Documenting Data for Secondary Analysis : the Primary Producer's Role and Responsibility

by Bridget Winstanley [1]
ESRC Data Archive, University of Essex, U.K

Background and Acknowledgements
This paper is based on a session and a round table lunch discussion on the same theme which took place at the IASSIST Conference held in Edinburgh in May 1993. The session was convened by Sue Dodd and Bridget Winstanley. Papers by Laura Guy, Joanne Lamb, Marcia Taylor, Paul Child and Kevin Schurer, as well as the numerous participants at the round table discussion have all contributed to the ideas presented here as have the members of a European Committee on Documentation Guidelines, set up by the ESRC Data Archive earlier this year. This committee has representation from the ESRC Data Archive, the Office of Population Censuses and Surveys, Social and Community Planning Research, the British Household Panel Survey and other areas of the British academic research sector and the Steinmetz Archive in the Netherlands.

The Need for Guidelines
We start from the basic premise that the person or persons best placed to document their data are the primary producers of those data. It is axiomatic that their knowledge of the data must be more complete than anyone else's. Yet many primary producers ofdata are reluctant to create documentation of a standard whichgoes beyond their immediate needs for their own analysis of thedata. The reasons for this reluctance, when it occurs, are obvious. The creation of documentation of a substantively and physically high standard is time-consuming and expensive. Thereare apparently few incentives to producing such documentation. The culture of data sharing is still largely in a state of infancy even after at least a quarter century of data archiving. And additionally, it is not always apparent to the primary producer what the secondary analyst requires in the way ofdocumentation.

The primary producers who fail to document their data to an acceptable standard must be balanced by some shining examples ofgood practice in this field. Some of the most recent of these include The British Household Panel Survey's two volume user manual(1) and the U.K. Employment Department's user guide to the Quarterly Labour Force Survey (2), both of which are available as machine-readable text files at a much lower cost to the user, as well as appearing in printed paper form. North America can show many examples of data which are well documented by the producer for public use, including the General Social Surveys produced at the National Opinion Research Center (3). A recent publication by the Steinmetz Archive in the Netherlands documentsa dataset put together from a time series of NIPO polls (4) with thoroughness and consideration for secondary users of the data.

Despite these fine achievements, and many others, by individual research projects, there is much more that data archivists and librarians can do to promote good documentation by primary producers of data. The arguments for doing so encompass both the promotion of good practice and necessity arising from financial and economic constraints facing disseminators. We have already stated what we take to be self-evident, that primary producers are capable of producing the best documentation because of the familiarity with the data. The further imperatives for persuading primary producers that they have a role and a responsibility towards the documentation of their own data lies in the decreasing resources and increasing material coming into data libraries and archives. Many can no longer afford to create documentation for all (indeed any) of the datasets which they distribute and in any case the upgrading of poor documentation after the original project is over is frequently painful and unsuccessful: memories have dimmed and in many cases the original investigators have dispersed. Yet datasets which are inadequately documented are of no use at all to the secondary users to whom the data are being distributed.

A further important incentive to the production of good documentation was described by W.J. Bradley at the IASSIST/IFDO 93 Conference (5). The sponsors of major data collection exercises, typically government departments and other policy-making organisations, expect more for their money than data. They expect information. According to Bradley, policy advisors are often quite desperate for timely, relevant information. Given their wide-ranging and often unpredictable requirements, advisors and decision makers are a prime target audience for easy, responsive secondary data analysis services that integrate and draw upon the broadest possible base resources. Bradley and his colleagues have created software which demonstrates how good documentation, when standardised and

structured, can integrate and front-end rapid and easy access to the data resources that have been documented in this way (6). They also describe how such documentation can actually serve to facilitate the creation of information and knowledge products which in turn can be integrated for re-use in information retrieval The development of documentation guidelines, together with associated methods of standardisation, are keys to the knowledge delivery process.

Strategies for Improved User Documentation

There are several lanes in the highway which leads towards the ultimate goal of improved documentation by producers of data. We need to convince data funders of the economic arguments in favour of improvements in the standard of documentation. We need to convince data producers of the value of good documentation to the organisation of their own research, as well as of the recognition of their work which will come from their work being re-used and acknowledged. We need to convince secondary users to afford this recognition to primary producers. Finally, we need to provide support to primary producers by developing and distributing guidelines on the production of documentation.

The case to be made to the funders of data is, as indicated in the previous paragraph, primarily an economic one. Many funding bodies are indeed aware of the wastefulness of funding projects with major data-collection components without ensuring that the data are made available for further research beyond its primary research aims. In many cases they are aware, too, that a major constraint on the re-use of data is the lack of adequate documentation. There is sometimes a perception, however, that the disseminating agency, usually a data archive or data library, will document the data, so the producer does not need to move beyond minimal standards. We must make the case that producers are better placed than archivists to create documentation of a high standard for their own data and that it is more cost effective for them to do so. A certain amount of data processing and standardisation will always be necessary in the archive or data library, but the better the incoming documentation, the better the outgoing data and documentation. Funders are in a powerful position to provide incentives in the form of additional funds for documentation procedures within the original project funding as well as penalties in the form of blacklisting for those who do not document their data adequately. The judgement as to whether the data are adequately documented for secondary research will probably be the archive's and for this reason we need minimum standards in the form of guidelines.

Data archives and librarians will rely largely on funding bodies to provide the penalties for inadequate documentation. But they have a major role to play in persuading their depositors or donors of the incentives for providing high quality documentation. Above all, the case has to be made for making their data widely usable. Why should they care? Because usage can be reported back to funding bodies as an argument for more funding; because when data are well-documented there is no need for the constant answering of queries from secondary users; and because usage will bring citation and recognition. Here we, the data librarians and archivists, have a task ahead to ensure that use of data which leads to publication also leads to the citation of the dataset. The rules of citation for datasets are well established (see Dodd (7)) but we can do more to ensure that they are observed. A scan of examples reveals also that there needs to be clarification on whether the documentation or the data, or both, are being cited. Of the examples given above, only the General Social Survey's documentation (3) gives guidance on both the citation of data with documentation and the documentation alone, although the ESRC Data Archive's citation guidance does make it clear that the citation shown is for data with documentation. The other two cases assume citation for documentation only. Guidance on citation should be included in all documentation, editors of journals should be approached to try to ensure their co-operation, and a constant stream of reminders published in newsletters and bulletins. Citation has its own rewards in the form of easier identification of data sources for those reading the citation, but also, of course, it ensures the recognition of the achievement of the producer of that dataset in making it publicly available. But citation can only take place when the dataset has a bibliographic identity conferred upon it by its documentation. Guidelines are required to show producers how to document their data in a way which will ensure this.

Existing and Future Guidelines

Guidelines already exist for creating the necessary elements for documentation. Two US examples are Carolyn Geda's Data preparation manual (8) and Richard Roistacher et al A style manual for machine-readable data files and their documentation (9). Other examples are the U.S. Bureau of Justice Statistics' Technical standards for machine-readable Data (10) and Patrick Collins and Jane L. Powers The preparation of data sets for analysis and dissemination : technical standards for machine- readable data (11). Excellent as they are, the earlier of these manuals are out of date and need revision while the latest (Collins and Powers) although providing a attractive introduction to the subject, focuses on the practices required by a particular archive (The National Data Archive on Child Abuse and Neglect at Cornell University) and is consequently short on general detail.

A new comprehensive set of guidelines, covering both

optimal and minimal standards, taking into account new media, new formats, new data collection techniques and a new archival environment, is urgently required. These should include a recognition of the fact that many social scientists are using and creating textual data, or mixed numeric and textual data in their research. It is important that new guidelines should recognise too, the considerable work already undertaken in the humanities and not to duplicate that work. The work of the Text Encoding Initiative should be brought to the attention of social scientists in a way which will be appropriate to their needs. Although the guidelines should deal with substance and content, format should not be forgotten. For many primary producers and the archives or data libraries which will be disseminating their data and documentation, the most convenient format in which to produce documentation will be machine-readable. In addition to providing a cheap and convenient means of disseminating documentation on the same medium as the data, machine-readable documentation opens the way to better information systems, allowing the prospective user to examine and compare documentation online before deciding on the appropriateness of a particular dataset for his or her particular research.

Once we have agreed on both optimal and minimal standards for documentation we need to think about how to get them accepted. If they have been developed in consultation with data producers and if they are attractive and easy to use, this will be easier. A printed paper version is indispensable but we must also develop software applications of the guidelines. Work in this area has already begun, notably by W.J. Bradley and his colleagues in the Social Environment group of Health and Welfare Canada. Their work on DDMS (6), a PC-based package for managing social science dictionaries and documentation takes into account the data elements recommended by Roistacher and provides an easy way to manage data as well as ensuring that these data will be well- documented. Such easy-to-use software in the hands of data producers will be an incentive to the production of complete documentation. The further work by Bradley, Hum and Khosla on DAIS (Data and Information Sharing) (12) shows how easy, end-user access to data can be provided by documentation that has been structured and standardised via DDMS. This system provides a vital incentive to the funders of data who are themselves able, via this system, quickly to locate relevant data items from a broad array of datasets and generate their own analyses using software of their own choice.

Other work on codebook software has been carried out by the Swedish Social Science Data Service and further work on codebook production is under way as an IASSIST Action Group led by Karsten Boye Rasmussen of the Danish Data Archives. While recognising the contribution this will make to the sharing of data through data archives, this paper, because it is concerned only with the primary producer's role and responsibility, does not aspire to enter into the current debate, conducted largely through the IASSIST listserver, on the desirability of replacing OSIRIS as a codebook tool. It is vital, however, that before we undertake the publicity and training required for the acceptance of software products, we are agreed on the substance of the guidelines for the documentation of data.

Conclusion

Penalties, incentives and support all depend upon the existence of guidelines for documentation. Funding bodies have to be persuaded (as many already are) that the provision of funds for research projects to collect data at great expense without making provision for the wider use of these data is intolerably wasteful. For some, such as large governmental organisations, good documentation is essential for sharing within their own organisations, and all that is required is some guidance on how to do it in a way which has a broader application outside their own spheres. Other types of funding organisations, who have traditionally seen a single report as the end product of their sponsorship, need to be made aware of how much further their money will go if many reports and analyses for different purposes and by different researchers can result from their investment. Their role with regard to the documentation of datasets which they have funded should be to withhold further funding if the data are not sufficiently documented for further research (stick) and to provide an element of funding sufficient to ensure that the data are documented (carrot). Primary researchers have to be persuaded (as many already are) that the creation of a dataset which can be used by others is worthy of recognition, acknowledgement and citation in the course of scientific research and public policy planning. Secondary researchers, those making public policy, and the editors of journals should be persuaded to provide the recognition, acknowledgement and citation. The wider use of data and the recognition of the primary producers is dependent on the quality of the documentation which accompanies the data. The quality of the documentation will depend on the guidelines which we, the data librarians and archivists whose task it is to facilitate the flow between primary and secondary researchers, can provide to primary producers.

1 Paper presented at IASSIST/IFDO'93 Conference, Edinburgh, Scotland.

(1) Taylor, Marcia Freed (ed.) (1992) The British Household Panel Survey user manual. 2v. Colchester: University of Essex.

(2) Great Britain. Office of Population Censuses and Surveys. Social Survey Division (1992) Quarterly Labour Force Survey user guide. Colchester: ESRC Data Archive [distributor].

(3) Davis, James Allan and Smith, Tom W. (1991) General social surveys, 1972-1991 : cumulative codebook. Chicago: National Opinion Research Center.

(4) Eisinga, Rob and Albert Felling (1992) Confessional and electoral alignments in the Netherlands, 1962-1992 : documentation of social background variables of 1,067 national surveys conducted by NIPO from 1962 to 1992. Amsterdam: Steinmetz Archive.

(5) Bradley, W.J., Diguer, J. and Ellis, R.K.E., Methods for producing interchangeable data dictionaries and documentation. Paper presented at IASSIST '90, Poughkeepsie, N.J. Social Environment Information Health and Welfare Canada, 1990.

(6) Bradley, W.J., Ruus, L., Ellis, R.K.E. and Diguer, J., DDMS : a PC-based package for managing social science data dictionaries and documentation : reference manual. 9th draft ed. Social Environment Information Health and Welfare Canada, 1991.

— DDMS [computer files]. Social Environment Information Health and Welfare Canada, June 1991.

(7) Dodd, Sue A. Bibliographic references for computer files in the social sciences : a discussion paper. IASSIST quarterly, v.14, no. 2, Summer 1990.

(8) Geda, Carolyn Data preparation Manual. ICPSR, 1980.

(9) Roistacher, Richard A style manual for machine-readable data files and their documentation. Urbana: University of Illinois, 1978.

(10) U.S. Bureau of Justice Statistics Technical standards for machine-readable data

(11) Collins, Patrick and Jane L. Powers The preparation of data sets for analysis and dissemination : technical standards for machine-readable data. Ithaca: National Data Archive on Child Abuse and Neglect, 1991.

(12) Bradley, W.J., Hum, J. and Khosla, P., Metadata matters : standardising metadata for improved management and delivery in national information systems. Paper presented at IASSIST/IFDO '93, Edinburgh. Social Environment Information Health and Welfare Canada, 1993.