Making Qualitative Data Fit
the "Data Documentation Initiative" or Vice Versa?

Arja Kuula

Abstract: The Finnish Social Science Data Archive is a newcomer in the area of data archiving for two reasons, firstly: it started its operation only in 1999 and secondly: from the very beginning it has had as an official strategy to enhance the reuse of available qualitative as well as quantitative data. Archiving and reusing of data has been a common and continuously expanding practise in quantitative research since the 1950s and 1960s. Qualitative research has thus far been almost invisible in this respect, except for a few successful cases like Qualidata in Essex, UK and Murray Research Center at Harvard, USA.

Questions concerning archiving and reusing of qualitative data are many. Here I will concentrate on a very practical but important issue in making qualitative data reusable, i.e. documentation of data. I highlight some reasons for making appropriate and adequate data documentation and give the Data Documentation Initiative (DDI) as an example of documenting social science data. The DDI was meant for quantitative data, but I claim that it can be used and elaborated for the special needs of qualitative data as well. Choosing the same documentation model for qualitative and quantitative data would be one step towards social science data archives which would have both quantitative and qualitative data. This would support their basic task of promoting a sensible use of all research resources.

Table of Contents

1. Introduction

The world is said to be moving fast into an era of computing—in Finland we even have an official government strategy to reach the level or state of an information society. If a move towards an information society has taken place in our everyday life, shouldn't we think that it already is the case also in the area of social sciences? Some might say that it has already happened. This is perhaps more evident if you work with statistics and quantitative data using advanced hi-tech. But as we know, there are more and more examples of digitised qualitative data which can be processed and analysed electronically. [1]

The form of the data is an important issue when we talk about moving to the world of computing. To me, there is something even more important about it: attainability. It does not matter whether the data is in digitised or paper based form if one is not aware of the existence of the data. The data should be catalogued and made accessible through electronic search. To create a catalogue, we of course need suitable software and electronic metadata, the documentation of the data. [2]

If we want to live in a common world of electronic data access, we need to have common principles and a common documentation model or standard. In the area of statistics and quantitative data the most progressive effort so far to attain a common documentation model is the so called Data Documentation Initiative (DDI). The Data Documentation Initiative committee started its work in 1995 and has succeeded in developing a specification for the content and structure of metadata describing empirical research in the social and behavioural sciences. This metadata model was planned and is mainly used in documenting quantitative data. So is it unsuitable for qualitative data? [3]

In this article I shall discuss the effort of documenting qualitative data in the Finnish Social Science Data Archive (FSD) using the DDI documentation model. First I briefly introduce some background information on the reasons why data archive information services in Finland will also cover qualitative data. After that I describe the reasons for documenting qualitative data using the DDI. Then I will give some practical examples of filling in the elements introducing a couple of very moderate modifications of the DDI-elements. In the end I recap the reasons for co-operative effort in making a move towards common archiving policies documenting qualitative data. [4]

2. A Need for Qualitative Database in Finland

The FSD started to operate as a separate unit within the University of Tampere in the beginning of 1999. As in other data archives, the main task of the FSD is to increase the use of existing social science data by disseminating it. The main functions include acquiring, storing and disseminating data for secondary research. In the beginning, the FSD concentrated only on storing numerical data but this year information services are extended to cover also qualitative data. That is due to our research culture, especially when it comes to the methods in social sciences. [5]

In Finland in the beginning of the last century it was typical in social sciences to use many kinds of data. Official statistics, newspaper articles as well as stories told by the people who were being studied could form the basis of analyses. But —as in many other countries in Europe—in the 1950s and the 1960s statistical methods were in the mainstream though qualitative methods had their small share, too. [6]

The 1970s marked a turning point in social sciences in Finland. It was an era of Marxism, i.e. particularly philosophical and theoretical studies in social sciences.

And this certainly made the gap yawn between theory and the empirical world. This discrepancy was one of the main reasons for a turn towards using qualitative methods in the late 1970's (LESKINEN 1995). Being extremely philosophical and theoretical, social sciences were not capable of producing any methods or instruments for empirical research. Though Marxism left its traces in the first empirical and qualitative studies in sociology, the increasing use of qualitative data and methods was an alternative both to theoretical Marxism and positivism. [7]

In the late 1970s there began in many ways a very successful period in the establishment of using qualitative methods in social sciences in Finland (see KUULA 2000). Today, qualitative methods have a remarkably established position in Finnish social sciences. For instance at the University of Tampere—the FSD's hometown—you have to take a compulsory course not only in quantitative methods but also one in qualitative methods when studying social sciences. For first year students there are introduction courses on such areas as theory of rhetoric, narratology, action research, discourse analysis, conversation analysis and ethnography. In doctoral studies of sociology, the majority of the method courses concentrate on qualitative methods. If one counts only method courses available it could be said that in social sciences qualitative methods constitute the mainstream in Finland. [8]

Research culture which is favourable towards qualitative research supports our strategy of promoting reuse of qualitative data at the Finnish Social Science Data Archive. A concrete plan for that is to develop and maintain a database of available qualitative data which can be reused. Of course our duties are also to develop, set and propagate principles of collecting, documenting, organising and storing qualitative data so that it could be used by other researchers afterwards. Here I will concentrate on the plans of documenting qualitative data, i.e. creating metadata by using the DDI. [9]

3. Why a Quantitative Documentation Model for Qualitative Data?

Why do we need metadata? Is it not enough to say a few pertinent words about the data in question? Metadata could be defined as data about data. It constitutes the information that enables an effective, efficient, and accurate use of datasets and data collections. Metadata is a crucial point of departure for every kind of discovery system—let the data itself be paper based, audio analogue or digitised. Of course the original collectors of the data have all the informal knowledge which would guide the analysis process, but metadata is needed for the re-user to understand the intellectual content, geographic and temporal coverage of the data and to understand the way the data was collected. A proper documentation is crucial also because the data might be used many years after its collection and very likely for purposes that are different from the original. Metadata could be described as a bridge between the original collector and the re-user giving the essential information for secondary analyses (RYSSEVIK 1999). [10]

Why choose the DDI for a documentation model? The DDI standard is based on Extensible Markup Language (XML). Among many other things, XML is

straightforwardly usable over the Internet, which is the key to discovery and dissemination. Extensible Markup Language (XML) is hardware and software independent and it allows writing special vocabularies, the DDI being one example. Software needs to understand XML, but it does not need to support tags relevant to social science data. Because markup is plain text, it is human readable and easier to preserve than non-text formats. Availability is also an important matter. XML specification is openly published on the net. (GRANDA & JOFTIS 2000) [11]

Besides the reasons mentioned above, we have our very own reasons at the FSD to use the DDI for qualitative data documentation. The FSD is a new archive: we started building up a quantitative database in 1999. Because of that it was an easy choice for us to start documentation from scratch using the DDI. Having the procedures and software programs for making html-documents and a database for quantitative data using the DDI, it is more than obvious that we would choose the DDI and XML-language also for the purposes of a qualitative database. [12]

4. The Structure of the DDI

The elements in the DDI are arranged in a hierarchical or tree-like structure. The DDI model contains five major components or sections. First one is (1) The Document Description. It describes the metadata document itself and the sources that have been used to create it. The second one is (2) The Study Description. It contains information about the entire study or more precisely, about the data collection telling the content of it, the methods used to collect and process it, the sources and access conditions of it. The third component is (3) The Files Description, which describes the files of the data collection. The fourth part is (4) The Variables Description. It describes each single variable in a quantitative datafile. The fifth component is called (5) Other Study-Related Materials. It includes references to reports and publications or other machine readable documentation that is relevant to the users of the study. (See the DDI homepages.) [13]

Each of these main components is divided into a finer hierarchy of sub-components and elements. For instance the Title Statement 1.1.1 of the marked-up document contains five sub elements: 1. Title—Marked-up Document, 2. Subtitle—Marked-up Document, 3. Alternative Title—Marked-up Document, 4. Parallel Title—Marked-up Document, 5. ID Number—Marked-up Document. (See the Tag Library in http://www.icpsr.umich.edu/DDI/codebook/codedtd.html; broken link, FQS, September 2003.) [14]

Altogether, there are around 300 elements in the DDI-tree that could be filled in when doing the documentation of a data collection. It is certainly not the purpose to fill in all the elements—in that case the documentation would be as time consuming as the original data collection process. I have found approximately 50 elements which could be suitable for qualitative data. I will give here only a few examples concentrating on the second part of the major sections of the DDI. That is (2) The Study Description, which gives—among many other things—

information on the content of data and methods used in collecting it. It might be possible to somehow use other components—especially (3) The Files description and (4) The Variables Description—in the case of electronic or digitised qualitative data, but that would be a different story. [15]

5. Same Elements Describing Social Science

I define the basic philosophy of the DDI as itemisation with detailed classification. It is combined to a strict structure which defines which dimension of the data can be expressed in which part of the hierarchy and in which element. Despite this strictness there is one aspect that helps to apply it to qualitative data: You fill in the elements by writing a text. Even though the DDI standard is developed mainly for quantitative data, there are lots of elements which already are suitable for qualitative data. Elements which are suitable for both without any special adjustments are, for instance, Title, IDNumber, Authoring entity, Other Identifications, Copyright, Depositor, Deposit Date, Bibliographic Citation, Keyword, Topic classification, Abstract, Time Period Covered, Date of Collection etc. [16]

Besides those 'ready to fit'-elements, there are those which can be interpreted in an appropriate way. One example to start with could be Sampling Procedure. If we were documenting quantitative data we would fill in the element by choosing either simple random sampling, systematic sampling, stratified sampling, cluster sampling, two stage cluster sampling, stratified quota sample, multistage probability sampling etc. But do we have sampling procedures when collecting qualitative data? Yes we do. They are just different, not as exactly defined as in quantitative data. If we had a research where women having their first child in their forties had been interviewed, the Sampling Procedure element could be filled in like this:

> <sampPorc>*the 35 women interviewed were drawn from a course organised by the maternity clinic of Kontula in Helsinki.*</sampProc> [17]

An other example could be an element called Mode of Collection. In the case of quantitative data, this element would tell, if the mode of collection was telephone survey, face-to-face interview, postal survey or an email survey. The basic idea of the element would be the same in the case of qualitative data. Only the options would be different. Here are a couple of examples:

> <collMode>*telephone interviews with audio recording*</collMode>
> <collMode>*face-to-face interviews with audio recording*</collMode>
> <collMode>*request to write by announcement*</collMode>
> <collMode>*video recordings on authentic situations*</collMode> [18]

Researcher using qualitative methods may think that it is certainly not enough to say that data consist of interviews of women drawn from a maternity clinic and which were done as face-to-face interviews with audio recording. But the very idea of the DDI is to itemise every dimension describing the data into different

elements. So there would be certain other elements containing, for instance, the universe, special characteristics of interview situation, extent of data collection, confidentiality issues etc. Itemisation makes sure that the dimensions are given that are needed to inform correctly and sufficiently about the data. [19]

6. Ideas of Moderate Modification of Elements for Qualitative Data

In addition to elements which can be interpreted in an appropriate way there are those which can be modified in a way that hopefully will not invalidate the basic structure of the DDI. But even very minor changes call for a suggestion to the DDI-committee. The committee will then tackle the issue and if the suggestion is well prepared and defined the committee might come up with a favourable decision. Of course a suggestion made by a group would be more convincing than a suggestion proposed only by a single person. [20]

One example of the 'better when modified' elements is one of the most informative elements in the case of qualitative data: Kind of Data. It tells the type of data, i.e. whether the data are interviews, interview notes, interview summaries, group discussions, thematically organised transcripts, field notes, participant observation field notes, observational recordings, summaries of observations, diaries, letters, life stories, newspaper clips, articles, advertisements, photographs etc. [21]

As such, Kind of Data element is an important and informative one. But exactness of this element would be much better, if someone searching for suitable data through the web would be able to define in advance the physical form of the data. Of course the physical form of the data would be also important for someone, who has found a list of datasets using for instance, a keyword search. Actually, The DDI elements may have attributes which are characteristics or properties that further define the element content. In addition to defining more precisely the content of the element, attributes are more easily understood by a software system—especially if they have controlled vocabularies. That adds to the capability of determining the search terms and constrains more exactly when looking for suitable datasets through the Internet. There is not any attribute in the DDI, that would indicate the physical form of the data. But in my dreams the future Kind of Data element could have a 'format' attribute with controlled vocabulary. [22]

If Kind of Data had a 'format' attribute, it would specify in which form the data are. Possible choices for the vocabulary could be, e.g., Machine readable, Audio analogue, Audio digital, Audio-visual analogue, Audio-visual digital, Paper-based. If we think again about the research on the women in their forties having their first baby, this element could be filled like this:

<dataKind format='audio analogue'>*27 interviews of women in their forties*</dataKind>
<dataKind format='machine readable'>*Transcripts of 25 interviews*</dataKind>

```
<dataKind format='paper based'>27 interview summaries</dataKind>
<dataKind format='paper based'>27 interview notes</dataKind> [23]
```

One other major advantage of the DDI is the possibility of linking different elements. Beyond that there is also an external linking mechanism permitting links from elements in the DDI to items outside the document. That happens by using URI-attributes. The possibility of external linking would be very useful in the case of the element called Type of Research Instrument. Going through the examples of quantitative DDI codebooks, there the information can only be found on whether the questionnaires were structured or semi-structured. In the case of qualitative data I would think this element could show the ways of guiding, focusing, advising and controlling the data collection process. If the DDI committee would be accommodating enough, in the future this element could have a URI attribute to enable links to pdf (or-what-ever)-versions of the research instruments mentioned. Until that day it is also possible to write in the element text, for instance, the full address of the pdf-version of the research instrument in question. Examples:

```
<resInstru>Interview schedule 'http://www.etc'</resInstru>
<resInstru>Topic guide(s) 'http://www.etc'</resInstru>
<resInstru>Diary format 'http://www.etc'</resInstru>
<resInstru>Questionnaire 'http://www.etc'</resInstru>
<resInstru>Observation checklist 'http://www.etc'</resInstru>
<resInstru>Codes used in data process (NUDIST, WINMAX etc)
'http://www.etc'</resInstru>
<resInstru>Interviewer instructions 'http://www.etc'</resInstru>
<resInstru>Coding instructions 'http://www.etc'</resInstru>
<resInstru>Writing competition announcement 'http://www.etc'</resInstru> [24]
```

When having this element as an html-document, one could link straight, for instance, to the observation checklist, to see whether this data collection contains the issues a researcher is interested in and needs complementary data. The research instrument contributes to and affects the content of the data collected and it would be essential for the re-user to get exact information on the instruments used. [25]

7. Concluding Remarks

In my opinion, the DDI is an opportunity for the qualitative research community to look for an application for documentation procedures concerning qualitative data. The advantages of the DDI outweigh the eventual shortcomings which are due to its original area of use, quantitative data. The structure of the DDI-hierarchical tree of elements is rigid in the sense that each change requires a new official DDI-structure. But it is possible to suggest that DDI committee would make minor changes in the elements. The official policy of the committee is to encourage the development of applications using the DDI (GRANDA & JOFTIS 2000). Developing controlled vocabularies for attributes to facilitate machine processing is one concrete goal of the DDI committee. So it is up to the international qualitative re-

search community which promotes the reuse and archiving of qualitative data, to embark a joint effort to attain an agreed-upon procedure for documenting data. [26]

Applications using the DDI enable importing text files and loading databases or library catalogues. A lot of qualitative research material is already in machine readable form and in the future Internet can be seen as a media for moving and exchanging also qualitative material. Knowing the possibilities of image scanning and digitising technologies one can only imagine the future prospects and possibilities of archiving qualitative data. This vision and its actualisation can only contribute to the main task of data archives: enhancing sensible use of all research resources. This target might be much closer if we chose DDI for the documentation standard in qualitative data. The language used in the DDI—Extensible Markup Language—is forecasted to become the mainstream technology for powering broadly functional and highly valuable applications in the Internet. That broadens up also the possibilities of archiving electronic qualitative data. [27]

Choosing a documentation model is not only an issue of pure rationality. Having the same documentation model for quantitative and qualitative data makes the possibilities of broadening the policy area of data archives towards qualitative research and data much better. So choosing the documentation model is also a political question: whether we stay in separate camps and continue to do things differently in the worlds of quantitative and qualitative research, or we take a chance and do not voluntarily miss the train taking us to the world of electronically accessible and processable social science data collections. [28]

References

Granda, Peter & Joftis, Peter (2000). *The Data Documentation Initiative (DDI)*. The keynote presentation in the CESSDA Expert Seminar, September 1-2, 2000 Tampere, Finland. http://www.fsd.uta.fi/cessda2000/materials.html; broken link, September 2002, FQS

Kuula, Arja (2000). *How to Make Qualitative Data Reusable: A case in Finland*. Paper presented at the IASSIST Conference, June 9, 2000 in the North-western University in Evanston, Chigago. (Forthcoming in IASSIST Quarterly http://datalib.library.ualberta.ca/iassist/iq.html; broken link, September 2002, FQS)

Leskinen, Jaakko (1995). Lyhyt katsaus suomalaiseen metodologiseen kirjallisuuteen [A short overview of Finnish methodological literature]. In Jaakko Leskinen (Ed.), *Laadullisen tutkimuksen risteysasemalla* [In the cross-roads of qualitative research] (pp.115-127). Helsinki: Kuluttajatutkimuskeskus.

Ryssevik, Jostein (1999). *Providing Global Access to Distributed Data Throughout Metadata Standardisation—The Parallel Stories of NESSTAR and the DDI*. Paper presented at the Conference of European Statisticians, UN/ECE Work Session on Statistical Metadata. Geneva, Switzerland, 22-24 September 1999. http://www.nesstar.org/papers/ [Broken link, FQS, August 2005]

Author

Arja KUULA is a Research Officer at the Finnish Social Science Data Archive. She has worked as a researcher at the Work Research Centre and at the Department of Sociology and Social Psychology at the University of Tampere. She has published articles and a book on methodological issues and the role of a researcher in research and development projects. Her areas of interest are research culture, qualitative data and its reuse and documentation. Her doctoral thesis in sociology deals with methodological issues of action research.

Contact:

Arja Kuula

URL: http://www.fsd.uta.fi/
E-mail: Arja.Kuula@uta.fi
DDI Homepage:
http://www.icpsr.umich.edu/DDI/

Citation