Documentation - what we have and what we want: Report of an enquete of data archives and their staff

by Karsten Boye Rasmussen[1]
Danish Data Archives

Abstract:
A report from an enquete. Data professionals have given their views to questions on a future codebook format ("should it be SGML?", "should it be supported by vendors?", etc.). This forms a description of "what we want". On the other hand archives have described their holdings with regard to levels of machine-readable documentation. This section focuses on the present and on the actual data thus: "what we have".

The paper presents the key figures from questionnaires sent out by "The IASSIST Codebook Action Group". Maybe we are moving in a direction demanding less knowledge about special formats from our users? Hinting to the conference theme the subtitle could be "Access for other than Partners?".

Background Information - History
At the IASSIST Conference in Edinburgh in May 1993 several sessions were centred around documentation, and some specifically with documentation at the variable level (codebooks) as opposed to documentation at the study level (study descriptions). The sessions concerned with codebooks were: "Roundtable on Codebooks" (Wednesday, May 12) "Poster session on production of codebooks" (Thursday, May 13), and the "Codebook session" (Friday, May 14). I chaired the first and last of these sessions and the DDA contributed to the "Poster Session". The willingness to present papers at the "Codebook Session" as well as the many arguments and the eagerness of discussions all pointed out that many professionals were interested in working in the area of codebook documentation. The papers[0] at the "Codebook Session" contributed to many discussions at the conference, and after the conference the paper by the ICPSR director Richard Rockwell created many e-mail discussions.

At the business meeting at the conference an action group was formed and named: "Codebook Documentation of Social Science Data". I was appointed the chairman or co-ordinator of this action group and during 1993 was joined by Lennart Brantgärde and Bill Bradley as formal members of the action group.

My intentions with the action group was broadcast - on the IASSIST listserver in late May 1993 - as follows:

> *During the last years there has been some confusion concerning who are making which changes to the OSIRIS codebook. The OSIRIS codebook format has for more than twenty years been the de-facto standard for archives around the world. The most obvious reasons for this standard are that the OSIRIS codebooks can store full text, are input to retrieval systems, and that the codebooks are easily converted to other formats (SAS and SPSS). I propose that the task of the working group is to remedy this confusion by*:
>
> *1) Identifying the tasks and areas that cannot be handled by the OSIRIS codebook in its present form. Two examples: A) Many archives are looking for a feature of presenting tabulations in the codebooks - not just frequencies. B) A less rigid print out of codebooks not limited by the original card image input format. It is important to note that the first is a problem of storing a type of information in the codebook which does not fit the present format. The second concerns the utilisation of the codebook, and could be changed without changing the codebook format (by flowing the text, and using a different font).*
>
> *2) Identifying the number of data sets at archives all over the world. The data sets should be grouped by the level of documentation: A) full text machine readable codebooks (what format?); B) abbreviated machine readable documentation (OSIRIS dictionary / SAS / SPSS ?); C) no machine readable documentation (paper / scanned information). Data sets that are originally stored at other archives (ICPSR etc.) are to be counted only at the original archive. Another question would be whether the archive is producing machine readable documentation at all.*
>
> *The rationale behind the second identification is that archives who have produced and stored machine readable documentation (e.g. OSIRIS codebooks) will be able to convert these to the new format (missing the special features of the*

new format). There will not be invented a new codebook format which automatically documents what has not been documented. The archives who now uses the same documentation format will be able to share software for making the conversion to the new format.

I must emphasize, that it is not my intention that the codebook working group should produce a complete proposal for "This is how a Codebook should look like". We have no way to enforce a new standard other than waiting for people and organisations to realise its superiority to older standards. The task of the working group is to identify and structure the problems: these are common problems; these are problems with historical data; these are problems with complex data; etc.I think one feature of the new codebook format can be revealed now: It has to be so flexible that new features do not require a new format.

Another reason for not putting forth a new codebook standard is that it is my belief that documentation at both the study description level and the codebook level (and levels in between) has to be integrated.

My plans for the actions group were not fully met. One of the things that things that changed was - not surprisingly - the time schedule. I had planned to report to the IASSIST conference in 1994, but due to other obligations this was postponed till 1995.

In the meantime several activities took place. In Europe a CESSDA seminar with the title "Variable Level Documentation" took place at the SSD in Göteborg in August 1993[1]; the participants list was not restricted to Europeans. In August 1994 a CESSDA seminar was held in Grenoble[2], this time the theme was "Networking and Internet?", again ICPSR willingly sent a participant so the perspective was broadened and more global than just European. Although the term "codebook" was not part of the seminar title the searching and availability of codebooks on Internet - and therefore also the format of codebooks -were discussed intensively during the seminar. In October 1994 ICPSR announced a commitment in the development of an SGML DTD for codebooks. This issue was addressed again in mid February 1995 when ICPSR announced an international committee.

The questionnaires
In the summer of 1994 I formulated two questionnaires. One questionnaire was individual and gave the opportunity to present personal views on present and future codebook formats. The other questionnaire intended to count the number of studies at different archives and to show the distribution of different levels of documentation. In late autumn I received comments and advices concerning the questionnaire from a group consisting of Charles K. Humphrey (Data Library of University of Alberta), Lennart Brantgärde (Swedish Social Science Data Archive in Göteborg), and William Bradley (Canadian Health and Welfare, Ottawa).

The population and the return rate
Finally on 29 November 1994 the two questionnaires were sent to the listservers consisting of the IASSIST membership (178 recipients), Official Representatives of the ICPSR (195 recipients), and to the small IFDO list (27 recipients). A number of individuals were on more than one of these lists. The guess is that individuals from around 200 institutions were given the opportunity to answer the questionnaires. All recipients had furthermore the opportunity to forward a copy of the questionnaires to other individuals whom might be interested.

A pull for the return of the questionnaires were made shortly after the announced deadline on January 15th 1995. Several questionnaires were received thereafter and the last questionnaire was received at the end of February. At that time the number of received individual questionnaires had reached 50, and the number of institutional questionnaires were totalling 20.

These numbers do not hold evidence of the representativeness or the missing representativeness of the collected data. The investigation was never intended to be representative. However it is a fair assumption that the individuals that took the time to answer the questionnaire are individuals that have an interest in the development of documentation formats. And it is a fact that the archives that have returned the institutional questionnaire are principally amongst the national social science archives. All though it is disappointing that not all IFDO members made the effort of answering the institutional questionnaire.

The individual questionnaire
As shown in the "Appendix 1" the individual questionnaire is mainly about 25 different assertions about codebook documentation that individuals rate with their level of agreement on a five level scale from "strongly agree" to "strongly disagree". The middle category was labelled "indifferent, don't know" and this presented a problem to some individuals as these two answers are not completely identical.

The methodology of exposing individuals to a battery of items is well know. Looking back a feature that limited the amount of points which each individual could distribute would have been effective. It is much easier to answer that a lot of factors are important than actual to rank the factors and point out that "these are most important". On the other hand the distribution via e-mail called for both a very simple layout and a simple question structure. This led to the concentration on the battery of items without any filtering structure or deepening sub-questions.

In "Appendix 2" the distribution of the answers in these 25 variables is shown together with information about the mean and the number of non-missing answers. The mean was simply calculated by appointing the values 1 through 5 to the five answer categories:

1 strongly agree

2 agree

3 indifferent, don't know

4 disagree

5 strongly disagree

The treatment of the items - calculating the mean - implies that the items are viewed as belonging to an interval scale. Lots of arguments can be put forth in favour of or against this decision. In this context I find that "keep it simple" will suffice as legitimisation of this manoeuvre, as a more appropriate measure like "mode" will lose information. If you are interested in viewing the actual distribution of the answer categories for each of the items you should take a look at "Appendix 2".

As the direction of the assertions differs a high mean on one variable and a low mean on another does not necessarily imply that these two variables do not support each other. The mean can range from 1 to 5. In order to compare two variables of different direction you should keep in mind that a mean of for instance 4.2 is equally strong as a mean of 1.8. Means around the number 3 implies that the community has no fixed strong feelings about this particular assertion.

In the following the themes of the 25 assertions have been boiled down to a few headings.

The need for standards

I do not intend to enter a long philosophical discussion about standards. I am sure that we are all aware of the benefits of standards. It is equally true that we all spend time getting from one standard to another, e.g. getting from one analysis package to another. The common sense meaning of standard in this context is "as a rule". We expect to be able to connect our electronic equipment when we move to a new house, the plugs are supposed to be "standard". But having been to IASSIST conferences we know that this is not true when moving between countries. Thus the less common sense and the more sophisticated the more standards are available. Or to put it more precise: the many standards are a painful fact of life.

So it is no surprise that the lead question among the assertions - "1. There is no great need for standardization of codebooks" - receives a very strong disagreement value (mean 4.2). There is a need for standards, and we can continue the search for the standards. The ultimate alternative to the codebook - namely "no codebook" - is considered a very bad solution. "2. A data user should be content with a study description and photocopies of relevant pages from the questionnaire", with a mean of 4.3 this item receives the strongest disagreement of all 25 items.

One of the currently used and widespread standards is OSIRIS[3] and this format is drawn into the discussion in "5. There is a great need for more structured information than is available in the OSIRIS codebook format" that shows a weak agreement (mean 2.6). Half the individuals answer "indifferent, don't know" and maybe they answer the latter simply because they don't know the OSIRIS format and its structure. Other formats - and these are more commercial and more available formats - are also given the same weak agreement (mean 2.6) in "10. Let us stick with commercially supported formats for social science data (e.g. SAS and SPSS)", but it should be noted from the distribution that there is a higher variance in this question.

The support of standards

Standards live because they have supporters. The supporters do not have to be very loudspeaking as the history of OSIRIS shows. In the 70'ies OSIRIS was rather widespread as a social science analysis package, but SPSS[4] and SAS[5] gained momentum and OSIRIS was abandoned by the researchers. But many archives continued to utilise the OSIRIS documentational format, and still OSIRIS is used by many archives. Often the researcher that receives archive materials does not know that the SAS or SPSS setup actually is an automated product created from an OSIRIS codebook. When standards

depend upon supporters in order to get a strong standard you would naturally want strong supporters.

Two items express the need for commercial support: “18. A new format should be supported by the analysis software industry (e.g. SAS and SPSS)” and “19. A new documentation format should be supported by major document software and applications (Word, WordPerfect, WWW)”. They both receive high agreement (mean 1.8). But does this mean that if we can not persuade SAS, SPSS, Word (Microsoft), WordPerfect (Novell) to support a new codebook format then we will have to give up? No, not in my opinion!

It would be nice with support from SAS and SPSS, but the history of data conversion shows[6] that the packages always almost do the job. That leaves you with problems concerning the character set and especially about missing data. Till now it has been much easier to write documentation conversion software that will reduce and format the documentation to the levels supported by SAS and SPSS. Another item indirectly addresses the commercial support: “21. A new documentation format will not be of any interest unless the data producers directly produce their documentation in this format”. Here we are not only demanding software to follow the standards, but humans and maybe persons we know are asked to follow suit. The mean for item 21 drops to 2.5. Then it is getting really hot: “22. A new documentation format is only interesting if all archives abide by the new standard”. The answer is close to indifferent (mean 2.7), but now we have moved from vendors to other people and finally we are trapped ourselves. The support of standards is a nice feature, and the less of one’s own work that is involved, the nicer the feature gets.

Support from word processing companies will depend upon the new format. But if the new format is going to be directly connected to the formats of the Word Wide Web (HTML[7]) there is no doubt that the format is going to be supported[8] at least when the word processor is used as a viewer. However we have to be careful here. HTML is moving and changing, new versions and new features are being introduced. The safe ground to build upon is SGML where definitions can be made. Furthermore a codebook defined as a document type in SGML and marked up accordingly is very easily converted or reduced to any HTML level.

Labels in the codebook

New formats for codebooks or not, everything will not change overnight. We are going to continue to support analysis packages that can only handle limited amounts of text. But there is harsh disagreement that users or archives should be content with this level of information. “7. A documentation format with short labels for variables and values is sufficient for the user” (mean 4.0) and “8. A documentation format with short labels for variables and values is sufficient for the archive where a study is deposited” (mean 4.2).

When the question is directed specifically towards the length of the variable labels it is no great surprise that a longer label is preferred to the shorter: “14. Variable labels composed of 24 characters is sufficient” is slightly disagreeable (mean 3.4) whereas the next item is found slightly agreeable (mean 2.7): “15. Variable labels composed of 40 characters is sufficient”[9]. The question of course is what the labels are “sufficient” for? I interpret the results as specific for the labels, and not that the codebook documentation should consist of nothing more than the variable labels.

Oldies but goodies of the codebook

Information about the marginals is considered one of the main benefits of the codebook. “3. Codebooks should contain marginal frequencies that will enable the users to check the data they have received” is heavily agreed upon (mean 1.6). New fashions are not considered that important: “4. Codebooks should contain cross-tabulations so the user has more information about how to analyse the data” only makes it slightly above the threshold of indifference (mean 2.8).

Most of the items about the layout and printing of the codebook receives the same low level of attention. “9. The presentation of a printed codebook is very important” (mean 2.5); “24. There is little need for a printed codebook if a machine readable codebook exists” (mean 2.8); and “25. I prefer to browse data documentation in files on my own computer” (mean 2.6).

New possibilities of the codebook

If a new codebook format is to be developed we could be talking about transferring information, or we could use the opportunity to expand the current format with new possibilities. Some possibilities are mentioned amongst the assertions, and the highest score of all items is received by “20. A new documentation format should include the study description” (mean 1.4). Once again this demonstrates the methodological problem of having unlimited resources (points) when filling out the questionnaire. We must conclude from the questionnaire that there is a craving for including the study description in the codebook. From our practical lives we must conclude that we have to solve one thing at a time. Redesign of the codebook has to consider - but not necessarily to solve - the implementation of the study description.

I assume that most people are aware of the new possibilities but these are not considered very important for the codebook. The answer to "11. A documentation format should be able to incorporate pictures and sound" is indifferent (mean 2.9). The item on scanned images "23. I would be content to receive scanned images of the questionnaires" receives a slightly less favourable score, but that is due to poor verbalisation on my behalf. I believe that receiving only scanned images will not satisfy the user, but the combination of the text based codebook with the actual questionnaire as presented through scanned images will be of interest to most users.

Efforts of internationalisation have been mentioned at several meetings on documentation. One of the main obstacles in reading other languages than ones own. The introduction of English labels is not greeted with much enthusiasm: "6. English variable labels should be used with all studies regardless of the language of the original study" receives only a mean between agreement and indifference (mean 2.5). With the syndrome of unlimited resources you would have expected this to be easily agreed upon. Furthermore people from English speaking countries (UK, USA, Canada) seem to be only a fraction more inclined towards agreement.

The institutional questionnaire
The intention of the institutional questionnaire is to determine:

A) the number of datasets amongst the social science data archives.

B) the number of individual datasets (not held as a copy from another archive).

C) the distribution of used archive formats.

Datasets in social science archives
One of the big subjects has been to make a clear definition and to answer the question about the presumed unit of analysis "what is a study?". As archives would tend to see their importance depending on the magnitude of studies, and especially as compared with other archives, the study unit seemed to be of crucial importance. This led to some correspondence over the issue as it was argued that some studies consist of many datasets, have a complex scheme, and are distributed on several tapes, other studies have only few variables and a couple of hundred cases.

However I found that the introduction of a "true" unit of analysis could not succeed within the limited time period for returning the questionnaire. It should be noted that the individual questionnaire demanded only the person to make up his mind. In the institutional questionnaire more questions are asked based upon the unit of analysis. Even though the unit of analysis (the study) is not the same at all archives, it was more important that the individual archive had the easiest possibility of answering the questions; it was hard enough anyway. At each archive they had to do some kind of stocktaking and to place the studies within the categories demanded by the questionnaire.

Categorisations and fundamentals of archives
The archives that have answered the questionnaire are not many. 20 in numbers. One returned questionnaire was afterwards disregarded because both a staff member and the director of the archive had returned a questionnaire. 19 is now the basis.

The reason for some people to fill in the individual questionnaire as the only person from their archive, and yet not fill in the institutional questionnaire must be that the institutional questionnaire involved much more work in order to be filled out. Therefore I want to express my sincere thanks to the persons who took the time to fill out the time-consuming institutional questionnaire.

When preparing to discuss what kind of codebook documentation format should be used in the near future it is interesting to know whether an archive produces documentation. The answer of 7 archives to the question "Does your institution produce original machine readable documentation for studies in your archive?" was "No - we are only storing". Most of these archives are university data archives in North America, but the English national data archive was also a member of this group. If you do not produce documentation you will end up having all kinds of documentation formats. The interest of these archives in a new documentation format must be for the utilisation of the splendours of a new format, more than how a new format will act as a standard and solve some of our current documentational problems.

Of the 19 archives 13 archives have English as their native language.

Paper documentation vs. machine-readable documentation
First the respondent from the archive is asked about the number of studies with only paper documentation (Q1_1) secondly about the number of studies with some kind (any kind) of machine-readable documentation (Q1_2). The 19 archives share among them more than 27,000 studies, of these close to 11,000 have some kind of machine-readable documentation. The ratio of machine-readable documentation compared to all studies is thus 0.4. But this ratio hides great differences amongst the archives, some archives go as high as 0.9 others below 0.2 (the MR Ratio).

It is important to note, that the majority of studies at data archives do not have any machine readable documentation at all. Even when the machine readable text information is very limited (e.g. only describes variable labels and few categories) it can be used as input for software that automatically will convert the information to a new documentation format. This is a qualitative difference compared to having no machine readable text information. The studies without machine readable documentation will demand much more work to process to a new documentation format.

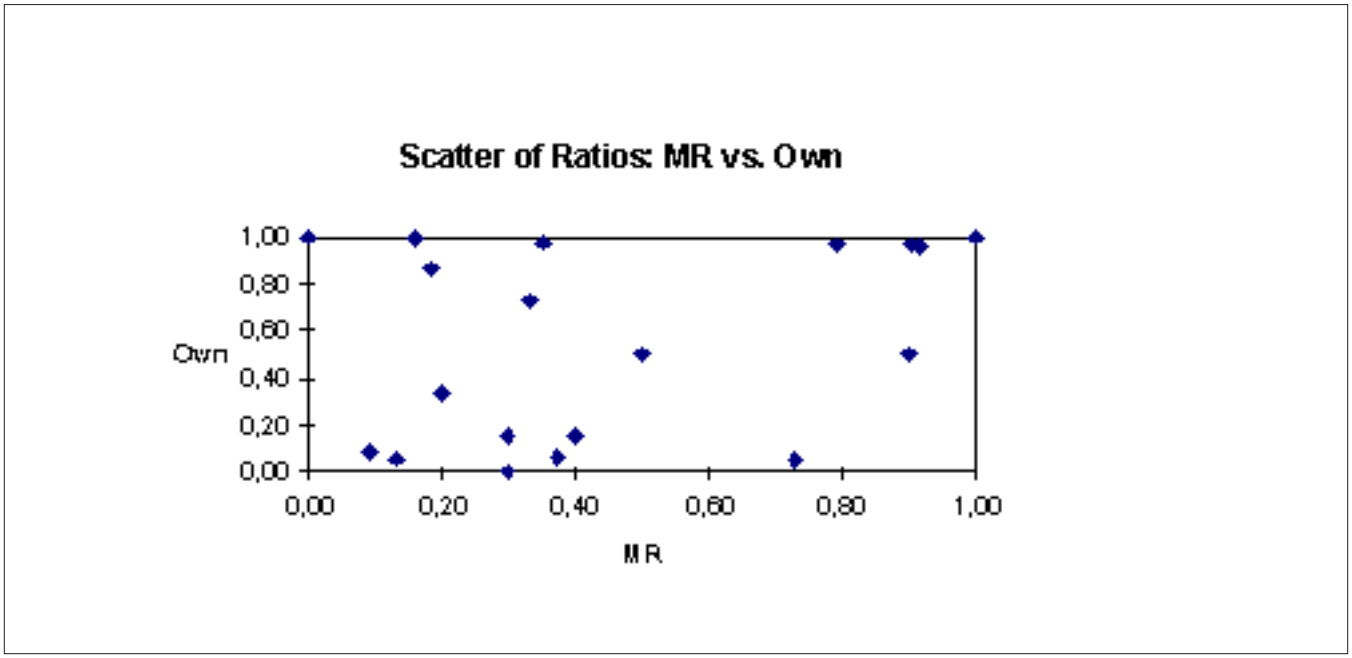
Own holdings vs. deposition from source archive
In Q2 the question is put about how many of the studies are actually stored and available from another archive (a source archive). The number of studies totals to more than 11,000. As many of these archives report their main source archive to be ICPSR, and as ICPSR is among the 19 answering archives we can decide to be bold and just subtract this from the total, leaving us with around 16,000 unique studies.

The ratio of own studies varies from 1.00 (all studies are own studies) to ratios close to zero (no studies that do not come from a source archive).

The scattergram shows the distribution between the ratio of own studies and the ratio of machine-readable studies in the figure below.

Types of MR
Then five questions about levels of machine readable documentation are asked. The respondent is asked to distribute the studies counted in Q1_2 having machine-readable documentation into five separate categories where the lowest rank is scanned images and the highest rank is a full codebook. The five categories are expected to sum to the total of machine-readable studies, but only half of the archives manage their stock data in such a way as to make the balance right. The sum of the studies in the five categories in Q3 totals to 12,627, whereas the total of machine readable studies in Q1_2 is 10,946.

Scanning as MR documentation
Scanning (Q3_1) of questionnaire pages as the highest level of machine-readable documentation is to close to 100 pct. only



found at the Amsterdam based Steinmetz archive. They have earlier shared their research and experience in scanning and made us aware of choosing TIFF-4 as the scanning format[10]. It is obvious that scanning is used at other archives both for security reasons as well as an easy deliverable - especially over Internet - but most studies will have some kind of higher level documentation as well.

One of the important things to remember when talking about scanning is that the scanned images should be referenced / pointed to / tagged from the character document that constitutes the codebook.

Format	Frequency
1. Scanning	505
2. Text	1943
3. Dict	4520
4. Dict+	3656
5. Dict + Codebook	2003
Total	12627

Plain text as MR documentation
After some editing[11] the distribution of question Q3_2 with the category "Unstructured and untagged text (text from OCR, questionnaire from WordPerfect, etc.)" ends with the following result:

At first it is surprising that the largest figure is given without any information about the format. On the other hand these formats are without importance because the formats are not directly related to the variables in the codebook or structured into elements of the variables. This category contains cases where the information is more like a stream of text for instance the non-edited result of OCR. When "ASCII" is mentioned this is not a reference to the irrelevant character format, "ASCII" here means "plain ASCII" which again means "no formatting information".

Dictionary as MR documentation
The category in Q3_3 "Machine readable dictionaries (information only about variable locations, labels, missing data in SAS, SPSS, OSIRIS or other format.)" means basic dictionary information on the variable level without information about the categories. In common database systems this is the dictionary information documenting the single fields of the database. However the introduction of information about missing data values and meanings is a social science product. A few formats are not widely used. For instance the Israeli archive mentions that they store their catalogue information in their ALEPH system, while the data is stored and documented in SAS.

Text Format	Frequency
ASCII	532
WordPerfect	299
Other	1112
Total	1943

Personally I do not find the differences in these systems interesting. As the dictionary level is defined in the question Q3_3 all these system will be very much alike. At the same time within a single product - for instance the much used SPSS system - the differences between different forms of SPSS can be as challenging to the user as receiving different products. SPSS can mean: Export files, System files (what platform), setups (what version) etc. 4,520 files are deliverable with the lowest level of machine readable variable level information.

Dictionary+ as MR documentation
This level is defined as the dictionary information plus information on values: Q3_4. "Machine readable codebook documentation (as 3. above but with the addition of explanation of the coding categories such as value labels in SPSS or user formats in SAS)". This is a restricted codebook format that does not have a lot of codebook levels and elements and does not support unlimited amounts of text.

Two interesting facts: DDMS is the system developed at Canadian Health and Welfare and the system is being used at different Canadian archives[12]. The NSD Stat is the format belonging to the statistical package developed by the Norwegian Archive[13].

Dictionary Format	Frequency
dBaseIV	20
OSIRIS	248
SAS	266
SPS S	3550
Other	436
Total	4520

Again the SPSS format is most often mentioned. And again my same views concerning formats between packages and within packages apply. Notice that the total number of studies archived drops when we demand value label information compared to the dict-level before.

Dictionary Codebook as MR documentation

In Q3_5 the level is defined as "Machine readable codebook documentation (as 4. above but including all questionnaire text and other information like in the OSIRIS format codebook)." This can give some difficulties if the respondent does not know the OSIRIS format.

DDMS	137
NSD Stat	100
OSIRIS	268
SAS	116
SPSS	2138
Other	897
Total	3656

Some are still mentioning SAS and SPSS, even though these packages cannot go above the level of dict+, but these packages are then combined with other text. Others mention different cataloguing systems and searching capabilities. All in all a conservative guess of the magnitude of fully documented studies can be as low as 1256 (DDMS plus OSIRIS). If we find this figure disappointingly low then on the positive side we can now remember that some archives have not answered the institutional questionnaire.

However these 2,003 studies plus the 3,656 studies belonging to the dict+ level are the studies with machine readable documentation available. These are the studies that the user will be able to analyse without consulting other material - except of cause the study description.

Dict Format	Frequency
DDMS	45
OSIRIS	1211
SAS	37
SPSS	37
Other	673
Total	2003

These 5,659 studies should optimistically be the studies that can be automatically converted to a new codebook format. If we are going to divide the task between us, I will personally choose to make the conversion of the OSIRIS codebooks because of their very simple format!

Conclusion

Future of the Codebook? Codebook of the Future!

We have seen that "what we want" is "everything".

"What we have" is pessimisticly close to nothing (1300 MR documented datasets).

We noticed that this high level of documentation should preferably be produced by "somebody else".

We know that only a small percentage of the studies are fully documented, and these can easily be converted to a new codebook format.

We know that more than half of the studies have no machine readable documentation at all.

We can conclude that we are facing a daunting big job.

However: If a new format can combine all (or almost all) our needs for structural elements in the documentation, the archivist will save time. I am very much looking forward to the work on a new format in the ICPSR "SGML Codebook Committee". We have a great opportunity in using the tools that are being offered.

Because the archives immediately will develop tools for processing the documentation for use in many surroundings like CD-ROM, Hypertext, Internet, etc. the possibilities of the new format will also be of interest to data producers. Laura Guy[14] talks about the archivist pleading with producers to ".. please .. document it properly?". With these modern add-ons there will be great benefits of doing proper machine readable documentation.

Paper presented at the IASSIST conference in Quebec City, May 1995

1 Report from SSD CESSDA seminar “Variable level documentation”, Göteborg 1993.

2 The report and paper collection from the Grenoble meeting is not yet published.

3 The OSIRIS format is documented in OSIRIS III, Volume I, ISR 1973, Univ. of Michigan.

4 ”SPSS for Windows, Release 6.0”, 1993, Chicago, SPSS Inc.

5 SAS has meters of manuals, I’ll mention 4 cm: “SAS Language: Reference, Verions 6”, 1990, Cary, SAS Institute Inc.

6 Poster session at IASSIST 92 and “Converting Data” in DDA-Nyt 62, Summer 1992.

7 ”Hyper Text Markup Language” is a document type definition (DTD) made in SGML (Standard Generalized Markup Language). HTML is used as the document format in WWW. A very usable SGML book is: “Practical SGML” by Eric van Herwijnen, 2. ed., 1994, Kluwer, Dordrecht.

8 In March 1995 Microsoft announced their HTML extension available for Word (but so far only for the US version 6.1).

9 As a curiosity I can mention that a cross tabulation of the two items (14 and 15) shows that some persons find 24 character labels agreeable but find 40 character labels disagreeable.

10 ”Exchange of scanned documentation between social scientists and data archives: establishing an image file format and method of transfer”.Repke de Vries and Cor van der Meer in IASSIST Quarterly Vol.16, number 1/2.

11 Apart from summarizing I have taken the liberty to evenly distribute figures connected to more than one format. If an archive gave the number 200 and mentioned the formats ASCII and WordPerfect both categories received 100.

12 William Bradley - mentioned in note 1 - is the leader of the group developing DDMS. It should be noted too that DDMS is a full codebook system, not a restricted system.

13 The documentation format is only mentioned at the Norwegian archive. But the use of the NSD Stat PC-package is much more widespread

14 ”The Need for Revised Data Documentation Standards: New Solutions for Old Problems”, Laura Guy in IASSIST Quarterly Vol. 17 Num. 3/4.

Appendix 1: The Questionnaires

In the following the questionnaires as e-mail to the listservers.

PART 1: DATA DOCUMENTATION PREFERENCES

Codebook Documentation of Social Science Data: an IASSIST Action Group

In this survey we are interested in your INDIVIDUAL opinions about improving data documentation and the formats used with data

documentation. The information obtained through this survey will become part of a report from an Action Group under the International Association of Social Science Information Service and Technology (IASSIST) about the current practices of social science data documentation and proposed standards for codebooks. Your thoughtful completion of this questionnaire is appreciated.

This first part is an individual questionnaire. If your institution stores or archives social science data we ask you to fill out the second part with information about your institution and archival format.

You are kindly invited to complete this questionnaire and return it to Karsten Boye Rasmussen, Dansk Data Arkiv, Islandsgade 10, DK-5000 Odense C., Denmark by mail, fax (+45 66113060) or e-mail (kb@dda.dk). If you are using e-mail please observe that you are not responding to the list server but directly to kb@dda.dk. Please return this questionnaire before the 15th of January 1995.

Q1. Below is a series of statements about data documentation. Please indicate for each statement the degree to which you agree or disagree with its content. Use the following five-point scale:

1 strongly agree
2 agree
3 indifferent, don't know
4 disagree
5 strongly disagree

__ There is no great need for standardization of codebooks.

__ A data user should be content with a study description and photocopies of relevant pages from the questionnaire.

__ Codebooks should contain marginal frequencies that will enable the users to check the data they have received.

__ Codebooks should contain cross-tabulations so the user has more information about how to analyze the data.

__ There is a great need for more structured information than is available in the OSIRIS codebook format.

__ English variable labels should be used with all studies regardless of the language of the original study.

__ A documentation format with short labels for variables and values is sufficient for the user.

__ A documentation format with short labels for variables and values is sufficient for the archive where a study is deposited.

__ The presentation of a printed codebook is very important.

__ Let us stick with commercially supported formats for social science data (e.g. SAS and SPSS).

__ A documentation format should be able to incorporate pictures and sound.

__ Missing data should always be coded as numeric values.

__ Coding of data fields using alphabetical and special characters should be discouraged.

__ Variable labels composed of 24 characters is sufficient.

__ Variable labels composed of 40 characters is sufficient.

__ Changing to a new documentation format would be very difficult to implement at our institution.

__ A new documentation format should be a specialized implementation of a general document format (e.g. SGML).

__ A new format should be supported by the analysis software industry (e.g. SAS and SPSS).

__ A new documentation format should be supported by major document software and applications (Word, WordPerfect, WWW)

__ A new documentation format should include the study description.

__ A new documentation format will not be of any interest unless the data producers directly produce their documentation in this format.

__ A new documentation format is only interesting if all archives abide by the new standard.

__ I would be content to receive scanned images of the questionnaires.

__ There is little need for a printed codebook if a machine readable codebook exists.

__ I prefer to browse data documentation in files on my own computer.

Q2. Your name : _____

Your position: _____

Institution: _____

Country: _____

Q3. Are you a member of IASSIST

( ) Yes
( ) No

Q4. Please comment further on any aspect of data documentation that is a concern to you.

PART 2: INSTITUTIONAL DATA DOCUMENTATION

Codebook Documentation of Social Science Data: an IASSIST Action Group

If your INSTITUTION stores or archives social science data we are interested in information about your institution and archival format.

You are kindly invited to complete this questionnaire and return it to Karsten Boye Rasmussen, Dansk Data Arkiv, Islandsgade 10, DK-5000 Odense C., Denmark by mail, fax (+45 66113060) or e-mail (kb@dda.dk). Please return this questionnaire before the 15th of January 1995.

This questionnaire is to be completed from an INSTITUTIONAL perspective. Individual opinions on social science data documentation are to be expressed in the first questionnaire. Several individuals from the same institution can fill out the individual questionnaire, but only one person needs to answer the institutional questionnaire.

The information obtained through this survey will become part of a report from an IASSIST Action Group about the current practices of social science data documentation and proposed standards for codebooks. Your thoughtful completion of this questionnaire is appreciated.

Your name: _____

Your position: _____

Name of your institution: _____

_____

Country: _____

Q1. We are interested in the variety of documentation that accompanies social science data and the number of studies with each type of documentation. Please indicate the total number of studies available with only paper documentation and those with some form of machine readable documentation. A study should only be counted once.

Number of studies archived with:

_____ 1. Only paper documentation
(no machine readable documentation)

_____ 2. Some form of machine readable documentation
(may also be available in print)

Q2. How many of your studies have you received from another archiving institution? Please specify number of studies:

_____ Studies from “source” data archive

Q3. Of the total number of studies with some form of machine readable documentation, please indicate how many studies are available in the following formats.

Please report a study only once at the highest documentation level (1=low 5=high).

Number of studies in machine readable format consisting of:

_____ 1. Scanned images
Please specify the most commonly used format:

_____ 2. Unstructured and untagged text (text from OCR, questionnaire from WordPerfect, etc.)
Please specify the most commonly used format:

_____ 3. Machine readable dictionaries (information only about variable locations, labels, missing data in SAS, SPSS, OSIRIS or other format.)
Please specify the most commonly used format:

_____ 4. Machine readable codebook documentation (as 3. above but with the addition of explanation of the coding categories such as value labels in SPSS or user formats in SAS).
Please specify the most commonly used format:

_____ 5. Machine readable codebook documentation (as 4. above but including all questionnaire text and other information like in the OSIRIS format codebook).
Please specify the most commonly used format:

Q4. Does your institution produce original machine readable documentation for studies in your archive?

( ) no - we are only storing
( ) yes

If "yes" please describe the process and format used for preparing machine readable documentation.

Q5. Does your institution have a policy about the format used for documentation at the variable level?

( ) no
( ) yes

If "yes" please describe the format used for describing variables.

Appendix 2: Distribution and mean of the individual questionnaire

	strongly agree	agree	indiffe -rent, don't know	dis- agree	strongly disagree	Mean	N non- missing
1. There is no great need for standardization of codebooks	0	4	3	18	24	4.2	49
2. A data user should be content with a study description and photocopies of relevant pages from the questionnaire	1	3	2	14	30	4.3	50
3. Codebooks should contain marginal frequencies that will enable the users to check the data they have received	23	25	1	1	0	1.6	50
4. Codebooks should contain cross-tabulations so the user has more information about how to analyze the data	4	17	16	8	5	2.8	50
5. There is a great need for more structured information than is available in the OSIRIS codebook format	6	12	23	5	1	2.6	47
6. English variable labels should be used with all studies regardless of the language of the original study	5	25	10	6	4	2.5	50
7. A documentation format with short labels for variables and values is sufficient for the user	2	4	4	22	18	4.0	50
8. A documentation format with short labels for variables and values is sufficient for the archive where a study is deposited	2	3	4	14	27	4.2	50
9. The presentation of a printed codebook is very important	10	21	5	10	4	2.5	50
10. Let us stick with commercially supported formats for social science data (e.g. SAS and SPSS)	11	12	13	6	6	2.6	48
11. A documentation format should be able to incorporate pictures and sound	5	9	22	10	4	2.9	50
12. Missing data should always be coded as numeric values	15	12	14	3	6	2.4	50
13. Coding of datafields using alphabetical and special characters should be discouraged	18	11	8	9	4	2.4	50
14. Variable labels composed of 24 characters is sufficient	1	11	12	15	10	3.4	49
15. Variable labels composed of 40 characters is sufficient	7	16	12	9	5	2.7	49
16. Changing to a new documentation format would be very difficult to implement at our institution	6	4	17	17	5	3.2	49
17. A new documentation format should be a specialized implementation of a general document format (e.g. SGML)	12	17	14	4	1	2.2	48
18. A new format should be supported by the analysis software industry (e.g.SASandSPSS)	19	23	6	2	0	1.8	50
19. A new documentation format should be supported by major document software and applications (Word, WordPerfect, WWW)	18	25	6	1	0	1.8	50
20. A new documentation format should include the study description	29	19	2	0	0	1.4	50
21. A new documentation format will not be of any interest unless the data producers directly produce their documentation in this format	12	14	9	14	1	2.5	50
22. A new documentation format is only interesting if all archives abide by the new standard	7	14	16	11	2	2.7	50
23 I would be content to receive scanned images of the questionnaires	3	10	10	16	11	3.4	50
24. There is little need for a printed codebook if a machine readable codebook exists	9	17	1	17	6	2.8	50
25. I prefer to browse data documentation in files on my own computer	8	16	10	12	2	2.6	48