Organising Survey Open-Ended Question Data for Analysis in CAQDAS packages: 'Document per Respondent' versus 'Document per Question'

This page discusses a preliminary aspect of preparing textual data derived from open-ended questions to surveys for analysis using selected CAQDAS packages[i] : ATLAS.ti, MAXqda, NVivo or QDA Miner. For further information on the specifics of data preparation of open-ended question data for each individual package, and other background information which contextualises this discussion, see the relevant pages accessed from the main Survey data webpage.

The choice has to be made at some stage:

When preparing data from open-ended survey questions for analysis in NVivo or ATLAS.ti, a crucial initial decision is whether to organise material on the basis of a 'document per question' (all of the responses to one question in a single document) or a 'document per respondent' (a separate document for each respondent which includes their answers to all of the open-ended questions).
Users of some other CAQDAS programs, such as MAXqda or QDA Miner, are not obliged to make this decision at the data preparation stage but will have to make a similar choice when analysis begins, as at this point a procedure for reading texts in a systematic way has to be selected. This discussion is therefore relevant to all analysts of survey open-ended question data, but specific software selections require it to be considered at different stages in a project.

It is important in NVivo and ATLAS.ti:

The following paragraphs consider the advantages and disadvantages of each approach. Some points are more significant in relation to certain programs than others. NVivo and ATLAS.ti both require textual data to be held in documents, which effectively freeze the layout decision at the point of data preparation. This is why practical considerations of data handling have to be considered at the outset. Both MAXqda and QDA Miner both use a flexible database structure which allows data to be displayed in a variety of ways during analysis independently of the form in which it was imported. Therefore when using these programs the choice of how to read the texts is more flexible and possibly less consciously made.

Some may find it hard:

The 'document per respondent' approach is likely to be the initial expectation of experienced qualitative researchers who are familiar with this way of analysing in-depth interview transcripts. It is often intuitive for qualitative researchers to think of the individual respondent or research participant as the basic unit of observation. In contrast, the 'document per question' approach may seem more natural to researchers with a quantitative background where a comparison across large numbers of cases according to the particular questions asked is often an important requirement. It is not meaningful to describe either approach as "right" or "wrong" but some mental adjustment may be needed for a researcher to be persuaded to use an unfamiliar

or seemingly counter-intuitive layout because it is more efficient from a software point of view. Here we set out aspects for consideration so that informed decisions can be reached at an early stage of work.

Consider any existing formatting:

An important factor to consider at the outset is the format in which the data is currently available to the researcher, because reorganising large quantities of data is likely to take considerable time and effort, as well as creating risks of error and data corruption. Ideally this issue would be considered before data is collected, so that subsequent processing tasks may be kept to a minimum. However if survey data have already been collected, and the responses to the open-ended questions have already been assembled in some particular form, then the first efforts should be directed to planning a way to use that form so far as possible.

A spectrum of situations in terms of data volume:

The next consideration concerns the quantities of data to be analysed. It is possible to imagine a spectrum of situations between (say) semi-structured interviews using 20 questions with 25 respondents towards one end of the scale and two open-ended questions included within a survey of 5,000 respondents towards the other. The former situation is likely best handled on a 'document per respondent' basis, while the latter would almost certainly utilise two large documents with one for each question. The problem is deciding at what point along the spectrum the switch-over between approaches should be placed, and this is a matter of judgement. The more significant factor is likely to be the number of respondents, rather than the number of questions, and a limiting condition is likely to be the time it takes to process each document for packages like ATLAS.ti or NVivo, which require data to be held as documents. In database programs, such as MAXqda or QDA Miner, a more relevant factor may be the number of mouse clicks or keystrokes required to display each successive group of texts.

For example:

To illustrate this point an experiment was carried out to test the extraction of texts on a 'document per respondent' basis from SPSS via a spreadsheet application (in this case, Microsoft Excel™) into a word processing application (in this case, Microsoft Word™) so that they could be saved for use in NVivo or ATLAS.ti. On average, two documents on the respondent basis could be created per minute. The full data set in this instance included 1,257 respondents and 8 questions so an extrapolation indicates that a minimum of 10 hours concentrated effort would be required to set up this data on the 'document per respondent' basis. The eight 'document per question' files were created from the same spreadsheet within less than 40 minutes in total. Beyond this there may be other time penalties with the 'document per respondent' approach when working within some CAQDAS packages, because operations that involve drawing data from multiple files for retrieval purposes may take an appreciable amount of time. Consider, for example, the time required for 1,257 separate files to be opened, read, and closed by the computer – even at the high speeds we expect of automated procedures. In the end this will be a judgement decision for each researcher to make but we would expect few to adopt the 'document per respondent' approach for open-ended survey data when the number of respondents is more than 100 times larger than the number of questions answered by each.

Research design and methodology:

Leaving practical issues to one side for a moment, it seems sensible to examine this from a theoretical point of view. The project design and analytical approach should be considered. Here we focus on a key methodological issue involving this type of data integration. This discussion is concerned with open-ended questions which have been asked within a survey context, and that context may have created some framework for the respondents, whether the survey was carried out with telephone or face to face interviews, self-completion on paper or on-line. The responses to any one question should have quite a lot in common with each other, because the question has probably been asked in the same way for each respondent, it should have fallen at about the same stage in each questionnaire, coming after certain questions and before certain others. In a survey the respondents should exhibit a spread of characteristics, however, if they are to be representative of a population. Thus the analytic strategy is likely to be to read all of the responses to a single question, in order to identify similarities and differences amongst them. It would make less sense to read the responses to all of the questions for each respondent in turn – unless it is felt that those responses are strongly linked together. This appears to provide some inclination towards the 'document per question' solution.

Separate questions or a conversation?

However, there may be other factors which affect the decision. For example, open-ended questions may have been spread throughout a survey questionnaire with the intention of gaining richer insights to illustrate the quantitative data being collected by other, closed, questions. In other situations a block of open-ended questions may have been compiled, perhaps creating a more in-depth conversation episode, quite possibly at the end of a structured interview. It seems likely that the former situation would lend itself better to the 'document per question' style as the open-ended questions are being used to add context to individual survey answers. However, in the latter situation the 'document per respondent' style of analysis may be more appropriate as it will likely be useful to keep all of the open-ended questions together because they relate to one another and to the particular experiences and opinions of individual research respondents.

Paraphrase or a full transcription?

A practical problem of data-collection technology may also have a bearing on this decision, concerning the method by which the response is 'captured' or recorded. Where an open-ended question is included in a fairly standard quantitative interview, then it is likely that the interviewer will be expected to type the respondent's answer into a free-text field in the database. However, where more emphasis is given to the open-ended elements of the study, an audio recording of the whole interview may be generated and subsequently fully transcribed. The former is likely to be represented by short statements and brief paraphrases of what the interviewer heard, the latter can be expected to yield richer and longer texts. Once again the former type of data would probably be more fruitfully analysed with a 'document per question' approach, and the latter might sometimes be better read the other way.

Relationship to other data:

A final matter for consideration will be to look at the whole research project and to think about how the analysis of these open-ended questions fits into it. The advantages of organising data in ways that are compatible with other parts of a project will probably outweigh the difficulties of working in an unfamiliar layout, using a work-around, or using software in a sub-optimal way. Additional factors may come into play in a longitudinal study, if you want conduct within-case as well as between-case analysis, for example to track differences in one respondent over time as well as to consider the similarities and differences between individuals at any given time point. Alternatively, in a mixed-methods study you may have some material which is clearly 'document per respondent' oriented and other material which is better handled using the 'document per question' approach. In such situations, some ingenuity may be needed to provide links between the two.

Overall objectives:

As with most decisions regarding the use of CAQDAS packages, it is important for researchers to make an informed decision with regards to individual processes, based on an appreciation of the most important factors for the project in any given situation. Either approach will necessitate a balance to be reached between the analytical needs of the project and the practical and technical efficiencies afforded by the use of particular software packages. Good planning and clarity of objectives should inform the decision-making process.

Practical Issues in CAQDAS packages:

It has been established that it is practically possible to analyse data that has been prepared in either of the two formats under discussion with the four programs reviewed in detail on the Survey Data pages of this website. Detailed guidance on data preparation and analysis strategies for each of these programs can be found in the materials accessed from the Survey Data main page. The 'document per respondent' approach may be seen as the most intuitive and common method of operation for these programs amongst many qualitative researchers, and where necessary specific work-arounds have been established to demonstrate ways of working with a 'document per question' approach in each program. However in ATLAS.ti and NVivo there are consequences that may impinge upon the analysis whichever decision is made and these should perhaps also be taken into consideration before a final choice is made. In QDA Miner and MAXqda this decision is only required when the researcher begins to read the texts and start the analysis. For all of these programs a limited outline of the suggested procedures is shown below in order to inform the decision-making process.

ATLAS.ti

In ATLAS.ti this issue is complicated. Variable information about respondents' characteristics is usually stored in "Primary Document Families", which only work correctly if the "document per respondent' approach has been adopted. In order to have some selective reporting options that use variable characteristics in the 'document per question' method, it is necessary to insert a structured alpha-numeric string next to each response and then apply autocode routines to allocate thematic-type codes to the texts associated with them. This work-around can work well, even for a dataset with a large number of respondents, but there is

subsequently no practical way to extract, report or work with the responses for just a single respondent other than by running a text search for their unique ID and using it to call those responses to screen one at a time. It is probably not really practical to use a thematic code and autocode routine for each individual respondent when there are very many of them.

Furthermore, because ATLAS.ti stores the data outside the main project file (or, in ATLAS.ti terminology, the Hermeneutic Unit), having a very large number of separate files representing a very large number of respondents on a 'document per respondent' basis could give considerable problems of data management during the period of active analysis if the implications of the external storage of files are not fully understood.

In addition, either choice made for ATLAS.ti has a consequential inflexibility to add further data during the analysis phase of the project. If the 'document per question' approach has been adopted it will be found that the selection of characteristics to be coded-in can only be extended with considerable labour, while if the 'document per respondent' approach is used it will be difficult to add data for an additional question at a later stage. In summary, ATLAS.ti can be used to work in either way with this sort of data but there are some inflexibilities and disadvantages with each method. For detailed information on preparing open-ended survey data for analysis using ATLAS.ti see this page.

NVivo

In NVivo, when using the 'document per question' approach, providing a unique identification string has been placed adjacent to each separate response, it is possible to use an autocoding procedure to link each response text to its appropriate respondent (or "Case" in NVivo terminology). It is then possible to extract all of the responses for any one case, thus creating the 'document per respondent' as a separate viewable item within the program. It is possible to allocate thematic/conceptual codes ("Nodes" in NVivo terminology) to sections of text on screen within this view, thus the full functionality of the 'document per respondent' can be reproduced giving the analyst the possibility of both views. However, users may experience problems when they need to switch between the list of cases and the list of thematic codes in the Listview Pane, so it would be more efficient to organise the data in documents that reflect the way it will be analysed in this package. Subsequently, by using the functionality of NVivo's "Casebook" which can hold variable-type data about each respondent, it is possible to extract texts satisfying any combination of variables and thematic codes for all or selected questions. For detailed information on preparing open-ended survey data for analysis using NVivo see this page.

QDA Miner

The other two programs considered here, QDA Miner and MAXqda, both use a fundamentally different approach for this type of data. These programs store each response to each question separately, and then present them to the analyst in any order or grouping requested, including sequenced by case or by question.

QDA Miner takes this slightly further than MAXqda in this respect. The main display screen only shows a single text at a time, i.e. one respondent's answer to one question. However, it is easy to generate a Text Retrieval report to show all of the responses for one question in a scrollable report window, which links interactively with the main data panel so that thematic coding can be carried out straightforwardly. So the 'document per question' view is quite simple.

However, it is not such a simple matter to extract or display all of the responses for one particular respondent in QDA Miner. It is possible to generate a Text Retrieval report for the entire dataset, showing all responses for all respondents, which will be sorted into case number sequence and so, by scrolling this report, it is possible to view the full set of responses for any particular person grouped together. Alternatively, by selecting the required person in the Cases panel and then clicking on each question's document tab in turn it is possible to view that person's set of responses one by one. For detailed information on preparing open-ended survey data for analysis using QDA Miner see this page.

MAXqda

MAXqda readily shows both the 'document per respondent' and 'document per question' formats within different 'browsers' simultaneously. The data has to be prepared and imported into the program in a specific way but, when this has been done, it is straight-forward to 'activate' simultaneously all respondents for one question in order to create the desired display. The full set of responses to the activated question code will appear in the "Retrieved Segments" panel as a single scrollable list, and when any item in this list is selected with a mouse-click the full set of responses made by that particular respondent are displayed in the Text Browser panel above. It thus becomes an almost unconscious decision for the researcher as to which panel is used and in which sequence the texts are read and coded. It is possible to alternate between the two, but for practical reasons, and in the interests of consistency, the choice will inevitably have to be made one way or the other for specific analytic purposes. For detailed information on preparing open-ended survey data for analysis using MAXqda see this page.

Conclusions:

Whichever software package you choose to use, you will have to make a further choice at some stage as to whether you will read and analyse the texts in question order within respondent groups or in respondent order within question groups. The more factors you can consider as you make that decision, the fewer surprises you will get as you put it into practice.

[i] This discussion is written in specific reference to the following versions of CAQDAS packages; ATLAS.ti version 6; MAXqda version 2007; Nvivo version 8; and QDA Miner version 3.2. However, the information presented here is likely to be relevant to both earlier and later versions of these packages, and may also be useful as the basis of consideration for other CAQDAS packages.