

Digitising and Providing Access to Socio-Medical Case Records: The Case of George Brown's Works

Louise Corti & Nadeem Ahmad

Key words: socio-medical research, digitisation, data access, archiving, George Brown, depression, qualitative data, secondary analysis, Qualidata

Abstract: In March 1999, Qualidata secured a grant from George BROWN, via the Medical Research Council to undertake the preparation, scanning, documentation and archiving of data and documentation from the lifetime research of Professor George BROWN. The research, comprising a very large resource of both qualitative and quantitative materials, dates back to the 1960s and is considered, internationally, to be among the most pioneering work on socio-economic aspects of mental health and the social aetiology of depression. It has already been in regular use for many years, both for ongoing research and for teaching purposes. In this paper we will describe: the logistics of conducting and managing the huge task of scanning the paper-based materials; tying this in with the quantitative data; and documenting the data in order to make them a useful resource for researchers. Because the data are very sensitive we also mention ways of enabling the most appropriate access.

Table of Contents

- [1. Introduction](#)
- [2. Origin of the Digitisation Project](#)
- [3. Nature of the Data](#)
- [4. Digitising the Collection](#)
 - [4.1 Which format for images?](#)
 - [4.2 Problems with digitisation](#)
- [5. Issues Regarding Access](#)
- [6. The Future: Opening up Access to Socio-medical Data](#)

[Acknowledgement](#)

[References](#)

[Authors](#)

[Citation](#)

1. Introduction

The archive of George BROWN's team's research data will include twelve collections, based on distinct projects dating from 1969 to the present. George BROWN's research dates back to the 1960s and is considered, internationally, to be among the most pioneering work on socio-economic aspects of mental health and the social aetiology of depression. The earliest and probably best known study to many social scientists and clinicians is the Camberwell Study, conducted from 1969-75 and providing the basis for the eminent book, "Social Origins of Depression", by BROWN and HARRIS (1978). The datasets we are describing in this paper are based on the role of psychosocial factors and "life events" in the onset, course and chronicity of and recovery from clinical depression, and the mediating role of self-esteem and social support. The team pioneered a data

collection instrument known as the Life Events and Difficulties Schedule (LEDS), used to record stressful experiences and significant life events. This is now an internationally acclaimed research tool and is used across the world. [1]

2. Origin of the Digitisation Project

In 1991, before Qualidata was set up, a survey was conducted which targeted qualitative researchers with Economic and Social Research Council funding. The survey aimed to find out what kind of qualitative data were out there, whether the data from key classic projects still existed, under what conditions they were stored, and whether the investigators were willing to consider sharing the materials (see CORTI [in this issue](#), and the Qualidata short description [in this issue](#)). The results revealed that many collections were not stored in suitable conditions, and generally were not available to other researchers to consult. In 1994, the data from George BROWN's notable studies of depression were identified as being based at his Medical Research Council (MRC) Unit at Royal Holloway and New Bedford College in London. At this time, George BROWN had no long-term provision for access to these data other than by his own research team. The BROWN collection comprises a very large resource of both qualitative and quantitative materials. It has already been in regular use for many years, both for ongoing research and for teaching purposes. [2]

In 1999 his unit was reduced in size and moved to St. Thomas's Hospital Medical School in London. In Autumn 1998 the MRC awarded a grant for the preservation of George BROWN's data and documentation for a number of his key projects. Commencing in March 1999, Qualidata has acted as the key operator in the preparation, scanning, documentation and archiving of these data, with three part-time data processing staff to carry this out. [3]

3. Nature of the Data

For each study, cases were identified by screening for a particular medical condition or treatment, either from local random samples or hospital/General Practitioner (GP or doctor) records. For the Camberwell Study, the sample consisted of

- a sample of 114 female psychiatric in- and out-patients living in Camberwell
- a random sample of 458 women from same community (with screening for depression)
- a small sample of male depressed psychiatric patients
- a small sample of women with depression attending general practitioners. [4]

The respondents were interviewed face-to-face up to three times over the course of 5 years. The methods used to collect the data included in-depth interview, clinical case note taking, observation and field notes. The interview schedules covered different facets of case history, lifestyle, life events, well-being and symptoms as well as biographical information. Symptoms of illness were

recorded using the "Present State Examination" (PSE) developed at the Maudsley Hospital, an instrument which has continued to feature throughout BROWN's work. Each individual interview schedule consists of around 250 pages of information for each interview. Responses were coded, with a mixture of pre-codes and post-hoc coding, but a substantial amount of information was annotated by hand by the interviewer at the time of interview. Furthermore, as the interviews were tape-recorded additional information was transcribed, in summary fashion by hand, onto the schedules. The richness of the resulting schedules means that the data can be used in both quantitative and qualitative ways. Whilst BROWN's team use the material in a predominantly qualitative way, by creating complex scales and establishing threshold measures, the material has great potential value for case studies. [5]

4. Digitising the Collection

When considering the qualitative, or case-study oriented, re-use of data of the kind found in BROWN's projects, it is important to ensure that researchers are able to view the data in a similar way to the investigators (i.e. keeping the paper experience). For each of the 12 studies making up the BROWN collection the paper schedules have a similar format in terms of the order in which particular sets of questions were asked. For each progressive study, changes, modifications and additions were made to the schedules, as the interviewing tools were developed. Each study has around 200-800 cases, each with several interviews yielding up 1000 pages per case. [6]

The nature and form of the raw materials leads us to consider a number of issues for digitisation. First, what digital format is suitable for preservation and is this digital format appropriate for re-use of the data. In fact, different digital formats are needed for these two distinct purposes. [7]

4.1 Which format for images?

Both the survey data and digitised hand-annotated original schedules are to be archived with the Data Archive at Essex (c). The format for preservation and access both need to be platform independent (i.e. can be used on different types of computer operating systems) and also they should be as standard and non-proprietary as possible. The [UK Data Archive](#) already has favoured file formats- Tagged Image File Format ([TIFF](#)) for document preservation and Adobe Acrobat Portable Document Format ([PDF](#)) for document dissemination. [8]

TIFF files are used because they can be read by the vast majority of graphics programs. It is a graphical equivalent of ASCII text, is platform independent and has published standards. Adobe Acrobat files (PDF) can be read on a variety of computers (Windows, Mac, Unix, etc.) and the viewing software (Acrobat Reader) is freely available from the Adobe web site.¹ Using PDF, the look of the original

1 See Adobe (<http://www.adobe.com/>) for free Acrobat Reader and Acrobat software, also available on their site are some tools to create a limited amount of free PDFs and also to convert PDFs back to text for accessibility.

paper can be preserved as they are similar in concept to a book displayed on screen. The documents can be made user-friendly with the addition of bookmarks (contents pages with headings with clickable links to pages) and annotations. Security can also be applied to the files where necessary, as well as the possibility of OCR (Optical Character Recognition) if required. [9]

It should also be noted that although Adobe created the PDF format (and distribute the free Acrobat Reader) they are not the providers of software capable of creating Acrobat files. Evaluation of several software programs led us to us the Davince suite of PDF programs (<http://www.davince.com/>), which proved to be a far more cost effective and faster method of creating PDF files. [10]

The scanning was carried out on a computer running Windows 95 and Powerscan software and using a high-powered Panasonic duplex scanner. A Plextor CD ReWriter was used for backup of files. The PDF creation, image file management and bookmarking was done on a variety of networked Windows NT computers using Adobe Acrobat, Davince tools, NameWiz (<http://www.softbytelabs.com/> for file management/[re]naming), [WinZip](#) (for archiving original TIFF files), specially written customised scripts and to a lesser extent [Adobe Photoshop](#) and Textbridge (<http://www.caere.com/products/tbpmill/>; broken link, September 2002, FQS). [11]

4.2 Problems with digitisation

First is the sheer quantity of paper to scan, an average of 250,000 - 400,000 pages per study (and there are 12 studies). As a result we require a vast amount of storage space, for scanning the images, backing them up and the network bandwidth used in transferring files from machine to machine. There is also the problem of managing millions of files, and creating an index system which enables us to track which patient and study each image file refers to. This can only be achieved by rigid naming and storage conventions and practices, as well as giving someone overall responsibility for ensuring that this is adhered to. [12]

Once these issues had been resolved we then needed to consider whether we should store the images as just images (i.e. a graphical picture of the paper) or whether we should store them as images with searchable text (i.e. should we use OCR). Due to the complex nature of the qualitative data—printed paper questionnaires and schedules with typed and hand-written comments—it was decided, after experimentation, that the data were not suitable for OCR. Even with new advanced OCR software around today it is extremely hard to translate hand-writing into machine-readable text. Furthermore, using a powerful computer it took on average three hours to OCR a basic PDF file (prior to it being bookmarked). By skipping the OCR stage it took on average 45 seconds to create the PDF file, clearly a considerable saving in time. The main disadvantage of not OCR'ing the paper schedules is that it is not possible to search for keywords of interest. Problem was overcome by scanning and OCR'ing a blank questionnaire. Thus it is possible to search only on keywords in the pre-printed part of the questionnaire. [13]

Further problems encountered for scanning in the format of the typed paper schedules were:

- the age of the paper—over the years (some over 25 years old) the paper had deteriorated;
- the paper colours used to differentiate between schedules—on average a set of schedules for one case in a study would have a mixture of single-sided and doubled-sided paper in around 15 different colours of paper;
- the hand-annotation onto the schedules—in some cases there was a huge contrast between the typed questions and the hand written answers, for example, red pen on red paper is hard to read at the best of times. [14]

These factors hindered the scanning process considerably as it was not possible to just feed the scanner batches of paper automatically—a lot of manual adjustments to the scanning software were necessary. [15]

5. Issues Regarding Access

In terms of providing access to these datasets, careful planning and negotiation is a prime requirement. For the Camberwell study, for example, there are a number of complex issues to be clarified. First, the investigating team did not gain permission to archive the data—indeed in the 1960s such projects were not required to gain permission for the study from a health services area ethics committee, which today is de rigour. Fortunately, because the team had planned a progressive follow-up study, they had sought permission to return for follow-up interviews. However, for the cases sampled via hospital or GP records, it would be illegal not to gain specific permission from the relevant health authority for archiving. [16]

After discussing the issues of access at length with a number of key players, we have arrived at the following solutions:

- anonymisation of key identifying information;
- permission for academic and clinical re-use through the local health authority Ethics Committee;
- checking on the bona fide research status of potential users by George BROWN's team, and in the future, the Institute of Psychiatry;
- adherence to the MRC Guidelines on "Personal Information in Medical Research" (see below);
- additional checking on the bona fide research status of potential users by a Data Archive nominee, where necessary. [17]

6. The Future: Opening up Access to Socio-medical Data

For Qualidata, this project presented an excellent opportunity to form stronger link with funders of socio-medical research in the UK. Previously the Medical Research Council (MRC) had no active interest in, or policy for, archival provision for data from research they sponsor. This initiative has helped break the mould, and the MRC has recently set up a feasibility study to establish the viability of long-term archival provision for a range of medical data sources (from clinical trials to epidemiological longitudinal studies). An excellent document entitled, "Personal Information in Medical Research" (MRC 2000, <http://www.mrc.ac.uk/PDFs/PIMR.pdf> -- the document is no longer available online; FQS, Jan. 2002), has been prepared by the MRC to provide the ethical and legal framework for archiving. [18]

Finally, this project has provided an in-depth feasibility study for the possibility of digitising large qualitative data collections. The overriding messages coming out of this project is that:

- the process is very labour intensive and therefore costly
- a good document management system is required, and
- a range of technical options are explored at the start of the project
- the project manager must liaise closely with the investigators [19]

Acknowledgement

We would also like to thank George BROWN and Tirril HARRIS for allowing Qualidata to take part in preserving this prestigious and historical material. Our thanks also go to Jane CADOGAN who devoted a year and a half to preparing and scanning documents for this project. This was a huge task, but one which did not daunt Jane. As a result, the superior quality and resolution of the images are a reflection of her expertise.

References

Corti, L. (2000). Progress and Problems of Preserving and Providing Access to Qualitative Data for Social Research-The International Picture of an Emerging Culture [58 paragraphs]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* [On-line Journal], 1(3), Art. 2. Available at: <http://qualitative-research.net/fqs-texte/3-00/3-00corti-e.htm>.

Brown, G. & Harris, T. (1978). *Social Origins of Depression: A Study of Psychiatric Disorder in Women*. London: Tavistock Publications; New York: Free Press.

Authors

Louise CORTI is currently the Deputy Director and Manager of Qualidata, the ESRC Qualitative Data Archival Resource Centre, based at Essex. In January 2001 she will be taking up the post of Director of User Services of the UK Data Archive, where alongside the duties of that role, she will retain an overall responsibility for qualitative data archives. In the past she has taught sociology, social research methods and statistics, and spent six years working on the design, implementation and analysis of the British Household Panel Study at the University of Essex. She is interested in both qualitative and quantitative aspects of social research.

Contact:

Louise Corti
Qualidata
University of Essex
Colchester CO4 3SQ
UK

Phone: + 44 - 1206 873058

E-mail: corti@essex.ac.uk

URL: <http://www.essex.ac.uk/qualidata/>

Nadeem AHMAD works at the UK Data Archive and Qualidata both based at the University of Essex, where he has taken a key technical role in the George Brown project. His current research interests lie in the digitisation of paper resources, the work of George Brown and children's rights (in fostering and adoption). He has previously co-authored the publications "The UK Record Industry" and "Rock Accounts" based on annual surveys.

Contact:

Nadeem Ahmad
Qualidata
University of Essex
Colchester CO4 3SQ
UK

Phone: + 44 - 1206 873058

E-mail: nadeem@essex.ac.uk

URL: <http://www.essex.ac.uk/qualidata/>

Citation

Corti, Louise & Ahmad, Nadeem (2000). Digitising and Providing Access to Social-Medical Case Records: The Case of George Brown's Works [19 paragraphs]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 1(3), Art. 6, <http://nbn-resolving.de/urn:nbn:de:0114-fqs000368>.

Revised 7/2008