees were all unfazed by the loss, convinced that the laptop’s contents were very likely to be wiped, before being sold on. Finally, we set about re-thinking our security procedures. One part of our re-think involved consulting a clinical colleague, who had developed some expertise on data security in relation to data he held about his patients, along with IT people in our own university with expertise on data security.

We had mistakenly believed when starting our research that holding our data digitally automatically afforded relatively easy and straightforward security procedures (e.g. the use of passwords), and would therefore almost inevitably be ‘better’ than paper-based data in this regard. We have discovered, however, that digitally-held data are just as vulnerable as paper-based data, and in some ways are more vulnerable. For example, digitally held data facilitates sharing data quickly within a research team (such as between fieldworkers and managers); this is an undeniably useful feature of the digital approach to team working. However, this in itself multiplies opportunities for insecurity. Procedures developed for a time when the qualitative research process was primarily paper-based (e.g. the ‘locked filing cabinet’), on their own, provide insufficient security for qualitative data in a digital age. Indeed, paper versions of ‘raw’ data increasingly may never appear, as all operations carried out by qualitative researchers – from collecting and transcribing data, to reading and analysing it – can be now be carried out, and increasingly are carried out, on-screen only, and without ever having to print data onto paper. It is therefore important to recognise that digitally held data does not automatically provide for ‘better’ security: both digital and paper-based approaches to holding and processing qualitative research data bring with them their own security problems that need to be incorporated into security procedures.

**Issues to consider**

Before presenting our guidelines, we outline a series of issues that need to be
understood in relation to carrying out qualitative research using digital data. Digitally held data allow multiple copies of data shared across people, locations and devices (e.g. to back-up data, to share data). As copies and versions of data proliferate, insecurity opportunities increase.

1. **Digitally held data is stored in numerous locations during research.** These places include: in ‘the field’; in the home of fieldworkers, interviewers, transcribers, and research managers, in the university offices of people with various roles, and ‘in transit’ between all these locations. *Security procedures developed for digital data must take into account the potentially numerous geographical locations in which digital data are held over the course of research.*

2. **Digitally held data is stored in numerous forms.** Digital versions of qualitative data are held in a number of forms: digitally recorded voice files (of interviews or fieldwork notes) held in MP3 format or similar; transcriptions of voice files held in word-processed documents; versions of work-processed text files held in CAQDAS (such as NVivo and Atlas.ti). *Security procedures developed for digital data must take into account that digital versions of data are held in a number of forms including voice files, text files for holding transcripts and notes, and text files for use in data analysis software (CAQDAS).*

3. **Digitally held data is used by numerous individuals with a variety of roles on a research team.** The many forms of digital data, held in many places, are also used by a number of individuals with different roles on a research team, and can be transferred back and forth between them, for example, in preparation for, and subsequent to, team meetings in which collected data are reviewed. The relevant individuals include: interviewers, fieldworkers, transcribers, research managers, and those responsible for data analysis. *Security procedures developed for digital data must take into account that digital data is likely to be passed back and forth during the data collection and analysis phases of research between members of a research team with different roles.*

4. **Digitally held data are stored on a number of different devices during data collection and analysis.** The devices on which digital data are held during fieldwork can include: the recording device on which voice recordings of fieldwork notes or interviews take place (increasingly, these are MP3-type recorders with in-built memory or hard-drives, and often are recorders with removable storage such as mini-disks); university office and home-based desk-top computers; laptop computers; and portable digital storage devices like memory sticks (aka USB drives, pen drives) and CDs. *Security procedures developed for digital data must take into account that digital data is usually stored a number of different digital devices.*
5. **Digital data are held over a range of times coinciding with the various phases of qualitative research.** These times or phases of the research process during which digital data are first collected and then manipulated and held include: initial data collection, adding to collected data, storing collected data, analysing data, and then finally, more long term archiving of collected data. *Security procedures developed for digital data must take into account the different phases of research and the length of time that data are required to be held.*

**Guidelines**

Our revised protocol patches up some of the holes in our original data security procedures, and further takes into account developments in dealing with data that we made during the course of our research. The issues identified above (where data are held, in what forms, by whom, on which devices, and over what period) have all been considered in these guidelines.

1. **Passwords to protect access to files, computers and devices are useful, and should be employed.** However the protection offered by many password features is flimsy, and can be relatively easily by-passed. Our initial use of passwords was limited to files on which these were already available as features (we used Microsoft Word for text files, and employed the password protection facilities available within Windows). We also used individual passwords to protect access to our desktop and laptop computers. As there was no obvious way to password-protect our voice files, at least prior to revising our protocol, we did not. After consulting our security experts, we learned that the kinds of passwords were using were easy to break, and that the software on which we relied did not incorporate security features that were adequate to our needs.

   *Our recommendation: use ‘good’ passwords that: combine letters and numbers, combine uppercase and lowercase characters, and are sufficiently long. Use different passwords for different purposes. Do not allow your computer to ‘remember’ a password for you; instead, enter this each time from memory. Note that some kinds of passwords can be fairly easy to break for a determined hacker.*

2. **Use encryption software to provide secure places to hold data on desktop and laptop computers, and on portable storage media such as memory sticks.** Our discovery that the passwords, and the features of the software programmes employing them, that we used to ‘secure’ our data were actually relatively insecure (see Guideline 1) led us to begin using encryption software to increase the security of on all the places we were storing data (including hard drives on our laptop and desktop computers, and on the memory sticks we used to both transfer and back-up these data). A number of different approaches to encryption of text-based data are available, and different products on the market reflect these approaches. After the advice of experts
and some initial research, we chose ‘Steganos’. This product allowed us to identify a portion of a hard drive that we wanted to secure; everything stored within it is referred to as an ‘encrypted disk’. When opened, work on files in the encrypted disk area carries on as usual. When closed, the encrypted portion of the disk itself is not visible. Access to the encrypted disk is controlled via a password, and Steganos assists users in creating highly secure passwords. Encryption is not permanent: when the encrypted disk is opened, files can be removed from it and placed elsewhere, thus removing the encryption. The use of encryption software was, for us, the single biggest improvement to our security procedures. All members of our research team used an ‘encrypted disk’ to store all files related to the project on each of the computers they used in their work (that is, all desktop and laptop computers). We found it easy to use the encrypted disk to store all the different kinds of files we used, including voice files, text files, and analysis files. Our use of encryption software meant that, in the event of loss or theft, such as happened to us early in our research, it would be very unlikely for a thief even to know that an encrypted disk was present; and if they did know this, they would be extremely unlikely to be able to access these files contained in it. Note: Like any other approach to security, however, encryption is not a magical solution, and must be employed in conjunction with a range of other security solutions. It is important to learn to use the software properly, for example by removing temporary files.

*Our recommendation: use encryption software to create digital space on computers and storage media for holding all qualitative data and related work.*

3. **Ensure that all members of a research team use the same security procedures.** This includes transcribers, and others not always considered to be core team members. Transcription is sometimes done by contracting the work out to external companies, and it is important not to assume that the security procedures they have in place are sufficient. Ask them to demonstrate their security procedures. If the procedures they have in place do not appear to be as robust as yours, ask them if they would be willing to adopt your procedures for the work they do for you.

*Our recommendation: insist that all people who will be dealing with your qualitative data, including those not directly employed by your institution, use security procedures at least as good as the ones you employ.*
4. **Consider storing data online using an anonymous password protected webpage.** At a fairly early stage in the research, we developed, along with a member of our IT staff, a way for our team members (initially three fieldworkers and two principal investigators) to share progress, ideas and data in a way that was both relatively secure and that reduced the need for creating multiple copies of our data for back-up. We created a set of linked password protected webpages (which referred as our ‘Research Journal’). The Journal consisted of online space for us to: place relevant field contacts, timetable weekly field activities, contribute relevant links and documents from the literature and describing important local events, and most important to this discussion, store data files (including voice files, and text files of fieldwork notes and interviews). The Journal pages were constructed so that no links to it existed anywhere in the university (or elsewhere), and web pages were stripped searchable ‘tags’ of HTML code of so that they could not be accessed through web-searching. This Journal initially provided us with a useful forum to exchange data and other kinds of information with one other quickly and easily. The primary advantage in terms of data security was that, once files were placed on the Journal (password protected where possible), originals could immediately be deleted from the devices of fieldworkers (that is, their fieldwork laptops and voice recorders). The data linked to the webpages was backed up daily by the university in its normal procedures. This therefore eliminated the need for having back-ups available in multiple locations. Both of these features of the online approach to data storage thus reduced the problem of insecurity through the proliferation of files. The Journal itself was password protected. **Note:** We only used our Journal to hold research data in the early phases of our research, and abandoned it after moving to the use of encryption for work we later carried out on local computers. In spite of being assured by IT staff in our university that the Journal was a secure place for data storage, we nevertheless remained nervous about this approach, worrying that a determined hacker could find the way to our data. In future team-working qualitative research projects, we might again consider the on-line approach to data storage after exploring methods for encrypting data on the server and remote access through a Virtual Private Network (VPN).

*Our recommendation: the creation of a suitably secure online location for the storage of data files, that cannot be accessed through web searching, may reduce or eliminate the need for security-problematic multiple copies of data to facilitate data sharing in research teams and back-ups. However, the use of shared networks/servers carry with them a different set of security risks to those posed by storage on local computers and portable media, and in future, we would explore additional security possibilities including encryption and remote access through VPNs.*
5. **Make back-ups of work carried out on encrypted computer disks to portable encrypted media (e.g. memory stick), and store this separately to the computer where original work is carried out.** Because both in-built and portable disks (e.g., memory sticks) can be encrypted using software such as Steganos, use of this product simultaneously helped us to secure our back-ups. During the analysis phase of our research, all data were held on an encrypted portion of the computers being used by each of the four of us involved in data analysis using Atlas.ti. Our individual coding progress was saved not only on this encrypted disk, but also backed up daily to the encrypted portion of a memory stick (pen drive, USB drive). For those working on laptops and carrying these back and forth between home and university offices, the laptop and memory stick were always carried separately, in case of loss or theft.

Our recommendation: **minimise multiple copies of data as back-up where possible through the use of online data storage (see Recommendation 4). Where this is not possible, ensure that backed-up data is stored on an encrypted disk.**

6. **Paper fieldwork notes should be destroyed as soon as possible after digital encryption.** Recording notes in the field can be done in a variety of ways, to suit both individual preference (e.g. digital voice recordings of observations versus hand-written notes on paper), and the demands of the research project itself. Some settings, for example, may be suitable for taking laptop/notebook style computers to type notes that immediately result in text files. In other settings, the use of visible electronic equipment may be inappropriate or even risky. Security procedures will vary depending on the approach to note-taking, but some principles for good practice apply. Handwritten notes should be translated into encrypted digital notes text files as soon as possible and destroyed with a cross-cutting paper shredder. Where recording ‘dictaphone-style’ digital voice files, use password protection on the device if available, translate into encrypted digital text files as soon as possible, and then delete from the recording device. Use of a laptop computer to record fieldwork notes in a way that is immediately encrypted has the advantage that initial parallel versions of the notes are not constructed to later be destroyed. This disadvantage is that, during use when the encrypted disk is open, its contents can be viewed by others.

Our recommendation: **Employ the style of field note taking that is most suitable to the research and that suits personal preferences; however, delete or destroy notes held temporarily in any medium that is not encrypted (paper notes, voice recording) in preference for more permanent storage on the encrypted disk of a computer that is not based ‘in the field’.”

7. **Minimise paper printouts of data.** In general, paper printouts of data should be minimised. All phases of the research process can (in principle) be carried out without recourse to paper, with work (transcribing, reading, coding, analysis) being viewed on-screen. However, even researchers adopting primarily digital methods sometimes prefer not to carry out some tasks on
screen, such as reading transcripts, or comparing documents. Other tasks, similarly, may be inherently more suitable to paper and pen approaches, such as field note taking in particular research environments where visible equipment may inappropriate. Wherever data are printed onto paper, documents should be kept in locked cabinets when not in use and shredded using a cross-cutting paper shredder immediately after use.

Our recommendation: in general, employ on-screen methods for reading and analysing data. Where you do print out paper versions of data (such as in meetings), shred using a cross-cutting shredder as soon as possible; store paper in a locked filing cabinet as a last resort.

8. Simply deleting files from a computer’s hard disk does not remove them permanently. An additional advantage of using the Steganos product we used for encryption (‘Steganos Security Suite’) is that it contained other useful security features. We used this programme to permanently delete files so that they cannot be recovered, and to permanently ‘clean’ the spaces on the disk where files have been deleted from. This is important, since computers can often be sold on or recycled within research institutions.

Our recommendation: employ methods to ensure that data files you want to delete are permanently and completely deleted, so that they cannot be recovered. Steganos Security Suite, amongst other products, provides this.

9. Anonymise interviews and fieldwork notes as early as possible into the data collection process. It is considered good practice to anonymise interview transcripts as early as possible into the data collection process to improve data security. This was problematic in our research for a few reasons. First, many of the individuals we interviewed would be recognisable to some others reading even anonymised transcripts because of the – usually well known – events they discussed. Second, although it is simple enough to assign pseudonyms to interviewees, it was not straightforward to do so with the many scores of references made during an interview to other named individuals, a difficulty further complicated by the fact that named individuals often had one or more ‘street’ names or aliases by which they were known. It was only as our research progressed that we began to match names with aliases; but this process could only ever be partial. For this reason, working out ‘who was who’ could never been done in a final way – essential when replacing names with unique pseudonyms. The only procedure which is consistent is the - admittedly minimalist - approach of replacing all individual names with ‘person name’. Doing so, however, reduces the analytic utility of the interviews. In the end, we were not only unable to complete anonymisation at the early juncture we had planned, but unable to do it in a sufficiently satisfactory way to retain data integrity. Although this was less than ideal from a security point of view, we made use of our updated security procedures to store non-anonymised transcripts for longer than originally anticipated.
Our recommendation: Early anonymisation of interview transcripts is good practice; if anonymised transcripts are lost or stolen, problems in relation to confidentiality are minimised. In our research, anonymising the names of individuals and places often resulted in individuals and events still being recognisable, and was in any case practically difficult because of the widespread reference by interviewees to named individuals and well known events. To compensate, we extended our use encrypting these data files over a longer period than originally anticipated.

10. If interview data pose confidentiality problems that prevent them being publically archived, consider alternative ways in which access to the data by other researchers can be negotiated in order to facilitate secondary analysis and comparative research. As described in Recommendation 9 above, our data were particularly difficult to anonymise for our own security purposes. The process of anonymisation that we attempted for the purpose of public archive was even more extensive, and rendered the data ‘left’ after anonymisation considerably impoverished, and in some instances actually misleading. In addition to removing specific person and place names, we also had to black-out sections of text that referred to recognisable events and people, even where specific names were not mentioned. We found that ‘what was left’ amounted mostly to non-specific generalities, and that this kind of talk often contradicted more detailed reporting about specific events elsewhere in the transcript. However, there are precedents for sharing data whereby specific use agreements can be individually negotiated. We believe it is important to allow access by other researchers to our data in order to facilitate both secondary analysis and comparative research. For these reasons, we will explore the possibility of negotiating agreements with individual researchers and research teams for them to use our research only under our the specific circumstances set out in the agreement. Such agreements would need to be negotiated with institutional ethics committees where the research was carried out, and specify conditions, such as (for example) that data are re-analysed on site.

Our recommendation: if the type of data produced within the research prevents it from being publically archived, consider putting in place the possibility for individually negotiated agreements allowing other researchers access to the data.

11. Former employees may retain possession or access to data. Take steps to reduce this. Not all those who work on a research project do so from start to finish as principal investigators usually do, because their contributions may be limited by the nature of their work (transcription for example) or as a result of their employment contract. It may be useful to put in place agreements limiting ongoing access to data with all those working on a project. The agreement should require these individuals to return data and media on which they are stored when they no longer require access to that data (either because they no longer require access to carry out their tasks, or because they are no longer employed on the project). Agreements like these might also specify
specific ‘good practice’ behaviour such as ‘deep cleaning’ hard disks where data have been stored but deleted. It is important to then ensure that these activities occur when contact with the employee is still in place. Another useful strategy is to change passwords that grant access to data that are held communally once the need for that access is finished.

Our recommendation: Put agreements in place with those employed to work with data to ensure that data are returned before the end of their employment, and that ongoing access to data is removed.