Towards standards for the description of machine-readable historical data

Reinke, Herbert

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

TOWARDS STANDARDS FOR THE DESCRIPTION OF MACHINE-READABLE HISTORICAL DATA

Herbert Reinke

Some years ago, the Center for Historical Social Research has started to archive and to disseminate machine-readable historical data for comparative and for secondary analyses. (1) This work has been accompanied by the development of a specific instrument for describing machine-readable historical data. This instrument has to meet the information needs of users of machine-readable historical data, information needs which are different to those of the users of machine-readable survey data. In this paper, standards for the description of machine-readable historical data are proposed which are also designed to become reporting standards for primary researchers in describing their data.

Comments on the proposals for the description of machine-readable historical data are highly appreciated.

THE STATE OF THE ART: SOCIAL SCIENCE DATA ARCHIVES AND THE PRACTICE OF DESCRIBING MACHINE-READABLE HISTORICAL DATA

The practice of archiving and disseminating machine-readable data is not that new to the social science data archive community: The Inter-University-Consortium for Political and Social Research has since its beginnings also been engaged in archiving and disseminating historical data. Various other social science data archives keep historical data in their holdings. Archiving and disseminating are also part of the activities of some social science data archives, e.g. the Norwegian Social Science Data Service, the Danish Data Archive, and the Survey Archive at the University of Essex. All in all, the social science data archives have already invested some work in machine-readable historical data. Nevertheless, archiving and disseminating machine-readable historical data cover only a small portion of the activities of the data archives. The majority of their holdings is of survey origin, the majority of their users come from the survey research community. This has "quite naturally" led to a situation, where only the information needs of secondary users of survey data have been met by the data organizations. The social science data archives are not to blame for this: Their users determined the priorities of the data archives, and until recently, the users have been predominantly survey researchers.

Address all communications to: Herbert Reinke
Zentrum für historische Sozialforschung, Universitätsstr. 20,
D-5000 Köln 41, West Germany

The Center for Historical Social Research had to fill a gap: to develop appropriate standards for describing machine-readable historical data.

STANDARDS FOR DESCRIBING MACHINE-READABLE HISTORICAL DATA (MRHD): THE STUDY DESCRIPTION SCHEME

In the past, almost all social science data archives developped some sort of data set description; which was to provide information about machine readable data files. The technically most advanced description applied at present is the "Study Description Scheme" which has been worked out in a joint venture by a group of European social science data archives concentrating on survey data: The Danish Data Archives (DDA), the Steinmetz Archives/Amsterdam, and the Zentralarchiv für empirische Sozialforschung at the University of Cologne. (2)

Four archives currently employ this study description scheme: The Danish Data Archives, the Zentralarchiv für empirische Sozialforschung, the Steinmetz Archives and the Leisure Studies Data Bank at the University of Waterloo, Ontario. The SSRC Survey Archive at the University of Essex and the Belgian Archives for the Social Sciences are presently assessing the scheme.

The purpose of this study description scheme is fourfold:

- It is to list all items of information about a data set a "producer" should record and report in its own publications in order to enable others to assess his work.
- It is to provide all the information about a data set "that a researcher needs to conduct secondary analysis".(3)
- It should be applicable as a retrieval instrument.
- It should be applicable as an intra-archival log book which provides information about the current state of data set processing.

The following listing of the headings of the sections included in the description scheme shows that it is designed primarily to describe survey data:

1 Identification and acknowledgement, origin of the data

This section currently includes the following information:

1.1 Title of study
1.2 Local (data) archive where the study is stored
1.3 (Data) archive where the study was originally stored
1.4 Depositor (donor) of data
1.5 Date of deposit of data in the archive
1.6 Principal Investigator
1.7 Data collected by
1.8 Research initiated by

PROPOSAL No. I:

A category "Data collected from" is to be included. Under this heading the sources should be referred to which were used by the primary researcher for his data collection. All archival sources, other unpublished sources, published source material, official statistics and other sources are to be quoted in detail. This category is essential for historical social research.

2 Analysis conditions

This section provides information on the data-gathering operations of the primary researcher. It allows the secondary analysist to evaluate the scope and quality of the data set concerned. Currently the study description scheme includes the following items:

2.1 Research topic (abstract)
2.2 Kind (types) of data
2.3 Units
2.4 Number of units
2.5 Dimensions of data set (e.g. variables per case, cards per case etc.)
2.6 Completeness of study stored
2.7 Time dimensions
2.8 Definition of total universe (universe sampled), target
2.9 Sampling procedures
2.10 Dates of data collection
2.11 Methods of data collection
2.12 Type of research instrument
2.13 Procedures to minimize losses
2.14 Data gathering staff
2.15 Characteristics of data collection situation noted
2.16 Weighting criteria and procedures
2.17 Other analysis conditions

Except for additional information where the items were too narrowly designed even for the description of survey data (2.3 units, 2.7 time dimensions, 2.9 sampling procedures, 2.11 methods of data collections, 2.12 type of research instrument), the following new categories should be included

PROPOSAL No. II:

A category should be included which allows to "judge" the correspondence between target and frame. This category reflects one of the core problems of the data gathering process in historical social research: Very often the relation between the target of a study, i.e. a group of people (or aggregates or something else) about which inferences are to be made in a study, and the frame, i.e. the approximation to the target population in terms of the units (cases) sampled, is very critical. Just one example from historical community

studies: The target is supposed to be the total population of a given community, the frame could be: All households in that community. The critical problem now could be: The listings of the households (e.g. tax registers) exclude the non-tax-paying parts of the population, thus excluding the extreme ends of the social hierarchy in the Ancien Régime: beggars and vagants, priests etc..

That there might be differences between targets and frames also within survey research is generally accepted. The decision to be rather short in this matter might have been a consequence of the existence of a specialized literature and professional standards within the field organizations. In the area of textual analysis and document analysis/analysis of process-produced data, the study description scheme has to cope with a still low level of standardization.(4)

PROPOSAL No. III:

In the description instrument currently employed, the category 2.13 (procedures to minimize losses) is "underdeveloped", at least in terms of the information needs of the secondary analysist of historical data. It is proposed to split that category: One category which describes procedures to minimize losses at the level of the units of observation: Record-linkage procedures, and another category which describes procedures at the variable level: Matching operations, i.e. fusing information on same units from different sources.

PROPOSAL No. IV:

Category 2.16 (characteristics of data collection situation noted) in its present form (the subcategories - not listed here - all refer to interview situations) have to be supplemented by a description of the data collection situation "at the sources" (document archives, organizations). Category 2.16 in the currently employed scheme is meant as an indicator of possible biases in survey data as a result of characteristics of the data collection situation. The additions proposed here are due to the knowledge that biases could also occur as a consequence of the specific data collection situation in historical social research. The categories proposed here reflect such possible biases. Specifically the following should be included:

- origins (Entstehungsbedingungen) of the sources concerned
- characteristics of the storage of the sources concernd in the document archive or in a system of administrative record-keeping
- access to and documentation of the sources
- data collection situation in the document archive or record-keeping organization

Knowledge about the "how" and "when" of the origins of a specific source is important: It gives an idea about the "administrative style" of source-producing persons and/or agencies, especially the way information was handled and compiled.

Characteristics of the information-gathering-process on the level of the "producer" of sources are very often known to archivists in document archives or researchers using process-produced data. To know about the information-gathering -process on the level of the data producer is very important because of its implications for the research use of the data: Sources can sometimes be taken at face value but more often they should be used as indicators for anything theoretically and empirically significant. For using information in sources as specific indicators you ought to know what was originally intended and actually measured.(5)

Distortions could occur as well as a result of specific storage conditions in a document archive or organization. To have knowledge of possible distortions because of storage peculiarities is also important and should therefore be included.

The same applies to information about the access to and the documentation of sources. A limited access to data or an incomplete documentation of data could have severe distorting effects, e.g. via limitations on accurate sampling procedures etc. In order to have information about such possible biases which could influence research results, background information should therefore be included.

3 Reanalysis conditions:

This section of the study description scheme is - besides a proposal for minor changes - to be employed by historical social research in its present form. The same applies - without the notation of minor changes - for the section of the study description scheme dealing with:

4 References to relevant publications.

The last section

5 Background variables included

originally provided for a "categorised record" of information commonly sought in social surveys: "Face-sheet variables" such as age, sex, occupation, income, household characteristics etc.

The present "state of the art" in historical social research requires that the compatibility of socio-economic variables has to be evaluated first: That is the task of the variable comment proposed

here for the documentation of variables in historical social research.

GUIDELINES FOR A DOCUMENTATION OF MACHINE-READABLE HISTORICAL DATA: THE CODEBOOK AND THE VARIABLES COMMENT

In survey research, a codebook is an instrument which provides information about variables in a data set such as:

- a variable name, text, or label
- the code value labels
- location of variable on the storage medium (card, tape, etc.)
- the marginals of the variables concerned.

It is understood that the above items give "all information" to "understand" the variable. Additionally, the inclusion of coding instructions is requested. It is commonly thought that to give the full wording of an interview question is sufficient for understanding what a variable means. The everyday-life-language in survey-questionnaire wording facilitates the understanding of the meaning of a variable. But to understand the "meaning" of a variable already becomes difficult if a secondary analysis of survey data from the 1950ies is to be carried out. The everyday-life implications of a variable constructed in the early fifties may no longer be compatible to today's everyday-life experiences: To understand a variable which was constructed about thirty years ago, you have to have information about the historical context in which the question was posed, that is you have to have information about the historically specific meaning of that variable. This holds true of course for all variables for time periods out of the range of the present-day-understanding of variables. As a consequence it is proposed:

PROPOSAL No. V:

to include in a historical social research data description detailed variable comments.

As this variable comment could turn out to be quite lengthy it is proposed to add to the variable name a variable comment as a further component of the data set description. This variable comment should include a number of explanatory items important for understanding the variable.

PROPOSAL No. VI:

Codebook:

1. No. of variable
2. Name and full text of variable
3. Code value
4. Card/column (if card image)
5. Width of information
6. Measurement level
7. Marginals
8. Variable comment
9. Source

Items Nos. 1 - 2 are to recognize the variable, items Nos. 3 - 7 give further information on the attributes and distribution of the variable, whereas item 8 is to contain the detailed variable comment which enables an understanding of the variable in a historical context.

PROPOSAL No. VII:

It is proposed to include a listing of the source(s) which provided information for the variable concerned.

Further proposals concern the specific problems encountered when aggregate variables are to be described.

PROPOSAL No. IX:

Give geographical and time references when "aggregate variables" are to be documentated.

PROPOSAL No. X:

Retain information on the original ("primary") measurement operations (tons of grain per acre, homicides per thousand of population etc.) when documentation is applied to aggregate data.

SUMMARY OF PROPOSALS FOR THE DESCRIPTION OF MRDH'S

No. I:

Include a category "Data collected from" in the study description.

No. II:

Give details of relations between target and frame.

No. III:

If applicable, give details of record-linkage and matching operations.

No. IV:

Give details of the specific data collection situation in historical social research.

PROPOSALS Nos. V - VI:

Add a detailed variable comment to the codebook.

PROPOSAL No. VII:

Include in the codebook references to the sources which provided information for the variable concerned.

PROPOSAL Nos. IX - X:

Include geographical and time references and retain information on the original measurement operations for aggregate data.

FOOTNOTES

1 QUANTUM INFORMATION No. 4, October 1977; Herbert Reinke, Archiving Machine-Readable Historical Data: Data Services of the Center for Historical Social Research, in: Historical Social Research, No. 12, October 1979, pp. 36 - 38.

2 For a detailed history of the beginning and the development of the study description scheme see: C. S. Brown, Access to Machine-Readable Social Survey Data, M. A., University of Sheffield, 1980.

3 C. S. Brown, p. 29.

4 Wolfgang Bick and Paul J. Müller, Probleme der Nutzung prozeß-produzierter Daten, Köln, Oktober 1980.

5 Wolfgang Bick and Paul J. Müller, The Nature of Process-Produced Data - Towards a Social Scientific Source Criticism, in: Jerome M. Clubb & Erwin K. Scheuch (eds.), Historical Social Research. The Use of Historical and Process-Produced Data, Stuttgart 1980, pp. 369-413 (= Historisch-Sozialwissenschaftliche Forschungen, vol. 6).