Eliciting Quantitative e-Social Scientists User Requirements

Florian Urmetzer and Vassil Alexandrov

ACET Centre, The University of Reading

Abstract

e-Research is a relatively new and expanding means of undertaking multi-disciplinary on-line research. As such, the needs and requirements of the e-Research community are not yet fully understood and it is evident that e-Research needs further investigation and study. This report aims at providing an in-depth insight to the computing requirements of the e-Social science community involved into quantitative analysis of secondary data. The investigation made use of qualitative research methods in order to gather information from the e-Social scientists. Focus groups and open ended interviews were applied as research methodologies. On these grounds, first, the report details the outcome of a number interviews held with e-Social scientists across the UK. This involved interviewing twelve people in six sessions. Second, the author makes a number of conclusions on the problems and needs of the interviewed e-Social scientists. Third, the information gathered during this research is condensed into a list of requirements for tools that may be developed and added into projects.

1. Introduction

The Colaboratory for Quantitative e-Social Science (CQeSS) is working to develop and provide a portal interface for Grid enabled resources. It has been stated that portal based interfaces will widen the usage of grids to new users. Further portals will serve collaboration between researchers working in similar scientific fields. Portal structures enabling the access of collaborative tools and grids will be named here as Virtual Research Environments (VRE). These portals are described to integrate all possible tools and services needed during the research lifecycle; including new and existing tools [2]. As a result, the CQeSS project is aiming to develop tools for the Sakai portal framework to support the electronic expression of the research cycle and to allow collaboration via the VRE.

Prior to the future development of tools, potential users were questioned to elicit information about the needs and prioritisation of tools in such infrastructures. The need for such communication was for example discussed in [1].

This report will detail the process of the information gathering which has been conducted by the authors.

In the first of the following parts of this report, it will specify the research methods used during the study to elicit information from potential users.

In the second part, the paper will detail on the outcome of the interviews conducted. In the third part, the authors will give a list of tools with a description towards their functionality and problem solving capability. In the final part, conclusions as well as a pointer towards future work will be highlighted.

2. Research Methods

There are several methods used for collecting information from a user base. They are typically split into two sub groups. The first group encloses quantitative methods, such as user surveys on the basis of multiple choice answers. The second group encompasses qualitative methods, such as user interviews [3][4]. This study has been using a qualitative approach to conduct the interviews during the research. A more detailed discussion on the choice of methods can be found in [16].

2.1 Methods for eliciting user information

Qualitative research is defined as a methodology that allows the interpretation of data, [3][5][8]. Qualitative methods may take a longer duration and/or involve more intensive contact with the study situation than other methods [6]. The task of the researcher during the process is to capture a situation in its full complexity and log the information found. The analysis then includes investigating the topics or subjects found and interpreting or concluding on their meaning from a personal, as well as from a theoretical point of view [3].

Where qualitative methods have many facets, the most applicable for the eliciting of user requirements are qualitative interviews and interviews conducted with focus groups.

Qualitative interviewing can be defined as a dialogue, where one of the participants (the researcher) is looking for answers from a subject [9]. This form of interviewing is often split into unstructured interviewing and semi-structured interviewing. When the interview is unstructured the interviewer is only equipped with a number of topics that should, but do not have to be, discussed in the interview. When the interview is semi-structured the interviewer has a series of questions, which should be addressed, but the sequence of posing the questions may be varied [5].

Focus groups are defined as a small number of people chosen to participate in an informal discussion guided by a moderator. The group's task is to discuss defined subjects, opinions, and knowledge [10][11][12].

The computing literature is showing many examples of positive use of qualitative interviews and focused groups as methods [13] [11] [14] [15].

3. Interview Analysis

This part will firstly indicate the background of the interviewees and their relation to quantitative analysis using computers. Secondly, interviewees' responses and problems in relation to topics will be discussed. Thirdly, the responses of interviewees for problems with computing needs will be discussed. Finally, the usage of applications and the comparison of newer and traditional computer infrastructures will be assessed.

3.1 Who were the interviewees?

Concerning this study, the subjects interviewed were all members of staff of UK universities, who are using quantitative statistical modelling in a social scientific background, excluding one interviewee. This interviewee was involved, at the time of the interview, in a study looking at medical data, but used quantitative computing tools and was experienced in social sciences.

The interviews conducted were normally held over lunch. There was no personal or work relation between the interviewees and the interviewer and the interviewees were recruited using a snowball system of recommendation. The selection was based on the involvement of the participants in using quantitative methods in social science.

The interviewees had different levels of computer abilities and were categorized in highly, medium and low computer literate researches. There were two single interviews conducted with researchers who were able to use high performance computers remotely and were confident in using programming languages such as FORTRAN and command line tools. These participants stated independently that they would be less likely to use FORTRAN and other programming languages for their daily work; however, they would be more likely to use STATA, which is a statistical analysis package. These users were seen as the highly computer literate.

On the medium enabled scale, there were three group interviews concerning participants who were involved currently or in the past in large-scale secondary data analysis. However, they would only be using PCs for processing of datasets. They would be able to cope with a command line interface, but then again they would not use a command line based interface on a daily basis or would prefer not to use it. In fact the interviewees stated that they would use packages like STATA and SPSS for their evaluations and would be using advanced features of these tools. However, they would not be able to write programs using high level programming languages.

Finally the lower end of the scale was one individual interviewee. The participant showed the use of SPSS; but pointed out that the work done would have been possible to be done using Excel.

3.2 How interviewees work

All the researchers' tasks were structured and their work was described as incremental research.

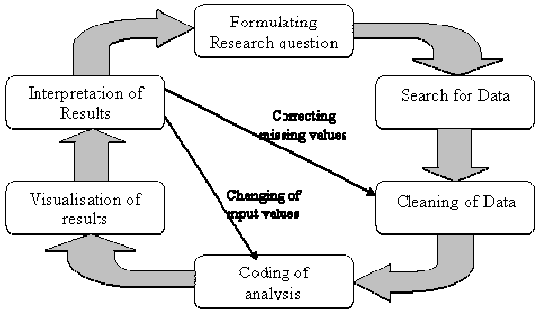

Figure 1: Research cycle for most e-Social scientists interviewed.

Their research tasks can be broken down into the following steps:

1. The research question is defined.
2. The data searched for is downloaded in SPSS or STATA format.
3. The data is then cleaned. The process of cleaning includes merging, filtering and collapsing of information. Merging means to join, for example waves of datasets. Filtering is

the process of, for example, extracting an age group out of a dataset. Finally collapsing means the rearranging of datasets from age to age groups, e.g. 1-10 year olds, 11-20 year olds and so on.

4. The detailed analysis is conducted using either the interface of a statistical analysis package or one of the available scripting languages.

5. The output is visualized and analysed in greater detail. This process is repeated depending on the success of the outcome.

Concerning the point 5 mentioned above, in the case of some of the more computer literate researchers, it was often noticed that the availability of a more detailed dataset was seen as important (see Figure 2).

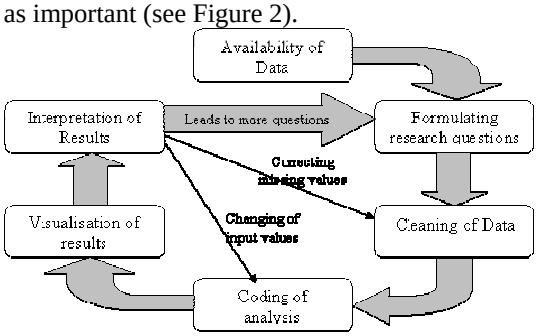

Figure 2: Research cycle for highly computer literature social scientists.

Therefore, research questions seemed to be set often on the basis of having availability of detailed datasets.

The interviewees pointed out there were several statistical software tools that were mentioned to be used in the field of study. Everyone interviewed was familiar with the software tools SPSS and STATA. Also mentioned, there were tools like EView, Stats, SAS, GEAS, Scross and R. The latter were not mentioned as major tools; however, they were used occasionally. The major governmental data providers are said to have made the decision, that their data is downloadable only in SPSS and STATA format. In fact, it was mentioned that it was rather complicated to get "only" an ASCII text file from the servers, which may introduce a bias towards the two main software tools.

3.3 Data

This section deals with responses concerning problems with data in general and will then detail them in several categories.

3.3.1 Problems concerning Data

There are several issues concerning data, which were pointed out during the interviews held. One issue pointed out was that the data available online is highly accessible. This would bring dangers of low quality analysis. However the participants quickly agreed that the data should stay accessible.

Also it emerged that specific knowledge about the datasets used is necessary, because the government publishes regularly updates and new datasets through the providers. This does not mean that data is being coded the way it has been before. Hence, when combining waves of data (different collection times of questionnaires) the researcher has to be aware of these changes to be able to counteract them.

Errors, meaning actual mistakes, are found rather rarely in datasets, so one participant stated that he found about ten to twelve errors in a dataset consisting of eighteen thousand different entries. He pointed out that this is a minor error rate, however, and he admitted that for a social scientist it is a fascinating puzzle.

Another interviewed researcher stated that the beauty of secondary datasets is that no single researcher can collect these vast amounts of data and multiple waves of data over several decades.

When looking at international datasets and their comparison, researchers have to be aware that there are different collection mechanisms in different countries. For example, some countries may collect household data every ten years. This data then has to be relied on for a decade. In this context as well, it was mentioned that organizational differences in data collection also have to be looked at to produce quality output.

3.3.2 Searching through databases

When searching for data, the data providers have search tools enabling the search for datasets. These tools are based on the web forms. One of the problems mentioned by the interviewees is that there is no tool to allow cross database searches. Therefore, it is not possible to search for data over multiple data providers offering a larger selection of data to choose from. Another issue arose is that there seems not to be the availability of searches at the item level on most databases. Some search engines allow simplistic analyses on the data prior to downloading the data. Nevertheless simplistic analyses prior to the downloading were mentioned to be useful.

3.3.3 Downloading of Datasets

Some of the datasets are neither searchable nor downloadable. This fact links with disclosure of information and confidentiality, which will be

discussed in more detail at a later point. Besides, there are services of classified or highly detailed data that are available in a descriptive format. Hence there is only the description of the datasets detailing which data can be found in what column to enable the researcher to write the analysis commands. An analysis is conducted by sending the code for the analysis to the data agency. The data provider then runs the analysis for the researcher. Another option is for the researcher to go to physical locations, which is described as protected rooms, where the researcher can run the analysis. From the interviews conducted for this project, it emerged that datasets that requires special protection are for example offenders' records.

Another issue of downloading is that, when looking at international data, there are differences in coding and, further, the legislations about data access differ from country to country; thus access is not always the same, which can cause problems.

3.3.4 Ways of keeping data

Traditionally data used to be in a printed form. There were also datasets kept on UNIX computers in Manchester. These would keep data depending on their topic in different directories. The UNIX computers used command line interfaces, which made them difficult to be access for non computer literate researchers. Currently the datasets are available through web interfaces. The computer literate users keep their datasets stored on servers provided for them by the university. For editing purposes datasets are kept on personal computer. Most interviewees indicated that they would rather like to keep the datasets on their desktop, from where it can be cleaned and analysed.

3.3.5 Cleaning datasets

Cleaning datasets refers to the process of enabling the dataset to be analysed in the software of choice. For this process most people said they use tools which are included in their analysis program, e.g. STATA and SPSS. It was said in the interviews that cleaning is needed before any form of analysis can start and it normally takes the most time during a project. One example of cleaning datasets, given by the interviewee, was a survey containing an educational variable. In fact, in a survey having educational variables, where there are similar small records, variables are merged to get larger statistical significance. However these categories may need to be collapsed into few groups to be able to execute them in the statistical model of choice.

For the cleaning processes of information sheets are needed about datasets which explain how the data was recorded. This would then explain, for example missing values that are dependent on an answer given by the respondent, such as 'if you are unemployed then continue with question eight'. The questions in-between would in this case stay empty, because they may be concerned with income and area of work.

To undertake cleaning remotely or in any form that is not on the desktop PC was seen by most participants as a rather complex process. Therefore, remote data access was seen as unpopular by researchers.

In conclusion of this section it can be noted that the process of cleaning datasets will need future attention in projects concerning quantitative e-Social sciences. The interviewees all agreed that the task of cleaning datasets is the most problematic and the most time consuming in their area of work.

3.3.6 Merging datasets

One process closely related to the cleaning of datasets is the merging of datasets. This was reported in the context of two needs. The first need is to merge different waves of data over time, e.g. British household data collection in 2001 and 1991. The second need is the interest in merging two datasets that have different types of information that the researcher is interested in, for example different datasets from one survey. These tasks have been described especially in the second case as rather trivial if one knows how to conduct them; however, newcomers in the field and students would have rather large problems with these tasks. It was also reported a problem for dataset merges concerning overlapping. This problem would occur, for example, in the case of the British Household Panel Survey which would come in waves and in different sets. Specifically the waves would come in five quarters. In detail, every quarter would have an overlap with the next quarter starting. On the problem of overlapping, one participant stated that he had to write his own code to join these together, even that he would believe that there is an official version somewhere; he was not able to find one.

It was stressed during the interviews that it is less complex to match data that have a common identifier after which the data can be sorted. One participant stated that he would only ever match databases after a common identifier. On

the contrary, other participants described the use of theories that allow the matching of databases not using unique identifiers.

3.3.7 Geography conversion

The common ground on which merging is done are geographies. This means that the data is coded, for example, after a postcode and another is coded after town districts. The postcodes can be collapsed into town districts and merged into one dataset. A problem is that there are different geographies. For example within the UK the police have different geographies to the health service, which is again different to postcodes. It was said that there are conversion tables, however, there are not known standard programs dealing with geographies.

3.3.8 Confidentiality

Resulting from the geographies is a problem concerning the protection of the single identity of the interviewee. If the geography is chosen to be a small grid, for example postcodes, a single person, or household may be identified with some prior knowledge. To find a household with five children in one postcode should not be a problem. Thus, confidentiality has been described as one of the biggest concerns of the profession and it has been mentioned that some of the data providers are providing less data or larger geographies to protect individuals from identification. One example that has been pointed out is the educational longitudinal data set in the UK. The protection mechanism is such that a researcher only gets a description of the data and then codes his programs in STATA or SPSS using the description. The program file for the analysis is then sent to the authorities and they run the program on their database. Another option described is to go to the physical location where the dataset is stored by the authority. The solution to leave the data on a server in a physical location and to only give remote access to the data is a logical alternative. However, it has been mentioned in several interviews that the users prefer to have the data on their machine, because of the physical availability and the option of being able to clean and to treat the data along with its availability. It was clear in all of the interviews that this is the future challenge for the profession.

3.3.9 Metadata

The final problem concerning data was the creation and maintenance of meaningful metadata. It was pointed out that there are no standards about metadata in the field; this gives rise to problems during searches on datasets.
The search on datasets was mentioned as one of the research challenges in the field in which many people are already working on. For this reason there was no further detail discussed during the interviews.

3.4 Computing issues

During the interviews, computing power in e-Social science research was introduced in the discussions. However, there seems to be some diversity of the requirements. Indeed amongst the interviewees there were high computer literate users who would not mind having larger computers, while medium and lower users were not that interested. Besides generally the high computer literate users seemed to be happy to work on multi processor machines, even that these are less often used than desktop PCs or to utilize single processors in a multi processor system. This was based on the argument of time and ability.
The interviewees pointed out that most of the calculations would only take small amounts of time, meaning one to several days. Therefore, most of the jobs would be started before the weekend or scheduled on a single processor.
Another relevant aspect of computing which emerged during the interviews is that the computer literate users would be happy spending more time on a single processor and therefore without waiting for a multiple processors being available on a multi processor system.
A final computing issue is that there is limited parallel programming knowledge in the social sciences. One interviewee mentioned he is interested in using parallel programming techniques to speed up analysis, but does not have the skills. This shows that research collaborations with computer scientists with experience in parallel programming may need to be established.

3.4.1 Towards a Grid infrastructure

It appeared to the e-Social scientists interviewed that Grid infrastructure would be a good way of making their databases sharable. One interviewee detailed that his experience at present involving grid technologies was more on the data provision rather than processing.
The focus on data services using the grid infrastructure may be related to the fact that the current programs using e.g. SPSS and STATA are not able to run on the grid. So, for most of the e-Social scientists there would be too much

change and new learning involved in changing their habits. The scientists interviewed seemed to be very much comfortable in using smaller scale HPC or local computers for their processing needs. Only one participant mentioned the use of batch processing and schedulers on a HPC.

3.4.2 Applications for data analysis

The computer literate users, who were interviewed for this report, insisted on command line interfaces and scripting languages for the process of analysing data. However, they recognized that there should be high level programming languages for data analysis rather than low level.
There seemed to be an interest to get standard applications to do certain cleaning and merging tasks. At present these tasks have to be described and there are no standard tools available. Currently every researcher has different ways of undertaking these tasks.

3.4.3 Views about Portals

The use of Web portals was received controversially. This is because every interviewee seemed to be protective about their datasets that they had invested time into. There seemed to be a mutual understanding of not cleaning and arranging data online or remotely. One interviewee argued this to be for self protection. The general idea of sharing was received quite positively by one of the focus groups which was interviewed. Indeed there was only the wording used "to encourage people to share different resources". One participant suggested that these resources could depend on a group of people sharing the material and having a protection towards a generic use. This means allowing exclusive access to material created by one member or others of the group. These collaborations were said to normally consist of three people as described by the interviewee.
A Web portal was mentioned to have advantages for example as some form of data management system, so that the data that is downloaded from the providers does not have to be replicated. A researcher, who has been involved in a collaboration that had used versioning systems, which he found useful in a collaborative setting, mentioned this. A versioning system was described to be a potential for datasets too. Also the interviewee shared information in form of blogs and files via the portal, which he described as being used instead of e-mailing the collaborators. The effect of annotating data and versions of data within a collaboration was described to be valuable, because the information could be shared between the entities in the collaboration as well as being accessible as a reminder to the individual. The researcher pointed out that if he would annotate and describe everything he was doing during his work, he would properly not be able to be productive anymore.

4. VRE Tools

This part of the report provides a list of tools, which are useful for the participants to conduct their research tasks through a portal. The authors suggest implementing the tools for the use in quantitative social science research portals, because of the scope of the wider research project, which is aiming at an implementation in a portal framework.

4.1 Tool for searching and downloading datasets

The interviewees have been describing the need for the development of a tool that allows searches across different data providers. This should allow national (UK databases) and international databases to be utilised. There was the need to visit multiple data providers to allow the gathering of data and to compare its availability. Therefore an interface allowing the entering of search words for data and a return of data availability from different data vendors would be helping the process.
After a successful search, respondents have pointed out usability problems with downloading data, especially where there are multiple waves of data. The problem is that most data provider portals are not enabling the user to download multiple waves of data with one click. Therefore next to the search result there should be a selection button. The download could then be structured like a 'shopping cart'. Hence the users would choose which datasets they want and then go to a download section, which has all the datasets chosen listed. Most importantly the datasets can be downloaded using one click. It may be seen as important that the download should be able to be directed to the portal server to allow further processing of the data online. However, many of the participants prefer to have the data on their personal computer and not on a decentralized server. Therefore there should be an option at any point in time to download the data onto the users' personal computer.

4.2 Tool for cleaning of datasets

The cleaning of datasets was seen as the most time consuming process by e-Social scientists and was seen by the interviewer as the most complex.
A part of the cleaning process may be automated is the collapsing of data. When the data is collapsed, it is for example sorted from individual ages to age groups. This process is normally done in preparation of a merge of datasets.
The process of cleaning data would be one of the most challenging to automate and to make easier but it seems as well the most complex. The detailed definition of this task should be the scope of further work.

4.3 Tool for merging datasets

There are issues with respect to merging datasets. Mainly new users to quantitative social sciences have problems merging datasets. In this context the British Household Panel Survey was mentioned as example of panel survey which uses different waves that overlap in time. The mechanics of a merging of these or similar datasets however are described as trivial e.g. unique identifiers or waves of data. Thus there should be a tool developed allowing the merging of datasets automatically. Additionally and as a more challenging option, geography conversions could be implemented into a merging tool allowing datasets with different geographies to be merged.
The process of merging should be attached to the selection and downloading (described as a similar to a 'shopping cart' earlier), so that the user would get the option to link unique identifiers or join different waves of selected datasets. The server could then join the multiple datasets into one dataset.

4.4 Tool for coding analysis and visualization

There should be a tool to allow the analysis of datasets with the option of running the analysis on the server. When asked, the participants mostly used statistical analysis packages like SPSS. These packages are mostly chosen on the basis of availability of training and personal choice. It is therefore clear that a tool for the coding of an analysis would have to reflect the interface and options of the mainly used tools, namely SPSS and STATA, for the tool to be adopted.
A tool for the coding and the analysis would be seen as advantageous to enable the upload of scripts from these tools to the server and therefore let the analysis be run on the server rather than using the home PC. The running of the code could then be done by the server or could be submitted to some form of grid computing infrastructure.
The returned outcomes should be available in textual form to be downloaded and should be able to be visualized in a predefined form.

4.5 Tool for the sharing of information

Web portals for information sharing were seen to be useful by the participants interviewed. In particular, web portals for information sharing which would allow the sharing of content to a defined user group, belonging, for example, to a research collaboration. This would include sharing data, source code as well as information on progress and annotation of data by using wiki and blogs. Collaborative video conferencing and chats should also not be excluded. Additionally a versioning tool that would maintain the stored datasets and code was seen as useful by the interviewed researchers.

5. Conclusions

In the course of this project the authors have investigated the methods and opportunities coming from a multidisciplinary background to elicit information from a potential user group. This investigation has shown that quantitative open-ended interviews are giving a good opportunity to discuss needs with the participants.
Once the needs were elicited from the user community the information was transferred into a list of tools. The further work of the project will therefore be the detailed definition of the tools and the production of the tools in order to make the tools available for e-Social scientists.
The interviews conducted in this study have demonstrated that the research cycle of e-social scientists involved in quantitative analysis using secondary datasets can be expressed in electronic means. The interviews have shown that problems concerning the handling of datasets are more constraining to the quantitative e-Social scientists than computational power.
Throughout the interviews the participants pointed out that there are standard applications, like SPSS and STATA, which are used for data analysis, which are widely adopted in the quantitative e-Social sciences field.
The tools emerging from the interviews are tools that can be implemented into a portal structure. These tools are namely a tool for the

search and downloading of data, a tool for the cleaning of data and finally a tool for the coding, analysis and visualisation of the data.

Additionally, the research conducted in completing this paper has shown that it is possible to elicit detailed requirement information from users. The interviews have given in-depth views of the needs and of the research process of e-Social scientists involved in the analysis of large datasets.

The most challenging task of this study was to get appropriate interviewees. The reason being, that most people are not prepared to participate, because of time constraints. Also it was seen as advantageous to approach research groups rather than individuals, because it was found that there was friendly relationship between the participants. This environment allowed an in depth discussion. Further, the transcription of the interviews was very time consuming. It would be advisable in other similar studies to avoid transcribing the interviews if the text format is not needed.

It was also interesting that detailed datasets are delivering the highest quality of research from the quantitative e-Social science point of view. The data providers are, however, becoming protective over detailed datasets on the grounds of confidentiality. In fact the participants have described data protection vs. accessibility of data as a future challenge. One solution to this challenge could be to move the research cycle online without making the data visible to the analyst. This would include the ability to clean and analyze the date online. Potentially, this could overcome confidentiality problems.

In addition it has to be highlighted that the research conducted for this report is only seen as a definition of the tools used by e-Social scientists. Finally research has to be undertaken to finalize the development and usability of tools as well as involving the user community.

6. Acknowledgement

The authors would like to thank the Colaboratory for Quantitative e-Social Science, especially Ties van Ark and Robert Crouchley.

7. References

[1] Marmier, A. The eMinerals minigrid and the National Grid Service: a user's perspective. The All Hands Meeting. 2005. Nottingham, UK.

[2] Fraser, M. Virtual Research Environments: Overview and Activity. [Internet] 2005 [cited 2006 01 May 06]; http://www.ariadne.ac.uk/issue44/fraser/].

[3] Bryman, A., Social research methods. 2nd ed. 2004, Oxford: Oxford University Press. xiv, 592.

[4] Creswell, J.W., Research design: qualitative, quantitative, and mixed method approaches. 2nd ed ed. 2003, Thousand Oaks, Calif.; London: Sage. xxvi, 246.

[5] Punch, K., Introduction to social research: quantitative and qualitative approaches. 2nd ed. 2005, London: SAGE. xvi, 320.

[6] Faulkner, X., Usability engineering. 2000, Basingstoke: Macmillan. xii, 244.

[7] Gillham, B., Developing a questionnaire. Real world research. 2000, London: Continuum. ix, 93.

[8] Seaman, C.B., Qualitative methods in empirical studies of software engineering. Software Engineering, IEEE Transactions on, 1999. 25(4): p. 557.

[9] Gillham, B., The research interview. Real world research. 2000, London: Continuum. viii, 96.

[10] Litosseliti, L., Using focus groups in research. Continuum research methods series. 2003, London: Continuum. vii, 104.

[11] Rose, A.F., et al., Using qualitative studies to improve the usability of an EMR. Journal of Biomedical Informatics, 2005. 38(1): p. 51-60.

[12] Nielsen, J., The use and misuse of focus groups. Software, IEEE, 1997. 14(1): p. 94.

[13] Holzinger, A., Rapid prototyping for a virtual medical campus interface. Software, IEEE, 2004. 21(1): p. 92.

[14] Kontio, J., L. Lehtola, and J. Bragge. Using the focus group method in software engineering: obtaining practitioner and user experiences. 2004.

[15] Urmetzer, F., Qualitative enquiry: the use of focus groups in The University of Nottingham VRE Implementation Evaluation. 2006, unpublished: Reading.

[16] Urmetzer, F., M. Baker, and V. Alexandrov. Research Methods for Eliciting e-Research User Requirements. in All Hands Meeting. 2006. Nottingham, UK: Nesc.