How to teach data producers "the noble art" of data documentation

Nielsen, Per

Per Nielsen

How to Teach Data Producers „The Noble Art" of Data Documentation

1. Introductory Remarks

In the good old days, preservation, storage and access problems in the social science area were successfully solved by the library and the document archive:

Before the arrival of the computer, these giant data-collection agencies cooperated without great difficulty with the data-storage institutions: the tables and the analyses published by the statistical bureaus were stored in libraries and the original data sheets (census sheets, register protocols, and the like) were with some regularity transferred to the established archives[1].

In the article on Data Services in Western Europe quoted above, Stein Rokkan claims that „the inertia of the traditional institutions" in adjusting to the storage and display demands after the computer revolution created an unsatisfied demand for mass data in computerized form; the Data Services tried to bridge the gap between production and distribution vis-a-vis the social science community.

In this paper, we shall touch on preservation, storage and access problems from the point of view of such data service organizations: What are the main obstacles to secondary analyses of the vast and ever-increasing holdings of machine-readable data, and which remedies can secure a better utilization in the coming decade.

[1] Rokkan, Stein, Data Services in Western Europe. Reflection on Variations in the Conditions of Academic Institution-Building, in: American Behavioral Scientist, Vol. 19, No. 4 (1976), p. 445. This issue of the ABS deals with the „data archive movement". The academic data archives, data services, data libraries etc. will in the following be referred to by the term „data service organization" or just „data organization". In Europe, 7 data organizations with national coverage cooperate in CESSDA (Committee of European Social Science Data Archives); internationally, a dozen data service organizations have just established IFDO (International Federation of Data Organizations).

2. What is Documentation

In this section, we shall postulate that a minimum requirement for closing the gap between the data producer and the secondary analyst is a high standard of data documentation. The documentation items are listed after the following outline of a subset of the obstacles to secondary analysis[2].

The *data producer* (in relation to the secondary user: *donor*) can be reluctant or even unwilling to place data at the disposal of secondary users. The donors may claim that analysis by „outsiders" is restricted to protect the individuals registered in a file, or that s/he is afraid that secondary analysts will misinterpret the data due to a lack of background information. The former argument is real (consider the Data Law issue) for files with information on individuals, but there are measures (aggregation, anonymization, scrambling) that soften the argument; the latter concern that the secondary user may misinterpret the data is hypocritical, making a virtue of a sin of omission: if the data were aptly documented the risk of misinterpretation would be negligible. Sometimes one has the feeling that a few unspoken considerations underlie the reluctance of data producers to disseminate data: fear of a critique of methodology or even challenges of reported findings; intentions to maintain an information monopoly; and so on. We shall return to these considerations below.

The *data user* (in relation to the data producer: *secondary analyst)* has a scientific problem area that s/he wants to investigate by means of quantitative methods. The search for relevant data is difficult due to lack of information about existing data holdings. Research libraries in Europe do not catalogue machine-readable data holdings, and neither public (e. g. statistical bureaus) nor private (e. g. individual researchers, market research organizations) data producers have been eager to catalogue their data-holdings. In addition, professional prestige seems to be higher if the social science researcher collects new data for a specific purpose rather than using data already collected. Secondary analysis research designs are complicated, and few researchers have the combined skills in data processing (computer use), quantitative methods (statistics), and one or even several substantive fields (interdisciplinary research) required to engage her/himself in such projects. In addition, the value for secondary analysis of a dataset is sometimes reduced with the „age" of the data.

The *data service organizations* (the mediator between data producers and data users) have collected and stored data with a considerable investment in the „processing" of each acquired dataset. Essentially, what the data organizations do is what the data producers ought to do: „Produce" complete *documentation for users.*

[2] For a more extensive discussion of obstacles to secondary analysis, see Hyman, Herbert H., Secondary Analysis of Sample Surveys: Principles, Procedures, and Potentialities, New York 1972. Chapter 1 deals with the issue.

This brings us back to the question: What is data documentation? In an International Association for Social Science Information Service and Technology (IASSIST) meeting in Copenhagen recently[3], the working group indicated that the following list of documents/facilities were desirable from the user point of view:

2.1. Library Information

Information on a data-file should be produced in a card catalogue form to be entered into the library system. Till this day, the inertia of the library systems in Europe has been so strong that data files are not considered for reference. In the U. S. the communication between data organization personnel and traditional librarians is better: Coordinated by *Sue A. Dodd* and following recommendations from the ALA Catalog Code Revision Committee's Subcommittee on Rules for Cataloging Machine-Readable Data Files, the Classification Action Group of IASSIST in North America has set standards that should be adopted also in Europe[4].

2.2 Archive Information/Study Abstract

Each data service organization has its own way of presenting its data holdings (Inventory). It would be very useful if other data holders (e. g. statistical bureaus, machine-readable divisions of traditional archives, research institutions) would publish guides/inventories to their holdings. Abstracting of data file contents (cp. the immense resources invested in bibliographic abstracting these years) is hardly seen in Europe outside the realm of the data organizations.

[3] IASSIST (International Association for Social Science Information Service and Technology) is an international membership organization for data organization and information center personnel, quantitatively oriented researchers, etc.
The user documentation items listed were defined at a workshop in Copenhagen, June 26–29, 1977. The Report appeared in the IASSIST Newsletter, Vol. 1, No. 4 (Fall 1977), pp. 7–10.

[4] For further details, see Working Manual for Cataloging Machine-Readable Data Files, compiled by Sue A. Dodd, Data Librarian, Social Science Data Library, Institute for Research in Social Science, Univ. of North Carolina, Chapel Hill – or write to Sue Dodd to have the IASSIST Classification Action Group working materials.

2.3. *Study Description*

The study description is the locus for all information on a data file necessary to secure correct interpretation of analysis results, i. e. a complete description of the research project or administrative process in which the data was originally collected and processed; this is probably the area where the sins of omission on the part of the data producer are most outstanding: Not only is it hard to find a full description when a data organization acquire a dataset, but, alas, we still see reports on quantitative research results containing incomplete technical report sections. Under the auspices of the former Standing Committee for Social Science Data of the ISSC a standard study description scheme was developed in 1974[5], and this scheme is presently being tested further by 6 data organizations[6]. The standard study description scheme has been designed in a flexible way to allow for a multi-purpose use: Besides containing all background information necessary for a reliable secondary analysis of the file described, the information can be automatically subsetted for abstracting purposes; furthermore, the sum of study descriptions in a data organization can be used for mapping and methodological research as well as for information retrieval and intra-archival logging purposes; finally, the study description is used for inter-institutional exchange of information on data holdings[7]. It should be added, however, that the standard study description in its present version is geared to survey files, primarily.

2.4. *List of Variables*

Each data organization should produce a list of variables for every file in their holdings; this list of variables should be printed as well as machine-readable. Like the data abstracts may be automatically generated from the study description, this list of variables may be automatically produced from a machine-readable codebook.

[5] See Report on Standardization of Study Description Schemes and Classification of Indicators and Study Description Guide & Scheme, both edited by Per Nielsen (Sept. 1974 and April 1975, respectively). Both available from the Danish Data Archives.

[6] Since 1974, the ZA, Cologne, and the DDA have been testing the standard study description scheme, and Steinmetzarchief, Amsterdam, has used an earlier version thereof. Now, also BASS, Louvain-la-Neuve; Leisure Studies Data Bank, Waterloo; and ICPSR, Ann Arbor, have agreed to test the instrument.

[7] Inter institutional exchange of study description is now being tested.

2.5. *Codebook*

At the recent IASSIST meeting, it was indicated that the ideal codebook should contain 15 major items (the first three items referring to file level, the last 12 to be applied variable by variable where applicable): (1) title of study (file and subfile names); (2) format of the data file; (3) comments at file level, e. g. concerning application of missing data codes, special weighting features, special precautions for use of file; (4) variable identification (number, label, short name, mnemonics; (5) variable source reference; (6) variable location and length; (7) variable type (alphabetic, alphanumeric, numeric, symbolic); (8) number of decimal places (scale of measurement); (9) source statements/texts/questions/scale description/introductory statements related to responses; (10) answer code values; (11) answer code descriptions; (12) comments originating in field work and coding experience: interviewer instructions and coding instructions; (13) variable contingencies (filter, skip, control); (14) variable consistency.(i. e. results of checking of variable contingencies); (15) reference to derived variables.

2.6. *Classification/Index*

Within the area of classification and indexing (on file and/or variable level) a number of different schemes are available, describing different dimensions. However, the testing of these schemes has not yet reached a level where recommendations can be set forth. For a discussion of variable level retrieval, see the description by Ekkehard Mochman[8] who is coordinating the European classification arena within IASSIST.

2.7. *The Data Matrix*

Even if the data itself can only indirectly be called documentation, it documents the results published by the primary investigator; the data matrix is the object that is described in the data documentation. It should be borne in mind that machine-readable data (unlike paper-carried information) can be read only of a correct de-

[8] Mochmann, Ekkehard, Information Access at the Data Item Level: Approaches to Indicator Retrieval from Survey Archive Data Bases, SIGSOC Bulletin, Vols.6,2&3 (1974–75). Edited by Alice Robbin, this issue of the Bulletin dealt with „The Data Library: Systematic, Structural and Process Problems of Data Access".

scription of the data carrier and the physical and logical characteristics of representation of data on the data carrier is supplied. Single punch data (if possible in card-image format) is preferable for archiving purposes.

2.8. Special Publications

Many data organizations produce publications concerning specific files. Items 2.3. to 2.5. above often make up several hundred pages for one survey. Such data documentation publications should, of course, be entered into the library system. Very often, all data documentation is machine-readable.

Even if the list of user requirements for proper documentation of data files may be biased in the direction of survey data, we shall claim that it is readily generalized to process-produced data files (administrative records). Therefore, per definition, the production of written (and preferably machine-readable) descriptions following the eight-item outline above[9] is data documentation. Thus we can proceed to the „why".

3. Why Data Documentation

In the preceding section, we postulated that data documentation is an indispensable prerequisite for secondary analyses. In this section, we extend the argument further by outlining some considerations which support extensive data documentation even in the primary analysis phase/administrative process.

3.1. Reliability Aspects

The late Sir Cyril Burt, an outstanding English psychologist for decades, seems to have involved himself in research reporting of a dubious nature (hostile people call it research swindle): In his old age he has reported statistical findings based on data that (presumably) have never been collected. In *New Scientist* a survey among

[9] The whole data documentation standardization issue is now being addressed also within IFDO.

readers soliciting known cases of „scientific" reports based on erroneous data resulted in 199 acceptable responses. Apparently, such examples of „research" on the continuum from swindle to more or less conscious but severe misinterpretations in quantitative social science research are legion, and in a number of cases political directions have been based on the research; and probably the known cases constitute only the tip of the iceberg[10].

Vis-a-vis the public, Danish newspapers and Radio/TV persist in their interpretation and speculations based on statistically insignificant ups and downs of political parties reported from opinion polls. In many countries, statistical figures reported seem to be dependent on power elite preferences rather than enumeration procedures.

Social scientists are responsible, and probably the hoped for impact of ethic codes is overoptimistic. In the „hard" sciences, research results can frequently be verified by repetition of the experiment; in the „soft-data" social sciences, measuring unique (i. e. non-repeatable) phenomena, control of results and conclusions are possible only if data and full documentation are readily available to all.

In most research projects, certain „make-up" processes are necessary to produce a decent report from less decent data; in this paper, we are too polite to discuss the issue in further detail. However, there are cases where a random number generator would be the cheapest and most harmless data source; the quality of the research is not determined only by the sophistication of the analysis programs applied.

3.2. Methodological Aspects

In effect, the data documentation requirements outlined in subsections 2.3. and 2.5. above are trivial check-lists only. All this information must be in the head of the primary investigator, the only problem being to get it down on paper (better: into the computer) in a structured way. The structuring of this methodological information may have a positive effect on research quality – as may the consciousness that secondary analysts may criticize. The reported failures and errors may have an unexpected cumulative effect on social science research quality over time.

It should be stressed that we are not statistical purists: We do not advocate that, for example, data cleaning be continued until the last invalid code has been corrected; we do advocate, however, that all known imperfections be reported, that warnings be provided of all lacunae. In short, we ask that all methodological considerations and decisions be reported at the time and place they are relevant. Computer memory is more persistant than human memory.

[10] For a fuller description of these events, see Science 26 (1976), resp. New Scientist, September 2 and November 25 (1976).

3.3. Economic Aspects

In the area of sample *surveys*, the Technical Section of the Danish National Institute of Social Research has found it cheaper to clean and document data files for general use before the primary analysis is started. The time and money savings in cases where reports on new issues can be based on existing well-documented files is considerable. The potential accumulation of new knowledge based on several well-documented files (longitudinally, cross-sectionally, cross-nationally) is, in the end, an economic benefit. This is an area that data service organizations are now moving into.

In the fields covered by *statistical bureaus* the production of printed statistical information is still considered the main function. Despite the fact that the statistical publications are produced by means of a computer, we know of several examples where, for example, researchers have punched the figures from the printed publications rather than requesting the machine-readable file. The level of servicing of machine-readable files from statistical bureaus to external users is relatively low.

3.4. Historical Aspects

The machine-readable divisions of the traditional archives acquire an increasing amount of tape reels containing process-produced and statistical information. With the automated text-processing revolution of the coming decade these information repositories will find themselves busy. Given the present documentation standards of public registers, the problems (from an archival point of view) in handling on-line real-time databases, and the scarce allocation of resources to the storage-display functions in National Archives, there is a risk that future historians will find problems in dealing with data from the seventies.

4. Who Produces the Documentation

Let us make explicit what has been implicit above: It is the data collector who produces the data documentation, not the storage-display agent (archive, service-section of a statistical bureau, or data service organization). The latter may develop standards and set directions, but the former has the information necessary for good data documentation.

Long articles could be written on the many cases in which it has proven impossible to re-use existing data. Storage of undocumented machine-readable data is organized waste of time and money.

5. Summary on the HOW

Having touched on the what, why and who, we shall now try to sum up how new attitudes and behavior on the part of the data producer may be implemented. The list of directions below by no means claims to be complete in any sense; rather, the directions listed are the ones that are envisioned or being implemented within data organizations.

5.1. Focussing on the Issue

A necessary condition for better documentation standards is to focus much more sharply on the issue than we usually do: Within social science periodicals, such „technical" discussions are rare; in the textbook area, we have many monographs on sociological method where the technicalities of computer-data-handling are only mentioned, we have other textbooks on advanced analysis techniques – but we have few books on computer usage for data entry, editing, cleaning, and documentation. Within the area of software the situation is similarly biased: There are many programs (packages) for statistical analyses, structure searching and even content analysis, but only a few programs for trivial data processing tasks – and generally poor facilities for integration of documentation with the data. (ZAR and OSIRIS are exceptions from this rule as both work on text-data; typically enough, these program systems have been developed within the biggest data organizations)[11].

Although it seems necessary that „data pushers" in the data organizations take the lead in this process of change of attitudes, some help is needed from the social science professionals. (In Europe, the two groups happen to overlap which is an advantage). Within the university world, student reports, doctoral theses and other reports on quantitative research the quality of methods applied in the data handling process should be considered in awarding merit. The collection of a solid data base accompanied by published data documentation is (even if the associated analysis report may be poor) a contribution to social science resources; a research report based on data analysis is nothing without proper technical reporting because interpretation of analysis results is impossible.

[11] The OSIRIS program package for data handling and statistical analyses is the only one among the larger analysis packages used in Europe that has unlimited and flexible built-in facilities for an integrated data documentation; the extended OSIRIS dictionary-codebook is input to the ZAR retrieval system used for text-retrieval by the ZA and DDA. (The two systems were developed by ICPSR/ISR, Ann Arbor, resp. the ZA, Cologne – in the later phases in cooperation with the DDA).

5.2. Giving up Data Monopoly

There is still the monopolistic attitude to overcome – and this attitude is seen in private organizations, among individual researchers, and in public research and statistical bureau alike. In the United States, data can be bought whereas in Europe the transfer of data is based on good-will. Even in Eastern Europe, the private-property-attitude concerning data prevails.

Explaining free data exchange as a quid-pro-quo arrangement from which everybody wins and nobody loses; providing that the anonymity of the respondents will be guaranteed; giving the data producer control of access – all of these good arguments do not always convince the data producer that he should deposit data with a data organization. Again, we end up with the data documentation obstacle: Either the data producer is afraid that the data organization will take too much of his/her time when processing the data file – or the data producer simply does not want to show dirty underwear in public; in the latter case it is a poor comfort that this grey shade is the rule rather than the exception.

5.3. Offering Services

From the point of view of data organizations, the most efficient acquisition policy may be to offer better services in terms of textbooks, software, technical aid, methodologically oriented courses, etc. Seen from the economic angle, it is probably less expensive to offer such services to primary investigators than to do a lot of documentation and processing of the data after having acquired the file from the donor.

5.4. Enforcing Free Data Flow

Science foundations and other research granting agents may make it a condition for grants that quantitative data be made available, e. g. by being deposited with a data organization. The Danish SSRC does this, and the clause applied adds that the data should be properly documented according to standards set by the Danish Data Archives (which itself is an SSRC initiative).

A different initiative has been adopted by the *Journal of Personality and Social Psychology:* The authors of articles shall at the latest five years after the publication give access to underlying data files.

5.5. Education

No doubt only the next generation of social scientists will learn fully how to use the computer. Consequently, it is very important that social science students be offered methodological courses where computer-use, statistics, and application of the theories of their discipline are integrated. The development of instructional data packages some of which can be used at the undergraduate level is an important contribution in this area[12].

5.6. Infrastructural Considerations

For future social scientists exploitation of the vast data bases of statistical and administrative nature will be of great interest. Data Laws may be a new obstacle to secondary analysis; in some countries, government commission reports reveal the fact that the commission has had a double target: To protect individual integrity and to secure an information monopoly for the public bureaucracy. We may be building a barrier that makes infrastructural developments in the sense of a more smooth flow between public institutions and the research community almost impossible, thus adding to the historically determined inertia of giant data collection agents in the public sector.

However, Data Laws may have a positive effect on the quality of public data as well; with strict rules for error corrections and data organization and management, the documentation standards of process-produced data may well improve considerably.

After Data Laws have been passed the communication and flow of data from public agents to the research community may improve; at present, data producers in the public sector are reluctant to engage themselves in regular data transfer arrangements partly because they do not want to contribute to even stronger data legislation.

[12] Development of instructional data packages was started with the SETUPS series being a cooperative effort of the ICPSR and the American Political Science Association. Now, the International Social Science Council is sponsoring the development and testing also European teaching packages.