Security and anonymity of qualitative data in a digital age: the experiences of ethnographic gang researchers, and thoughts on the feasibility maintaining anonymity in digital archive

Judith Aldridge, Juanjo Medina & John Churcher

Introduction

This paper has two overlapping aims. The first is to describe some strategies we developed to improve the security of our digitally created and held qualitative data (involving interviews with gang members) after the theft of a laptop computer – containing highly sensitive data – from the home of a fieldworker. After a thorough review of our data security procedures, we found them to be lacking. We reflect in this paper on our experience of our recently completed ethnographic research *Youth Gangs in an English City*[1]. Our improved guidelines take into account what we refer to as the 'principle of proliferation': versions and copies of qualitative data held digitally proliferate quickly as a result of being held: in different forms (voice, text), in various physical locations (office, home, in the field), on different storage devices (voice recorders, laptops, desktops, memory sticks), by individuals with particular roles on a research team (managers, fieldworkers, interviewers, transcribers), and during different phases of the research (data collection, data analysis, data archive).

The second aim of this paper is to contribute to the emerging literature on ethical issues related to the archive of qualitative data, again using the exemplar of our ethnographic gang research. We describe the characteristics of our interview data with gang members and others that limit the possibility of maintaining confidentiality for our research participants if these data were to be publicly archived, even when identifying names and places have been stripped away. We draw on some psychoanalytic literature where similar issues have been debated in relation to the publication of case study material. We argue that preparing and lodging data in a public archive should be seen as another phase of the research into which data security procedures for a research project should be extended. We propose an alternative to the public archive in instances where achieving anonymity in archived material is not possible.

Background to the protocol

Our research involved collecting highly sensitive data in the form of interviews and fieldwork notes from gang members, 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, over a period of more than two years, about their own lives and those of others, including in relation to often very serious criminal events. Should these data – in the form of, for example, transcribed interviews – be made public, this could result in danger to interviewees themselves or people they discussed, from others in the community or from the police.

[1] ESRC funded research 'Youth Gangs in an English City' REFERENCE No. RES-000-23-0615

Only a few months into our fieldwork, an unexpected event put our our original security procedures to the test. A laptop containing a handful of 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 our university 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 in discussing other people, or highly sensitive events such as those we succeeded in getting data about in later stages of the 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.

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. We had 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 feature in itself multiplies opportunities for insecurity. 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.

The 'principle of proliferation'

The principle of proliferation can be illustrated by counting the versions and copies of some qualitative data in the form of an interview conducted in the field, that are produced on its 'journey' over the course of a research project. This is illustrated in Table 1 below, with a running count of the copies and versions that result at each turn. Simply keeping track of all these copies of data in its various forms and physical locations is a complicated data and personnel management issue. What is certain is that as data proliferate, so do opportunities for insecurity.

Table 1: The proliferation of digital data: the journey of one interview

Activity	Copies of data files produced (cumulative)
An interview is conducted by a fieldworker and recorded as an MP3 file on voice recorder	1
This file is later transferred to the fieldworker's laptop computer	2
This voice file, via a memory stick, is transferred to the research manager who stores the file on a university desktop computer…	3, 4
…and later transferred again, perhaps after a few interviews have accumulated, to a transcriber, again via a memory stick	5
The transcriber stores the voice file on a laptop, where it remains during the process of transcription into a text file	6, 7
The text file is transferred back to the research manager via a memory stick, where it remains on the university desktop computer	8, 9
This transcribed text file of the interview is shared in a research team meeting after it is emailed…	10, 11, 12, 13
…to all four staff members participating in the meeting, and after these team members have saved the text file of the interview to their desktop/laptop computers…	14, 15, 16, 17
…the text file of the transcribed interview is 'backed up' for the sake of security and to protect against loss/damage	18
All transcribed interviews, including this one, are transferred to the locations in which analysis using CAQDAS by the three team members carrying out data analysis will work	19, 20, 21
Finally, once research and analysis are complete, the text file of the transcribed interview will be stored permanently and securely for future use on university premises, perhaps on a desktop or burned to a CD…	22
…and then prepared for public archive…	23
…and finally archived	24

Table 2 below provides an analysis of the underlying sources of proliferation for qualitative data that include its various forms (voice, text), physical locations (office, home, in the field), storage devices (voice recorders, laptops, desktops, memory sticks), individuals with particular roles on a research team (managers, fieldworkers, interviewers, transcribers), and different phases of the research (data collection, data analysis, data archive).

Table 2: The sources of proliferation for digital qualitative data

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 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.*

3. **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.*

4. **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).*

Principles of data security

An emerging literature over the past few years on digital data security for researchers is dominated by the technical contributions of data archive institutions such as ESDS (the Economic and Social Data Service) and the UK Data Archive in Britain. The recommendations of ESDS suggest that data security should be addressed from the point of view of both IT systems and physical security (see Table 3), and are relevant to any kind of data stored digitally on a computer.

Table 3: ESDS recommendations for data security

Network security

- ensure restricted access to files
- confidential data should not be stored on servers that host internet services (web or email)
- especially sensitive material should be stored on computers that are not connected to a network

Upgrades and patches

- apply all relevant security-related upgrades and patches to operating systems and applications as quickly as possible

Physical security of systems

- locking rooms when staff are absent, limiting access to rooms where computers or media are held to a few individuals, logging computer media or hardcopy material that are removed from store rooms, recording who holds keys, etc.

Viruses

- all project computers should have regularly updated virus detection software

Source: http://www.esds.ac.uk/aandp/about/preservecollect.asp

Although all these ESDS security recommendations are useful (and see other similar approaches in the emerging literature on digital data security, e.g. RELU Data Support Services 2006), even following all of these recommendations would not have prevented the kind of data loss we experienced. Indeed, most of this literature does not specifically address the particular issues of relevance to qualitative researchers (although for a discussion of some of these, see Corti 2008). These contributions generally function to provide researchers with technical solutions to data security problems, and often do not take into account the 'lived experience' of the research process and its management. Our approach has been to develop a set of concrete strategies and procedures that researchers embarking on their research can employ. Our resulting recommendations are grounded in how we contended with the challenges of maintaining good data security practices in a team-working context, and so we pay particular attention to the problems for data security posed for researchers working in teams. In addition to discussing what we found that 'worked' for us, we discuss the procedures we had originally employed that we discovered *not* to work. It is ultimately only after a security breach that procedures are put to the test; moreover, the initial procedures we employed certainly looked good on paper: they passed through our University's Ethics Review Committee, and we'd have employed them again had we not experienced the loss of data we did.

Our guidelines for qualitative social researchers

We developed the following guidelines to improve the security of digital qualitative data that take into account the principle of proliferation discussed above, and through learning from our successes and failures in our gang research.

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 (e.g. Windows passwords used to 'log on' to a PC). Use 'good' passwords that: combine letters and numbers, combine uppercase and lowercase characters, and are sufficiently long (at least eight characters, but more is better). Use different passwords for different purposes. Do not allow your computer to 'remember' a password for you.
2. Use encryption software to create 'encrypted space' on computers and storage media (like memory sticks) for all research data. A number of different approaches to encryption of text-based data are available. Use of encryption software means that, in the event of loss or theft of a device, it becomes extremely difficult for someone to access the encrypted files. Encryption is not permanent, and only exists as long as files are held within encrypted areas on devices. This latter point is important for qualitative researchers who require data, for example in the form of text transcriptions of interviews, to be accessible outside, as well as inside, of the encrypted environment (e.g., for analysis or to be archived).
3. Insist that all people who will be dealing with your research data, including those not directly employed by your institution, use security procedures at least as good as the ones you employ. Transcription work is sometimes contracted out to external companies. Do not assume their security procedures are sufficient: ask them to demonstrate their procedures, and ask if they are willing to adopt your procedures for the work they do for you.
4. Consider creating a suitably secure online or networked location for the storage of data files. The use of shared networks/servers, 'Virtual Private Networks (VPNs), carry with them different security risks to those posed by storage on local computers and portable media, but simultaneously address the problem of 'proliferation' of digital data by reducing or eliminating the need for security-problematic multiple copies of data. VPNs also facilitate the sharing of data important to team working by allowing remote access, again reducing the problem of proliferation.
5. Make back-ups of data from encrypted computer disks only to portable encrypted media (e.g. memory stick), and store/carry these separately. Where the only two copies of data (e.g., laptop and memory stick) are being transported together (e.g., digital coding work being transported daily between work and home) the laptop and memory stick should be carried separately (i.e., not all in one bag) in case of loss or theft. Even though, in the event of loss, data stored on the encrypted portions of these devices may not be easily recovered by others, the loss of work (for example, during the coding process) will be detrimental to the research project.
6. Employ the style of field note taking (on paper, or digitally via voice or text files) that is most suitable to the research and that suits personal preferences; however, delete and permanently destroy notes held temporarily in any medium that is not encrypted (paper notes, voice recording) as soon as

possible, and in preference for more permanent encrypted storage away from the field setting.

7. Simply deleting files from a computer's hard disk does not remove them permanently. Employ methods to delete files permanently and completely so that they cannot be recovered, and to 'clean' the spaces on the disk from where files have been deleted. Various software products are available to accomplish this (e.g., Steganos Privacy Suite). This is important because computers can often be sold on or recycled within research institutions.
8. Employ on-screen methods for reading and analysing data if possible. However, even researchers adopting primarily digital methods sometimes prefer not to carry out some tasks on screen, such as reading transcripts, comparing documents, or in meetings. When paper copies of data are produced, these should be shredded (using a cross-cutting shredder) immediately after use, in preference to storing them.
9. Former employees may retain possession or access to data. Put agreements in place with those employed to work with data (e.g., fieldworkers, transcribers) requiring these individuals to return data and media on which they are stored when they no longer require access, alongside 'good practice' activities such as 'deep cleaning' hard disks that remain in their possession 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.
10. Early anonymisation of interview transcripts and fieldwork notes is good practice; if anonymised transcripts are lost or stolen, problems in relation to confidentiality are minimised. Each name, place, and organisation should be replaced with *unique* identifiers (e.g., a pseudonym), rather than anonymous placeholders (e.g., 'Person Name'). Without this, ongoing analysis is compromised by the loss of meaning as the links between individual names are severed. However, 'unique identifier' anonymisation may be impossible at an early stage. In our research, interviewees spent considerable time talking about other known individuals, eventually numbering in their hundreds, and working out 'who was who', particularly given the proliferation of nick-names and street names, was difficult at the start – indeed, the process was never complete.
11. The problems of data security continue beyond data collection and analysis into the final stage of data archive. If interview data pose confidentiality problems that prevent them being publicly archived (for example, extensive anonymisation can sometimes result in impoverished or misleading data), consider putting in place the possibility for individually negotiated agreements allowing other researchers to access your data, in order to facilitate secondary analysis by other researchers and comparative research.
12. Digital text files (e.g., Word documents containing transcriptions of interviews) will be transformed during the process of anonymisation as identifying names and places are removed. Previous versions of these documents can, however, be 'recovered' because of hidden code contained in the document, thus allowing, for example, earlier non-anonymised versions to be reconstructed. This can be prevented by copying the text of the final version of the transcribed interview and pasting into a new document as 'text only', a format unable to store hidden code. We discovered that our first

> strategy of electronically 'blacking out' text did not destroy the text, it just made the text and the background colour the same; the original words were still available. There are also dedicated 'redaction' (data hiding) tools available, such as in Microsoft Word.

Our journey to the brink of the public archive

ESDS distinguishes between 'data management' and 'digital preservation', the former occurring within the life of the research project, and the latter, carried out by a public archive when research has been completed (ESDS 2007). Our experience with the *Youth Gangs* research suggests that the process of preparation for digital preservation that needs to be carried out by researchers themselves (prior to work undertaken by public archivists) can involve considerable time and resources. We had not anticipated exactly how complex the process of attempting to anonymise our data would be, and although our original research proposal involved time set aside for this, we discovered that this was not nearly enough. Our agreement with our funders (the ESRC) required us to prepare a sample of our data for presentation to the UK Data Archive. The preparation of one 40-page transcribed interview took a full day's work, and so completion of all our interviews would have taken about four months. We had only set aside a few weeks for this activity.

We first replaced the names of people, organisations, neighbourhoods and areas, gang names, the city in which the research took place, all with suitable placeholders (e.g. '[Gang Name 6]' or '[Person Name]'). Our strategy here was to retain as much specificity as possible, whilst still retaining the requirement to preserve anonymity, but doing all this in a way that was practically feasible. For us, however, it was difficult – indeed impossible – to anonymise early in the data collection in our research (see Recommendation 10 above). 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 (e.g., was 'Thomas' the same person as 'Tom', as 'Tommy', as 'Tommo', as 'TJ'?). 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 that was feasible was 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. After replacing all names with [Person Name], we were able to take a more differentiated strategy in replacing gang names, neighbourhoods, areas and organisations, although overlapping structures and substructures for gang names and organisations made even this problematic. This process of replacing names of people and places with pseudonyms and placeholders was a tedious task, and fairly time consuming – but nowhere near as time-consuming as the remaining work involved in anonymising the interview transcripts – and it is to this problem that we now turn.

Events described in our interviews included notorious and serious events, including unsolved homicides and other crimes. Many of these events are well known because of reporting in the media, or because of the grapevine for unreported events in very tight communities. One individual, for example, spends a large proportion of his interview describing the events that led up to his arrest and lengthy imprisonment for drug dealing. These actually quite singular events will be well known in Research City (the term we use in all public and published references to our research). Another individual talks about the implications for her of being the girlfriend of someone suspected of being the responsible (although yet to be charged) for the death of a very high profile victim. Falsifying demographic details in the transcripts could in these instances result in inappropriate interpretation of the facts of this person's life – or as problematic, the misattribution of the identity of the people involved. These examples illustrate (as best we can without betraying confidence) the problem with some transcripts – we'd be forced simply to remove much of what was said, leaving comparatively trivial information, and not in its context. We found that 'what was left' amounted mostly to non-specific generalities, and – even more worryingly – that this kind of talk often contradicted more detailed reporting about specific events elsewhere in the transcript. 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. Moreover, this whole process was incredibly time-consuming, involving not only making decisions about the notoriety of events being described, but discussion of them amongst the research team, alongside occasional consultation with key actors in the field.

For these reasons, and after consultation and advice from the UK Data Archive, it was considered inappropriate for our research data to be publicly archived. Nevertheless, we understand the importance of allowing access by other researchers to our data in order to facilitate both secondary analysis and comparative research. In fact, there are precedents for sharing data whereby specific use agreements can be individually negotiated with interested researchers, and whereby we can still play the important role of helping those writing about our data 'draw the line' in relation to which events can be written about in publications without putting interviewees at risk. Therefore, rather than publicly archive our data, we will negotiate agreements with individual researchers and research teams for them to use our research only under our the specific circumstances set out in the agreement. Thus, our thirteenth recommendation:

13. 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. Such individually negotiated 'specific use' 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.

Archiving for whose use?

Some recent contributions to the psychoanalytic literature have relevance to debates considered here (e.g. Gabbard 2000 and Stanjer-Popovic 2001). The reporting of ‘case studies’ in the professional psychoanalytic literature has some important similarities to both the publication of the results of qualitative data analysis, and to the lodging of qualitative data in public archive. In relation to our gang research in particular, we found some of the questions and problems addressed within this psychoanalytic literature to be relevant to our anxieties around maintaining confidentiality and anonymity for our respondents where issues of public archive are concerned. Both communities (the psychoanalytic community and its patients; and ‘gang’ communities) are relatively small, and often contain individuals (patients, gang members) who are well known both within their communities, and often outside of them (‘notorious’ criminals or victims; well-known or ‘celebrity’ patients).

Gabbard (2000) discusses both solutions and problems offered by the strategies of ‘thick disguise’ and obtaining patient consent in the publication of psychoanalytic case studies. In particular, the question arises: from whom is the identity of patients to be concealed? (Is it the patient? The patient’s family? The patient’s friends? The analysts colleagues?) A similar issue arose in relation to the question of lodging our gang research data in public archive.

Who are the people likely to have access to publicly archived research data? It is convenient and reassuring to think that the academic researchers who have access to these archives in no way overlap with the world of research participants. In Manchester (the city in which our academic institution, the University of Manchester, is based) many of our own undergraduate and postgraduate students studying criminology degrees: (1) come from communities with gang and other crime problems (indeed these socially excluded communities are often targeted by ‘widening participation’ strategies); (2) are involved in similar grassroots community organisations to the ones we encountered in Research City; (3) are from the police, the probation service, and Youth Offending Teams. These kinds of students, particularly postgraduate students, are likely therefore to have access to data held in public data archives. Moreover, a gang member’s status is usually transitory (gang members ‘grow out’ of gang activity, just as most criminals mature out of crime), and a not uncommon ‘destination’ for them is involvement in community groups and youth work, and sometimes further and higher education.

Some researchers have argued that research participants may increasingly wish to be identified in research, rather than hidden (Grinyer 2002; Wiles et al 2008), but note the complexities and difficulties of doing so. We concur, and question whether obtaining the consent of research participants to make their identities public is sufficient criteria for actually doing so. We certainly encountered in our research individuals who were eager to talk to us about very sensitive issues, and who brushed away our reassurances of confidentiality as unimportant. However, the consent of a 16 year old gang member, perhaps at that age enjoying their notorious status, may be a decision that is regretted a few years later when gang activity is left behind.

Conclusions

The 'principle of proliferation' of digitally held qualitative data makes these data difficult to manage, especially in a team-working context, and raises particular issues for data security that are not solved by the 'locked filing cabinet'. Some recommendations are proposed in this paper that build on the more technical contributions provided by data archives themselves. However, these recommendations should not be seen as an exhaustive list. Data security can never be guaranteed, even using the most rigorous standards. Moreover, it was only the experience of one research project that gave rise to these. We suggest that the research methodological literature should pay attention to issues of data security more than it does, and that other researchers who develop useful data security strategies should write about these in publications. Finally, the digital world is a very fast-changing one, and new possibilities for storing and protecting data emerge regularly. Our recommendations are likely to be superseded quickly as new technologies for dealing with data arise.

The qualitative research data arising from our gang research study did not lend themselves to public archive because thorough anonymisation of interview transcripts was not possible. This was the case because of a number of features of the data, involving research participants in talking regularly about others in relatively small communities, and about events that were widely known in those communities and even outside of them. For our data, thorough anonymisation not only involved the stripping away of identifying names and places, but also redacting the text of interview transcriptions where events were described that were likely to make individuals identifiable, even without identifying names being present. We suggest that the extent to which data that remain after anonymisation are significantly impoverished or misleading can be used as one criteria in decision making about whether data from a research project are suitable for public archive. Another criteria concerns the extent to which populations of academic researchers and the communities we research overlap.

References

Corti, L. (2008) Data Security. In (L. M. Given, Ed.) *The SAGE Encyclopedia of Qualitative Research*. London: Sage.

ESDS (2007) *Data Management – Preservation*. Economic and Social Data Service. [http://www.esds.ac.uk/aandp/about/preservecollect.asp]

Gabbard, Glen O. (2000) 'Disguise or consent: problems and recommendations concerning the publication and presentation of clinical material' *The International Journal of Psychoanalysis*, 81: 1071-1086.

Grinyer, Anne (2002) 'The anonymity of research participants: assumptions, ethics and practicalities' *Social Research Update*, 36: Spring.

RELU Data Support Services (2006) Guidance on Data Management. Colchester: UK Data Archive, University of Essex.

Stanjer-Popovic, Tamara (2001) ‘Disguise or consent: problems and recommendations concerning the publication and presentation of clinical material: Glen O. Gabbard and the Editorial by David Tuckett’ *The International Journal of Psychoanalysis*, 82: 415-424.

Wiles, Rose, Crow, Graham, Health, Sue and Charles, Vikki (2008) ‘The management of confidentiality and anonymity in social research’, *International Journal of Social Research Methodology*, 1-12.