ESRC Green Paper on Data Policy and Archives 2001

Extracts from memoranda sent by John F Hall to Martin Boddy

Subject: ESRC Green Paper on Data Policy and Archiving

First submission 9 March 2001

My credentials:

From 1965 to 1992 my entire career was spent in empirical social research, both as practitioner and teacher, particularly in the design, computer processing, statistical analysis, reporting and documentation of questionnaire surveys. Long before SPSS, I wrote a suite of computer programs in KDF9Algol (???!!!) which were used at Salford Univ right up to 1979. A Lectureship at Birmingham University (1968-70) was followed by a Reader level Senior Research Fellowship at the then SSRC Survey Unit under the late Dr Mark Abrams (1970-76) and a Principal Lectureship in Sociology at the then Polytechnic of North London (1976-92). My Survey Research Unit was established there in view of the external funding and consultancy I was bringing in over and above a full teaching programme. I have extensive and varied publications in the fields of social science, computing and statistics (rare in the UK, but common in the USA), and have a long list of surveys designed, documented, analysed and reported by me or under my supervision, particularly from 1970 onwards. I have dozens of satisfied clients. and. hundreds of grateful students, especially the ones who made a complete balls-up of their computer files. I and my team of researchers specialised at times in rescue jobs, value for money and documentation. Given our practical experience in fieldwork and programming we were particularly adept at user-documentation and SPSS. For the record, I was privately offered the Directorship of the Data Archive before Ivor Crewe took over, but for family reasons was at the time unable to move to Essex. I was particularly pleased when Denise Lievesley became Director (in my opinion the first with anything like sufficient entrepreneurial and technical expertise for the job), but by then I had already decided to take early retirement.

Your Consultation Paper

I shall be sending under separate cover (unless I succeed in transferring and sending files from another desktop (all written in WordStar4) various examples of documentation and teaching materials developed by me over 20 years or so at SSRC Summer Schools and subsequently on part-time evening courses at the Polytechnic of North London (and eventually incorporated into 2nd yr undergraduate courses prior to 3rd yr placements). My aim in everything was to prove that the decision to close the SSRC Survey Unit in 1976 was short-sighted and premature, but above all to demonstrate that it was possible, never mind necessary, to include practical field and laboratory based training in undergraduate courses in Sociology and related subjects in the UK and not to leave it to postgraduate level.

The reasons adduced by students on my evening courses more than bear testament to the totally inadequate training offered then, and to some extent even today, at undergraduate and even postgraduate levels. My graduates on the BA and BSc Applied Social Studies degrees were getting jobs and funded post-graduate studentships ahead of Masters and Doctors from prestigious universities because they could actually do the things others could only read about or criticise. Indeed, I had two second year undergraduates who helped me out as teaching assistants in 1990 and 1991 who were terrified of the qualifications of the postgrads on the course, but the postgrads were equally impressed by their technical and integrated theoretical skills.

The underlying philosophy of my courses was always geared to the logic of survey analysis, the use of data sets from large general population surveys and hence the use of adequate documentation, not to mention technical laboratory skills (in our case SPSS on a Vax mainframe) before any student was allowed loose on the general public. (Practical fieldwork skills were taught separately, but this is not within your remit.) The final version before I retired in March 1992 had refined the course to take account of the huge change in initial skills (when they started in 1976 many students could hardly even type: when I left in1992 virtually all students had PCs), but had an assessment element which required not just technical skill in data retrieval, management and analysis, but also genuine secondary analysis requiring an account of theory applied to or derived from the selected data-set (usually, but not always, the British Social Attitudes series).

ESRC Paper para by para

1 Why archive?

At the risk of repetition: No-one should be allowed ESRC funding for primary data collection unless they have demonstrated competence in the secondary analysis of previous data sets (even in an unrelated area). This should apply even where there is no relevant previous data (which I doubt). In the event that ESRC funding is forthcoming, part should be withheld until adequate arrangements have been made for documentation and archiving (I think this is now a condition. but it was not always so). One of our first cases at SSRC Survey Unit was to review a request for a very large amount of extra funding for a project where there was no discernible survey or computer expertise: we advised against. The grant was awarded anyway, yet it was ten years before the book came out and always the excuse was "computer problems". Had the Unit been asked to prepare and document the data set, the excuse would have been exposed as transparent. The book may well still have taken ten years, but at least the data set would have been in the public domain. (Perhaps someone else would have written it quicker!)

2 What to archive?

At SSRC and PNL I generated a large number of survey data sets, most of which were deposited at Essex, but some of which were queried or refused. The problem seemed to centre on small local studies in the local govt or voluntary sectors, or in my case, some student projects (overlap here anyway given our course structure and community involvement) Not sure how this should be solved as criteria could be policy area or geographically organised, but some quite interesting and original surveys conducted to professional standards were done by both units and by students which may often be unique. They have to be kept somewhere, but who will know what is in them and how?

For example, I did a survey in St Paul's Girls' School whilst with the SSRC which included elements from Himmelweit's study of political awareness and a scale from the Inst of Psychiatry: parts of this were replicated in a student project at PNL with all fifth formers at a North London comprehensive school. Both studies also replicated parts of the Quality of Life surveys we did in collaboration with ISR at Ann Arbor. I still have box full of tapes and wallet folders of documentation, but who will keep them? At least at PNL I developed or at least insisted on questionnaires and documentation which could be directly related to the data set even if there was no proper user manual. This system was eventually followed by SCPR (now NATCEN) for the BSA series.

Coupled with my (SPSS) variable nomenclature standards this enabled most surveys to be followed up even if the only documentation available was the original questionnaire (where data column positions are indicated on the actual questionnaire). For their own reasons SCPR adopted a system for (SPSS) variable names tied to longitudinal analysis, but this was easily overcome by writing a renaming program to convert SCPR names to positional names when using only one of their annual surveys. (Graham Farrant of NATCEN was one of my undergraduate students and wrote such a RENAME programme for the 1989 data)

There should be earmarked funding for genuine secondary analysis, possibly in the form of scholarships or bursaries tenable at known centres of excellence or under approved independent supervisors. Another possibility would be as a compulsory element in postgraduate training courses (shouldn't be necessary if undergraduate courses were good enough)

I sometimes wonder whether unprepared and undocumented data sets should be refused deposit altogether - hopefully likely culprits would be refused funding anyway. There really is no excuse these days for data of the poor quality the archive had to deal with in its early days. Ideally deposit should be of specified system files with full documentation, including original facsimile questionnaire and interviewer briefing notes. In my case these would normally be SPSS, but BMDP, Quantum, Osiris or other proprietary package capable of being read by these should be de rigeur. Compliance with such standards should be a requirement for ESRC funding and encouraged for other public funding. There are plenty of examples of machine-readable data-sets with reasonable or very good documentation for would be applicants to follow (eg British Attitude Survey and my own user manuals for the First British Crime Survey and the SSRC Quality of Life Survey, all available fromthe Archive at Essex)

3 Future needs

Earmarked funding should be available for training and studentships in approved centres or under approved independent supervision.

We should concentrate on variables rather than questions (question wording is a poor indicator of the sociologically interesting underlying variables which are the subject of the original research). For example, altruism, or lack of it, is interesting, but there are unlikely to be any survey questions with the root altruis.... in them. Many survey reports include compound or derived variables without explaining how these were derived. Documentation should be clear on this as well, including the original (SPSS) source program which generated them.

Need for commissioned research to produce manuals along the lines of Robinson and Shaver in the US with examples of attitude scales developed and tested on general population samples, with frequency counts, derived variables and statistical properties, for replication in future research. (I think a start was made on something like this by Bridget Taylor, working on BSA data, but I wrote to Roger Jowell encouraging this well over ten years ago)

4 Priorities

Problem here is to ensure none-loss of important data. ESRC can insist for ESRC funded research, some commercial research is routinely deposited (eg National Readership Survey, NOP and Gallup political opinion surveys. Years ago Liz Nelson offered to discuss with Ivor Crewe deposit of the Taylor Nelson surveys which included batteries of personality items, but minus any commercially sensitive brand information. Nothing seems to have come of this except that TNS-BMRB is now a listed company.

5 Balance

ESRC should continue to require deposit. I would go so far as to claim ESRC ownership of all data to be placed in the public domain with unrestricted access the day (well, maybe a year and a day!) after expiry of the grant. If the grant-holders can't get their act together in the time they estimated, then tough. Let someone else do the work who knows what they are doing.

6 Data currently missing

See above re Taylor Nelson monitoring surveys, but there are surely other similar sources of data from large scale general population surveys which need to be safeguarded. Need to discuss this with commercial companies or their representatives in MRS and AMRO

7 Europe

No comment here, but perhaps my variable-centred approach would be more productive than a question-centred approach. I doubt if anyone in the UK has even looked at the stuff I did with Dave Phillips and Steve Harding on European Values in spite of two books and various papers. Again, ESRC might think about earmarked funding for comparative research requiring use of European (or come to think of it USA and other) data. Roger Jowell started doing something like this with items in the BSA surveys. but I've yet to see any serious analysis and reporting.

8 Architecture

Much of this is covered by comments above, but I would emphasise again the necessity of top quality machine-readable data-sets accompanied by equally top quality documentation. Although it should not be necessary to check that data have been edited and cleaned, it is sadly the case that many data sets contain errors or omissions that the original depositors have failed to spot, sometimes intentionally, but usually through incompetence. I once found a data set which had been the subject of a report, but which had a complete line of data duplicated, thus throwing out every subsequent line of data by one line. The researcher had not noticed this. In another case I found that a data set which supposedly contained a double sampling proportion of young people aged 18-34 had actually had the data for each such respondent duplicated. When I pointed this out to the client, the managing director of the company which did the fieldwork was furious. When preparing the User Manual for the First British Crime Survey I even found a code in a single case which was not in SCPR's Coding Manual: fortunately Colin Airey remembered the case and the reason for the code, otherwise their data set was perfect!

Models of documentation as examples of best practice would include all my stuff from SSRC Survey Unit on the Quality of Life surveys (improving as the years pass and the software becomes available) and the stuff from PNL Survey Research Unit (with the exception of the Quality of Life of the Elderly in Residential Care which I did not prepare) on Quality of Life in Urban Britain, the First British Crime Survey 1982 and SCPR's technical manuals for the British Attitudes series. I found the latter a little too complicated for teaching use as parts of them were unnecessarily duplicated, but they are easily available. In fact with my SPSS variable naming conventions all you really need is the questionnaire as it has line and column positions marked for the data. Without exception all my student and other surveys have a questionnaire, coding frame and SPSS source files and frequency counts as the bare minimum documentation to go with the data and system files. All my staff and students were trained to think that someone must be able to understand what they have done and to carry on where they left off. With very few exceptions, this is what they did and why they have perhaps been more successful than others in their subsequent careers.

As data sets explode, clearly some thought needs to be given to satellite archives at local authority or local university level. There must be hundreds of local government surveys for planning and other purposes which have not been archived, but for which the data sets and documentation may well remain intact. Another criterion might be subject-based (eg Social Services or Housing or where there is a large resident academic user-base) but quite how this could be organised and funded is another question.

11 Charging

If the job has been done properly in the first place there should be no need for any additional costs for preparation and documentation.

General comments (as if the above weren't already enough!)

If I could emphasise anything out of all the above, it would be on the importance of first hand training in data handling and data analysis, preferably apprenticed to, or taught by, genuine practitioners, as I think I successfully implemented at undergraduate and postgraduate level at the Polytechnic of North London. It is gratifying to see the names of my ex-students cropping up in professional publications and newsletters and emerging as practitioners after my own heart. We need a new vocabulary for such people, equivalent to literate and numerate (dataphiles, data-ate?) who demonstrate integration of data skills with theoretical and professional skills.

Documentation is crucial for any analysis, but especially for secondary analysis.

Genuine secondary analysis is rare in survey research in the UK. This is why I instigated the Mark Abrams Prize with the remaining funds of the Quantitative Sociology Newsletter and which is now run by the Social Research Association. It was to be awarded annually for the best paper linking social theory and/or social policy with survey research, particularly if this resulted from secondary analysis. Since the SRA took it over it has been usurped by the qualitative researchers, not at all what Mark intended!

There was an attempt by SSRC Survey Unit to encourage secondary analysis via a series of seminars chaired by the late Prof Morris Janowitz. Proceedings were edited by Tim Leggatt and published, but most of the papers were from primary analysis. This is an area which ESRC could encourage by earmarked funds for studentships and bursaries for genuine secondary analysis developing or applying social theory to classic data sets (eg Affluent Worker or Symmetrical Family), perhaps challenging original findings, or by using new statistical techniques (eg LISREL) to search for patterns and structure not originally apparent. Had I continued at PNL (now UNL) this would have been the next step in development of our postgraduate training programme.

You will judge from variations in typeface and the enormous length of lines that I am not yet fully conversant with Windows and Outlook Express. I think it safer to send documents by post rather than attempt to copy or attach them to this.

These comprise:

List of data sets available for teaching and research at PNL
List of research clients
Course outline for SR501 (Survey Analysis Workshop - Postgrad part-time evening)
List of students on SRT501 for 1990 and 1991 intakes
List of publications arising from my contracts at PNL
List of survey-based reports by PNL students

If this looks a lot, it gives some flavour of what I was trying to achieve, and indeed did achieve, for British social research. The mantle has now shifted from UNL to places like Surrey and Southampton or to NATCEN and SRA events, but there is still nothing quite like what we offered at PNL or for such low fees. However I eventually became so tired of being baulked in my research efforts and continuous harassment by senior management of PNL and the Faculty that I ensured I had a full teaching and administrative programme for my final year and took the first chance that came my way to retire early. Having had my "garden" destroyed twice by politically motivated hooligans, I now restrict myself to a garden they can't come near.

I would be interested to see what becomes of the Green Paper and would be grateful to be kept abreast of events. I would be happy to be of service to any future developments provided my expenses were met. Please convey my regards to senior colleagues at ESRC who may remember me from the good old days.

Follow-up submission 14 March 2001

I have today sent by land post the papers I promised, plus an abstract from my publications listing various papers on computing, data-processing, statistical computing and user-documentation. There are also one or two reports based on secondary analysis, but I'm not sure if I can still lay my hands on some of them.

I refer to, but do not include, the booklets and worksheets relating to my post-graduate evening course Survey Analysis Workshop. These were developed over many years, starting with the SSRC Summer Schools in Survey Methods which we used to run at Oxford and latterly at Reading. When the SSRC closed the Survey Unit in 1976 I continued the course at PNL in the form of a part-time evening course. I taught the computing side and John Utting the statistics. When we started in October 1976, we were sending hand-written sheets over to computing for punching up and running, so the students didn't get their results until a week later. (It was incidentally the very first evening course offered from the Ladbroke House site: we had to buy and make our own coffee etc. You can check with Peter Glasner at UWE what the place was like: it was his evening B.Sc. Sociology which forced PNL to open a coffee and sandwich bar in the evenings several years later) When I left we had our own computer lab with 16 terminals with four very fast servers to the Vax, two local line-printers and a very user-friendly interface to SPSS written by Jim Ring (SPSS didn't delete its log files so students immediately ran out of space) which enabled us to run semi-interactively with helpful prompts and which allowed us to correct errors instantly in the current line without having to start all over. The effect on student motivation and progress was amazing.

Although the course looks deceptively simple, it did in fact enable complete beginners to progress rapidly to quite advanced analysis and left them in very good stead for progress to more advanced courses and to work independently. Many of them got (better) jobs on the strength of their training at PNL. Whilst academia (ie "proper" universities) was loath to acknowledge this, the voluntary and not-for-profit sectors, local and central government and even commercial research firms were more than ready to recognise the value of the training received at PNL (extending to undergraduate students who received a scaled down version of the same course) which the Market Research Society accepted as a qualifying component for their Dip MRS.

I think the main advantage for students was being taught by senior practitioners. Our sister course Survey Research Practice was taught entirely by practioners from not-for-profit and commercial research organizations (eg Nick Moon, NOP; Barry Hedges, Bridget Taylor, Patten Smith, Jean Morton-Williams, Jane Ritchie, SCPR [now Natcen]; Jean Martin. Roger Thomas, NCS Survey Division; Alan Marsh, PSI). It may be significant that they refused to carry on after I left. I like to think that they only came to PNL on wet Tuesday evenings when Arsenal were playing at home because I was there, but I suspect they also couldn't wait to leave because of PNL's appalling management reputation. It was certainly the case when, although I was instrumental in setting up the first British Crime Survey, for which Roger Jowell and I submitted a joint bid, the Home Office gave the fieldwork contract to SCPR, but refused PNL (on political grounds from "upstairs" the data-processing and analysis contract. (There was a sublime irony when Mike Hough later came cap-in-hand to ask me to prepare the user documentation!) Professional and disciplinary jealousies later intervened to prevent PNL involvement in later waves, but by then I was past caring.

Another factor, which may have implications for ESRC policy, was that most of my colleagues elsewhere were specialists in sociology, statistics, computing or survey research, but rarely in more than one of these. Things may have improved since my time, but then it was rare. There were a bunch of people from the old Quantitative Sociology Newsletter, Study Group on Computers in Survey Analysis (now Association for Survey Computing) and UK SPSS Users' Group plus odd centres at Surrey, Southampton or LSE, but they rarely covered more than two of these specialties. Exceptions were Peter Halfpenny at Manchester (a convert from Chemistry) the late Cathie Marsh (to whom I gave her first job) Colin Brown at PSI (ditto), Tony Fielding (Birmingham) and Randy Banks (not quite sure where he is now, but it was the Archive at Essex). Others may well have covered the fields and been equally at home, or at least able to hold their own, with single field specialists, but most of the ones I met had their own (careerist or political) agendas rather than a vocational disposition.

In my experience, in the UK, physical scientists share things (mainly), but social scientists seem to spend most of their time shafting each other, especially when they referee grant applications, whereas my old USA colleagues (on sabbatical at the SSRC Survey Unit or on vacation, but visitors to PNL), the late Angus Campbell (ISR Ann Arbor, Michigan), Bernard Blishen (ISBR, York, Ontario) and Jim Davis (Harvard) were invariably open and helpful. I like to think I trained my students towards the USA rather than UK model in respect of their professional practice.

To return to the topic of training and documentation: had I stayed on at PNL there were a number of developments I should have liked to see. Some were in the development of further modules in the post-graduate training courses (eg Attitude Measurement), but given the rapid expansion of the Internet and remote learning resources, I always hoped to make my teaching materials available to others, not simply on paper, but also on-line. Although they were developed for use with a mainframe, I had always hoped to develop them for the PC version. (We did once try SPSSPC, but students don't seem to be able to keep their fingers off the keyboard and in next to no time they were all in different parts of SPSS with no hope of return to base and, being a new implementation with the usual complete lack of staff retraining time, no-one to help them get back either!)

At that time the PC version had severe limitations and the commands were not standard between the mainframe and PC versions, but what I did hope to do was to have my teaching notes in some kind of file that was not specifically geared to SPSS, but would have the same underlying logic and structure so that worksheets could be substituted for alternative software (eg BMDP, Quantum) and data sets through the computer equivalent of interleaving or loose-leaf files. I don't know if anyone has subsequently tried this, but I would be more than willing to be involved in the attempt and to make my materials available. subject to copyright protection.

Most of them are in WordStar4, but my brother-in-law Raymond McDowell, Lecturer in Marketing at UWE Business School r.mcdowell@uwe.ac.uk has managed to get some of them into MSWord format. Parts of them were facsimile copies of bits of questionnaires or of raw data files, but I'm sure it would not be an insuperable problem to get them into manageable format for PC use, although it might be better to use more recent data than the 1989 BSA series. Do not underestimate the time needed to prepare a major survey data set with associated documentation and to find appropriate and interesting examples to illustrate substantive and statistical properties: I remember one of my evening sessions (3 hours class contact) took 19 hours of preparation when the Dean's draconian formula allowed only half an hour.

I should be grateful to receive any feedback or observations in due course.