BID - Bringing Integration to Data

by Karsten Boye Rasmussen [1]
Dansk Data, Denmark.

Abstract
The purpose of data archives is not only to store data materials for posterity, but also - and equally important - to fulfil the needs of the present users. The growing demand for studies along with a number of other factors force the data archives to change their strategy for retrieval and dissemination of studies in order to serve their customers in the best - and cheapest - way possible. This calls for effectiveness and new thinking in the retrieval and dissemination procedures.

This paper outlines the state-of-the-art at the Danish Data Archives (DDA) by describing the format of machine-readable records and the present utilization of machine-readable documentation at the archive. A projection of the growing service rate shows the pressing need for a more effective system for direct servicing of the users, the idea being to integrate the study description, the variable documentation and the data and as the distributing medium use CD-ROMs, as the production costs are reasonable. The remaining problem is obtaining or financing the development costs of a suitable retrieval system. These will most probably be rather high, and as probably a lot of archives have the same need for a retrieval system, it could be a good idea for the archives to cover the development costs between them.

Background: Machine-Readable Records
To ensure the right perspective I shall shortly outline the records kept at the DDA[2]. The data are social science quantitative investigations (typically surveys). After the processing at the archive of the raw data and the (typically) paper documentation (questionnaire, coding instructions, reports etc.) each survey consists of three files:

Documentation of the study (study description)
Rectangular (flat) character data file
Documentation of the variables
The files of the finished survey

Study description
I shall not go into details about the study description. Discussions concerning item numbers, subitem numbers and especially the introduction of new item numbers have been the subject for discussion at many earlier IASSIST conferences. The "Standard Study Description Scheme" has gradually been improved[3], but it is still close to the original scheme agreed upon more than ten years ago[4]. The scheme is used at several archives[5]. At the DDA all study descriptions are maintained in both a Danish and an English version.

The actual format and the many items makes the Standard Study Description scheme a rather complex instrument. The design is specifically intended to be a machine-readable record at the study level. The study description format contains coded as well as text information. It is of importance that there is no limitation to the amount of text in the text fields.

```
TITLE    Title of study
CLASS    Ready-made for analysis or just deposited
ACCESS   There may be restrictions
YEAR     The year the survey was carried out
CASES    The number of cases in the survey
NATION   Which country or countries
DONOR    The depositor of the data
INSTITUTE    Place of employment of the depositor
SPONSOR      Finance (Social Science Research Council)
COLLECTOR    Who carried out the field work
ABSTRACT     Description in free text
KEYWORDS     Controlled vocabulary
```

Areas of the study description

The data file

Data stored at the DDA are kept as rectangular character data files (not as system files). One record per case. Interlinked datasets are either divided into separate files (like tables in a relational database) or patted into one (greatly redundant) file. The data for each observation is placed as a record in the data file. And each record holds information of the variables stored in fields (fixed columns). The data per se will not be of any use without the documentation.

```
02091203060203100390030056602011061103902010
02081210101010102002005110801102110310 3100
02063203030102010330062034408011051103302000
02071203060203030330050052301011041103103001
02071201050202010130020051308011011103306000
01044204050204040440022021208011011103101000
01071201050203010190010054308011051103103001
02051299101010040490023031405011061103203001
02072301010102010190110063208011011103302000
02043204050203030430030012208011041103202010
02042312050203121290011022108011021103203000
01034199101010999910030011105011061103201000
02051204020103010430010012408011051102303000
02061204020103020430010053408011011103303000
01073203010102030390990054408011011103103000
```

Character data file

Documentation of the variables

The character data file needs documentation in order to be of any use in its own right. Without the proper documentation the user is not able to distinguish one flat file from another. In order to exploit the information in the data file we need variable documentation. The variable documentation can be thought of as a relational table where each field describes certain characteristics of the variable. This idea is already in production in several databases. In IBM's SQL database[6] several system tables keep track of the information stored in the database. An SQL database may thus be viewed as a complete data archive, where each study is a separate table. The available tables are then stored in the SYSTABLE-table and the documentation of the fields of each table (data) is stored in the SYSCOLUMNS-table.

The project of making a complete description of the fields in the variable documentation lies outside the scope of this paper, so the list below shows only a selection of possible fields. The list shows that the documentation provided by the most widely used social science data packages (SAS and SPSS[7]) does not have the necessary facilities for a complete documentation. Both packages support only the first part of the field list.

```
NAME    The variable name or number
LABE    Short text description
PLACE   Extract the data from these columns
MISSING     Definition of missing data
FORMAT      Output format / categories

QUESTION    The complete questionnaire text

STUDY   Identification of the study
QID     Identification in original questionnaire
FILTER  Reference to filtering
CODINGSpecial coding instructions
PROCESSING  Unofficial comments
CHECK   Checking performed
CHECKLOG    Generated notes on the checking
...

Fields for a variable definition
```

The most important field for the user is the QUESTION field consisting of free text description exactly as it appeared in the questionnaire. This field provides opportunity for performing full text retrieval. Neither SAS nor SPSS are capable of storing and utilizing the complete questionnaire text. The only field for giving a text description of the variable is the variable label. The label in both packages is in practice[8] limited to 40 characters resulting in definitions like:

```
LABEL  INCOME =

'PRS MON AVE INC MAIN-JB -TAX DK 894-904' ;

The variable label, 40 characters
```

40 characters is not much! The variable INCOME actually covers "Personal average monthly payment on main job, after taxation, in Danish Kroner, April 1988-April 1989". This is certainly not a lot of text for a questionnaire text or a created variable, but even then this cryptic label presented above is a common result when the text is restricted to 40 characters.

This is one of the reasons for the DDA to use the data-archaic format provided by OSIRIS[9]. The format was developed by the ICPSR. The virtues of the OSIRIS codebook format is that there are no limitations to the amount of description for each variable. There can be several lines describing the variables and the categories. There can even be references to notes.

The format uses a sort of tag in the first column of each "card" - yes, it is that old - to identify which portion of the variable documentation is described:

T	**Label, columns, missing**
Q	**Question**
X	**Explanation**
J	**Unofficial comments**
F	**Frequency**
C,B	**Category text**
G	**Reference to note**
S,E	**Introduction**
M	**Note text**

The OSIRIS codebook types

Using the format will result in a variable description looking like the following. This is certainly not as straightforward as the syntax used by SAS or SPSS. But normally this lay-out is not written by human beings. A machine-readable format of this complexity and rigidity is best written by machines. In the actual production of machine-readable documentation at the DDA a pre-processor software is used.

```
T0093 EEC VOTE TODAY          018400020   0100000090000009
Q00930093     If there was a referendum on joining the EEC today would
K0093    you vote yes or no?
X0093    In the Danish Election Study, 1975 "Don't know" is
K0093    included in "4. Don't want to answer".
J0093    VARIABLE BASIS
K0093    PRE 1971: -
K0093    POST 1971: -
K0093    DES 1973: 157
K0093    DES 1975: 90
K0093    DES 1977: 211
K0093    DES 1979: 166
K0093    DES 1981: 259
X0093  2     1971-1 1971-2  1973   1975   1977   1979   1981
K0093  ------------------------------------------------------
K0093    1.    -      -     45     37     30     41     41
K0093    2.    -      -     44     38     50     41     40
K0093    3.    -      -      1      3      3      2      2
K0093    4.    -      -      0     13      2      1      1
K0093    5.    -      -      9      -     15     13     16
K0093    9.    -      -      1      9      1      2      1
K0093    11.   100    100    -      -      -      -      -
K0093  ------------------------------------------------------
K0093    WGT N = 1302   1302   533    1600   1602   3192   1500
K0093
C0093    255801. Would vote yes
F0093    UNWEIGHTED:          2558 WEIGHTED:          3231
C0093    280702. Would vote no
F0093    UNWEIGHTED:          2807 WEIGHTED:          3542
C0093     15003. Would return a blank ballot paper
F0093    UNWEIGHTED:           150 WEIGHTED:           189
C0093     28104. Don't want to answer
F0093    UNWEIGHTED:           281 WEIGHTED:           297
C0093     68805. Don't know
F0093    UNWEIGHTED:           688 WEIGHTED:           931
C0093     20809. Not ascertained
F0093    UNWEIGHTED:           208 WEIGHTED:           237
C0093    260411. The variable is not included
F0093    UNWEIGHTED:          2604 WEIGHTED:          2604
```

Example of OSIRIS codebook[10]

The OSIRIS software itself has not been used for many years at the DDA, only the format lives on. But around this format a great many procedures and programs have been developed for checking the data against the documentation, for presenting the documentation, and for further utilizing the documentation.

Utilization of Machine-Readable Documentation

The main reason for the production of machine-readable documentation is the reuse of information. Safely stored documentation can be used in a variety of ways. This chapter describes the current activities at the DDA. The present status implies

some inexpediencies that will be clarified in details in the next chapter.

Printout
Conversion to other formats for analysis
Retrieval
Dissemination

The utilization of machine-readable documentation

Most of these evident virtues of machine-readable documentation goes for the study level (the study description) as well as the variable level (the codebook).

Printout
In compiling a complete documentation a printout of the study description in a readable format (for human beings) is placed as an introduction to the codebook. Printouts of several studies form a catalogue. As the study descriptions are rather voluminous, a catalogue of study descriptions normally includes only the most important items, but the entries are heavily indexed (by persons, subjects, and keywords). However, it is expensive to produce printed catalogues, and it is difficult to keep up with the immediate obsolescence of the catalogue.

To make a printout at the variable level is rather straightforward. The problem is to make the documentation as accessible as possible to the user. Most of the inconvenience of the rigid OSIRIS format is overcome by getting rid of the redundant information in the codebook[12]:

DDA-0658 Danish Election Studies, Continuity File 1971-1981

VAR. 93 EEC VOTE TODAY

start pos. 184, missing data: = 9 or >= 9

If there was a referendum on joining the EEC today would you vote yes or no?

In the Danish Election Study, 1975 "Don't know" is included in "4. Don't want to answer".

	1971-1	1971-2	1973	1975	1977	1979	1981
1.	-	-	45	37	30	41	41
2.	-	-	44	38	50	41	40
3.	-	-	1	3	3	2	2
4.	-	-	0	13	2	1	1
5.	-	-	9	-	15	13	16
9.	-	-	1	9	1	2	1
11.	100	100	-	-	-	-	-
WGT N=	1302	1302	533	1600	1602	3192	1500

unweight MD

marg % %

2558 28 39 01. Would vote yes
2807 30 43 02. Would vote no
150 2 2 03. Would return a blank ballot paper
281 3 4 04. Don't want to answer
688 7 11 05. Don't know
208 2 . 09. Not ascertained
2604 28 . 11. The variable is not included

9296 100 99 Weighted respondents: 11031

Conversion to other packages (SAS and SPSS)
Very few users - if any - make their analyses on delivered datasets using OSIRIS. Instead most users in Denmark use SPSS or SAS. To the user the OSIRIS codebook produces only a printed codebook. The ability of software programs to read system files from other packages has appeared to be quite unstable. With every new release and every new operating system platform this conversion often proves to be the tricky part.

However, with the OSIRIS codebook being a machine-readable variable documentation it is possible to convert the documentation file without affecting the data file. The rigid format of the codebook makes this a relatively straightforward process. Presently the DDA uses a micro computer program called OSI-SPC[13] for doing the conversion. The idea is to leave the data file unaltered. The same data file can be described by both OSIRIS, SAS and SPSS. The user then receives the OSIRIS documentation and the data file, and with the help of the conversion program, the user is able to convert the documentation to the preferred package. And the conversion will produce a SAS-program or SPSS-controlcards, which the user would otherwise have had to write himself.

```
TITLE   'Danish Elections, Continuity file 1971-1981' .
DATA LIST           FILE='L:\U0658\MINI.DAT'
                    FIXED TABLE /
                VAR3 8 VAR93 184-185 .
VARIABLE LABELS VAR3 'SURVEY ID
                VAR93  'EEC VOTE TODAY
VALUE LABELS
        VAR3 1      'Danish Pre-Election Study, 1971'
             2      'Danish Post-Election Study, 1971'
             3      'Danish Election Study, 1973'
             4      'Danish Election Study, 1975'
             5      'Danish Election Study, 1977'
             6      'Danish Election Study, 1979 (no. 20)'
             7      'Danish Election Study, 1979 (no. 21)'
             8      'Danish Election Study, 1981' /
        VAR93 1     'Would vote yes'
              2     'Would vote no'
              3     'Would return a blank ballot paper'
              4     'Don t want to answer'
              5     'Don t know'
              9     'Not ascertained'
             11     'The variable is not included' .
SAVE                FILE='L:\U0658\SPSSMINI' .


SPSS/PC+ generated program from OSI-SPC
```

```
TITLE 'PRSETUP  Two variables from DDA-0658' ;
TITLE2 'Danish Elections, Continuity file 1971-1981' ;
LIBNAME LIBRARY 'L:\U0658' ;
PROC FORMAT LIBRARY=LIBRARY ;
VALUE  VAR3L      1 = 'Danish Pre-Election Study, 1971'
                  2 = 'Danish Post-Election Study, 1971'
                  3 = 'Danish Election Study, 1973'
                  4 = 'Danish Election Study, 1975'
                  5 = 'Danish Election Study, 1977'
                  6 = 'Danish Election Study, 1979 (no. 20)'
                  7 = 'Danish Election Study, 1979 (no. 21)'
                  8 = 'Danish Election Study, 1981' ;

VALUE  V93L       1 = 'Would vote yes'
                  2 = 'Would vote no'
                  3 = 'Would return a blank ballot paper'
                  4 = 'Don''t want to answer'
                  5 = 'Don''t know'
                  9 = 'Not ascertained'
                 11 = 'The variable is not included' ;
        RUN ;
        LIBNAME  U0658  'L:\U0658' ;
        FILENAME DATAIN 'L:\U0658\MINI.DAT' ;
        DATA U0658.MINIDAT ;
        INFILE DATAIN LRECL=185 ;
        ATTRIB     VAR3       LABEL = 'SURVEY ID          '
                              LENGTH = 3   FORMAT = VAR3L. ;
        ATTRIB     VAR93      LABEL = 'EEC VOTE TODAY     '
                              LENGTH = 3   FORMAT = VAR93L. ;
        INPUT    VAR3 8 VAR93 184-185  ;
        RUN ;

SAS setup generated by OSI-SPC
```

The development of OSI-SPC is part of the archive policy. The OSIRIS codebook format is used because of its unlimited capabilities for storing text description. With the help of conversion program(s) the needs of the users of the future can be fulfilled as well. To support new analysis packages we need only to make adjustments to the software (contrary to developing a new conversion program) to stay in pace with the development of new software packages.

Documentation retrieval

I reckon that all archives make use of some kind of retrieval system (possibly many systems) in order to search and identify datasets of interest to the users. At the DDA we have made several systems at the dataset level as well as at the variable level.

The system used for searching study description excerpts (DDAGUIDE) has the most extensive documentation and it is the

only system available to outside users. DDAGUIDE is running on a mainframe host[14], and uses a dialect of the Common Command Language (CCL, promoted by the EEC). Retrieval at the study level is the first step in searching for a suitable dataset:

```
FIND  YEAR>1981 CASES>1000 WORD=TV. INDEX=ELECTION

YEAR>1981              Study carried out after 1981

CASES>1000         With more than 1000 cases

WORD=TV.               Words beginning with "tv"

INDEX=ELECTION         "Election" found as a keyword
```

DDAGUIDE retrieval in study descriptions

This request can be regarded as a construction of 4 sets, which are all combined (with logical AND) in a resulting set (possibly empty). Normal Boolean logic (AND, OR, ANDNOT) can be applied for building new sets. The result is a vector of study numbers, and then the codebooks for these studies must be searched to secure that they contain variables that come close to the user's need.

Dissemination of studies
At present studies are distributed to the user by means of many different media: mainframe tapes, diskettes, and electronic network as preferred by the user. Compared with the situation only a few years ago a lot of the analysis is now taking place on micro computers, and consequently many users prefer the data to be delivered on diskettes.

The Pressing Need for more effective Servicing
The growing demand for delivery of studies forces the DDA to make the retrieval and selection as well as the practical dissemination more effective. In this chapter I shall take a look at the present technical obstacles that make the service less effective.

Both the retrieval and the dissemination are carried out at the DDA and require a lot of manpower at the archive. About 25 pct. of the work of the archive is devoted to servicing[15], and because of the growing rate of data deliveries this number is expected to increase. Because no extra funding is available the success of data deliveries will drain the potential for processing new studies to becoming available in fully documented form. It is thus a necessity to make the retrieval and dissemination procedures more effective.

More effective retrieval tools
The present retrieval tools suffer from the following drawbacks:

Dispersed retrieval facilities

DDAGUIDE searches only study descriptions

Updating is irregular and cumbersome

Retrieval only available on one specific mainframe

No thesaurus (actual words searched)

Line mode user interface

Drawbacks of presently used retrieval tools

Having several poorly integrated and sometimes individual retrieval facilities makes it more than a one-man-job to select the appropriate studies. Using DDAGUIDE will narrow the number of studies to be investigated further, but these studies will have to be scrutinized at the variable level. This means that the codebooks must be searched, but at present no system is available for searching the codebooks, therefore the most obvious choice is to ask the people at the archive responsible for the selected studies. Secondly, some studies (series of studies like Gallup polls etc.) have some extra machine-readable indexing of variables, but these indexes are not integrated into the codebooks and are made by a third person. This shows poor integration and a massive use of manpower, which eventually introduces the risk of performing erroneous retrieval.

The solution is first to integrate the index with the codebooks. The retrieval of codebooks must then be integrated with the retrieval of study descriptions. It ought to be possible to select a set of studies at the study level. Furthermore the variables of these studies should be searched within the same system. Superficially there is not much difference whether the unit of retrieval is a study or a variable. But this proves to be wrong, as this example illustrates[16]:

SET1: Housing

SET2: City

1 AND 2: Variables with both "Housing" and "City"

VAR: found within the same single variable

STUDY: found 2 variables within same study

Combination (AND) of retrieved variables

A retrieval system for codebooks has to have a kind of extra parameter to distinguish whether the combinations of variables take place at the variable level (both subjects found within the same variable) or at the study level (two variables covering both subjects found within the same study).

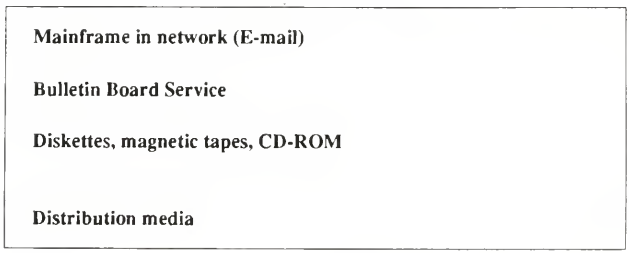
Updating text databases presents major problems. Adding new studies or variables is no problem, the obstacle exists in replacing a text entry. This is caused by the most used algorithm where the actual text is subdivided and scattered throughout the data base. The simplest solution is to build the data base from scratch every time replacement or updating is required. This demands computing power, which in turn is becoming still less expensive. Earlier retrieval data bases were typically placed on large mainframe hosts. The performance of recent microcomputers makes this kind of machine a perfect choice for the integrated retrieval system.

Unfortunately an integrated system for retrieval at both the study and the variable level is not available at the DDA at present. Some of the design - mostly in the form of wishful thinking - is presented in this paper, but the actual development requires funding. We are not so stubborn as to insist on making this development. Presently we have not come across a system fulfilling our demands, but we try to experiment with systems and we are looking forward to the perfect system being presented. Also a lot of archives must be investigating into the same problems, so maybe a more integrated and common effort is required to reach the goal.

Bringing the Data to the User

If we assume that the user has decided on which studies he wants, the distribution of studies to the user can take place in a lot of different ways.

Presently the DDA places the studies on the requested medium (diskettes, mainframe tapes or using electronic mail). This demands active work from the employees at the DDA. But in addition to sending data to the user there exist several other distribution methods that require a more active role of the user.

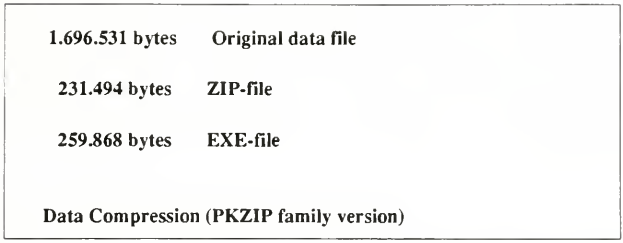


Mainframe in network

If the users are all connected to the same machine the solution is simply to place the archive files on this computer. This is the solution of campus-archives, but it is seldom efficient for archives with national coverage. However, as mainframes are being connected with networks this solution has turned out to be a big success. Viewed from the ICPSR, access through network (CDNet) in 1989 "accounts for almost three quarters of the ICPSR data orders"[17]. The success is totally depending upon the users' access to, knowledge of and familiarity with the networking procedures. Although the DDA is the national data archive for Denmark, in a lot of respects it would be a great mistake to compare the DDA and the ICPSR. In this context the big difference is that the ICPSR is serving trained staff at member institutions. This staff is then handing out the data to the users, which on their side are accustomed to using the campus computer. At the DDA we are serving the user directly. The users are using a lot of different mainframes and are seldom aware of nor interested in the networking techniques.

Bulletin Board Service

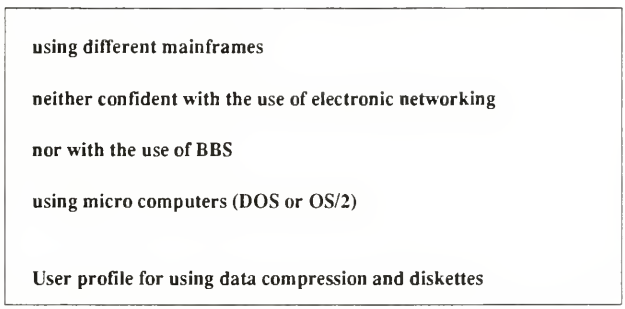
All the users can be expected to be using some kind of micro computer. It would seem rational to make the data archive holdings available on a remote basis for PC-users. The most common form to be utilized is then running a Bulletin Board Service (BBS) with download facilities. However, most of the users can not be expected to possess the necessary technical facilities (especially modem connections). As a side effect, the investigations into BBS have drawn our attention towards data compression techniques. In a recent BYTE article comparing data compression software[18] the virtue of data compression is seen as a saver of space on the hard disk, but for long the greatest potential for data compression has been the dissemination of data via BBS by modem (normal modem without built-in data compression facilities). However, the data compression technique is fully applicable to diskettes. This example shows what is saved by using the PKZIP[19] data compression:



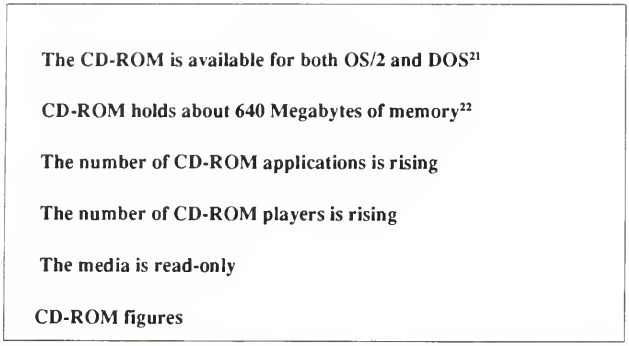
Due to the limited variation of the bytes in the data file the compressed data file gains 86 percent of the space required by the original file. This means that a file that is too big for a single one of the largest diskette formats available (IBM 1.44 Megabytes) will now fit an old floppy disk (360 Kilobytes). The ZIP-file is converted back to the original format by an unpacking program. To make sure that the user will be able to unpack the ZIP-file - even if the user does not possess the UNZIP program - the ZIP-file can be extended to a self-unpacking program (an EXE-file produced by the ZIP2EXE program). This will cost around 30.000 bytes. Furthermore there exists a family version (a program that is capable of running both under DOS as well as OS/2), so the produced EXE-file can be unpacked in any of the two environments.

By using the most common baud rate (2400 baud) it would take approximately 20 minutes to download the EXE-file.

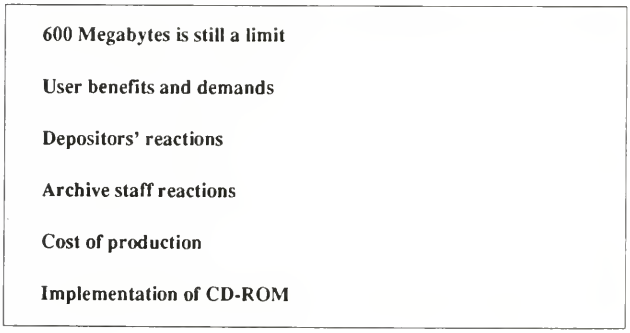
Diskettes with data compression
Presently data compression combined with the mailing of diskettes seems to be the best alternative for the dissemination of a minor collection of studies to users:



The Data Archive on a Disk (CD-ROM[20])
The real potential of a data archive lies in the user's opportunity to perform comparative analyses on more than a single study. This means that most secondary analyses will need several studies, and that a single diskette will prove insufficient. A radical solution would be to put the complete archive on a disk which could be distributed. This has become relevant with the marketing of optical disks. Also WORM-disks exist in a lot of different and incompatible formats. But with the growing number of CD-ROM players the CD-ROM medium seem to be the standardized and perfect solution for bringing out the archive to personal computers.



Normally read-only media are thought of as some kind of second class media, but read-only is the perfect attribute for the stable archive data. In this chapter I shall present some storage considerations as well as some different viewpoints on the implementation of a CD-ROM solution.



Is 600 Megabytes enough?
Lets take a look at the size of data stored at the DDA. In the following table the unit is files archived at the DDA. But each file is in addition weighted with the number of bytes that it occupies. Totally we are talking about 5398 files occupying 2953 Megabytes or 3 Gigabytes. The main archive is placed on mainframe tapes.

Each file belongs to one of the categories: Finished, Process (being processed, i.e. an intermediary file that is stored for extended security) or Original (the file received). Likewise each file also belongs to one of the package-categories (OSIRIS, SPSS or SAS) and finally the file contains either DATA or documentation (DICT/DICB/MISC)[23].

Of these files only a small fraction of the 3 Gigabytes are of interest to the CD-ROM project namely studies finished and stored in an OSIRIS version: At present 586 data files (341 Megabytes) together with the documentation files (491 DICBs (dictionary-codebooks) totalling 94 Megabytes and 546 DICTs totalling 6 Megabytes[24]). At present these approximately 500 finished studies will occupy around 450 Megabytes. A CD-ROM will hold from 540 to 640 Megabytes.

		STATUS						All	
		Process		Finished		Original			
		N	SUM	N	SUM	N	SUM	N	SUM
OSI									
	CDBK	21	4 Mb	.	.	20	4 Mb	41	9 Mb
	DATA	480	412 Mb	586	341Mb	1412	1417 Mb	2478	2171 Mb
	DICB	21	12 Mb	491	94 Mb	125	41Mb	637	148 Mb
	DICT	302	3 Mb	546	6 Mb	227	6 Mb	1075	16 Mb
	DIV	448	158 Mb	40	11 Mb	80	48 Mb	568	218 Mb
	ALL	1272	590 Mb	1663	455 Mb	1864	1519 Mb	4799	2565 Mb
SAS									
	DIV	2	0 Mb	9	0 Mb	1	0 Mb	12	0 Mb
	ALL	2	0 Mb	9	0 Mb	1	0 Mb	12	0 Mb
SPSS									
	CTRL	30	2 Mb	90	9 Mb	104	9 Mb	224	21 Mb
	DATA	13	35 mb	83	133 Mb	146	184 Mb	242	353 Mb
	DIV	7	0 Mb	5	0 Mb	109	11 Mb	121	12 Mb
	ALL	50	38 Mb	178	142 Mb	359	206 Mb	587	387 Mb
ALL									
	CDBK	21	4 Mb	.	.	20	4 Mb	41	5 Mb
	CTRL	30	2 Mb	90	9 Mb	104	9 Mb	224	21 Mb
	DATA	493	447 mb	669	475 Mb	1558	1602 Mb	2720	2525 Mb
	DICB	21	12 Mb	491	94 Mb	125	41Mb	637	148 Mb
	DICT	302	3 Mb	546	6 Mb	227	6 Mb	1075	16 Mb
	DIV	457	158 Mb	54	12 Mb	190	60 Mb	701	231 Mb
	ALL	1324	629 Mb	1850	598 Mb	2224	1725 Mb	5398	2953 Mb

At least 100 Megabytes of the space on a CD-ROM would be left free. And with compression techniques the required space would be less than 100 Megabytes. So a CD-ROM disk could hold approximately 2000 surveys and still have plenty of free space.

Bringing retrievable documentation to the user

For each study the documentation should be available to the user as well. But the user - and the archive staff - need a retrieval system integrating the documentation at the study level as well as at the variable level as mentioned above. The retrieval demand is the reason for setting aside some free space on the CD-ROM for the distribution of retrieval software, indexed files, and other assisting materials.

Integration of study and variable level documentation

Subsetting of variables

Conversion to other packages

Logging of search criteria

Inclusion of search criteria

Human interface

User needs for integrated retrieval

Even though a CD-ROM can hold a lot of information there is no guarantee that the user has similar abundance of space available on his hard disk. Often the user will need only selected variables from a study. In addition to this type of selection, the user should have the option of transferring the background variables. This calls for the possibility to mark off the background variables in the documentation. Apart from hardware limitations there may be some software limitations[25] too. A lot of patience is required when building datasets with a great number of variables on a PC[26].

Retrievals may become quite complicated, so it is a necessity that the user is given the possibility of tracking down the searches and combinations. A log will provide documentation of the actual retrieval for documentation. Similarly a session should be able to start off where a previous session was left.

The line-mode CCL search syntax is developed for dumb terminals. But the increased use of micro computers has accustomed users to more intuitive search systems. Many search systems now employ pull-down menus as well as windowing systems[27]. The graphic user interface standardized in IBM's CUA[28] uses radio buttons, push buttons, check boxes, list boxes etc. as well. At present I have no knowledge of the CUA standard used in an actual implementation of retrieval software, but probably it has already been marketed.

As most users are conservative about learning new computer languages for analysis it should be possible to convert the documentation to the user's preferred package as exemplified earlier in this paper. An integration of a (new) analysis package could be counter productive for the user and would definitely involve some further costs.

No answer to questions like:

"When was the proportion of Social Democrats among men higher than among women?"

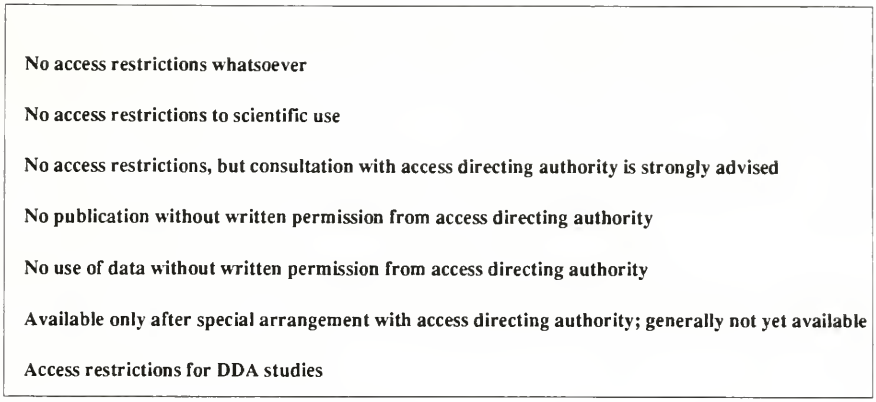
Analysis can be carried out in other processes (OS/2)

With the analysis package preferred by the user

Analysis package separated from retrieval

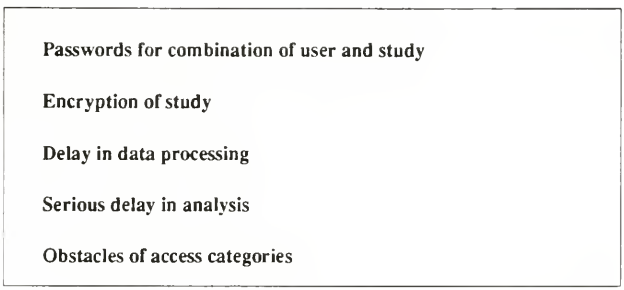
Depositors reaction

The studies stored at the DDA are not all directly available to the user. Although the data are placed at the archive, the depositor still has the formal rights to the material, and the depositor has the possibility of assigning access categories to the dataset. The access categories are assigned to the data only, the documentation is always available without special permits.



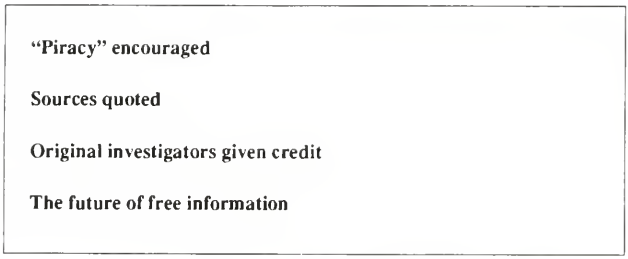
One solution to this problem would be to incorporate only studies without access restrictions (the first three categories) in the BID-project. As this would be a dramatic solution it is not advisable. Another solution would be to work for a transfer of all studies to the category of free access. But this work has already been carried out in so far that most studies after a while are transferred to a less restrictive category. But even though the data are freely available, the depositor (the access directing authority) is presently being informed of whom the datasets are disseminated to. All in all this implies some kind of password, so that it will be possible for the user only to access files that he has been positively assigned to.

The password ought to be distinct for every combination of user and dataset. But user information can not be placed on the CD-ROM, and the password would then have to be distinct only for the dataset. As dataset protection passwords are not a standard feature of the operation systems for micro computers, the passwords would be implemented in the form of a key for the encryption of the data file. A user would then be able to pass-on the key to other users. This shows that the security is not perfect, but then it does not have to be perfect. The same thing is happening now, when a few users are actually distributing the data they have received from the DDA. This is in contradiction with the agreement between the archive and the user and therefore illegal. But unless the analysis system is an integrated part of the retrieval system the spreading of data can not be prevented.



But every restriction introduces drawbacks for the user. First of all the data would have to be both decompressed and decrypted unless a program would be able to do this in a single pass. Secondly - and most seriously - the user would have to contact the archive to receive permission. This would delay the actual analysis by several days for the studies placed in the most restricted categories that requires the interaction of the depositor.

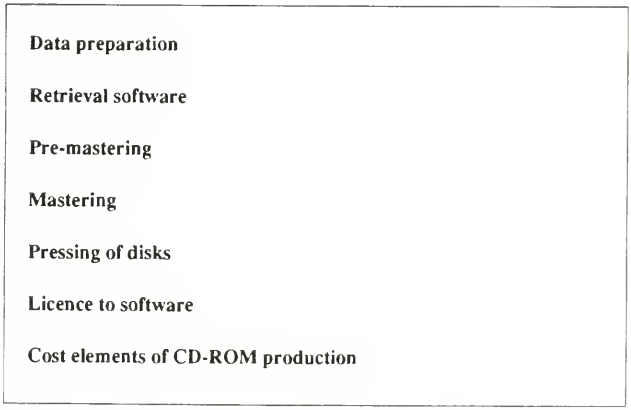
As proven above it is impossible to totally prevent the "pirating" of data. And in my opinion it should be legalized and encouraged. The data and documentation of social science research - which in Denmark is most often funded by the State - should be regarded as public property. Any use of machine-readable material should imply the same consideration as the use of other sources of material. This means that all sources should be quoted, and at the same time the original investigators given credit.



Archive staff reaction
The BID system is intended for the user. But the system should be used at the archive as well. At the archive it would be possible to maintain a more updated version in order to give access to the newest studies.

It must be foreseen that some staff work would be transferred into giving advice on the use of the BID CD-ROM system. But a major part of the present service work at the archive would disappear and leave more resources for bringing up studies to the highest documentation level.

Cost of production
The production costs of CD-ROMs have gone down very fast over the last year. Recent announcements mentions prices as low as 30.000 DKr[29] for the total production of 300 CD-ROMs. The cost would be expected to be even less in the US. But recently a figure of USE 10.000 (70.000 DKr) for the complete process has been mentioned[30].



The difference in pricing is caused by the exclusion or inclusion of some steps in the process. The low costs are obtained if you simply have some files and want them available on a CD-ROM. The receiving company will then do the mastering and pressing of the disks. So this is comparable to the actual printing costs.

If all data and documentation are available the expensive process will be to develop or apply retrieval software. Ready-made retrieval tools for CD-ROM production are available. For companies using computers the software can be acquired, indexes constructed, and the system tested locally. The technical requirements are a medium-sized PC, lots of disk space (maybe optical) and some sort of medium for copying the large quantities of data to the mastering company (tape).

Do not forget that software is under the law of copyright. This means that although you can locally set up a nice system using the software, you are not allowed to put the retrieval software on the CD-ROM. A licence or royalty fee is required in order to distribute the retrieval software to the users.

Conclusion
A lot of different companies and software packages are available to chose among when deciding on the retrieval software. But the "plastic"-CD-software I have seen have all been limited to typical library systems. They could only handle the retrieval from one level of information (Eg. bookcards) and not the hierarchy of studies and variables

Retrieval on two levels (study and variable)

Start of other external processes:
- **conversion to data analysis packages**
- **decrypting the data file**
- **decompressing the data file**
- **subsetting the data file**

Storage of search profiles

The need for open data archive software
***is* The price of integration**

This need for retrieval software that can be further developed is arisen from the idea to integrate the study description, the variable documentation and the data. The conclusion of this paper is that the total integration of these parts makes it even more pressing to make data, documentation and the retrieval software freely available without bureaucratic hindrance.

This paper has been an advertisement for a retrieval instrument open for further development and not the introduction of the product from a concluded development. But the need must be very similar at many data archives, and there should be a potential for covering the development costs between the archives.

Footnotes:
[1] Presented at the "IASSIST 90" Conference held May 30 - June 2 at the Radisson Hotel, Poughkeepsie, New York, U.SA.

[2]. The DDA documentation standards were earlier presented in my paper "Data on Data", in "SUIGI'89", Proceedings of the SAS European Users Group International Conference, SAS GmbH. The BID project has in an earlier version been described in "Beskrivelsens Integration med Data" in DDA-Nyt 48 (in Danish).

[3]. "Standard Study Description Scheme". Latest update 1988, available from the DDA (DOC00364)".

[4]. "Study Description Guide and Scheme" by Per Nielsen, Copenhagen, DDA, 1975.

[5]. The use of the standard study description - and similar vehicles - among European archives described in: "From Localization to Cataloguing of Data Sets". (Workbook of the First CESSDA Expert Seminar). Ed. Astrid Bogh Lauritzen, 1987, Danish Data Archives, Odense, Denmark.

[6]. "IBM OS/2 EE 1.1. Database Manager Programming Guide and Reference", 1988, IBM. (Appendix C). The Database Manager is part of the Extended Edition for OS/2.

[7]. "SAS Language Guide for Personal Computers. Version 6 Edition". 1985, SAS Institute Inc., Cary, NC, USA. "SPSS/PC+ version 2". 1988, SPSS Inc., IL, USA. The Mainframe versions of SAS and SPSS do not differ significantly with respect to documentation facilities.

[8]. "SPSS-X User's Guide" 3rd edition, SPSS Inc., Chicago. SPSS-X supports up to 120 characters, SPSS-PC+ (now version 3) places the limit at 60 characters, but most procedures print out only the first 40 characters.

[9]. "OSIRIS III. Volume 1, System and Program Description" 1973, University of Michigan, USA. (Appendix D describes the OSIRIS dataset).

[10]. This codebook is untypical for studies at the DDA. First of all it is translated into English. The codebooks at the DDA are mostly in Danish. Secondly the study actually consist of 7 studies that have been merged, this accounts for the tabulations showing response- percentages at different points of time. The study is now available through ICPSR (DDA-0658 "Danish Election Studies, Continuity File 1971-1981", ICPSR-8946).

[11]. The latest versions of the catalogue of holdings at the DDA are: "Danish Data Guide 1986" and "Danish Data Guide Update 1988". They are both in English.

[12]. The reason for the atypical layout is explained in note 9.

[13]. OSI-SPC runs under DOS and OS/2 and is available on application to the DDA. The program is capable of subsetting variables from an OSIRIS documentation.

[14]. The DDAGUIDE retrieval database runs under IBM VM/CMS.

[15]. "DDA Annual Report 1988", in DDA-Nyt 49 (in Danish).

[16]. I shall not mention the standard retrieval problems: How to ensure that the employed retrieval terms actually cover the subject searched for; and how - at the same time - to obtain both a high level of precision and a high level of recall. These aspects are covered in "Information Retrieval Experiment", Karen S. Jones (ed.), 1981.

[17]. "ICPSR Annual Report 1988-1989". The ordering of datasets is free but the use of the CDNet database SEARCH is a charged-for service.

[18]. "Saving Space" by Steven J. Vaughan-Nichols, BYTE march 1990.

[19]. PKZIP (PKWARE Inc.), Ver. 1.01 DOS and OS/2 family mode. The latest and much faster version is 1.10 for DOS only.

[20]. The CD-ROM "bible" is: "CD ROM. The New Papyrus", Microsoft Press, 1986. In this collection of early papers Leonard Laub's "What is CD ROM?" is recommended as an introduction to the media.

[21]. IBM has just introduced SCSI-interface in their PS/2 family and also marketed a CD-ROM player supporting this format.

[22]. The CD-ROM maximum capacity is from 540 Megabytes to 640 Megabytes depending on the software used for processing and the software (and hardware) used for reading.

[23]. The study descriptions are left out of this calculation, as their size is relatively unimportant.

[24]. A few of the finished studies do not include a codebook, only the dictionary (variable location and label etc.)

[25]. The SPSS PC+ DATA LIST can not read more than 200 variables nor read ASCII files with a record length of more than 1024 bytes. ("SPSS/PC+ version 2". 1988, SPSS Inc., IL, USA. page C-38). In practice PC SAS (DOS) has limitations to the number of variables too.

[26]. Both SAS and SPSS are just now being released in OS/2 versions that do not have the memory problems of the PC-DOS version. Thus software limitations concerning the number of variables will disappear as the operating systems demand more hardware.

[27]. The ready-to-use software guides from Norton and Microsoft also have the potential for setting up new user defined retrieval systems. WordCruncher from Electronic Text Corporation is especially designed to search great quantities of text information.

[28]. "SAA Common User Access Advanced Interface Design Guide", IBM, 1989 (SC26-4582-0).

[29]. "Compact Data Nyt", April 1990 (Newsletter in Danish).

[30]. "Do-it-Yourself CD-ROMs" by Wayne Rash Jr., BYTE May 1990