Towards a standard for the description of historical datasets

Hall, Ninette van

Empfohlene Zitierung / Suggested Citation:
Hall, N. v. (1989). Towards a standard for the description of historical datasets. *Historical Social Research*, *14*(1), 89-117. https://doi.org/10.12759/hsr.14.1989.1.89-117

Historical Social Research, Vol. 14 — 1989 — No. 1, 89-117

DATA SECTION

Towards a Standard for the Description of Historical Datasets

*Ninette van Hall**

Abstract: Several countries are planning a data-archive facility for historical datasets. In the archives the data will be extensively documented. A standardized way of describing the machine readable datasets will facilitate the exchange of information. The Standard Study Description Scheme used at the long existing social science data-archives is - in unaltered form - not applicable to historical datasets.
There are many similarities between social scientific datasets and historical datasets. It is useful for historians to use a documentation standard that can be exchanged with the standards in use at social science data-archives. At the same time the specific demands of the historical datasets should be taken into consideration.

1. Introduction

The object of this paper is to convince everyone of the need of a standard for the description of historical datasets. With the ever growing amount of historical datasets, that are being created and hopefully will become available for secondary analysis, it is high time a standardized way of describing these datasets is developed.

In order to reach the largest population of historians it is important that these descriptions are made available from a centralized point. If descriptions of machine readable historical datasets are available to researchers

* Address all communications to: Ninette van Hall, Department of History, University of Leyden, Doelensteeg 16, NL-2311 VL Leyden, The Netherlands.

who want to do their research using those datasets, the access to those files is facilitated and thus the whole historical community can benefit.

In order to get the best information on available datasets one should aim at similarity and comparability in shaping its description. An overview will be given of developments that have contributed to the realization of a proposal for standard for the description of historical datasets. An answer will be given to questions such as what should be in it and why.

Structure of the Paper

First general information on the architecture and the functions of the Study Description is given. This is followed by a very brief history of the Study Description with special attention for developments in the field of historical datasets. The last part deals with differences between social scientists and historians and the resulting differences between social scientific datasets and historical datasets and their respective descriptions. Afterwards a proposal for the description of historical datasets will be presented on this topic. To start the actual presentation of this paper some background information is given on the Netherlands Historical Data Archive.

The Netherlands Historical Data Archive

Last april (1988) a study was started on the possible foundation of a historical data archive in the Netherlands. The initiative for this study was taken by the Steinmetz Archive, in cooperation with the Dutch Working Group on History and Computing. The research is funded by the Ministry of Education and Sciences.

The increasing use of computers in historical research is causing a rising number of datasets based on sources from the past in machine readable form. Recently a preliminary inventory of historical datasets in Netherlands was made (Historische databestanden in Nederland, P.F.A. de Guchteneire & S.C.M. de Schaepdrijver, Steinmetz Archive 1987). The diversity of the available material appears to be considerable. The state of preservation of the datasets however, is deplorable. Sometimes the files have simply been lost or discarded once their immediate purpose has been served. As yet, there is no central place where historical computer-readable files are catalogued, preserved and made available.

The central question of the feasibility study is in which way a proposed Netherlands Historical Data Archive could be formed. Several possibilities are to be evaluated. The research deals with the delimitation of the tasks of such an archive, the desirable organizational structure, the personnel and financial means needed, the information-technical aspects and the juridical and copyright aspects of archived historical data.

An other aspect that has to be investigated is that of the possibilities of national and international cooperation. In some countries machine readable historical data are already archived, mostly in either social science data archives or text archives. This may lead to a fall of value, since the way in which the data are archived is not primarily aimed at historical research. On the other hand there is no wish to invent a totally new way of archiving as the similarities are far greater then the differences. Obviously a lot can be gained by working together. In several countries, similar initiatives as in the Netherlands are being developed. Together with researchers in the U.K. and in Germany plans are developed to make a cross-national inventory of machine readable historical data sets. Hopefully other countries will join. International cooperation is also needed in order to develop standards for the exchange of historical data sets across the borders.

Nationally, support is sought in the conventional historical institutions, as well as in the circles of the traditional historical archives and the Steinmetz Archive (Dutch social science data archive). Interest has been expressed also from the side of (historical) museums.

So far reactions have been very encouraging. It is expected that in a years time, when the results of this project are due, a Netherlands Historical Data Archive can be founded.

Archiving means storing and making available for secondary use. One of the things that has to be dealt with is describing the datasets. As it looks as though the Netherlands Historical Data Archive will be the first purely historical data archive, it is the designated place to set the pace in describing historical datasets.

One of the problems Netherlands Historical Data Archive has come about - and this is a serious one that can not be dealt with easily - is that everything is still very theoretically. The first dataset has still to be archived.

2. Study Description Scheme

Architecture and Contents of the Study Description Scheme

A study description is a detailed account in machine readable form of information on a machine readable dataset. It contains information of different kinds:

- bibliographical:
  the title of the dataset, information on people or institutions that are responsible for the creation of the dataset (primary investigator, the depositor, the funding agency);

- administrative:
  its accessability, the publications based on the dataset, other datasets and publications that bear a relation the dataset described;
- concerning the contents:
  what the data deal with, what the research topic is, it tells you where the data stem from, it tells you about the relation between the real world and the universe depicted by the dataset;
- technical:
  how the dataset is organized, the size of it, what can be done with it in terms of the software that can be used in analyzing the data.

Functions of the Study Description

With the Study Description an instrument is created to get information on machine readable datasets. It is meant to be used by two kinds of people. On the one hand the historian, who wants to analyze an existing dataset and on the other hand the archivist, who wants to have knowledge about his holdings, both for internal - administrative - uses as well as for optimising his service to the clientele. The way these people can make use of the Study Description is partly overlapping. To understand its functions the Study Description can be looked upon from several points of view.

On the first ('lowest') level, the description of one dataset, the Study Description is a necessary prerequisite for doing secondary analysis. It is no use having a tape or a floppy and not knowing what is on it. The researcher should at least know what the data on the tape or floppy deal with: what kind of data, what period, what geographical area, what source material is used in creating the dataset (parish registers, death certificates, medieval charters?), what publications have been based on the dataset. On ground of this information the analyst may decide how to use the dataset concerned, or not to use it at all. The availability of the description in machine readable form will facilitate the retrieval of information.

The usefulness of keeping track of what is in the archive in such a way as before described is even more clear if you step one level higher, the description of all the holdings in the collection of the archive. The potential secondary analyst and the archivist who wants to lend assistance to the researcher can both profit from a well described collection of holdings, that gives the opportunity to get a clear view on the whole of the collection. If someone is doing research on the population in the 19th century, he may want to know whether or not he can make use of an already existing dataset dealing with the same kind of data, the same sort of research topic. He himself, or the archivist assisting him, can now easily retrieve information on the kind of data, the units of observation, the time period covered of several of the holdings. He then can decide whether or not he may want to use and can use an available dataset.

Beside these ad hoc uses of the Study Description database one can also think of a more systematical use. With the Study Description as a basis one can compile a catalogue of the holdings of the archive. In several data-archives data catalogues are more or less automatically produced using data from the Study Description database as input for phototype setting programs.

To take one step further, one level higher, one reaches the what can be called 'intra archival' level. This means dealing with more than one data archive. What is true on the level of one archive becomes even more true on this level: the effectiveness of properly describing datasets multiplies if you -in this way- not only give information on a dataset to the potential users of your own data archive, but to the entire universe of all researchers of all data-archives in the whole world. And, vice versa, if your usergroup can have access to descriptions of datasets stored at other data archives. It is clear that the optimum in exchanging this kind of information will only be reached if the archives will use the same kind of description. The exchange of dataset is only a small step further.

For the Netherlands Historical Data Archive it is of the utmost importance to be able to exchange information on available datasets with - for instance - the Steinmetz Archive. They keep quite a few datasets that are most interesting to historians. In fact part of the archive's collection is looked upon as historical datasets. Not only by us, but also by them. But as up till now there has not been a special historical archive, the designated place to store datasets of historical interest has so far the Steinmetz Archive. With the term 'historical datasets' is meant in the first place datasets dealing with events from the period before 1945. Examples of such datasets at the Steinmetz Archive are 'Prices of food products in the 19th century' and 'Dutch census, labour, land and emigration data', a study on emigration in the 19th century.

In the future datasets that originally have been collected by social scientists will be getting more and more of interest for use by historians. One can in such a case think of postwar census material, or a study as 'repatriates from Indonesia' which dates from 1955.

Another reason to consider linking up with the existing data archives, so far as study description is concerned, is the fact that in the past 25 years in those institutions a lot of experience has been built up in archiving datasets and hence in describing the datasets. So it can be taken for granted, supposedly, that they have a fair idea on how the description of a dataset should look like, both from an archiving point of view - what is the best way of describing in order to facilitate intraarchival use, to compile a data catalogue, to advise researchers - as well as from a user point of view - what does a secondary analyst need to know about a dataset.

But this 'long history', that the social scientific data archives can boast upon, does have a dark side as well. Technical developments in database construction have been quite considerable in the past view years, which has had its consequences in terms of data descriptions.

Another reason why the social scientists should not be blindly followed is the fact that the object of a social scientific study is not always the same as that of a historical one. Their attitudes are apt to differ slightly. Later on more attention will be paid to this.

3. Study Description in the Past

Study Description for Social Scientific Data Archives

In the social science data archives cooperation in the field of designing a standardized way of describing datasets dates from a long way back. Contacts between people from the Central Archive in Cologne and those from the Steinmetz Archive resulted in the late sixties to the first drafts for a common way of describing datasets. Further developments led in 1974 at an international conference held in Copenhagen to an agreement between several data archives on the layout of the Study Description (1). In 1981 at another conference, this time in Grenoble, the Danish Data Archive presented a proposal for a Standard Study Description Scheme, based upon the earlier agreements (2). This has been in use (with minor adaptations) since then at several archives. Other archives have used it as a starting point for their Study Description.

Developments in the Field of History

The social science Study Description scheme has also been a starting point for discussions on the description of historical datafiles. Maybe the impression has been given that everyone thinks of historical datasets as something completely different from social scientific datasets. This view is certainly not commonly shared. Especially not at ICPSR, the data archive that is - certainly in its first few years was and still remains - as much a social science as a historical data archive.

It was started by historians, they have archived a lot of historical datafiles, and still a large part of the staff is historian by training. At ICPSR the integration between historical and social scientific datasets is in a stage that both are regarded as equal. Historical datasets are looked upon as social scientific datasets that deal with society in the past. In their Study Description hardly any attention is paid to typical historical aspects. That is why it should not be copied.

Some of the other 'traditional' data archives, however, have adapted their Study Description and made room for typical historical items . The Steinmetz Archive for instance has - as early as the beginning of the eighties - added information on the time period covered by the dataset to all Study Descriptions. At the Danish Data Archives even more items have been added.

In the last couple of years lots of discussions and publications by european historians regarding the description of historical datasets have taken place. One major event should be included here: The establishment of Historical Social Research Centre in Cologne with roots in the late seventies. Though not in first instance a proper historical data archive, it united data archival roots and thinking with historical training. Such a combination should lead to something beautiful. Both here, in Cologne and elsewhere, where social scientific and historical universes met in the world of computerized research, people started thinking about historical datasets and what should and could be done with it. One of the things that can be done is making available for secondary use. In order to be able to do something with someone elses datasets, it should be documented properly. Realizing this discussions started on materialization of this idea.

One major development which is worth focussing upon was started in July 1986 in Göttingen during the International Workshop on the Creation, Linkage an Usage of Large-Scale Interdisciplinary Source Banks in the Historical Disciplines. Here a working group was formed by Hans Jorgen Marker (Danish Data Archive), Herbert Reinke (Center for Historical Social Research) and Kevin Schurer (Cambridge Group for the History of Population). Their aim was to draw up standards for describing historical datasets, especially dealing with quantitative data. These three have since then extensively published on the state of the art, on their findings, on what a secondary analyst should know about a dataset before he can start using it (3). But beside applause for the good job they have done criticism needs to be added. This is caused mainly by different point of view.

The triumvirate has taken as a starting point that the principal investigator himself makes his data available to other users for secondary analysis. The advantage for the principal investigator is of course that he can control who touches his data. Advantageous for the secondary analyst is the fact that he gets the best possible information on the dataset. The disadvantages however, are far greater.

- The secondary analyst should have the disposal of the same kind of hardware, and very often even software, the principal investigator has used.
- If the principal investigator does not take care of maintaining his

datasets, the chance exists, that within a few years the data can no longer be used. Either because of new releases of the software that should be used, or because the old fashioned floppy no longer fits in the drive, or even because the dataset is simply lost. Most of you will be able to sum up at least half a dozen of other reasons.

- One of the biggest disadvantages is that, if datasets remain with the one who originally created the dataset it is hard to get an up-to-date overview of the datasets that are available for secondary use.

When data are stored in a data archive value may be added to it by upgrading, cleaning, or merging with similar datasets.

All this speaks in favour of archiving datasets centrally and, as a consequence, describing them at the archive according to archival standards. Even if datasets are stored this way the secondary analyst can still get in contact with the principal investigator if the latter wishes so. In the Study Description will be registered the name of either the principal investigator or the person who should be contacted. So if the secondary analyst wants to get in contact, the possibilities are still there.

Putting data at an archive does not necessarily mean that the principal investigator will lose control. If he wishes so he can stipulate that his dataset can only be used after explicit written permission, so he can still decide who may have access to his data:

If a dataset is stored in a data archive, where more datasets are kept and archived, it has its consequences for the way the dataset should be described. If you are not thinking of one dataset, but of a whole bunch of them, it makes sense to draw a distinction between precise detailed information on the micro level, which could be called data documentation, and more general information, that can be put together in machine readable form. If you keep a lot of datasets, and all of them bring about a small drawer full of documentation it is hard to get a general view of your holdings, let alone let other people get insight into it.

Now it becomes more and more clear that it is not useful to distribute all the minutely detailed information on a dataset. It is indispensable for doing secondary analysis, performing the act of analyzing. Then you should precisely know what's what, but if you just want general information, to help you decide whether or not you want to use a specific dataset, then less specific information will suffice.

It is far more sensible to have less information, but still that information that one needs to get an idea about the contents of a dataset, and to have that stored in a way that it can easily be retrieved. Once the secondary analyst has made his choice, he can get the whole of the data documentation, maybe complete with oral explanations by the principal investigator, of the dataset of his preference.

To which extend one can call information essential for general use, and when the line is crossed that makes something a detail, that is only of use for those analyzing, is not always easy to say. That the one should be limited, and the other can hardly be extensive enough is clear.

In most of the articles on the description of datasets so far no distinction is made between the description of a study as a whole and documentation on the data level.

Another critical note also stems from a different opinion on where datasets should be stored. So far no manual has been produced on how datasets should be described, something like a simple straightforward checklist. In this paper an attempt is made at conceptualizing such a standardized manual.

4. Social Science Versus History in Terms of Dataset Description

The 'traditional' Study Description is not applicable to historical usage. Historical data are an other kind of data than social scientific data and historians are other kind of people than social scientists, their research questions differ, they stem from an other tradition. In brief: their datasets require different descriptions, both seen from the side of the secondary analyst as well from the side of the archivist.

Technical Differences

The 'traditional' social science data file is a rectangular file, that can be analyzed using software packages like SPSS and SAS. The social sciences have a long tradition in using computers as tools for doing research. Started in a time that computers could only work with figures and rectangular files, this has resulted into the fact that the bulk of social science data files is still organized this way. As a consequence the Study Description has been constructed to meet with these kind of files.

In the historical field people started applying computerized methods on a larger scale only recently. In the Netherlands it wasn't until a few years ago that the art of using computers started being taught to history students, while in the social sciences these kinds of courses were obligatory in the late sixties, early seventies.

As a result of this rather young history of computerized research, historians don't suffer from the - often self imposed - restrictions social scientists have to cope with, like working with rigid file structures, fixed field lengths, etcetera. Historians tend to work more with relational databases, mixing alphabetic and numeric fields, using the most advanced techniques, just because they don't know anything else. This is of course exaggerated, but it remains a fact that the use of relational databases is far more widespread with historians. One can come across such files in the social scientific world more and more, but it is still relatively new and not widespread. Which has its consequences for the Study Description.

Historical datasets are for the larger part of a relational kind, in which are included hierarchical and network files as well. That is why more attention should be paid to this kind of file structure, than currently is being done in the 'traditional' Study Description.

Any rectangular file can be re-analyzed using standard statistical packages, but relational databases often require specialized software. So it is rather important for a potential secondary analyst of these kind of files to know the technical ins and outs of the file structure. Maybe the contents of a dataset are of great interest, but if technically it can't be analyzed, it is no use ordering the dataset. Unless of course the archive takes care of preparing the data, restructuring the file, in a way that it can be of use to this analyst. One can think of merging tables, cutting and pasting entities etcetera. One has to be very careful in doing this, or one may lose information, or create disturbing redundancies, or both, or just get rubbish and nonsense.

Not everything should be done at the archive. The responsibility of the archive is to create the possibilities of doing secondary research, especially in a technical way. It is up to the analyst to do his own recoding, according to his own needs.

Another possibility is archiving the whole lot in one. This will often mean that the secondary analyst should have the same hardware and software at his disposal. There are more possibilities. The database can be exported, stored as separate tables. The source code of the software used will have to be archived as well as a separate textfile.

Apart from reformatting complex data structures, or archiving the software with the dataset, as outlined before, one can think of another solution, which also implies calling in a data archive. The original datafile results in more than one archived datafile; the archive can store two or more versions of the same dataset; the contents are more or less the same, but the technical structure varies. One version is the original one, say a binary or systemfile in combination with the corresponding software, an other contains exported tables and the source code of the software used, a third version consists of a newly created rectangular file etcetera. All files are kept, but for the records it is one dataset.

Whatever solution is chosen it will have to be documented. It can be imagined that something like described before will often be the case with historical dataset. That is why all this needs a lot of attention in the Study Description for historical datasets.

Differences in Content

Let us get back to the differences in the contents of historical and social scientific dataseis. Generalizing there is only one difference, but that is a big one: the object of research, the source of the data. Historians very often take written sources as a starting point whereas social scientists use survey material. This has its consequences for the description of the studies.

Background Information on the Source

In Study Description for historical datasets more attention should be paid to the sources upon which the dataset is based; the ideas behind the original recording of the source, what the purpose was of the source when it was created. And, very important, what has happened to the source since it was first recorded. Has any large scale rearrangement taken place, or have parts of it been lost? What are the results of this in terms of the value of the source?

Information on the Transformation of the Source in Machine Readable Form

Something else that also needs attention is in what way the source is reflected in the dataset. Is the whole, integral, complete, source transformed in machine readable form? How have indistinctnesses been treated. For instance if part of the source can no longer be deciphered, what has been done to the missing information. What if the person who composed the source has made a mistake? For instance if a taxcollector in a tax register has miscalculated the total amount of tax assessments for individual records. All the individual payments are recorded as well as this miscalculated totalized amount. The actions the primary investigator has taken, the way he has solved this problem in his dataset should be documented. A comparable problem can be found in textfiles in case of misspellings.

If the source is not completely transformed into machine readable format which parts been left out and why has it been left out and what are the consequences etcetera. Has some sort of coding taken place or normalization? All these things have to be accounted for everything needs proper documentation.

Besides these action, taken by the principal investigator in order to create a dataset out of traditional sources, it may occur that the data archivist does some additional cleaning of the dataset. All the changes this results in should also be documented. This however is not something new to datasets. The same procedure is also known for social scientific datasets. Although in the latter more routine, more standard techniques can be applied. A simple check on coding will very often not be fit for use.

These are the main items that makes describing historical datasets differ from describing social scientific datasets.

5. Summary

The social science data archive have a long history of describing machine readable datasets in a more or less standardized way. As it is very useful for historical data archives to be able to exchange datasets and information on available datasets with these institutes, it is very important to describe historical datasets in a compatible way. It is, however, impossible to completely copy their Study Descriptions, because a lot of differences exist between the two kinds.

The differences can be summarized as on the one hand regarding the technical format and on the other hand regarding the contents of the dataset.

The technical differences focus on the format of the data and on the structure of the dataset as a whole. Historians far less then social scientists limit themselves to the use of numeric fields only. Social scientists more often build rectangular files, whereas historians as a rule create some sort of relational database. As historians tend to use more complex datastructures, different ways of describing are required.

Concerning the differences in the contents the main differences stem from the information historians need on the sources the dataset has been based upon and how these sources has been transformed into machine readable form. And then of course what adaptations are made by the archivist.

Although historians are not the same as social scientists and historical datasets differ from social scientific datasets, and hence the descriptions of both require different focussing points a lot of similarities can be seen as well. Especially as far as bibliographical and administrative information is concerned.

There is no need for changing the way this information is described in the social scientific Study Description. No doubt the traditional data archives have enough experience as to what needs to be described concerning these topics.

As to how historical datasets should be described a lot of discussion has been going on amongst historians. A mix of outcomes of these together with a subset of the 'traditional' Study Description have resulted in the proposal for a Study Description for historical datasets. The idea behind it is that it should be applied and tried by those who work with historical datasets. Both the historians, who can tell what specialized topics are indispensable for doing secondary analysis, and the data archivists who can take a look from an archival point of view.

Notes

(1) P. Nielson, Study Description Guide and Scheme, Copenhagen: DDA, 1974.

(2) K. B. Rasmussen, Proposal Standard Study Description, Odense: DDA, 1981.

(3) e.g.: H. Reinke, K. Schurer, H. J. Marker, Information Requirements and Data Description in Historical Social Research, HSR 42/43, 1987.

APPENDIX: Study Description Scheme for Historical Datasets

0. Introduction

All original data files of the Netherlands Historical Data Archive will be documented according to the study description scheme. The information will be stored in machine readable form. Retrieval of information on studies will be possible with a query system or with hard-copy indices and a catalogue.

1. Identifications and Acknowledgements / Origin of Data

101 title of study

note:
- Original research title in original language.
- Do not use subtitles.
- Do not use articles at all and numbers in front.
- In case of odd title, describe study in short with rational title.
- If no research title available, use title of report based on primary research.
- Keep title within 110 characters.

example: »The woman in the 19th century. An investigation of attitudes etc., etc.« will be: »Woman in 19th century«.

111 local archive where study is stored

note: Indicate the name and complete address of the archive that will make the file available to users.

01 name of archive
02 address of archive
03 id-number of archive
04 number of study description
05 number of study (data-set)
06 numbering of substudies
99 additional information

112 archive where study is originally stored

note: Indicate the archive that originally created the file described here.

01 name of archive
02 address of archive
03 id-number of archive
04 number of study description
05 number of study (data-set)
99 additional information

121 depositor(s)

note:
- Mention 'institutes' here instead of persons, if applicable.
- If acronyms present, start with full name of institute (no abbreviations), followed by acronyms.
- State full name of institute from larger to smaller unit.
- Use subitems 01 through 03 for the first institution, 11 through 13 for the second and so on.

example: Rijksuniversiteit Groningen, Documentatiecentrum Nederlandse Politieke Partijen (DNPP).

01 institute, name(s)
02 address
03 additional information
11 institute, name(s)
12 address
13 additional information
99 additional information

122 date of deposit

note: Indicate the date that the file has been deposited with the archive listed in item 122. Missing information in the date should be coded 99.

examples: 8 September 1988 - 19880809
September 1988 = 19880999

01 day/month/year
99 additional information

131 principal investigator(s) -research organisator(s)-

01 name(s), institute
02 address
03 additional information
11 name(s), institute
12 address
13 additional information
99 additional information

note:
- Mention name(s) of researcher(s) participating in the study, followed by name(s) of institute(s).
- In case of different institutes use subitems (01), (11), (21) etc.
- First mentioned researcher of every institute: start with family name, followed by initials.
- Do not use academic titles.
- Do not mention more than three researchers, rest is indicated by »e.a.«.

example: (01) Vries, J.A. de, A. de Jong, Universiteit van Amsterdam.
(04) Berg, B. van den, Rijksuniversiteit Utrecht.

132 data collector(s)

note:
- Idem as item (121).
- Details must be mentioned in item (235).
- The data collector may be identical with the primary investigator.

01 institute, name(s)
02 address
03 additional information 11,12,13 and further as subitems 01,02,03
99 additional information

141 research initiator

note: - idem as item (121).
- initiator can be funding agency, but also individuals or f.i. working groups.

01 institute, name(s)
02 address
03 additional information
11,12,13 and further as subitems 01,02,03
99 additional information

142 funding agency

01 institute
02 address
03 additional information
04,05,06 and further as subitems 01,02,03
99 additional information

149 rationale for the study

note: in short an explanation of purpose of data-collection written towards future use.

199 other identifications/acknowledgements/origin of data

note: In this item indicate any other persons or institutions who have not been included in the above items, but who nevertheless had a role to play with respect to the research project.

100 bibliographic reference

title, year(s) (of machine-readable data file). Principal investigator: name(s) * address. Data-collector: name(s) * address. Funding: ña-méis) * address. Depositor: name(s) * address (producer). Distributer (archive). Study nr: number.

note: The bibliographic reference is written in free text. Much of the information can be obtained from other items in the SD.

2. Original Research Design / Analysis Conditions

Items 201, 202, 220 and 221 contain information on contents of the dataset: what kind of material has been collected.
Items 203, 237, 238 and 211 contain information on the sources on which the dataset has been based.
Items 212, 214, 222, 223, 224, 231, 232, 233, 234, 235, 236 and 321 give information on how the original sources have been used in the creation of the dataset.
Items 213 and 219 supply technical information.

201 theme of study

01 scientific discipline

note: - scientific discipline applies to the subject of study and not to the scientific background of the researcher(s).
- Use only one discipline. In case more disciplines are eligible, choose the one most applicable.

* Art History
* Archeology
* Demography
* Economic & Financial History
* Language & literature
* Law & Criminality
* Medicine & Health
* Political History
* Social History
* Trade & Industry

02 research topic: keywords

note: - Use keywords in sequence of importance as much as possible.
- Use at least 2 and at most 6 keywords.

examples: administration, africa, age, agriculture, aristocracy, arts&culture, asia, bibliography, biography, birth, building, cartography, childrearing, children, cities, citizenship, commerce, communication, crime, death causes, economy, education, elite, emigration, employment, expansion, family, family-reconstruction, farming, feudalism, film, finance, geography, health, housing, immigration, industrial revolution, industrialization, industry, international/relations, law, living- conditions, marriage, marine, measures&weights, migration, militaryaffairs, minorities, mobility<social>, monarchy, monetary, nobility, object-registration, occupation, parliament, peace, peasantry, personal-data, politics, prices, profession, property, prosopography, psychology, religion, shipping, slavery, stratification, taxes, theatre, time-series, topography, trade, transport, verdicts, villages, violence, wages, war, women, work, youth

03 abstract: themes of dataset / research instrument
note: - The abstract indicates major themes of the study for which data are available in the file.

99 geographical area covered
note: - Indicate geographical area to which the dataset refers.

220 time period covered

note: - This item reflects the time period covered by the data.
- Missing information in the date should be coded 99.

01 start of time period
02 end of time period
99 additional information

221 time dimensions

note: if applicable: specify more then one term

01 one time study
02 one time study - partial replication
03 panel study (specify number of waves)
04 trend study (specify number of waves)
05 time-series (specify number of observation points)
99 other (specify)

202 kind of data

* survey (both sample and total population)
* longitudinal (if same questions are measured at different points of time)
* international (if more than one country or nationality can be identified)
* medical (medical or psychiatric data)
* graph (for GRADAP-files)
* reference source (use only together with other keyword f.i. census data, statistics, regional, textual. Only for files which are intended to facilitate research, f.i. a bibliography)
* census data (individual attributes of a total population gathered from

the people themselves. May or may not be aggregated)

* statistics (data from administrations. F.i. data from a Central Bureau of Statistics about a country)

* regional (use together with statistics if explicitly a subnational level is indicated, f.i. municipality or COROP-area)

* textual (use only together with other keyword f.i. literary, juridical, religious)

* teaching package (instructional datasets)

* computer program

* legislative roll calls (voting in f.i. Parliament, UNO)

* literary/linguistic

* juridical (verdicts, sentences, notarial deeds)

* religious

* administrative (personal data, tax registers)

* business registrations

203 data sources

note: This item deals with 'bibliographical' information on the sources that are used in creating the dataset. It contains information on the whereabouts of the original sources and on publications concerning the sources.

01 unpublished source material

note: give detailed information including where sources are stored, how they are documented and how they can be accessed.

02 published source material

note: specify bibliographic entry if the whole source or parts of it have been published previously.

04 other written material (specify)

note: indicate relevant publications on the source.

99 other (specify)

237 origins of the sources

note: Specify the rules in establishing the original sources. Specify the purpose of compiling the original sources. Use free text format.

238 characteristics of storage of sources noted

note: - indicate level and manner of preservation for every source
- specify changes in the source that may influence its value and meaning.
- subitems 01 04 deal with the completeness of the source. Does the source have the same contents it had when it was originally established? If not, indicate what has been lost.
- subitems 05 07 deal with changes in the structure of the source. A source may have been reorganized and thus the meaning and value of its data may have changed.

01 completely preserved.
02 partially destroyed by personnel of document archive according to systematic criteria (specify).
03 partially and unsystematically destroyed by personnel of document archive (specify).
04 partially destroyed or damaged for other reasons (specify).
05 organization of source material in document archive or at the producer of the source concerned.
06 reorganization in the document archive or at the producer of the source.
07 reorganization in the document archive or in the administration by record linkage procedures.
99 other (specify).

211 units of observation

note: give all applicable terms of units of observation in the sources, which are identifiable in the data-set.

examples: (1) text units, charters
(2) individuals, families/households, birth- certificates

01 individuals
02 families/households
03 groups
04 institutions/organizations
05 administrative units (geographical, political, economic)
06 text units (documents, chapters, words)
99 other (specify)

212 number of units of observation (cases)

note: Indicate what adaptations, what operations have been performed in order to create the dataset out of the original (sample from the) sources in 99.

01 number of units in original sample ('target')

02 number of losses

03 number of replacements

04 number of cases (unweighted)Cobtained')

99 additional information

214 completeness of study stored

note: This item indicates the relationship of the data collected to the amount of data stored in the file.

99 additional information

222 definition of target universe

note:
- indicate which criteria were used for defining the universe from which the sample was selected.
- specify all applicable terms separated by slashes for a complete description of the population; f.i.: name of country: the Netherlands / age limits: 15-75 years / location of units of observation: national.

223 sampling procedures

note:
- next to sampling method, indicate the type of element which is sampled (individuals, municipalities, etc.) and the type of register the element is selected from (f.i. population register, housing register, tax register).
- Basically, samples are either probability samples or nonprobability samples. Probability sampling implies an equal and known probability of selection for each population element, and are also called 'representative samples'. Nonprobability samples are accidental (f.i. first fifty people passing a gate), purposive (f.i. experts purposely selected for their knowledge) or quota samples.

NO SAMPLING (total universe), complete count, a population census or the whole parish register.

SAMPLE, representative part of group of units.

MASTER SAMPLE, large sample drawn for future investigations from which samples are drawn when required.

STRATIFIED SAMPLE, sample consisting of sub-samples drawn from subgroups (strata) into which the population has been divided.

DISPROPORTIONAL STRATIFIED SAMPLE, stratified sample in which equal number of cases is drawn from each stratum.

PROPORTIONAL STRATIFIED SAMPLE, stratified sample in which the same percentage of cases is drawn from each stratum.

SIMPLE RANDOM SAMPLE, sampling without replacements.

TRUNCATED SAMPLE, sample omitting cases having values outside certain limits.

ACCIDENTAL SAMPLE, nonprobability sampling involving nonsystematic selection of cases which happen to be available. See note also.

AREA SAMPLE, sampling involving the selection of sub-areas of a larger area, the population of sub-areas being stratified or unstratified.

CLUSTER SAMPLE, sampling in which each sampling unit consists of more than one element, f.i. classrooms or schools rather than individual pupils.

DOMAL SAMPLE, a type of area sampling involving a systematic selection of houses in an area with a specification as to which persons in each selected house are to be included in the sample.

EXPERT CHOICE, may be by snowball method.

GRID SAMPLE, a type of cluster sampling in which a grid is placed on a map and grid areas are selected as clusters.

MULTI-PHASE SAMPLE, sampling in which some data is obtained from the full sample and additional data is obtained, either at the same time or later, from sub-samples, the same type of sampling unit being used at each phase.

MULTI-STAGE SAMPLE, sampling in which different types of sampling units are sampled at different sampling stages.

PROBABILITY SAMPLE, sampling in which every population element has an equal, known and nonzero probability of selection.

SITUATION SAMPLE, sampling in which the population elements are types of situation.

SNOWBALL SAMPLE.

SPACE SAMPLE, sampling over space as a protection against unknown sources of variation.

STRUCTURAL SAMPLE, sampling in which the sampling units are connected by specified relations, (sociometric, dominance, communication, interaction).

SYSTEMATIC SAMPLE, drawing sample by selecting from a file or list after random start from I to K of every Kth sampling unit.

TIME SAMPLE, sampling of periods for observation.

UNITARY SAMPLE, sampling in which the ultimate units are selected directly from the population.

QUOTA SAMPLE, investigator selected sample of a certain number (quota) of people meeting certain criteria, f.i. women with and without children.

224 major deviations from sample design

note: Indicate in free text format if it is suspected that the sample is not representative for the purpose.

231 dates of data collection

note: missing information in the date should be coded 99.

01 first date of data collection (year,month,day)
02 last date of data collection (year,month,day)
03 data collection period in weeks (code actual number of complete work-weeks spend on data collection)
99 additional information

232 method of data collection

note: indicate one or more techniques of data collection. Free format elaboration on the method is recommended.

01 analysis of administrative records (specify)
02 document analysis
03 content analysis
09 object description (f.i. coins)
99 other (specify)

233 type of research instrument

note: Indicate the type of instrument used.

99 other (specify)

234 actions to minimize number of losses (specify)

note: In narrative form, summarize briefly the actions taken. F.i. matching of information from different sources, estimation, reconstruction, information from historical literature.

235 data gathering staff

note: indicate status of employees collecting the data.

01 regular staff of institute (specify)
02 special staff (specify)
03 principal investigator(s)
99 other

236 characteristics of data collection situation noted

note: indicate information regarding the situation surrounding the collection of the data

data collection person identified (specify)
characteristics of data collection person (specify)
physical situation, location (specify)
special events (specify)
conditions of observation (specify)
instruments applied (specify)
other (specify)

321 control operations performed by original investigator

none
check on coding (verified)
check on coding (specify)
data checkpunched (verified)
check on wild, missing, unspecified codes (specify)
check on logical inconsistencies (specify)
other

213 dimensions of dataset

note: - Subitem text has to be written down.
- Use 01, 02, 03 for rectangular files.
- Use 11, 12, 13 for relational/hierarchical and network files.

01 number of cases (estimate)
if weighted: state -weighted- here

02 number of variables per case (estimate)
03 number of cards per case: (if card-image data set)
11 number of entities
12 number of attributes per entity for every entity (estimate)

13 number of records per entity

99 additional information: in case of text files specify record length and number of records for every file.

219 original mode of storage

note: - indicate for every specific data file.
- in case of more then one file or non-rectangular file(s) specify relationships between files or entities.

numeric file
text file
graph file (specify)
combination file (specify)
sequential file(s)
- fixed record length
- free record length
- fixed field length
- free field length
- indexed (specify)
relational or hierarchical database (specify)
exported tables, data model and data documentation
SPSS file
- SPSS file, not labeled
- SPSS file, labeled [dutch/english]
- SPSS-X file, not labeled
- SPSS-X file, labeled [dutch/english]
number of systemfiles
number of subfiles
number of files
no data available
unknown: specify problems
others (specify)

299 deficiencies of data-set: (specify)

note: specify curious, unexplainable or insoluble matters in dataset.

3. Reanalysis Conditions

301 present data representation -mode of storage-

note: - indicate for every specific data file the most recent mode of storage.

note: - in case of more then one file or non-rectangular file(s) specify relationships between files or entities.

01 original mode of storage (see item 219)

99 others (specify) most recent storage mode for every specific data file

numeric file (specify)
text file (specify)
graph file (specify)
combination file (specify)
sequential file
- fixed record length
- free record length
- fixed field length
- free field length
- indexed (specify)

relational or hierarchical database (specify)
exported tables, data model and data documentation
SPSS system file
graph file (specify)
number of systemfiles
number of subfiles
number of files
SPSS file, not labeled
SPSS file, labeled [dutch/english]
SPSS-X file, not labeled
SPSS-X file, labeled [dutch/english]
no data available

note: subitem 301.99 should be updated with every change in storage mode of the file.

302 applicable analysis packages

OSIRIS
SPSS
SPSS-X
SAS

GRADAP
ABC
Dbase
Oracle
Ingres
other (specify)

306 status of the study

01 data are not yet available from the NHDA, contact the primary investigator
02 data are available in acquisition form
99 additional information

311 language(s) of written material

01 study description
02 research instrument
03 codebook
04 SPSS-labelling
99 additional information

322 control operations performed by archive

01 none
02 check on coding (specify)
03 check on wild, missing, unspecified codes (specify)
04 check on logical inconsistencies (specify)
99 other

331 accessibility

01 no restrictions
02 no restrictions to scientific use
03 consultation with donor before use of data is advised
04 written permission of donor required for publication
05 written permission of donor required for use of data
06 special arrangements to be made with donor
99 other conditions
- results of analysis to be brought to the attention of donor
- special arrangements to be made with donor, including written permission

332 access directing authority

01 name(s)
02 institute
03 address
99 additional information

399 other reanalysis conditions (specify)

note: indicate information about the file(s) that does not appear in items 300 through 398, and that will be helpful to potential users.

4. References to Relevant Publications / Results / Studies

401 publications/reports by the primary investigator

note: the first publication is numbered 01-06; use for further publications subitems 11-16, 21-26 etc.

01 author(s)
02 title
03 city, publisher
04 journal, vol., pages
05 year
06 availability

411 other publications (secondary analysis)

note: the first publication is numbered 01-06; use for further publications subitems 11-16, 21-26 etc.

01 author(s)
02 title
03 city, publisher
04 journal, vol., nr.
05 year
06 availability

421 unpublished papers/reports of interest

note: the first publication is numbered 01-05; use for further publications subitems 11-15, 21-25 etc.

01 author(s), institute
02 title
03 year
04 pages
05 availability

431 results of analysis (databanks, indices, recoding, etc.)

01 results of analyses not stored with archive
02 results of primary analysis stored with archive (specify)
03 results of secondary analysis stored with archive (specify)
99 other (specify)

441 references to related studies

note: indicate in free format the studies that are related.

99 additional information (specify)

499 other references (specify)

Background Variables Included

501 basic characteristics

Indicate which standard data are included.
F.i prices, wages, occupation.

999 year(s) of study

note: - alphanumeric item: possible to mention from/to in case of substudies dates/years are separated by an asterisk.

01 year

999 english title:

note: rational title of data set maximum of 65 character positions !

999 study number