

---

# Data Infrastructure For The Social Sciences In The Netherlands

by Gerton Heyne<sup>1</sup>

IVA, Institute for socia research, University of  
Tilburg, The Netherlands

## INTRODUCTION

Computers have become commonplace during the past few years. There has been an obvious increase in, on the one hand, the computerization of data files, and on the other, the analysis and processing of machine-readable data files by computer. The pressing need, particularly by the state, for reliable and up to date policy information and the desire to raise the level of efficiency have been factors in this development. Data stored in computerized form offers greater possibilities for analysis. As a result, scientific research is also able to tap into a growing reservoir of available machine-readable files.

Technical innovations have moreover made it possible for researchers to collect and process research data more quickly and efficiently in order to obtain suitable material for analysis. In part because of innovations in research methods and techniques, researchers are able to carry out complex data analysis not possible before. The technical infrastructure is growing steadily because researchers are increasingly utilizing PCs and mainframes to set up and conduct research and to analyze machine-readable data files. In addition, they also have access to a growing number of networks which can be used to exchange data: for example, the various local university networks and the national SURFnet.

Until a few years ago, when researchers wished to conduct data analysis one of their most important sources was data published in written form. The most important data files were compiled for the state by the Central Bureau for Statistics (CBS). The availability of machine-readable data files on the one hand and the development of infrastructural facilities such as PCs and networks on the other are raising the demand for this information in machine-readable form. The possibility of conducting data analysis grows when researchers can access machine-readable files instead of written publications. In addition such files serve increasingly as "background information." Researchers have access to ever greater amounts of data which can serve as important sources of secondary information for designing and conducting research, because access to files already in use can also, in principle, be made simpler and easier.

The importance of a well-designed infrastructure is

evident. Researchers must have easy access to statistical material. Research in the social sciences makes it possible to gain systematic and detailed insights into society. Such insights serve two purposes that are, in fact, directly related. Firstly, research based on statistical information is important for both basic scientific research and for more policy-directed research. Insights gained in this way are of scientific value in and of themselves, but they are also important for efficient management and confident policy-making. Secondly, research serves a fundamental, democratic purpose. A balanced process of policy-making in a democratic society requires that all individuals and parties involved possess relevant information. Such a balance can only be strengthened by a vigorously independent research capacity autonomous of the expertise of the state. To achieve this, however, requires sufficient access to statistical information such as that compiled officially by CBS and other institutions.

In recent years a number of reports have put forward the idea that the data infrastructure used by the social sciences in the Netherlands needs to be augmented (De Bie, 1989; De Guchteneire & Timmermans, 1990). The authors conclude that researchers spend a great deal of time collecting and converting data files before being able to start analysis. They observe the following problems:

- the high cost of purchasing files;
- the restrictive conditions under which files are made available;
- the uneven quality of both the files and the accompanying documentation.

The assumption is that these problems prevent researchers from making adequate use of the data present. Delivery of data files is not optimal. The service provided by suppliers in general, and CBS in particular, is insufficiently geared to the growing demand by researchers for data files and statistics in the form of machine-readable numerical material.

The technical innovations described above have transformed desires and wishes on both the supply and demand side of data files. In addition, priorities have shifted. While the transformation process was still underway, the various parties formulated new goals and

interests that did not always turn out to be complementary. Researchers insist that data files be made available at minimal cost and under the least restrictive terms possible. The policy of CBS, the most important producer of data files, is not entirely self-determined but depends partly on political considerations. This supplier is obliged to comply with political agreements concerning the cost of field work and development and with statutory regulations concerning privacy. Part of the criticism that is being heard, therefore, concerns political leaders rather than CBS.

It looks as if technical and social innovations have set a process of change in motion that has caught the parties involved inadequately prepared. Measures have been taken to secure the interests of both producers and consumers of data files as well as of the individuals whose personal information has been collected. However, these measures often seem to seriously impede the work of researchers.

#### DESIGN OF THE PRESENT STUDY

Data infrastructure is becoming increasingly important to researchers. Improving this infrastructure is extremely important for the quality of social scientific research. The present study considers the current state of affairs with respect to the availability, accessibility and use of statistical data files<sup>2</sup> in the social sciences<sup>3</sup>. Which files are actually being used in research? Where do the bottlenecks occur in use or loan?

The study comprises an inventory of the use of external<sup>4</sup> statistical data files in 1989. The files had to meet a number of criteria formulated beforehand.<sup>5</sup> In order to gain as great an insight as possible into the use of external data files, both the **suppliers** and **consumers** of such data files were approached. This also offers insight into the way in which data files are made available for research in the social sciences. This is the first time that the above-mentioned suppositions concerning the functioning of data infrastructure have been subjected to systematic and relatively large-scale testing by those actually involved in this issue in practice: the researchers.

A number of **suppliers** of data files were requested to provide information about files either sold, lent or offered in some other fashion to researchers in 1989. A number of questions concerning the price, the size and the consumers of these files were added.

A number of users of data files received a questionnaire concerning files they had used in 1989. This also included questions concerning the price and size of the files, time of delivery, problems, quality, etc. Given the large number of research institutions, both commercial and non-profit (including faculties, research groups,

institutes and individuals), it was almost impossible to ask all of them to participate in the study. The goal was in any case to approach all relevant faculties. For this reason a number of relevant university faculties and research institutes were selected. An attempt was made to assign each research group or institute a contact person in charge of coordination. Only in a small number of cases did this contact person have a full understanding of every aspect of external data file use in his or her research group or institute. The rest of the time, contact persons were requested to distribute the questionnaires to those researchers who might answer questions concerning the use of external data files.

This paper will provide a short description of

- a. the most important information as provided by two institutions that supply data files to third parties (CBS and the Steinmetz archive);
- b. the most important results of the survey distributed to a number of data file users.

#### USE ACCORDING TO TWO SUPPLIERS

CBS and the Steinmetz archive, the Dutch data archive for the social sciences, were asked to answer a number of questions concerning the files that they supplied to researchers in 1989.

#### CENTRAL BUREAU FOR STATISTICS (CBS)

CBS provided an outline of a two-year period, specifically the microfiles and publication files<sup>6</sup> delivered in 1988 and 1989. The reason given was that in this way chance fluctuations would be less likely to misrepresent the information.

The following deliveries took place in 1988 and 1989:

- 91 deliveries of 20 different microfiles to research institutions;
  - 130 deliveries of 10 different publication files, of which:
    - \* 25 went to research institutions;
    - \* 66 went to businesses;
    - \* 39 went to the state.
- Two publication files were delivered 105 times.

Two publication files were by far the most popular, being delivered 66 and 45 times respectively. In addition, eight micro-files were made available more than five times in the period concerned; the Socio-Economic Panel was the most frequent (16 times).

CBS additionally supplied aggregated data<sup>7</sup> concerning the prices paid for various files by showing proceeds per file. It was impossible to deduce from this information how much individual deliveries yielded, or rather, what consumers must have paid for them. For this reason an outline of average prices<sup>8</sup> must suffice. Averages should

be interpreted with the necessary degree of caution, but the average amounts can serve as an indication. In 1988 and 1989 for individual deliveries of eleven microfiles average amounts have paid above 10,000 Dutch guilders (approx. \$ 6,060). The data shows that most microfiles are quite expensive. The highest (average) amount for one of these microfiles was 58,800 guilders (approx. \$ 35,600). Publication files, on the other hand, cost significantly less. Prices of the publication files supplied most often averaged 630 guilders (approx. \$ 380) and 340 guilders (approx. \$ 205) respectively.

Deliveries yielded 1.75 million guilders (approx. \$ 1.05 million) in total; 1.59 million guilders (approx. \$ 0.96 million) came from microfile deliveries. Publication files yielded 0.16 million guilders (approx. \$ 0.10 million). The amounts mentioned do not include compensation for additional field work or development costs.

#### THE STEINMETZ ARCHIVE

The Steinmetz Archive distinguishes between research files for the social sciences and weekly public opinion research (weekly questionnaires). The following will deal only with the first category. In 1989 the Steinmetz Archive sold<sup>9</sup> or made a file available 286 times. A total of 111 different files were offered; 42 remained after various volumes or waves belonging to the same file were combined.

The table below indicates how often Dutch and foreign files are used in the Netherlands and how often Dutch files are used in foreign countries.<sup>10</sup>

**Table 1. Use in and outside the Netherlands of the Steinmetz archive**

	use in the Netherlands	use outside the Netherlands	
Dutch files	110	105	215
foreign files	71	—	71
<b>Total</b>	<b>181</b>	<b>105</b>	<b>286</b>

There are 110 Dutch files and 71 foreign files in all used by Dutch researchers. Almost half of the Dutch files were sent outside the Netherlands.

Three distinct groups of foreign users can be distinguished: universities, data archives and research institutes, comprising respectively 81%, 13% and 6% of the foreign consumption. In the Netherlands users are universities and research institutes, which received 87% and 13% of the files respectively.

#### USE ACCORDING TO THE USERS

In the survey 70 institutions were requested to complete one or more questionnaires concerning the use of external data files in 1989. The response was low.

**Table 2. Response**

	approached response		
university research groups	44	18	41%
- social sciences	31	11	35%
- economics/business	13	7	54%
research institutions (non-profit)	23	8	35%
research institutions (commercial)	3	0	0%
total	70	26	37%

Of the 44 institutions that did not return a completed form, ten stated that they did not work with external data files and can thus be categorized as non-users. The remaining institutions failed to respond at all. The controlled response was therefore upwards of 51%. The 26 institutions, research groups and faculties returned 80 questionnaires in all with information concerning 121 files.

Because of the high level of non-response we do not consider the data ultimately received as statistically representative. The data does not provide an exhaustive view of the use of external data files. In our opinion, the results are nevertheless important for gaining insight into the use of files and the issues related to this use, in view of the number of completed questionnaires and their distribution across various institutions.

#### INTENSITY OF USE

The table below provides numerical data concerning the intensity with which the data files supplied were used. A distinction is made between CBS files and files coming from other supplying agencies.

**Table 3 Intensity of use**

	CBS	Rest	Total
1. A few analyses shortly after receipt	2%	4%	3%
2. Many analyses shortly after receipt; no others after this	10%	11%	10%
3. A few analyses over a long period of time	16%	37%	26%
4. Many analyses over a long period of time	62%	45%	54%
5. Unknown	10%	4%	7%
Total	100%	99%	100%

Intensity refers to use in time and the number of analyses performed on the file. The figures indicate that the files were generally used intensively. Approximately (54+26=) 80% of the files were used over a long period of time, and many analyses were performed on about (10+54=) 64%. For well over half (54%) of the files, many analyses were performed for long periods of time. Use of the CBS files is proportionately more intensive, in the sense that in general many analyses were performed.

**TYPES OF RESEARCH**

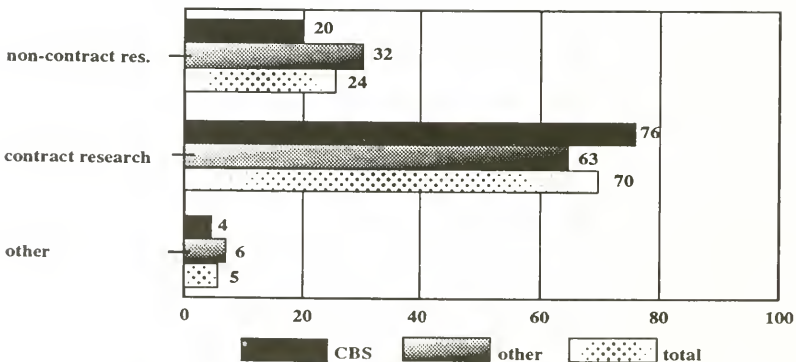
The study asked what type of research the files were used

for. The respondents were asked how often the file concerned was used for contract and how often for non-contract research. The results showed that files were chiefly used for contract research, in particular the CBS files (see figure 1). Indirectly this was actually a query as to who subsidizes the research (and the files). This information made it possible to indicate which commissioning agencies were directly or indirectly involved in the delivery terms for data files. For non-contract research this is the state. The respondents were asked to indicate who commissioned contract research. The state was named as commissioning agency in 75% of the cases. The conclusion is that the vast majority of research in the social sciences is commissioned and subsidized directly by the state. All in all 33 percent of this was charged to the ministry responsible for scientific research, the Ministry of Education and Science.

**DELIVERY TIME FOR FILES**

Questions were posed as to the amount of time that passed between the first formal contact with the supplier and the actual delivery of the file ordered. Delivery of CBS files took a long time in comparison with delivery of the remaining files. Delivery of a CBS file can take well over seven months on the average; delivery of the remaining files averaged little more than a month. Two-thirds of the "remaining" files were delivered within a month; delivery for the rest of these took longer than three months. Only 21% of the CBS files were at their destination within four weeks. Most of the CBS files were delivered within three to twelve months; 21% did not arrive for more than a year.

**Figure 1: Use of data files**





### PROBLEMS AFTER DELIVERY

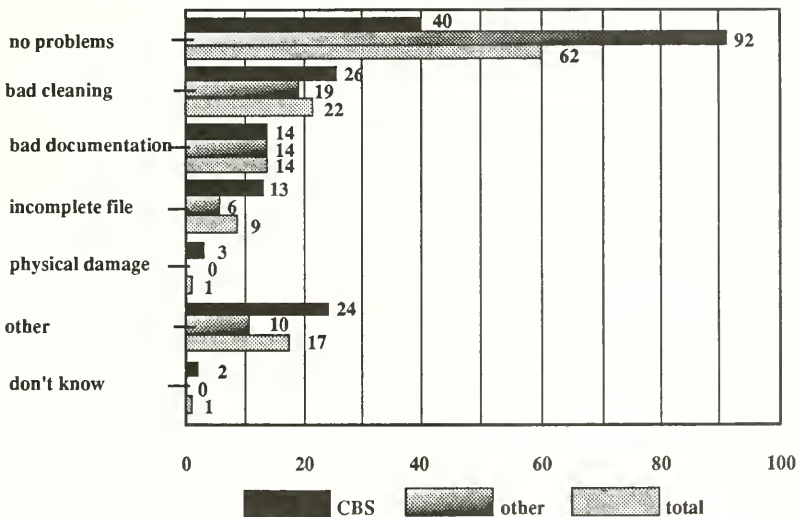
Problems involving costs, privacy, etc. come up before delivery. The questionnaire included a number of questions about the sorts of problems that came up after files were delivered. Of the CBS files, 60% caused researchers problems after delivery (see figure 2). This applied to 40% of the "remaining" files. The CBS files caused more problems for users with respect to cleaning and completeness than the remaining files: 25% as opposed to 19% and 13% as opposed to 5% respectively.

widely diverging scores with respect to these important problems in particular. Those who use CBS files find almost every aspect more problematic than those who use other files.  
4. On the other hand, CBS scores relatively high on familiarity.

### ARCHIVING

A data infrastructure that functions well presupposes facilities not only for efficient delivery of files but also

**Figure 2: Problems with files after delivery**



### QUANTIFYING GENERAL PROBLEMS

At the end of the questionnaire the respondents were asked to give their opinion on six different statements describing the same number of problems. This was an attempt to discover to what degree the respondents found specific hindrances problematic. The results are given in figure 3. The data led to the following observations:

1. Almost all scores are negative. This means that all of the aspects illustrated in the statements were seen as more or less problematic.
2. The most important problems concern the availability and delivery of the files; files are too expensive, considerations of privacy hinder or prevent delivery, and when delivery does take place, it takes too long.
3. CBS file users and those who used other files had

for file storage and management. The study asked what happens to files after use. Remarkably, only 80% of the files were archived in any way (see figure 4). The fact that 8% of the files are returned to the supplier and 6% are destroyed may be related to the terms under which the files were delivered; a supplier can, for example, require that files be destroyed or returned after a project has been completed.

Nothing was done with 4% of the files. As this study focuses on files delivered by others, one may assume that the original files remained in the possession of the supplier. Nonetheless, if the same practice applies for files put together by the institutions themselves, then there is an irrevocable loss of material. However, the study gives no definitive answer as to whether this

actually does occur. Only 1% of the work of archiving is contracted out to others, and this although a national archive exists for storing all files that in theory may prove relevant in future social scientific research. Once again, this study concerns files whose original versions are in all probability still with the supplier; copies therefore did not necessarily have to be archived by others. However, by the same token it was by no means strictly necessary to archive files locally, which is what happened in 80% of the cases anyway. For this reason the balance between "archiving" (80%) and "archiving by others" (1%) remains remarkable.

The next question concerned the way in which archiving, management and registration take place. No uniform archiving method appears to exist. Archiving and management take place at different levels; sometimes the individual researcher does it; then again it may be left to someone within the research group, the faculty or the computer center. The lack of guidelines or agreements for storing data systematically hinders access to files that might be suitable for secondary analyses in the future.

There is no clear survey of the various locations where files are available.

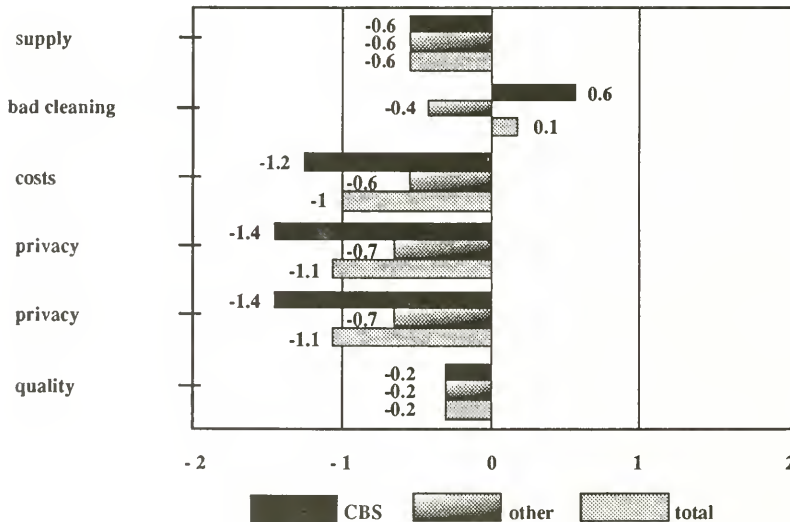
#### CONCLUSIONS

We must first mention that this study focuses largely on the use of external data files and the problems researchers encounter in this use. The most important points of this issue are taken up in this article. Suppliers of data files, on the other hand, are not given an opportunity in this study to express their grievances concerning file users. Although this article says nothing about their complaints, this by no means suggests that they do not have problems with users' unclear, unreasonable or badly formulated wishes.

The study has led us to formulate the following conclusions.

1. Poor response is one of the most important reasons why this study did not result in an exhaustive inventory of the use of data files in the social sciences. It is remarkable that researchers who depend so much

**Figure 3: Quantification of problems**



The respondents have given their opinion on a number of statements included in the questionnaire. Scoring ranges from 1 = "Agree completely" 3 = "Indifferent" to 5 = "Disagree completely". For clarity the results have been converted into a figure whose minimum score is equal to "problematic" and whose maximum score of 2 is equal to "not problematic".

on the cooperation of others in their work could respond so poorly.

2. The various institutions and research groups within the social sciences have not kept systematic records concerning the use of external data files. Perception of how files are used is diffuse. In this respect, there is a lack of order in the way social scientific research is organized and conducted. In no sense does it fit the image of a professional, well-oiled machine. The data highways see very little orderly traffic.

3. CBS is actually the most constant factor on the data file market of supply and demand. As the most important supplier of data files, CBS has, firstly, a wide range of files to offer and secondly a number of clearly stated delivery terms. This gives CBS a well-defined policy concerning the delivery of data files, making it the cornerstone of data facilities. The demand side, or rather the research world, has little to offer in return. As mentioned before, it is a diffuse and unorganized field in which mutual interests and viewpoints are difficult to formulate. This unequal situation hinders coordination between the parties concerning delivery terms for the files.

4. The information provided by CBS makes clear

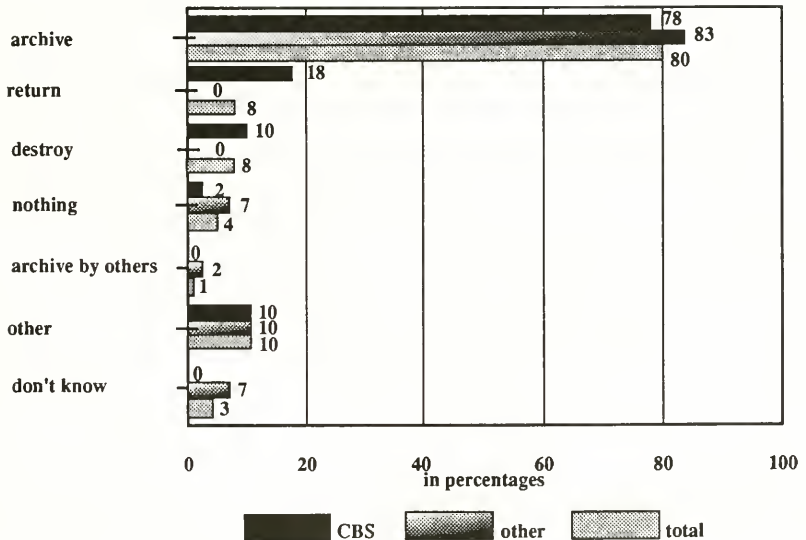
that a demand with great purchasing power exists for a select number of files in the social sciences. CBS's files are better known than files offered by other suppliers. The study also confirms what has been observed in the literature: that researchers find the limited availability of microfiles due to cost, considerations of privacy and delayed delivery problematic. These problems are particularly in evidence with respect to CBS files.

5. The demand for Steinmetz archive files is quite specific. In 67% of the cases, the Steinmetz archive delivers files containing voting research data. Remarkably, almost half of the Dutch files go to foreign countries, and an important portion of the files used by Dutch researchers are from abroad.

6. The survey reveals that 70% of the research is on a contract basis, and that almost all of this is financed by the state. Given that the state also finances the remaining research either directly or indirectly, we can conclude that the state subsidizes almost all social scientific research for which data files are purchased.

7. The fact that most of the data files are used in contract research, together with the observation that non-contract research is much less likely to use CBS

**Figure 4: Archiving**



data than data from other files, indicates the difficulty, if not impossibility, of acquiring costly files given the limited financial resources of the various research groups.

8. Because most of the files are ordered for contract research, it may seem acceptable to pass the costs of the files on to the users. However, it is also important to note that while certain costs may be included in the price of the file, other costs should not be.

9. Local storage and management of data files is conducted at various levels. There are no guidelines or agreements concerning systematic data storage. This hinders access to files that might theoretically be suitable for secondary analysis.

#### RECOMMENDATIONS

Technical innovations have made the processing of machine-readable statistical files increasingly important for social scientific research, a development requiring reflection upon the supply of files and the use of new technical facilities.

1. This reflection can take shape in the form of coordination between the suppliers and consumers of data files. Segers (1990) has already observed that the research world is too unorganized to set up a structural dialogue with suppliers; it must first gather its forces and negotiate with suppliers over new delivery terms and a broader range of user options. Segers recommended establishing an independent organization to represent the research world. As part of its task this institution would coordinate the desires and wishes of the users on one side and CBS on the other. On the basis of this coordination a number of standard agreements could be formulated.

2. Clearly CBS must play a vital role in improving the data infrastructure. Its cooperation is necessary in implementing a number of possible adjustments concerning:

- first of all, the **delivery policy** of CBS. Too much time is lost bargaining over terms of delivery. Delivery of microfiles is often laborious, partly due to the disclosure risk; as a result, delivery is expensive and irregular. Prices increase because files have to be custom-made every time. Both researchers and CBS would be better off if they switched to making standard deliveries of the most important data, perhaps in the interim stage in the form of publication files. These would be "stripped" microfiles.
- second, the **cost of microfiles**. Besides the risk of disclosure, this is the most important barrier to delivery of files. The standard delivery of data mentioned above, in the form of inexpensive publication files, may help to solve this problem, but the cost of microfiles must be reconsidered further.

The main idea would be to generate a buyers' demand within non-contract research; in other words, to take a number of measures that will facilitate the acquisition of files by university research groups. The recommendation made by De Guchteneire & Timmermans to establish a central budget for the purchase of data files might serve as a point of departure.

- third, the **use of other media** than what has been used to date. This involves creating facilities destined for long-term use, for example CD-ROM. This initiative clearly supports the standard delivery of important data. CBS has assumed that for the time being such media are more suitable for the dissemination of aggregated data than for the distribution of individual level data, for example all statistical data concerning municipalities. Mention must be made here that university libraries will no doubt be unavoidably drawn into this issue. When libraries purchase and make such information available, it becomes available to a much wider audience of users.

3. Data infrastructure would profit by improved "signposting". Users or potential users must have better information about which files are to be found where, and how and under what conditions they can be acquired. At the moment most researchers do not know which paths to follow to get this kind of information; often they may not even know that such paths exist. Consequently, policy should aim at improving information on user options; in other words, documentation, etc. with respect to finding and acquiring the files.

4. Researchers should have a service organization available to them for information and assistance in finding and acquiring data files. In the United States in particular an increasing number of universities are beginning to set up and develop local data libraries. In most cases these libraries still form part of a "regular" library. Such libraries perform the following tasks:

- \* systematically storing used data files;
- \* entering bibliographical information on these files into a computerized catalogue;
- \* providing services to researchers or students, including:
  - searching for particular files
  - using the files
  - analyzing the files.

Besides offering advice, other administrative duties can be included as part of the service:

- performing the tasks mentioned above,



- finding and retrieving data files present elsewhere<sup>11</sup> upon request.

5. De Guchteneire & Timmermans proposed establishing a service organization which would perform the following tasks:

- \* gathering and developing expertise directed at making sensitive material anonymous<sup>12</sup>
- \* developing techniques through which sensitive material can still be made available (for example, by working with test files),
- \* advising researchers.

To these tasks may be added:

- \* designing sound rules concerning liability and other organizational terms concerning the safety of the files.

As in point 4, the issue here is service to researchers. We advise describing the possible functions and tasks of a service organization in a follow-up project, and exploring which of these tasks should or could be performed locally or nationally.

6. To summarize, in shaping a new or adapted data infra-structure attention must be given to the following points:

- \* organization and representation from the field;
- \* conditions for making data files available;
- \* the way in which the files are made available;
- \* guidelines for data file storage and management;
- \* service for researchers and students.

An important consideration in describing the infrastructure is the current and future position of the remaining supply agencies, in particular the Steinmetz archive.

#### INCONCLUSION

Technical innovations have made it increasingly possible to produce computerized data files, make them available and analyze them. Data can be analyzed more often and from various perspectives. For this reason, we recommend that CBS files be made available without many restrictions to as broad a sector of researchers as possible, and that different types of analysis be conducted by various institutions. Because CBS produces statistics, it must process the collected data in a number of ways. There are so many different manipulations that can be performed on files that it would be impracticable for CBS to carry out them all. This is not to say that CBS would no longer have to perform manipulations, on the contrary. However, microfiles should be made more readily available to a larger group of researchers. Wider delivery of CBS files would serve not only the interests of researchers, but of society at large.

Adequate construction and efficient use of the data infrastructure requires the involvement of builders, managers and users. Technology on the one hand and sound agreements on the other must lead to the development of an infrastructure that allows researchers to make high-speed use of the available "data highways." At present, however, unexpected roadblocks and no trespassing signs have been thrown up. Regulations concerning availability must be amended in order to give data files the right of way. The interests of all parties involved must be kept in mind. It is time to take measures and come to agreements: a well functioning infrastructure is a necessity in a modern and democratic society.

The respondents have given their opinion on a number of statements included in the questionnaire. Scoring ranges from 1= "Agree completely," 3= "Indifferent" to 5= "Disagree completely." For clarity the results have been converted into a figure whose minimum score of -2 is equal to "problematic" and whose maximum score of 2 is equal to "not problematic."

#### BIBLIOGRAPHY

Bie, S.E. de, Wetenschap en statistiek, standpunt van de vereniging van Onderzoek Instellingen inzake de levering van CBS microdata ten behoeve van sociaal wetenschappelijk onderzoek. Leiden, VOI, 1988.

Guchteneire, P.F.A. and J.G.M. Timmermans, Wetenschappelijk gebruik van overheidsdatabestanden. Amsterdam, Swidoc, 1990.

Segers, J., De data-infrastructuur in Nederland: snelweg of zandpad? Presented during the "Statistics Day" conference, 1990.

Steinmetz archive, Data catalogue & guide. Amsterdam, 1990.

<sup>1</sup> Presented at the IASSIST '90 Conference held in Poughkeepsie, NY, United States, May 30-June 2, 1990.

<sup>2</sup> These files were set up for statistical analysis and can theoretically be used for planning and policy. As opposed to data in an administrative file, data stored in a statistical file is not altered again except for aggregations and scale or index constructions. For example, data included in the "Population statistics" file undergoes no further change once they have been collected, whereas data in an administrative file such as the register of births, marriages and deaths is altered whenever a resident whose personal information is included in the file moves or receives a new passport or driver's license. In statistical files, just as in administrative files, informa-

tion is often collected at the level of the individual. In the analysis of statistical files, however, the object is manipulations at the aggregate or case level. Data from administrative files, in contrast, is used at the individual level (De Guchteneire & Timmermans, 1990), as the above example makes clear.

<sup>3</sup>The social sciences encompass a wide range of fields, and the borders are not always sharply drawn. Even the Dutch term “gamma sciences” does not solve the border issue. In the context of this study the following categories were used:

- Economics (microeconomics, macroeconomics, business economics and econometry) and business administration;
- sociology;
- psychology;
- education;
- political science;
- public and policy administration;
- environmental planning, geography and related studies.

<sup>4</sup>These are files whose data was not collected at the initiative of the institution itself, but were acquired or purchased from others. Not included are files whose data

- was collected by the institution’s own field workers;
- was collected by others at the request of the institution.

<sup>5</sup>The files had to meet the following criteria:

- \* the data was in a file; the study specifically did not focus on tables that were delivered;
- \* the files were stored in such a way (tape, diskette, CD-ROM) that they could be accessed by computer (that is, as a datamatrix);
- \* the files were in the possession of the researcher, research group or institution; on-line external files were not included in the study;
- \* microfiles (data at the level of persons or households) had to include at least 1000 research units; mesofiles or macrofiles (data at the level of companies, institutions, sectors or countries) needed at least 100 research units;
- \* the files were used for research in 1989;
- \* the files were used for research in the social sciences.

<sup>6</sup>In **publication files** the research data, anonymous or not, is reworked in such a way that disclosure of individuals is almost entirely ruled out. **Microfiles** are also anonymous data files at the individual (or household) level, but they include such detailed information on variables that by making intelligent combinations,

disclosure may be possible. The ethical code maintained by researchers does not permit such disclosures. Microfiles are only supplied under the terms of an agreement; publication files are offered on the open market. Unlike publication files, microfiles are supplied neither to government agencies serving a public administrative function, nor to business, but only to research institutions, meaning universities and the Planning Bureaus.

<sup>7</sup>For example, the response states that ten deliveries of a particular file yielded 375,000 guilders. It is not clear whether this is two large deliveries yielding 300,000 guilders and eight others yielding 75,000 in all, or another combination.

<sup>8</sup>CBS was unable to meet our request for a survey of purchase prices of the most complete files. Because negotiations are underway concerning policy adjustments with respect to data file availability, CBS found it an inopportune moment to provide price information which would become outdated in the near future.

<sup>9</sup>The weekly questionnaires were offered 226 times in all. These were 226 different questionnaires.

<sup>10</sup>These are files delivered by the Steinmetz archive.

<sup>11</sup>Various developments in the United States have taken place in close collaboration with the International Consortium for Political and Social Research (ICPSR).

<sup>12</sup>CBS comments here that “anonymous” is hardly a viable term. It is not only a matter of sensitive data files but also of data that cannot be identified at the individual level, and of non-disclosure, according to the CBS. The Bureau states that a high level of expertise has been focused on this issue and that the Bureau, affiliated institutions outside the Netherlands and certain foreign research institutions are continuously developing greater expertise.