# Preparing Machine Readable Codebooks at the Zentralarchiv Actual Situation and General Perspectives

*by Rolf Uher and Rolf Rontgen [1]*
*Zentralarchiv fur Empirische Sozialforschung at the University of Cologne*

**Abstract:**
In the first part of the paper we will show, how the different parts of the original documentation at the ZA will be processed to end up in a full-text machine-reada- ble codebook on disk and on paper. The second part presents a "checklist" of tasks, which should be considered if we think about a new - ideally IFDO-wide - production-line for codebooks. This "checklist" is also meant to be an offer for the ongoing discussion, and should be expan ded by ideas from interested colleagues.

**A. Introduction to the First Part**
In the mid-seventies the ZA began to process the machine-readable codebooks along the OSIRIS format using most of the original OSIRIS tools. This production-line substituted a system of programs developed at the ZA in the late sixties. The "philosophy" however standing always in the center of this strategy was:

1. to have a complete documentation for a study which includes all the information which a secondary analyst might need to interpret the data.

2. to start a text base for information retrieval purposes, regarding the fact that there will be an immens growth in the amount of relevant information (here: full-text retrieval on variable-level within a pool of studies/codebooks).

The OSIRIS format might be called "old-fashioned" or "unflexible" - and it is true for some cases - but there is no alternative format for it right now, in which complete question-text, complete answer-categories, archive-comments, general introductions, notes etc. can be processed. Generally spoken the "unflexible" format has the advantage that you can transfer an OSIRIS-codebook to nearly any other format which considers the variable- structure of a flat data-file.

Until there is no other solution which - at the same time - can convince us that it really works, it is a matter of reason to apply to a reliable tool.

**B. ZA Codebooks**
The ZA produces (machine-readable) codebooks as a standard documentation for some ongoing projects: ALLBUS (Germany's General Social Survey), ISSP (International Social Survey Programme), German Election Studies. The discussion about the EUROBAROMETER-codebooks is not finished yet. Besides that codebooks will be produced for single studies (e.g. Wohlfahrtssurvey, Health-surveys, Youth-surveys ...) or cumulations of study-series (e.g. Politbarometer), according to capacity in the codebook-department at the ZA and according to the expected user-demands.

**C. Processing a Codebook - First Step**
The first step producing a codebook is to make the que- stionnaire machine readable. The general tool for this step is a text-processing software on the main-frame. The structure of this "raw-codebook" is already very near the OSIRIS format.

At the same time we are running tests with scanner and OCR software. This is a promising perspective for the future, thinking of a more "machine"-supported approach in producing full-text documentation.

**D. Processing a Codebook - Further Steps**
The second step producing a codebook goes along with the processing and cleaning of the data. The end of this step is a data-proofed documentation, which describes the (completely processed) data-set (and which is not meant to be a "data handbook" or something else). The tools used in this second step are partly original OSIRIS programmes (FBUILD, FMRG, etc.), some are IBM utlities and software, some are from the DDA (MERGET, SLABGEN etc.), and some are ZA software (CDBK-PRT etc.). The ZA tool CDBK-PRT can cope with tasks, which OSIRIS cannot solve:

- merge SPSS CROSSTABS output into the machine-readable codebook-file (e.g. for international comparative studies, cumulative files etc.)

- merge frequencies with more than 4-digits per answercategory

- can print as well crosstabels as univariate frequencies in one variable

- considers layout-parameters like bold-printing, underlining, page-formatting etc.

**E. Codebook Output**

The "traditional" output from a codebook (OSIRIS format or whatever) was on paper. But the amount of workload, time, and paper for producing all the codebooks requested grew and grows beyond reasonable borders. On the other side more and more users ask for machine-readable codebooks on floppy- disks. Thus the (above mentioned tool) CDBK-PRT allows to output a codebook in different ways:

- print a codebook on the mainframe line printer (for the purpose of proof reading the documentation during the production process)

- print a codebook on a laser printer using the PRESCRIBE language (for the purpose of duplicating it on a photocopy-machine internally or at copy shops outside the ZA)

- print a codebook on a laser printer using the POSTSCRIPT language (for the purpose of duplicating it on a photocopy-machine internally or at copy shops outside the ZA)

- copy the codebook output to a disk file in the POSTSCRIPT language (for external users who have access to POST-SCRIPT and want to reproduce the whole or parts of the codebook at their place)

- copy the codebook output to a disk file in ASCII-format only with carriage-control characters for page-feeds. (for external users who want to reproduce the whole or parts of the codebook using a simple printing routine or who want to import the machine-readable documentation into a text processing software.)

ZA's experiences with the distribution of codebooks on disks and with the responses from the side of the users are not yet very systematically. But the fact that people who once received a machine-readable copy start to request codebooks on disk also for other studies is a promising development.

**F. Changing Profiles**

One critical point must be mentioned at this place: Once we have started to distribute codebooks on disks the feedback about the number of usages of the documentation will probably be decreasing even though more people might have access to the codebooks (e.g. in PC pools) than before. This means however that we need other arguments for the legitimation of the archival work to the funding organisations.

The profiles seem to change, as well the profile of our services as the profile in the structure of the demands of our users. (The question: Which side influences the other? should not be discussed right here.) Additionally services and demands become more and more international, which means at the same time that international (inter-archival) cooperation becomes more and more important.

**G. Introduction to the Second Part**

The second part of this paper now tries to define tasks and demands for "the codebook" in general (processing and formats), considering "tradition" and perspective. This list is, in the present form, a rather subjective collection of items but it is meant to be a help for the ongoing discussion and an offer for additional ideas.

Collection of items for a (new) codebook production-line (Questions: What is a codebook? What is it used for?);

*information on the variable level: technical desription of a data-set:*

in terms of a "raw-data file" (position, width, deck etc.);

in terms of an SPSS (or other) system file (Variable name, - label, definition of missings, decimal places, variable type etc.);

questionnaire information (question text, answer- categories, other information from the questionnaire, comments from the archive) da:a information (frequencies - weighted or unweighted -, crosstabs, other statistics, graphics);

less redundancy in the codebook example: a list of "dummy-variables" could be arranged in a way that redundant information is left out and only the relevant fre- quencies are documented in formats like: tables, graphical presentation etc. to allow readers of the codebook to look at the relevant information at the first sight;

*general information about the study:*

preface (study desription, index, list of variables, copyright information, how to cite a codebook)

appendix (notes explaining special variables, copy of the original questionnaire)

*purpose of a codebook:*

information for the (secondary) analysts, who work with the data

input for information retrieval

data-handbook for people, who don't use the survey data but only the codebook???

*media:*

codebook on paper (in-house production, printed at an editing house, copy-shop)

lay-out (fancy looking - or - plain printing routine?)

codebook on tape/disk/other media/via file-transfer (which printing-routine, form feed characters, comparability problems)

*exchange format:*

codebook on tape, disk, other media, via file transfer (with the same format in each participating archive, special software for processing, formatting, and producing/reproducing output on paper, on disk etc.)

*codebooks in a general concept:*

codebooks (as defined above) plus data plus general background information (aggregate data, maps...) plus retrieval-system plus print and other output options plus bibliography

on CD-ROM or whatever can be thought of as a "service package" containing everything which a user might need

*producing codebooks:*

tools for mainframe, PC (MS/DOS, OS/2) or workstations(UNIX, AIX..) to produce the codebooks in the exchange format

1. Paper presented at the 1993 IFDO/IASSIST Conference Edinburgh, May 1993