

## CAQDAS, Secondary Analysis and the Coding of Survey Data

### Nigel Fielding

First off I want to apologize for my presumption. You will see just underneath my name there that it says Institute of Social Research. Bizarrely enough around 1990 my department established an Institute of Social Research, just blatantly ripping off the name of the august institute here at Ann Arbor. Because I am now the Dean for Research and we are undergoing the stripping out of redundant titles and entities I have abolished the Institute of Social Research so the one here in Michigan can revel in its name all on its own. We are simply going to revert to the Department of Sociology at the University of Surrey.

Here is the outline of the material I am going to cover. I am going to begin by characterizing the open-ended question problem. Apologies if that is massively redundant which it probably is. I want to talk in terms a little bit wider than the ANES. I'm going to talk about the way qualitative researchers understand what is called the audit trail requirement in the field of qualitative software. Then I am going to talk about this term CAQDAS. It simply stands for Computer Aided Qualitative Data Analysis. You get the S in the end of the final word analysis. This really is the central term. Me and Ray coined the term called CAQDAS. We ran a conference in 1989. We sat around in my office. There was interesting literature on qualitative software, a form of software that lately appeared in the last 4-5 years and was beginning to get written up in the journals. We wanted a tag for qualitative software. There were some potted cactus plants in my office. Ray said to me that getting qualitative researchers to use the computer, to use software is a thorny issue. It came to us in a flash. I actually still have this cactus. It is quite a bit bigger now but this cactus plant still sits in my office.

I am going to outline some basic features of qualitative software. Then I am going to hook out issues of data integration about quantitative and qualitative data being worked together in the same software. Then I'm going to make some comments about secondary analysis which in fact, although I wrote them in the brief, we haven't really said very much in the proceedings about this issue. But I do want to save a little time for that. Then I will draw a kind of conclusion that brings it back to ANES.

I think what we have been hearing in the couple of days of proceedings, at the software level, is very much a move to machine based content analysis where the coding operation is largely outside of your hands. The dictionary is there, you can customize it but you are stuck with a dictionary once you have defined it. That is what you are using. In the talk that Fabrizio just gave us we have an interesting, in my mind, hybrid. A step between content analysis and qualitative software of the CAQDAS type in so far as the training staff and the user was fundamentally involved in testing and the decision rules.

With qualitative software, what this is about is very much enabling researchers by providing them with basic digital support for fundamental operation ['operations'] in qualitative analysis but there is very little automation. Things rest almost entirely with the user's control. It is very much an approach, that is distinct from content analysis from that respect, and most of the automation is reviewable. It is always very quick to get back to the data and back to the context.

Let me characterize the problem – letting respondents express themselves in their own words on certain items is a practice that is as old as survey research itself. Open-ended questions are undeniably hard to handle. They are out of step with the other responses and sometimes surveys have simply left them unanalyzed. Open response items often times involve an intermediary, the interviewer who semi codes the data when recording or transcribing it. Ultimately there is a limit like 256 characters. They have to translate what respondents are saying into a gloss which is actually what gets written down. Their sense of what is worth recording is only going to be as good as their understanding of the purposes of the project and the kind of information it is going to want to code. The same applies when the staff reads transcripts against lists of coding categories to match data segments to categories. Decision criteria may remain implicit, and even where conventions have been adopted as we have heard they may not be archived and form part of survey documentation.

Nevertheless, we hopefully all recognize that we get very different answers to research questions according to the methods we use. I want to show you an example from one of Mike Patton's books on Using Qualitative Methods in Evaluation. This I think is a particularly good example. The term there, accountability, is what at least in British management circles is called staff development reviews. I don't know if you undergo these here but at British universities they are pretty general. Once a year you fill out a form that reports your doings over the previous year. You then have an interview around the form with your line manager. We do actually use the term line manager at British universities. At any rate, here we have the closed item at the top – accountability as practiced in our school system creates an undesirable atmosphere of anxiety among teachers. Our teacher respondent has indeed strongly agreed. Then they have the invitation to fill in any personal comments. This is actually heavily abridged compared to Patton's original book. There is about 3 times as much words as this. The extract reads "Fear is the word for accountability. Accountability is a political ploy to maintain power. The bitterness and hatred in our system is incredible. What began as noble has been destroyed. You wouldn't believe the new layers of administration to keep this monster going. The finest compliment around our state is that other school systems know what is going on and are having none of it. Lucky people. Come down and visit hell sometime."

Those responses do go in the same direction but they obviously offer different degrees of insight. Views like come visit us in hell aren't captured by strongly agree but we also have the extra dimensions which is presumably why we are all here talking about these things. Accountability is seen as being about power. That is an interesting thought. It is divisive - that points to conflict. It could be interesting to analyze. It is bureaucratic. So those who do engage with open-ended items, and I obviously think it is worth it, have sought coding solutions. These can be slippery because as the late Mathew Miles said, qualitative data are an attractive nuisance. He was a major figure in educational evaluation. He was very much known as someone that was involved in the visualization, the visual representation of qualitative data.

For qualitative researchers open-ended data are not just an attractive nuisance. But they are central to our inquiries. A couple of people mentioned as I pointed out in my abstract, all data analysis – qualitative and quantitative- is necessarily a process of data reduction. However, we might contrast qualitative and quantitative research in approaches to coding. As the organizer's position paper says, coding implicitly contains a theory of meaning. In the qualitative field it is not as coherent around what constitutes appropriate coding as the quantitative field. There can't be many quantitative researchers who doubt the appropriate outcome of coding is a variable and considerable consensus is likely on what the dimensions the variables relevant to a particular topic will display. Qualitative research lacks that consensus. Relevant dimensions more narrowly reflect theoretical orientation so given two minutes of interview data a conversation analyst may see several days work using the Jefferson notation to time pauses and describe turn taking, Q/A sequences, adjacency pairs and so on. A grounded theorist may take an hour or so on open coding and axial coding. And a narrative analyst may say 2 minutes means nothing, where are the other 10 hours. That is of course overdrawn. But the point is to discuss analysis for a given data set qualitative researchers have to start by stating their theoretical approach, the concepts that are in play and their position on epistemological matters like criteria of validity. I'm not saying that doesn't happen in survey research but in qualitative research there is a big problem of people talking past each other.

So, if qualitative research is to generate any kind of cumulative knowledge researchers have to be explicit about their starting assumptions. Getting explicit is like having a good intimate relationship. It isn't just about being charming on the first date; it is about making yourself accountable as long as the relationship lasts. Being clear about starting assumptions is helpful but there also needs to be an audit trail that lets others see how you reached your conclusions, what data you selected out and why, what refinements you made to your coding scheme, and what puzzles emerged during data analysis that was so compelling that you had to go back to the field to get more data and then accommodate that in new codes. I have emphasized that audit trail devices are necessary if qualitative research is going to be cumulative and to build on the findings of others.

These are the traditional ways that qualitative has approached that task. I'm calling it manual methods here. These are very long established procedures – field diaries and logs, coding commentaries, analytic memos. Field diaries have their roots in the anthropologist's notebook. They go back to the 19<sup>th</sup> century to the Cambridge and Manchester schools of anthropology. In the work of methodologists like Severyn Bruyn they employ systematic conventions. They record information in comparable depth between the project's different stages. Bruyn makes the distinction between objective dimensions and subjective dimensions and he requires of the researcher that on each occasion of field work that they compile an index of the objective dimensions and subjective dimensions that were applicable to that episode in the field. The subjective could include – had a terrible headache that day, not sure I was paying attention fully to my respondent. Just to give you a flavor of the kinds of things they were trying to record. It is a kind of quality control for each episode of field work.

That distinction between objective and subjective dimensions recognizes that who has recorded the data and worked with it is important. And that individuals vary between rounds of analytic work. Coding commentaries like survey research code books offer extended definition of codes and criteria for distinguishing them from adjacent codes. They can vary from just a few words about the thinking behind the code up to many pages rehearsing the elements of the final analysis such as in Grounded Theory procedures, the analytic memo. Some of which can be in effect chapters in the final analysis. That is preamble to characterize some of the characteristics of qualitative research and qualitative analysis.

Now I want to say a bit about where the software comes in. From the mid 1980's increasingly sophisticated qualitative software has been available. Most of it in fact in the first round of development initiated in the United States but there was a prominent example from Australia and two from Germany. Latterly the German ones and the Australian one seem to be come more or less the market leaders. That is kind of the background.

One thing that qualitative software has done has been to provide audit trail features including hyper linking, code maps and outputting date-stamped meta-data tracking the sequence of assignment of codes so you can recover at any time the sequence you worked in assigning codes and adjusting codes and so on. Also several packages also now include integrated coder reliability tools that provide a statistical measure of agreement between codings. But qualitative software is more than the audit trail feature. As the S in the word analysis in the acronym suggests, its major function is to provide support for analytic operations.

Here are some of the key operations that qualitative software can support: data entry and coding. I have done a bunch of user experiences research and we have found that through the mid 1990's the primary use of qualitative software was a glorified electronic filing cabinet. So the data entry support needed to be pretty good. As we got drag and drop in the general software, we got it for coding operations in CAQDAS. Increasingly sophisticated support for analytic memoing. Like the grounded theory activity I mentioned. Then the principal analytic strategies, code and retrieval which is in either single retrieval or multiple retrieval formats. Finally the use of Boolean operators to do pattern searches across the coded data.

Just to say a little bit about data entry. Data can be organized in a variety of ways in CAQDAS. One is by data type such as keeping survey open response items separate from structured interviews. An important issue about data entry has to do with the partial or verbatim transcription. Partial transcription is a problem because most of the analytic procedures rely on the assumption that you have a complete verbatim transcript of the material. That is particularly important if you are going to use the data integration feature to bring qualitative and quantitative information together. Writing analytic memos to provide a narrative relating to the data – involves anchoring a text box to the data segment and entering your commentary into the text box. In some of the better packages you can treat the text box itself as a code and manipulate that as an object along with the segments of data and the codes applicable to them.

A couple of packages – MAXqda and Atlas now also allow you to geo reference text data such as the location where the interview took place or indeed you can geo-reference anything that can be

manipulated as data in the package so you can geo reference photographs, you can geo reference digital video and so on and so forth. Interview abstracts can be tied to points on a Google Earth view or coordinates in Google Maps and be linked to visual and audio visual about the location.

Here are the three basic types of qualitative software: text retrievers, code-and-retrieve packages and theory-building software. I have also given the reference to two of the principal authorities that employ this typology and you won't be caught out by it. It is pretty much the general typology in the field.

Text retrievers – these are most like the basic or fundamental content analysis applications. Text retrievers recover data pertaining to each category where specified key words appear in the data. For example when you search for president, wherever this word appears, the software will extract it. Other character strings and combinations can be retrieved from selected or all files. Things that sound alike, mean the same, or have patterns like the alphanumeric sequences in social security records.

Analytic memos can be linked to retrievals and the specialty of text retrievers is simply retrieving the data held in large number of documents for example transcribed survey interviews. Examples include Metamorph, WordCruncher, ZylINDEX and Sonar Professional. These programs are fast, simple and basic in features. They don't go much beyond the familiar search feature in a word processor though they can cope with an enormous volume of data and retrievals are very fast.

Code and retrieve packages are a bit more interesting. They focus on dividing text into segments, attaching code to the segments and retrieving the segments by code or combinations of codes. Data can be retrieved by individual codes or groups of codes or on how codes relate to each other. For example where data coded for recession coincides with data coded fiscal security. Organizing data to known characteristics such as socio-demographic variables allows retrieval based on combinations of conceptual codes or features. For example enabling a search to recover only data where 2 particular characteristics apply but not a third. An example would be 'data from MALE respondents with LOW PAY who are NOT registered voters'.

Code and retrieve packages focus the analyst's attention on relationships between code and data. The theory building software emphasizes relationships between the codes. It can help users develop higher order classifications, analytic typologies and data representations other than those derived directly from data. Such as constructing propositions and testing their applications using hypothesis test features. They can also visually represent connections between categories showing code names or other objects like memos as nodes in a graphic display. Users can link them to other nodes by specified relationships like causes, leads to, is a kind of, and so on.

Here are some advanced features of particular packages. I'm going to say a bit about these. There is the network view feature in Atlas/di, the hypothesis test feature in the HyperRESEARCH package that comes out of Boston University. It is one of the few that is cross platform so it can be operated on a MAC. Then the system closure feature in N.Vivo that is the Australian package. Artificial Intelligence features in the package from the University of Missouri called Qualrus. I will also say a bit about data integration and you can see the names of the software that enable that.

The theory building package, Atlas/ti, lets users view linked quotations within a network that represents data relationships. You've got the network down there on the bottom left; Network View has full interactivity with other parts of the project. In this case the quotations are being derived from multimedia documents so it is manipulating some text images in the middle of the lower left window. Up to the extreme left is code that applies to them and then there are some photographs of the town being described in the data. At any rate, in Network Views it's got full interactivity with the other parts of work project. If you change a code name in the Network View the program will change the code name in all the relevant segments. So this is why I am saying that the relationships are operating at the code-to-code level and it does the dirty work of adjusting the data underneath for you when you are working with the Network View. As it shows you can use that for textural, visual and graphical source material.

I will just mention N.Vivo. I don't have a picture but it does provide another way to link categories using relationship nodes which enables the relationship that has been set to act as a code. The relationship automatically enters the code set in a process called system closure that enables automated searches for the relationship. With this feature the results of every analytic query are banked and added to the data set and the composition of a complex retrieval is retained. It can be repeated each time you load new data, so a step towards automation.

Using CAQDAS, you can assign any number of codes to the text segments. There is no practical limit. The size of the data segment is defined by you, the researcher. The same data segment or overlapping segments can be given as many codes as you want. Support is provided to define codes as they are generated via the memoing function and you can modify definitions as the meaning changes and you get more sophisticated about your code.

Here is a picture of a screen from Qualrus. This is about an artificial intelligence feature that Qualrus has which is called suggestive coding. It is slightly risqué but anyways, here you are. Qualrus will suggest codes based on a combination of strategies including case based reasoning. It reviews segments that have already been coded. It identifies those most similar to the current segment and suggests codes applicable to it based on those applied to similar sets.

Verifying coding patterns and relationships requires comparing code assignments in different subsets of data using the programs search tools. Search operators differ between these packages but are pretty standard with Boolean operators AND/OR, XOR, NOT and proximity operators like NEAR and CO-OCCUR. The whole database or subsections of it can be searched for the position of two or more codes in the data or, for example, where certain codes occur within certain types of data like amongst respondents with particular socio-demographic characteristics or data from a particular empirical context.

Frequency information shows the prevalence of selected codes across different parts of data set. Word frequency tools that are content analysis features in these packages may differ between packages. Atlas/ti for example generates a spreadsheet file readable by an external application, counting every word by document or by groups of documents. We have already met earlier on the KWIC search and retrievals – that is a picture of a screen in MAXqda with the MAX Dictio add-on module. It is enabling an index of one word to locate finds against their context.

QDA Miner which we heard about from Paul earlier on with Wordstat and SimStat modules provides quantitative content analysis and text mining that supports with lemmatization (dictionary definition), stemming, stop lists, hierarchical categorization of words, word patterns and phrases, KWIC lists, co-occurrences and adjacency pairs, and links to lexical databases, vocabulary and phrase extractors. Tabular information in most of these packages can be outputted to a spreadsheet application or to SPSS.

There are features to support disambiguation, particularly in QDA Miner. But the principal way these packages support disambiguation is to enable you to get back problematic passages instantly. So really it is something that vests disambiguation work in your head. It simply produces the relevant data rather quickly.

A further point that may be relevant to the discussion in this meeting is the support that CAQDAS offers for data integration. CAQDAS has always included basic support for quantification by offering ports to export data to SPSS and import quantitative data tables and to count hits from specified retrievals, for example all female interviewees who voted for McCain.

Recent editions enable statistical information to be imported to inform coding of text such as divisions within the sample that have emerged from survey response followed by integration of coded information back into a stats package. That can enable comparative integration of large amounts of data such as open-ended survey response. It has to be said that for its part SPSS has expanded support for this kind of work. They have a routine called SPSS text analysis for surveys. It is called Stats. I mentioned it this

morning. It provides semi-automatic coding and natural language processing for work with survey open-ended responses.

Essentially CAQDAS packages are customized database functions. Any database can sort text by pre-categorized response variables. What CAQDAS does is to offer more flexible coding systems to help the text material be "coded on" to emerging categories or themes. Bazely used this to combine analysis responses to closed and open questions on attitudes to organ donation amongst patients in Australia. The reasons for choosing to donate were coded on into categories representing altruism, pragmatism and anxiety about bodily integrity. The categorized responses were then related to variables like grief resolution. These procedures can be used as an anchor point for scaled responses. CAQDAS enables responses to be systematically matched to respondent's socio-demographics or information from rating scales or survey response. Some programs enable the export of more complex associations between variables as a quantifiable matrix like a similarity matrix.

Beyond the sorting of qualitative responses by categorical or scaled criteria, Bazeley (2006) suggests two main kinds of data integration: 'combination' of data types within an analysis, for example, using categorical or continuous variables both for statistical analysis and to compare coded qualitative data, and 'conversion' of data, such as converting qualitative codes to codes used in a statistical analysis. Where codes derived from qualitative data are recorded as the presence/absence of the code in each case or as a frequency of the code's occurrence, a case-by-variable matrix can be derived. Statistical techniques like cluster analysis, correspondence analysis and multidimensional scaling can then be used. Software like QDA Miner supports multidimensional scaling, heatmaps, dendrograms, and proximity plots.

But to do that you have to be pretty confident about the caliber of the data and the quality of the coding of the data. You need enough cases to satisfy the statistic. You need to decide whether you are using an ordinal or simple absence/presence. You need to accommodate non-directional codes that don't have an outcome. That might involve recoding such as court cases that haven't eventuated in a verdict. And you need to be confident that the code assignment was done in a commensurate way from one part of the data set to the next.

Sadly I am going to skip the stuff about secondary analysis. I was going to tell you a joke about Clintons but it was a very stale joke anyway.

I want to finish by saying a little bit about the Rehnquist example. The highlighted issue is that a response could mention that he was on the Supreme Court without mentioning Chief Justice; mentioning Chief Justice without Supreme Court; to referring to him only being a federal judge and so on. Using Boolean operators would help researchers refine their criteria for rating a response as satisfying these questions. Text with the words "court, federal, judge and justice" might satisfy the criteria and discounts on full satisfaction could be given text with less than the full set. A straightforward matter using Boolean operators and retrievals could examine linked features using code terms indicating certainty or hesitancy, for instance, phrases like 'I think' or 'I would guess maybe'.

Most work with code and retrieve selective sifting of data and the researcher reading the retrieved data to develop an interpretation. There are more formal methods available. HyperRESEARCH enables a hypothesis test which builds a set of if/then rules and determines whether they are satisfied. This involves several steps towards the proof of the hypothesis. The second package of interest operates in more formal terms, Qualrus, the one with AI features. It lets users design rules that have to be satisfied for a code to be correctly assigned. If that rule isn't satisfied it will report that and invite you to either adjust the code or adjust the assignment.

Those things seem to me to be promising steps towards more rigorous ways of handling open-ended responses. Qualitative analysis mostly involves using multiple retrievals to get a sense of the patterns in the data. But that doesn't mean that qualitative software can't be used to pursue responses to questions that actually have factual answers.