Information requirements and data description in historical social research: a proposal

Marker, Hans-Jörgen; Reinke, Herbert; Schurer, Kevin

DATA NEWS

INFORMATION REQUIREMENTS AND DATA DESCRIPTION IN HISTORICAL SOCIAL RESEARCH. A PROPOSAL

Hans-Jörgen Marker, Herbert Reinke, Kevin Schurer (•)

Abstract: Description and documentation is one of the major prerequisites for disseminating and exchanging machine-readable historical sources and data. This proposal for description and documentation items is an attempt for standardizing information requirements in historical research; it does not give recommendations for all possible description and documentation needs, but instead attempts to set essential parameters for describing and documenting machine-readable historical sources and data. In the appendix to this proposal, an example using the 1851 Census of England and Wales is given to provide a detailed illustration of the proposal.

1. Introduction

The proposal should be used as an annotated checklist for items to be included in the appendix to a research publication as well as a data archiving scheme which allows data archives to obtain from the primary researcher those items of information necessary for secondary uses of the data. In addition, in its fairly general form, the proposal reflects the broad lines of information requirements in quantitative historical research. In the appendix to this proposal, an example using the 1851 Census of England and Wales is given to provide a detailed illustration of the proposal.

This proposal has an specific history: As an offshot of the workshop on standardization and exchange of machine-readable data in Göttingen 1985, a working group was formed by Kevin Schurer, Herbert Reinke and Hans-Jörgen Marker with the aim of proposing standards for the description of machine-readable historical data sets. Later that year this working group was recognized as a subcommittee of the International Commission for the Application of Quantitative Methods in History. A draft of the proposal for minimum standards for the description of historical data sets was presented at the "International Workshop on Standardization and Exchange of Machine-Readable Data in the Historical Disciplines" Graz 1986. This contribution

(•) Address all communications to: Hans-Jörgen Marker, Danish Data Archives, Campusvej 55, DK-5230 Odense M; Herbert Reinke, QUANTUM, Greinstr. 2, D-5000 Köln 41; Kevin Schurer, Cambridge Group for the History of Population and Social Structure, 27 Trumpington Street, Cambridge CB2 1QA.

focusses on presenting the description and documentation example. For explanatory comments on the description and documentation items, the authors refer to the contribution in the proceedings at the Graz workshop, which is obtainable by writing to them (1).

2. Description and Documentation as Standardization Efforts for Computer Applications and Quantitative Approaches in History

Combining historical investigation and the use of the computer is now a well established and broadly recognized approach in historical research. The late 1960's and the early 1970's saw the first - mostly anglo-american - attempts to introduce the computer into the art of history (2). During the seventies and the early eighties, historical research concentrated primarily on exploring and developing further the potentials of quantitative approaches and computer applications.
The major publications on the subject of the earlier period were primarily textbooks which tried to collect the available knowledge about "the ways to do" quantitative historical research or presented research results with the aim of demonstrating the possibilities of quantitative historical research (3). In contrast, the textbooks of the later period, published in a number of European countries and in the United States, were mainly concerned with applications, excluding debate attempting to legitimize their approach to the subject, a characteristic of many of the earlier textbooks (4).
The 1980's has witnessed a third phase of development, orientated primarily toward standardizing the products and processes of computerization. Standardization normally occurs in the history of scientific innovations, when a certain maturity is reached. This is the "normal science" level in terms of Thomas Kuhn's history of science approach (5). The maturity of computer applications and quantitative approaches in history is indicated by a number of factors:

- Quantitative history is more or less integrated into the historical curriculum.
- Articles on the computerized study of history are not only accepted by journals which concentrate on these approaches but by traditional historical journals with a broader coverage.
- Computerization is not restricted to certain research questions but covers almost all subject matters of historical research which allow quantitative investigations.
- The growing availability of machine-readable historical data and sources has resulted into considerable efforts for establishing specialised archives for machine-readable historical data and sources.

This establishment of a data-infrastructure for historical research has increased demands for standardization. These demands result from increasing data exchange and transfer activities between researchers. Consequently, investigations of machine-readable data may occur without either the involvement of the primary researcher who created the respective data set, or the manuscript archive from which the data set was made. A prerequisite for this secondary use of machine-readable historical data and machine-readable sources is an appropriate description and documentation of the available material.

3. Outline of the Description and Documentation Scheme

The following parameters for descripting and documentation machine-readable sources and data should be met:

1. Description and Documentation of Source Material

1.1 The normative administration of records

- The purpose of the source: Why was the source established? (1.1.1)
- The scope of the source: To whom was the source applicable? (1.1.2)
- Content of the source: What was to be recorded? (1.1.3)
- The time dimensions: When was the information in the source to be recorded? (1.1.4)

1.2 The actual administration of records

- Manners of recording (1.2.1)
- Styles of record-keeping (1.2.2)

1.3 Archival Procedures/History

- Results of rearrangements/transformations of the original record structure (1.3.1)
- Effects of intentional partial destruction of the record, carried out by archive personnel (1.3.2)
- Effects of unintentional (accidental) partial destructions (1.3.3)

1.4 Accessability of sources

1.5 Bibliographical Reference(s)

2. Description and Documentation of Machine-Readable Historical Data

2.1 Entities received in a free-format data set

2.2 Entities selected in a standardized data set

2.3 Attributes per entity

2.4 Values per attribute

Appendix

The Machine-Readable 1851 Census of England and Wales. An Example for the Proposed Description and Documentation Scheme.

The 1851 Census of England and Wales is the source which has been converted to machine-readable form more than any other in England. This description would be accompanied by a printout displaying the first few hundred lines of the file, the attribute values and the frequency of their occurence (6).

1. Description and Documentation of Source Material

1.1 Normative administration of records

General note: much has been published on the background, administration, useage and analysis of the mid-nineteenth century censuses. The major works are listed under section 1.5 below and users should consult these for further information.

1.1.1 The purpose of the source: why was the source established?

To enumerate the population of the country. The schedule that householders were required to fill in carried the statement that, "The Return is required to enable the Secretary of State to complete the Census: which is to show the number of the population - their arrangements by ages and families in different ranks, professions, employments, and trades - their distribution over the country in villages, towns and cities - their increase and progress in the last ten years".

1.1.2 The scope of the source: to whom was the source applicable?

The entire population of England and Wales, including all persons on board vessels in English coastel waters. (Yet, see section 1.4 below). Persons on vessels and those resident in institutions (ie. schools, prisons, barracks, workhouses, hospitals, etc.) were enumerated using special schedules and recorded in separate enumeration books from the normal resident population. Similar enumerations were also made in Scotland, Ireland and the Islands in the British Seas.

1.1.3 Content of the source: what was to be recorded?

For each individual:

(i) Surname and Christian Name
(ii) Relationship to the Head of Family
(iii) Marital Condition
(iv) Sex
(v) Age last birthday
(vi) Rank, Profession or Occupation
(vii) Where Born (if in England or Wales, name of county and parish; if elsewhere name of country, and if overseas state if British Subject)
(viii) If Deaf and Dumb or Blind

This information was organised into 'households' and 'houses' with the address of each house also being given. Information on disabilities (viii above) was not retained on this dataset.

1.1.4 The time dimension: when was the information in the source to be recorded?

The census was to record all persons on the night of Sunday 30 March 1851. In this respect it should be noted that the population was enumerated at the actual place they slept on the night of 30 March, regardless of whether or not it be their usual place of residence.

1.2 The actual administration of records

1.2.1 Manners of recording

The country was divided in Registration Divisions; Registration counties; Registration Districts; Registration Sub-Districts and Enumeration Districts. A single enumerator was appointed for each enumeration district. In the week ending 29 March 1851, the enumerator was required to deliver a schedule form to each householder in his district, the maximum size of each district being restricted by the number of householders the enumerator could visit in a single day. In rural areas, an average sized parish would

equate to one enumeration district. The schedules were then to be collected on the morning of Monday 31 March and upon collection examined to ensure that it had been answered correctly. Then during the week up to 8 April, the enumerator had to enter the information from the householder's schedules into an 'enumeration book'. This was to be carried out using standard contractions for relationship, condition and occupation, yet was often the source of much error (see Tillott, 1972, section 5). The enumerator then sent his book to the local Registrar for inspection who in turn sent it finally to the census office in London for checking and analysis. The householder's schedule forms were later destroyed, thus the 'enumerators' books' are now the source, not the original census schedules.

1.2.2 Styles of record-keeping

Whilst in the London census office, the enumerators' books were marked with a series of ticks and annotations by the checking clerks who compiled the information for the official published returns abstracted from the 1851 census.

1.3 Archival procedures/history

No relevant information.

1.4 Accessibility of sources

Microfilms of the census are available at the Public Record Office, London, and can be inspected once a hundred years has elapsed since the taking of the census. The original documents can only be publicly inspected under special circumstances; this can create a problem since the annotations and indices of the source have been made by local and family history societies.

See: Census Indexes and Indexing, (ed) J. Gibson and C. Chapman, Federation of Family History Societies, 1981.
Marriage. Census and Other Indexes for Family Historians, (ed) J. Gibson, Federation of Family History Societies, 1984.
Census Returns 1841-1881 on Microfilm: A Directory to Local Holdings (ed) J. Gibson, Federation of Family History Societies, 1984.

All obtainable from J.S.W. Gibson, Harts Cottage, Church Hanborough, Oxford, 0X7 2AB.

1.5 Bibliographical Reference

Guide to Census Reports: Great Britain 1801-1966. Office of Population Censuses and Surveys, HMSO, London, 1977.
This deals mainly with the scope of the enquiry and the nature of the official publication of findings.

Nineteenth-Century Society. Essay in the use of quantitative methods for the study of social data (ed) E.A. Wrigley, Cambridge University Press, 1972. See especially essays by M. Drake (Chapter 1), P.M. Tillott (Chapter 3), and W.A. Armstrong (Chapter 6).

The Census and Social Structure (ed) R. Lawton, Frank Cass, London 1978.
A Guide to nineteenth-century census enumerators' books. Historical Sources and the Social Scientist, The Open University Press, Milton Keynes, 1982.

For a bibliography of research using the census see: Census Enumerators' Books: an annotated bibliography of published work based substantially on the nineteenth-century census enumerators' books (ed) C.G. Pearce and D.R. Mills, Faculty of Social Sciences, Open University, Milton Keynes, 1982.

An updated and extended version of this bibliography is available in computer print-out form from: The Data Editor, Cambridge Group for the History of Population and Social Structure, 27 Trumpington Street, Cambridge, CB2 1QA.

2. Description and documentation of machine-readable historical data

2.1 Entities received in a free format dataset

The data are written in standard 'record-card' image, ie 80 character length record, in fixed block format.
With the exception of the special case of two embedded blanks (see section 2.3 below) each record-line starts with a two character line type code (numeric) in columns 1-2. The meaning of these line types codes is as follows:

Type	Meaning	Frequency of occurrence	
10	File Header Record	1	
20	Census Data Type Record	1	
31	Cluster Identification Record	1)	One per cluster
32	Cluster Address Record	1)	
41	Registrar General's ED Classification Code Record	2))	
42	Public Record Office ED Classification Code Record	) 2))	One per ED
43	ED Description Record	2)	
50	Household Address Record		One per household
61	Person Record #1	)	
62	Person Record #2	)	One per person
63	Person Record #3	)	
64	Person Record #4	)	
70	End of House Record		One per house
90	End of File Record	1	

The historical structure of the file and the way in which the individual entities relate to each other is as follows:

The file consists of:
one FILE HEADER RECORD (type 10) followed by
one CENSUS DATA TYPE RECORD (type 20) followed by
one FILE BODY (a composite entity) followed by
one END OF FILE RECORD (type 90).

The FILE BODY composite entity consists of:
1 or more occurrences of DATA CLUSTER.
Each DATA CLUSTER consists of:
1 CLUSTER IDENTIFICATION RECORD (type 31) followed by
1 CLUSTER ADDRESS RECORD (type 32) followed by
1 ED FILE (a composite entity).

Each ED FILE consists of:
0 or more occurrences of an ED (a composite entity).

Each ED consists of:
1 REGISTRAR GENERAL'S CLASS. RECORD (type 41) followed by
1 PRO CLASS. RECORD (type 42) followed by
1 ED DESCRIPTION RECORD (type 43) followed by
1 ED BODY (a composite entity).

Each ED BODY consists of:
0 or more occurrences of HOUSE (a composite entity).

Each HOUSE consists of:
1 or more HOUSEHOLD (a composite entity) followed by
1 END OF HOUSE RECORD (type 70).

Each HOUSEHOLD consists of:
0 or more occurrences of HOUSEHOLD BODY (a composite entity).

Each HOUSEHOLD BODY consists of:
1 HOUSEHOLD ADDRESS RECORD (type 50) followed by
0 or more occurrences of PERSON (a composite entity).

Each PERSON consists of:
0 or more occurrences of PERSON RECORD #1 (type 61) followed by
0 or more occurrences of PERSON RECORD #2 (type 62) followed by
0 or more occurrences of PERSON RECORD #3 (type 63) followed by
0 or more occurrences of PERSON RECORD #4 (type 64).

2.2 Entities selected in a standardized dataset

Not applicable.

2.3 Attributes per entity

	Starting Position	length	Numeric/ Character
FILE HEADER RECORD (type 10)			
Attribute 1: File Title	3	25	$
Attribute 2: Data Collector's Name	30	30	$
Attribute 3: Date of Data Collection	60	8	N
CENSUS DATA TYPE RECORD (type 20)			
Attribute 1: Data Type	3	30	$
Attribute 2: Date of Census	35	20	N

	Starting Position	length	Numeric/ Character
CLUSTER IDENTIFICATION RECORD (type 31)			
Attribute 1: Cluster Identification Number	4	5	N
Attribute 2: Cluster Type	10	30	$
Attribute 3: Community Type	40	40	$
CLUSTER ADDRESS RECORD (type 32)			
Attribute 1: Cluster Country	3	15	$
Attribute 2: Cluster County	20	20	$
Attribute 3: Cluster Community	40	20	$
REGISTRAR GENERAL'S ED CLASSIFICATION CODE RECORD (type 41)			
Attribute 1: RG's enumeration dist. classification code	4	15	$
PRO ED CLASSIFICATION CODE RECORD (type 42)			
Attribute 1: PRO enumeration dist. classification code	4	20	$
ED DESCRIPTION RECORD (type 43)			
* Attribute 1: ED Description	3	1-n	$
HOUSEHOLD ADDRESS RECORD (type 50)			
Attribute 1: Household Schedule Number	3	6	$
Attribute 2: Household Address	10	70	$
PERSON RECORD #1 (type 61)			
Attribute 1: Forename	3	20	$
Attribute 2: Surname	23	25	$
Attribute 3: Relationship to Head	48	20	$
Attribute 4: Marital Condition	68	12	$
PERSON RECORD #2 (type 62)			
Attribute 1: Male Age	4	6	N
Attribute 2: Female Age	10	6	N
• Attribute 3: Occupation	20	1-n	$
PERSON RECORD #3 (type 63)			
Attribute 1: County of Birth	3	20	$
Attribute 2: Community (parish) of Birth	23	30	$
Attribute 3: Country of Birth and Nationality	53	27	$
PERSON RECORD #4 (type 64)			
Attribute 1: Person Identifier	4	4	N
Attribute 2: Household Identifier	9	4	N
Attribute 3: House Identifier	14	4	N
Attribute 4: Person Sequence Number	21	2	N
Attribute 5: Number of Persons in Household	24	2	N
Attribute 6: Code of Relationship	27	3	N
Attribute 7: Code of Marital Status	31	1	N
Attribute 8: Sex	33	1	$
Attribute 9: Age	35	4	N
Attribute 10: Code of Occupation	40	3	N
Attribute 11: County of Birth Code	44	8	$

	Starting Position	length	Numeric/ Character

END OF HOUSE RECORD (type 70)
No Attributes

END OF FILE RECORD (type 90)
No Attributes

- NOTES: Attributes of this kind have no maximum length. If the entry extends over the 80th character column a plus sign (+) is recorded in the 80th column position and the attribute is continued onto the next record line, starting in column position 3.

2.4 Values per attribute

The values of all attributes are listed out on the accompanying printout, with the exception of the Person identifier, household identifier, and house identifier. These three attributes are sequential counters, the first relating to each PERSON entity, the second to each HOUSEHOLD entity and the third to each HOUSE entity. Each of these counters is set to one at the start of a new ED BODY.

The given attribute names on the printout are as follows:

TITLE	File Title
COLLECTOR	Data Collector's Name
DATECOL	Date of Data Collection
TYPE	Data Type
DATE	Date of Census
CID	Cluster Identification Number
CTYPE	Cluster Type
COMTYPE	Community Type
CCNTRY	Cluster Country
CCNTY	Cluster County
RGCLAS	RG's ED Classification Code
PROCLAS	PRO ED Classification Code
EDDES	ED Description
SCHEDNO	Household Schedule Number
ADDRESS	Household Address
PNAME	Forename
SNAME	Surname
RELAT	Relationship to Head
COND	Marital Condition
MALAGE	Male Age
FEMAGE	Female Age
OCC	Occupation
BPCNTY	County of Birth
BPCMTY	Community of Birth
BPCTRY	Country of Birth
PERSONID	Person Sequence Numbers
MAX	Number of Persons in Household
RELA	Code of RELAT
MAR	Code of COND

SEX	Sex
AGE	Age
OCCODE	Code of OCC
CNTY	Code of BPCNTY

On the printout coded attributes are crossed against their parent value to enable researchers to note the classifications used. The occupation code is derived from the scheme used in the official publications of the 1861 census and modified by Booth and Armstrong. The scheme is discussed in W.A. Armstrong, (Chapter 6) Nineteenth Century Society (ed) E.A. Wrigley, Cambridge University Press, 1972.

NOTES

1. Published as: Hans-Jörgen Marker, Herbert Reinke, Kevin Schurer, Making Sense out of Historical Information. Towards Standards for the Description and Documentation of Machine-Readable Historical Sources and Data, in: Friedrich Hausmann et al. (eds.), Data Networks for the Historical Disciplines, Graz 1987.

2. William O. Aydelotte, Allan G. Bogue (eds.), The Dimensions of Quantitative Research in History, Princeton 1972; Charles M. Dollar, Richard J. Jensen, Historian's Guide to Statistics. Quantitative Analysis and Historical Research, New York 1971; Val R. Lorwin, Jacob M. Price (eds.), The Dimensions of the Past. Materials, Problems and Opportunities for Quantitative Work in History, New Haven 1972; Don Karl Rowney, James Q. Graham (eds.), Quantitative History Data, Greenwood 1969; Edward Shorter, The Historian and the Computer, Prentice Hall 1971.

3. Michael Drake, The Quantitative Analysis of Historical Data, Milton Keynes 1974; Heinrich Best, Reinhard Mann (eds.), Quantitative Methoden in der historisch-sozialwissenschaftlichen Forschung, Stuttgart 1977; Jerome M. Clubb, Erwin K. Scheuch (eds.), Historical Social Research. The Use of Historical and Process-Produced Data, Stuttgart 1980; Roderick Floud, An Introduction to Quantitative Methods for Historians, London 1973; Konrad H. Jarausch, Gerhard Arminger, Manfred Thaller, Quantitative Methoden in der Geschichtswissenschaft, Darmstadt 1985.

4. Jean Heffer, Jean-Louis Robert, Pierre Saly, Outils Statistiques pour les Historiens , Paris 1981; Norbert Ohler, Quantitative Methoden für Historiker, München 1980; Manfred Thaller, Einführung in die Datenverarbeitung für Historiker, Köln/Wien 1982; Konrad H. Jarausch, Gerhard Arminger, Manfred Thaller, Quantitative Methoden in der Geschichtswissenschaft, Darmstadt 1985.

5. Thomas S. Kuhn, Die Struktur wissenschaftlicher Revolutionen, Frankfurt 1967.

6. Much of this example follows the ideas put forward by John Wetford of the University of Edinburgh in his development of a Public Format for the ESRC backed project to make machine-readable a two percent national sample of the 1851 Census. See: J. Wetford, The Establishment of Portable Interchange Formats for Genealogical Data - Can We Hope to Reach an Acceptable Standard?, in: Computers in Genealogy, 1(7), 1984, S. 178-187, and Kevin Schurer, Census Enumerators Returns and the Computer, in: Local Historians, 16(6), 1985, S. 335-342.