

**STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE**

**COMMISSION OF THE  
EUROPEAN COMMUNITIES**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROSTAT**

**Joint ECE/Eurostat Work Session on  
Statistical Data Confidentiality**  
(Thessaloniki, Greece, 8-10 March 1999)

Working Paper No. 17  
English only

Topic (iii) - Administration and policy of statistical data confidentiality

## **STATISTICAL CONFIDENTIALITY AT THE EUROPEAN LEVEL**

Submitted by Dr. Jan Holvast<sup>1</sup>

### **Invited paper**

#### **I. INTRODUCTION**

1. In the mid 1980s new awareness arose about the possibilities of disclosure of statistical data with the breach of promised confidentiality to the data providers as a possible consequence. Although the cause of this renewed discussion can be ascribed to several factors, without any doubt the development in technology and especially in information technology has played a very important role. Computers are becoming more and more powerful and have proven to be a powerful instrument for matching or record linkage, one of the ways by which statistical data can be disclosed. The developments in (information) technology are at the same time a major cause of the continuous discussion on data protection and privacy in most European countries and beyond.

2. Reacting to this discussion the statistical institutes are very careful in releasing statistical data for secondary use outside the institutes. Technical measures are taken in order to prevent simple (re)identification of data. For data users these technical measures are sometimes unacceptable since they make the data almost useless for the kinds of analysis they need. All this has resulted in a renewed interest in the question how to balance the demand for data dissemination and the need for confidentiality.

3. From time to time the balance needs to be controlled. This is of special interest for the national statistical institutes which play a dominant role in the dissemination of statistical data. On the one hand, they have to satisfy the data users by supplying them with the data they require. On the other hand, they have promised their data providers that this information will never be published in any form which makes identification possible.

4. On behalf of the European Commission we have studied the balance for the national statistical institutes in the European Union and, for comparison, some countries abroad. A questionnaire was sent to all statistical institutes of the Member States and Eurostat and to some institutes beyond in order to

---

<sup>1</sup> Holvast & Partner, Privacy Consultants, Landsmeer, Netherlands.

get empirical information about the ways these safeguards are used in practice. Additionally in some countries interviews were held with key persons involved in confidentiality and disclosure and relevant literature was studied. Some of the results are presented in this article.

## II. THE PROBLEM

5. The problem of confidentiality and disclosure is not a modern one which began only at the start of the so-called information age. Confidentiality has been an issue almost as long as data have been collected for statistical purposes. Study of the various forms of legislation which existed in the past, such as Scherff<sup>2</sup> in 1952, clearly show that the main arguments for maintaining this confidentiality were very similar to those of today: lack of confidentiality would cause mistrust on the part of the data providers and therefore result in poor response and in less reliable data. Therefore, in most countries it is forbidden to disseminate statistical data to other persons or for other than the statistical purposes for which the data have been collected and processed. In some cases it is stated explicitly that the data should not be used for fiscal purposes, for tax control, or by the police. In almost all countries statistical data may not be passed on to the government, unless provided for by law, by individual consent, or when the data are necessary to prove that the Statistical Act or other agreements were broken.

6. In the discussion, some important concepts are used which need some clarification: (re)-identification, confidentiality and disclosure. Keller and Bethlehem<sup>3</sup> use a general accepted definition of *identification*: "Identifying a record is to establish a one-to-one correspondence between the record and a specific individual." This definition has proved its usefulness in an experimental situation conducted by Müller, Blien and Wirth<sup>4</sup>. In this experiment an anonymous Micro Data File (MDF) was compared with an Identification File (IF), in which some variables (key variables) were the same. In this way they tried to identify unknown persons in the microdata file. In their view identification takes place "when a one-to-one relationship between a record in the MDF and the IF can be established through linkage of the information contained in the key variables and when it can be established that the information pertains to the same individual"<sup>5</sup>.

7. According to Dalenius<sup>6</sup> privacy is a concept which applies to data subjects, while *confidentiality* applies to data. Like Davis he defines the concept as follows: "It is the status accorded to data which has been agreed upon between the person or organization furnishing the data and the

---

<sup>2</sup> Scherff, Goetz E., Die Rechtsgrundlagen der Statistik, Eine international vergleichende Darstellung unter besondere Vorhebung der Volkszählungen und der statistischen Geheimhaltungspflicht. Stuttgart, Baden Württemberg, 1952.

<sup>3</sup> Keller, W.J. and J.G. Bethlehem, Disclosure protection of microdata: problems and solutions, in *Statistica Neerlandica*, vol. 46, nr. 1, 1992, blz. 5-19. See also Keller, W.J., Disclosure protection of micro data, in proceedings of the Seminar on openness and protection of privacy in the information society. Voorburg: Netherlands Central Bureau of Statistics, 1987, 92-99.

<sup>4</sup> Müller, Walter, Uwe Blien, Peter Knoche, Heike Wirth u.a., Die faktische Anonymität von Mikrodaten. Metzler Poeschel, 1991. See also Blien U., H. Wirth, M. Müller, Disclosure risk for microdata stemming from official statistics, in *Statistica Neerlandica*, vol. 46, nr. 1, 1992, blz. 69-82; Uwe Blien, Walter Müller, Heike Wirth, Needles in Haystacks are hard to Find: Testing Disclosure Risks of anonymous individual Data. Eurostat (1992), 391- 406; Müller, Walter, Uwe Blien, Heike Wirth, Identification Risks of Microdata, Evidence From Experimental Studies, in *Sociological Methods & Research*, Vol. 24, No. 2, 1995, 131-157.

<sup>5</sup> Blien, Wirth, Müller (1992), o.c. 393.

<sup>6</sup> Dalenius, Tore, Controlling Invasion of Privacy in Surveys. Statistics Sweden: Department of Development and Research, Statistical Research Unit, 1988.

organization receiving it and which describes the degree of protection which will be provided." This promise of confidentiality is important for the statistical institutes in order to ensure the quality of statistical data that the respondents provide. Concerns about confidentiality have been heightened by the decline in response rates for censuses and surveys over the past two decades. There is a definite relationship between confidentiality and *privacy*. Breach of confidentiality can result in disclosure of data which harms the individual. This is an attack on privacy because it is an intrusion into a person's self-determination on the way his or her personal data are used. This self-determination, however, never can be absolute, since it is part of the balance of interests. This brings us to another important concept used in this study: *disclosure*.

8. Cox<sup>7</sup> has given the most universal description. In his view, disclosure is breaching the pledge of confidentiality by revealing confidential respondent data. Sometimes the data of the exact respondent to whom the data are related are divulged. Cox defines this as *exact disclosure*. Keller and Bethlehem give three ways in which disclosure can occur which have become generally accepted<sup>8</sup>:

- \* disclosure by matching<sup>9</sup>: with high resolution keys disclosure can be accomplished by matching the data set with a register which contains the keys and names and addresses;
- \* disclosure by response knowledge, i.e. the knowledge that a person was interviewed for a particular survey; if an investigator knows that a specific individual has participated in the survey, and that consequently his or her data are in the data set, identification and disclosure can be accomplished more easily;
- \* disclosure by spontaneous recognition, i.e. recognition of rare persons; disclosure may occur by accident<sup>10</sup>.

9. The first factor is a technical one, the second and third are based on prior knowledge. In fact there are more factors which contribute to disclosure risk. Greenberg and Zayatz<sup>11</sup> give five factors which effect the ability of an investigator to accurately link a record in a public use file to a record on an internal data register. They are all related to the key variables.

10. Based on these concepts, the central question of this study can be described as follows: *How must the balance between the interests of the data user, in being provided with usable data, and the interest of the data provider, in assuring privacy and confidentiality be judged?*

---

<sup>7</sup> Cox, Lawrence H., Solving Confidentiality Protection Problems in Tabulations Using Network Optimization: A Network Model for Cell Suppression in U.S. Economic Censuses, in Eurostat (1992), 229-245.

<sup>8</sup> They call it 'types of disclosure'. As others are also using the word 'types' for their distinctions, we will use the more applicable 'means of disclosure' here. See Keller (1987) and Keller and Bethlehem (1992).

<sup>9</sup> As we have seen, this method was used by e.g. Müller et al in their reidentification experiment.

<sup>10</sup> Bing (From footprints to electronic trails: Some current issues of data protection policy. 17th International Conference on Data Protection and Privacy Commissioners, September 7th, 1995) mentions the example of 'a female rector of a Norwegian university', as the identification of the current rector. In the Report on Statistical Disclosure Limitation Methodology. Statistical Policy Working Paper 22. Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, May 1994, the term 'high visibility' is used, with as examples movie star and federal judge.

<sup>11</sup> Greenberg, B.V. and L.V. Zayatz, Strategies for measuring risk in public use microdata files, in Statistica Neerlandica, vol. 46, nr. 1, 1992, blz. 33-48.

### III. THE PROBLEM ACTUALISED

11. The balance is coming under pressure for several reasons. The first, because statisticians have stimulated the discussion on identification by showing that prior knowledge in combination with indirect identifiers can lead to identification and disclosure. Duncan and Pearson also give five technical factors in contemporary concern about (re)identification of individual records:

- \* identification is easier, because sophisticated and more widely available computational and analytical technology is accessible;
- \* more microdata exist in government and business circles;
- \* the consequences of disclosure are greater: those who collect data are increasingly concerned that the technology and the detail of the records will diminish the public's trust and cooperation;
- \* the motivation for identification may have increased in an information society in which the incentive to gain advantage through intelligence-gathering techniques is increasing;
- \* microdata are harder to disguise prior to their distribution, without also degrading the scientific value of the data.

At the same time there is a growing demand for statistical data. Computers and advanced statistical programs at universities and research centres make it both possible and attractive for (social) scientists to manipulate and analyse data in the way they require, based on their own formulated research program.

12. A second important element is the role of information technology and especially of the computer and communication technology. Access to powerful computers is necessary in some special cases of disclosure. These computers are not only available but are also becoming more and more powerful. It is therefore doubtful whether anonymity still exists now. Since 1980, the unreasonable amounts of time, money and manpower which have been expended to protect de facto anonymity as a result of developments in hardware and software has become an issue in itself, particularly now the technical problems for matching or record linkage seems to have been solved. The computer must be seen in combination with communication, and especially Electronic Data Interchange (EDI). The advantages of EDI over paper-based documents are speed, avoidance of ambiguity, ease of data capture, reliability and cost. It is obvious that statistical institutes will also use EDI for collecting and disseminating statistical information. This interchange of data is an important element in the new problems of protecting natural and legal persons with regard to data processing.

13. A third, and probably most important element, is the continuous discussion on data protection and privacy in most European countries and beyond. Most discussions started in the early 1970s, sometimes in response to a census (Netherlands, Sweden)<sup>12</sup>. In the mid-1980s the subject came under discussion once again in some countries, namely in Germany in 1983 and 1987 in reaction to the proposed census, and in Sweden in 1987 as a response to the Project Metropolitan<sup>13</sup>. The discussion focused especially on scientific research and the possibility of (re)identification<sup>14</sup>.

14. All these elements are responsible for a new interest in the balance between the demand for data dissemination and the right to privacy, and in general the need for confidentiality. Central in the discussion is the consequences of a disbalance. Special conferences on this topic, organised by Eurostat, have shown that almost all national statistics institutes are concerned about the consequences

---

<sup>12</sup> Holvast, Jan, *Op weg naar een risicoloze maatschappij?, De vrijheid van de mens in het informatie-tijdperk*. Schoonhoven: Academic Service, 1986.

<sup>13</sup> Dalenius, Tore, *Controlling Invasion of Privacy in Surveys*. Statistics Sweden: Department of Development and Research, Statistical Research Unit, 1988.

<sup>14</sup> Campbell, Dennis and Joy Fisher [Eds.], *Data Transmission and Privacy*. Dordrecht: Martinus Nijhoff Publishers, 1994.

of this discussion on the willingness of data providers to cooperate with the statistical institutes<sup>15</sup>. When respondents are under compulsion, completeness and reliability of data may be affected. If participation is voluntary, it is feared that the level of non-response will rise unacceptably. In both cases statistical offices can be expected to suffer. Bunk<sup>16</sup> paints a dramatic picture of the likely impact on the Central Bureau of Statistics in the Netherlands (see Table 1).

15. Analysis of the elements responsible for the renewed interest in discussing problems of confidentiality and disclosure show that there is possibly more behind the discussion: considerable public distrust of statistics and statistical institutes in some countries.

#### IV. PUBLIC DISTRUST

16. This distrust is not always of statistical data itself, but of the collection of data in general, or perhaps, even more generally, of the information society, with its potential power, and the increasing distance between respondent and data. It seems that the general discussion of confidentiality and privacy taking place in most countries in turn influences the debate on confidentiality and disclosure. This privacy debate voices the concern of most statistical offices especially since statistical institutes seem to be the object of more general critical discussion.

17. Hölder<sup>17</sup> gives us an overview of the background to protests against censuses. First, there is a general uneasiness with modern information technology. Secondly there is a lack of confidence in the extent of separation between statistical institute and other agencies. Thirdly, in his view, there is a general rejection of the State as seen in the efforts to boycott the census in Germany. David Flaherty<sup>18</sup> partly endorses this view. Statistical agencies should be constantly aware of how their data collection activities may be perceived by the general public and by politicians. The existence of massive amounts of personal data used for research and statistics concerns many people. They ask legitimate questions about promises of confidentiality and statistics.

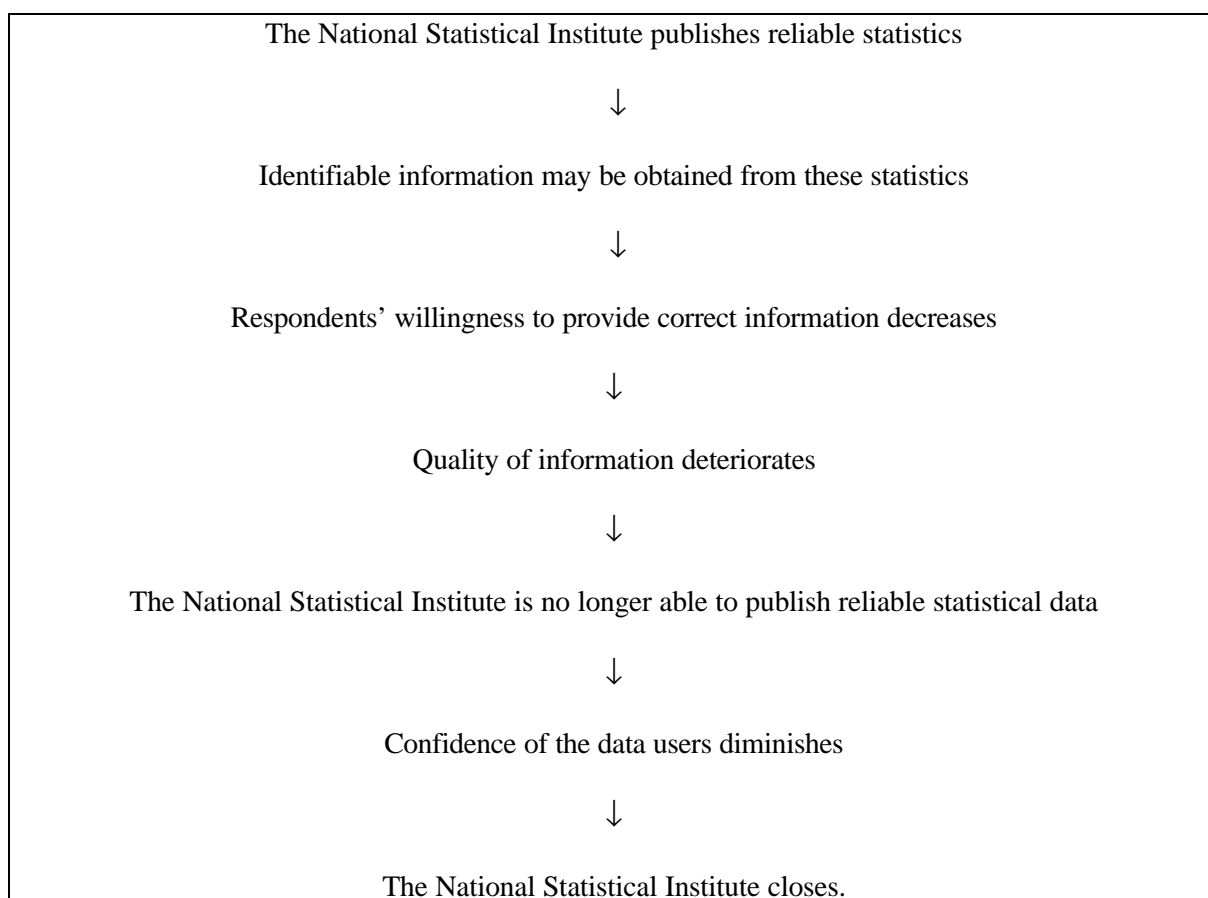
---

<sup>15</sup> Eurostat: Proceedings of the International Seminar on Statistical Confidentiality, 8-10 September 1992, Dublin, Ireland. Eurostat, 1993; Eurostat, Proceedings of the International Seminar on Statistical Confidentiality, 28 to 30 November 1994 in Luxembourg. Eurostat, 1995; Eurostat/ Statistical Office of the Republic of Slovenia, Statistical Confidentiality. Third International Seminar, 2-4 October 1996, Bled, Slovenia.

<sup>16</sup> Bunk, J.L.S., *Beveiliging van bedrijfsgegevens: wet en beleid*. Voorburg: Centraal Bureau voor de Statistiek, 1996.

<sup>17</sup> Hölder, Egon, Why are people refusing to participate in surveys, in *Statistics and Privacy*, Report from a conference in Stockholm, Sweden, 14-26 June 1987, 3-4.

<sup>18</sup> Flaherty, D.H., Data protection and the statistical community, in Eurostat (1986), 239-259.

**Table 1.: The downward spiral of a statistics institute**

18. This issue was the focus of much attention at a conference in Stockholm, Sweden in 1987<sup>19</sup>. Dalenius suggests that the problem of public distrust is more a sign of a lack of confidence in government than in statistics. Although, in this view, matters of policy become a major concern, this does not mean that questions of technique are unimportant. The statistical bureaus should have a good range of tools at their disposal, even if the public is not convinced of their importance. At the same conference, Flaherty expressed the opinion that too much time is spent in discussing the risks of identification, de-identification and re-identification. This is not what the public is really worried about. "They are not concerned about the technical chance of identifying someone in a statistical publication. It is the 'burning issue' of today that should demand our attention: the political disputes, the public misunderstanding of the need for statistics, the public scepticism of the role of statistics and research in society, and the monopolistic position of statistical agencies".<sup>20</sup> Georges Als reaches a similar conclusion in his thorough comparative study of the organization of the - then - 12 Member States of the European Community "Despite all the guarantees, the population distrusts the statistical services, and the responses it gives are no fuller and truer than those supplied to the Revenue or other authorities."<sup>21</sup> Recent discussions on the use of statistical data appear to justify the viewpoints of - Dalenius, Flaherty and Als.

<sup>19</sup> Statistics and Privacy, Future Access to Data for Official Statistics - Cooperation or Distrust? Report from a conference in Stockholm, Sweden, 24-26 June 1987.

<sup>20</sup> Statistics and Privacy (1987), o.c. 27.

## V. SAFEGUARDING CONFIDENTIALITY AND DISCLOSURE

19. When studying the enormous body of literature on tensions between confidentiality and disclosure, it is striking that most safeguards appear to be technical; legal and organisational measures are rarely mentioned. Tore Dalenius<sup>22</sup> and Duncan et al<sup>23</sup> are exceptions.

20. In our study on the scope of personal data within (socio-) scientific research<sup>24</sup> we used a subdivision which proved useful to describe all theoretical and practical safeguards for confidentiality and disclosure:

- \* legal and ethical safeguards;
- \* organisational and administrative safeguards;
- \* technical safeguards.

21. The legal and ethical safeguards consist of a legal system of agreements, arrangements (laws, codes of conduct or ethics) and penalties in case of breach of confidentiality. The most important difference between the other two categories of safeguards is that the organisational and administrative measures in place are intended to reduce the risk of disclosure (working on site, trusted third parties who participate as gatekeepers). Technical safeguards, on the other hand, prevent disclosure by modifying the data in such a way that identification and disclosure is limited. These technical safeguards attempt to achieve de facto anonymity.

22. Several authors emphasize that disclosure never can be totally eliminated; elimination, therefore, can never be the aim of safeguards. For that reason they prefer<sup>25</sup> to speak of limiting or controlling, rather than avoiding or eliminating disclosure. We thoroughly endorse this view. The only way to prevent disclosure is not to release statistical data at all, which would mean that neither the data user nor the data provider would be served.

---

<sup>21</sup> Als, Georges, Organization of Statistics in the Member Countries of the European Community, Volume I: Essays on the 12 statistical institutes, Comparative study. Luxembourg: Office for Official Publications of the European Communities, 1993. o.c. 164.

<sup>22</sup> Dalenius, Tore, Controlling Invasion of Privacy in Surveys. Statistics Sweden: Department of Development and Research, Statistical Research Unit, 1988.

<sup>23</sup> Duncan, George T. , Thomas B. Jabine, and Virginia A. De Wolf [Eds.], Private Lives and Public Policies: Confidentiality and Assessibility of Government Statistics. Washington, D.C. : National Academy Press, 1993.

<sup>24</sup> Holvast, Jan, Persoonsgegevens of niet: dat is de vraag, in Jan Holvast e.a., Nationaal Programma Informatietechnologie en Recht (ITeR), vol. 2. Alphen aan den Rijn: Samsom BedrijfsInformatie, 1996. See also: Nanopoulos, Ph., Discours Inaugurel, in Eurostat (1993), 19-21; and Wieland, Ulrich, Information Security and Statistical Confidentiality, in Eurostat (1995), 221-224).

<sup>25</sup> See e.g. Dalenius (1988); Duncan, G.T. and D. Lambert, The risk of disclosure for Microdata, in Journal of Business and Economic Statistics, 7, 1989, 207-217; Duncan, George T., Virginia A. de Wolf, Thomas B. Jabine and Miron L. Straf, Report of the Panel on Confidentiality and Data Access, in Journal of Official Statistics, vol. 9, no. 2, 1993, 271-274; Report on Statistical Disclosure Limitation Methodology. Statistical Policy Working Paper 22. Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, May 1994.

## Legal and ethical safeguards

23. Since the beginning of the processing of statistical data, a wide range of recommendations, council regulations, directives, forms of selfregulation and national data protection acts have been formulated and accepted in which the dissemination of (statistical) data is regulated. Some general conclusions can be drawn from these directives and recommendations:

- \* data shall be disseminated so that they remain anonymous, which means they can be reidentified only with an unreasonable amount of time, cost and manpower;
- \* statistical information shall only be published or made accessible to third parties if measures are taken to ensure that the data subjects remain unidentifiable;
- \* the use of personal data for statistical or scientific purposes should not be incompatible with the original purpose of the data collection;
- \* in such cases the European Directive (95/46/EC) stipulates that Member States provide suitable safeguards, in particular to prevent data being used for measures or decisions affecting any particular individual;
- \* all these regulations have been set out in national legislation which is already in force in many Member States of the European Union.

24. The relationship between statistical and administrative data is especially important. It is clear from the statistical process that the output of statistical institutes is different from the information of administrative offices. For this reason, a distinction is made between statistical and administrative data depending on the purpose for which they are used. Nanopoulos and Kioussis<sup>26</sup> distinguish these as follows:

- \* The *administrative* purpose applies when the primary use of the data is to permit a legal or mutually consented interaction between the collector and the respondent. This is the case of all public administrations, professional or social associations, regional organisations and even individual persons.
- \* The *statistical* purpose applies when the primary use of the individual data is to calculate characteristics of populations of units where the individual units cannot be identified.

25. In contrast to administrative data, the purpose of statistical data is not concerned with a course of action which affects a particular person or business, but to produce an aggregate description of a group of persons or businesses, e.g. the average income in a region or the percentage of employment in a town. A clear functional separation needs to be made between the statistical purpose and administrative and intelligence purposes<sup>27</sup>. Although administrative and intelligence data may be used for statistical purposes, the reverse is never allowed. Statistical data may not be used for either administrative or intelligence purposes<sup>28</sup>.

---

<sup>26</sup> Nanopoulos, Photis and Leonidas Kioussis, Orientation of Future Work in Eurostat on Statistical Confidentiality Issues, in Eurostat (1995), 15-22.

<sup>27</sup> See Duncan, George T. , Thomas B. Jabine, and Virginia A. De Wolf. [Eds.], Private Lives and Public Policies: Confidentiality and Assesibility of Government Statistics. Washington, D.C. : National Academy Press, 1993.

<sup>28</sup> In 1992, the investigation into gas and heating oil pricing in the US by The Antitrust Division of the Department of Justice requested access to individually identifiable company data collected by the Energy Information Agency. Although EIA refused, referring to its confidentiality policy, the Department of Justice Office of Legal Council responded that EIA was required to provide the requested information. See Fienberg, Stephen E., Conflicts between the Needs for Access to Statistical Information and Demands for

## Organisational and administrative safeguards

26. To assist the Panel on Confidentiality and Data Access<sup>29</sup> Thomas B. Jabine prepared a summary of restricted access procedures used by US statistical agencies to make data available<sup>30</sup>. For the purpose of the paper the most important attribute of data access is whether access is restricted or unrestricted. “Unrestricted access occurs when aggregate data or microdata are released to anyone who wants them, with no restrictions of any kind. Access is restricted whenever any conditions other than payment or fees are imposed.”<sup>31</sup> He describes especially the restricted access procedures. His experience is that there is a wide variety of formats and conditions associated with restricted release or restricted access.

27. The most important role in the process of release or access is that of data-recipient. Data recipients can range from being a regular employee of the agency to someone with no formal relationship at all. Access by regular employees usually falls outside the scope of data release or access. Nevertheless, a special arrangement should be noted here which is taken by some statistical agencies, e.g. US Census Bureau for *special sworn employees*. These employees are given restricted access to confidential data in a manner similar to regular employees. In most cases, this happens *on site*. This privilege is temporary for the duration of a grant or fellowship in case of academic research. The obvious advantage of this requirement is that data use can be closely supervised by the agency. All data released as a result of such activities must meet the usual agency restriction. A comparable form of restricted access is the legal condition which stipulates that data can be used only for research carried out exclusively by a university or other independent research institutes.

28. This particular form of access is sometimes combined with other safeguards, although these are also used separately from sworn employees: a licensing agreement, signing a contract or undertaking and an evaluation by a review panel. A *Licensing agreement* is a permit, issued under certain conditions, for researchers to use confidential data for specific purposes and for specific periods of time. This agreement consists of contractual and ethical obligations, as well as penalties for improper disclosure of identifiable information. These *penalties* can vary from withdrawal of the license and denial of access to additional data sets to a deposit paid prior to the release of a microdata file, which will be forfeited if the user fails to fulfill the provisions of the licensing agreement. A licensing agreement is almost always combined with *signing a contract*. This contract includes a number of requirements: specification of the intended use of the data; not to release the microdata file to another recipient; prior review and approval by the releasing agency for all user outputs to be published or disseminated; the terms and location of access and enforceable penalties.

29. Requests for access to confidential data are sometimes evaluated by a special *review panel* or *review committee*, which is regarded as part of a broader group of independent *Trusted Third Parties*. They may also have to monitor draft publications before permission for release is given. They are comparable to what Duncan and Pearson<sup>32</sup> call a ‘gatekeeper’. These intermediaries control access to data, especially when remote access is required.

---

Confidentiality, in *Journal of Official Statistics*, vol. 10, no.2, 1994, 115-132 and Duncan et al (1993), 185-188.

<sup>29</sup> Duncan et al (1993).

<sup>30</sup> Jabine, Thomas B., Procedures for Restricted Data Access, in *Journal of Official Statistics*, vol. 9., nr. 2, 1993, blz. 537-589.

<sup>31</sup> Jabine (1993), o.c. 538.

<sup>32</sup> Duncan and Pearson (1991), o.c. 225.

## Technical safeguards

30. Judging from the considerable body of literature and the frequent conferences, much attention is given to the problem of data control or restricted data solutions. Substantial articles by George T. Duncan in cooperation with Diane Lambert<sup>33</sup> and Robert Pearson are of special interest for an overview of technical safeguards<sup>34</sup>. This discussion has been further augmented by the publication of issues devoted to this topic in two influential journals: Volume 46, nr. 1, 1992, of *Statistica Neerlandica* and, above all, Volume 9, no. 2, 1993, of the *Journal of Official Statistics*. Some of the articles were issued previously in the publication presented at the International Seminar on Statistical Confidentiality in Dublin, Ireland, on 8-10 September 1992<sup>35</sup>. This seminar was followed by the International Seminar on Statistical Confidentiality in Luxembourg on 28-30 November 1994<sup>36</sup>, the 3rd International Seminar on Statistical Confidentiality in Bled, Slovenia, on 2-4 October 1996<sup>37</sup> and the Seminar on Use of Administrative Sources for Statistical Purposes, in Luxembourg, on 15 and 16 January 1997<sup>38</sup>. All seminars were organised in close cooperation with Eurostat, which has brought out its own publications on this topic; of these the *Manual on Disclosure Control Methods*<sup>39</sup> deserves special mention. This manual gives an overview of existing techniques and introduces a list of criteria to evaluate and compare methods. Last, but not least, two other influential reports should be mentioned. The first is the *Report on Statistical Disclosure Limitation Methodology*<sup>40</sup>, an update of the previous subcommittee's report on the same topic, which describes and evaluates existing disclosure limitation methods for tables and microdata files. The second is the *Report of the Panel on Confidentiality and Data Access, Private Lives and Public Policies, Confidentiality and Accessibility of Government Statistics*<sup>41</sup>. This Panel was commissioned by the Committee on National Statistics and the Social Science Research Council to make recommendations to federal statistical agencies to support supervision of data for policy discussion and research. All these publications provide their own

---

<sup>33</sup> Duncan, George T. and Diane Lambert, Disclosure-Limited Data Dissemination, in *Journal of the American Statistical Association*, March 1986, Vol. 81, No. 393, 10-18, with comments from Lawrence H. Cox, Ove Frank, Joseph L. Gastwirth and Harry V. Roberts, 19-27; Duncan, G.T. and D. Lambert, The risk of disclosure for Microdata, in *Journal of Business and Economic Statistics*, 7, 1989, 207-217.

<sup>34</sup> Duncan, George T. and Robert W. Pearson, Enhancing Access to Microdata while Protecting Confidentiality: Prospects for the Future, in *Statistical Science*, Vol. 6, No. 3, 1991, 219-239, with comments of Lawrence H. Cox and Sallie Keller-McNulty.

<sup>35</sup> Eurostat: Proceedings of the International Seminar on Statistical Confidentiality, 8-10 September 1992, Dublin, Ireland. Luxembourg: Eurostat, 1993.

<sup>36</sup> Eurostat, Proceedings of the International Seminar on Statistical Confidentiality, 28 to 30 November 1994 in Luxembourg. Luxembourg: Eurostat, 1995.

<sup>37</sup> Eurostat/Statistical Office of the Republic of Slovenia, Statistical Confidentiality. Third International Seminar, 2-4 October 1996, Bled, Slovenia, Collection of Papers.

<sup>38</sup> Eurostat, Proceedings of the Seminar Use of Administrative Sources for Statistical Purposes, in Luxembourg, on 15 and 16 January 1997, Eurostat, 1997.

<sup>39</sup> Eurostat, *Manual on Disclosure Control Methods*, Theme Research and Development, 9E. Luxembourg: Eurostat, 1996.

<sup>40</sup> Report on Statistical Disclosure Limitation Methodology. Statistical Policy Working Paper 22. Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, May 1994.

<sup>41</sup> Duncan et al (1993).

overview of existing methods. Different disclosure limitation methods are presented for macrodata (tabular data) and microdata files.

### **Some important research findings**

31. Since it is impossible to present all the results of our study, we will give the most important ones, starting with the answer to the central question.

**On the whole, we conclude that Member States of the European Union and Eurostat have achieved a reasonable balance between respondents' interests in privacy and confidentiality and those of data users in obtaining useful data.**

32. The central question in this study was how the balance must be judged between the competing demands for providing data users with usable data on the one hand and fulfilling pledges of confidentiality on the other. We are of the opinion that national statistical institutes achieve a good balance, with the support of national and communal legislation and regulations, as well as their own self-regulation. Without diminishing protection of confidentiality and privacy, they provide data users with macro- and microdata in sufficient detail to be of use for research and policy purposes.

33. In this conclusion, one has to realise that some researchers will only be satisfied when they have access to all anonymous data with no preventive technical measures except the removal of direct identifiers. On several occasions, these researchers have indicated that they should have access to these data as no serious breaches of confidentiality have occurred and that therefore the scientific community can be trusted. However, it is not only the attitude of national statistical institutes towards researchers which is important, but above all dependency on respondents and how far they trust the scientific community. If respondents provide statistical institutes with less (reliable) data then not only the institutes are harmed but also the researchers themselves. Therefore statistical institutes have to give higher priority to confidentiality and privacy than to free access to statistical data.

34. This conclusion also applies to Eurostat. Political and economic integration of Europe has resulted in a pressing need to improve access to statistical microdata on European level. But here too a balance must be sought between these demands and the need for confidentiality. Eurostat has so far succeeded in striking the right balance.

**All EU national statistical institutes are convinced of the importance of protection of statistical data relating to natural persons and businesses. This is further confirmed by the literature, interviews and many reports and notes on the subject.**

35. In the questionnaires, there is a 100% score in the affirmative for questions on the importance of protection of statistical data of natural persons and enterprises. This picture is confirmed by the replies to the second question, which covers the degree of importance national statistical institutes attach, respectively, to legislative and administrative, mathematical and computational, and organisational aspects of confidentiality. Although legal and administrative aspects score highest, the others are seen as almost equally important. The results for each EU Member State can be seen in Table 2.

**Table 2.: Importance of different aspects of confidentiality for national statistical institutes.**

	very often/often	sometimes/never
Legislative and administrative aspects	14	0
Mathematical and computing aspects	10	4
Organisational aspects	14	0

36. These data are confirmed by the results of interviews with representatives of 11 countries. In these interviews, confidential internal reports were often revealed on how staff and on-site researchers were expected to treat confidential statistical data and the different sanctions applied in case of breach of confidentiality. This is also reflected in the many conferences on confidentiality and the numerous articles in scientific journals all emphasize the importance of confidentiality.

**There is little difference in the treatment of personal data and company data. When there is a difference it is due to the nature of the data and the company and only in some cases to fundamental difference in policy.**

37. Confidentiality protection of statistical data of natural persons and of companies is of equal importance. In practice, however, there may be differences in treatment arising from the nature of goods produced and/or of the companies themselves. In small countries like Ireland and Luxembourg, for example, some products are only produced in one factory while for some products (textile, metal) there is only one manufacturer. In such cases, publication of statistical data will lead to identification, as may also happen in countries where only two or three companies make a particular product e.g. light bulbs and cars. Consequently, additional limitation techniques are taken as the dominance rule. These state that when the number of companies is below a certain figure (usually 3) or when a certain product of a company dominates the market (by e.g. 80%) then the content of the cell should be suppressed. All countries which disseminate business statistics data use this technique. In some countries such as France, however, specific company data are made public in the form of a register.

38. The extra care in the treatment of company data may be attributed to company attitudes which are in general more suspicious and negative than those of natural persons. The interviews showed that the reason for this negative attitude was not so much fear of disclosure as the burden of answering all the surveys. Some representatives reported that there were considerably more complaints from legal persons than from natural persons. The same is true of compulsory surveys, where there are hardly any refusals from natural persons, but a large number from companies.

39. In general we can say that data protection is respondent-oriented with no distinction between the kind of respondents.

**Data Protection Acts are in force in all Member States of the European Union. In addition there are special Statistics Acts at all national statistical institutes to provide for secrecy, confidentiality and finality. At the Community level harmonization is facilitated by Directive 95/46/EC and Regulation No. 322/97 which together provide a reliable legal base for protection of personal data used for statistical purposes.**

40. Italy (February 1997) and Greece (August 1997) were the last Member States of the European Union to enact Data Protection Acts. In addition all countries have special Statistics Acts under which the national statistical institute was sometimes established and where secrecy, confidentiality and the principle of finality is regulated. These state that data obtained for statistical purposes may not be used for any other purpose, with possible exemption of the police or intelligence services in exceptional cases. Release of personal data for statistical purposes is permitted in most Data Protection Acts as well as the release of personal data from national statistical institutes for research purposes. However, this is usually restricted by 'lex specialis', i.e. Statistics Acts.

41. Data Protection Acts have less in common than the Statistical Acts for two reasons. The first relates to the time they were brought into force. Some of the Statistical Acts (e.g. Sweden, Luxembourg) have been in force since the 1970s. The second difference is one of scope. Some data

protection acts apply only to natural personal data, others to natural and legal persons. Although in practice personal data include small companies e.g. one person companies, the difference is considerable as in some countries company data fall under law. The same can be said about the manual data files, which are in some under the act, whereas others only apply to automatic data processing. This has resulted in a difference in the treatment of statistical data.

42. Fortunately, the European Directive 95/46/EC on the protection of individuals and the processing of personal data and on the free movement of such data came into force in 1995. It states that at the latest, on 24 October 1998, all laws, regulations and administrative procedures in all Member States shall apply with this Directive. Although there is some degree of freedom, protection of personal data for statistical purposes will be harmonized. The difference in scope between natural and/or legal persons, however, will remain.

43. Regulation No. 322/97 on Community Statistics will also have a harmonizing effect, as there is a uniform definition of statistical confidentiality. Exchange of confidential data between institutes and between an institute and Eurostat is also regulated as well as access to data for scientific purposes.

**Although there is not always consistency between the probability of disclosure and intruders' interests in disclosing confidential data and the amount of measures involved, these measures are vital because of the vulnerability of national statistical institutes. These institutes are too dependent on the reliable information provided by the respondents.**

44. The study of Müller *et al*, which is confirmed in several other analyses, shows that the probability of disclosure by using matching techniques is very small. It also emerges that the probability of intruders disclosing confidential data is also very low. In discussions amongst experts on confidentiality this inconsistency frequently leads to the question of whether these measures are not disproportionate in number. It is also pointed out that, in the past, disclosure has never taken place. Consequently, it is sometimes said that this is a case of cracking a nut with a sledgehammer. There is a large body of opinion holding that the effort and resources invested in the protection of secrecy and security are far beyond what is required according to reasonable estimates of the risks to private individuals.

45. While this belief is well founded, we have also seen in the past that seemingly negligible incidents can influence attitudes towards social research in general and statistical research in particular. Further, we have seen that this attitude frequently leads to a drop in response. Given their dependency on these respondents statistical institutes are in a very vulnerable position. A loss of public confidence could result in increasingly unreliable data and an end of statistical institutes. Therefore it is an absolute necessity that the institutes retain the confidence of the respondents as well as of the general public.

**A more active campaign by the statistical institutes is required to inform the public of the need for statistical information and of the kinds and amount of measures which are taken to protect confidentiality and prevent disclosure.**

46. In the proceedings and reports of conferences on confidentiality and disclosure, the need to inform the public is almost always emphasised. The public should be informed of the importance of statistical information, especially with the growth of the information society. There are important differences of interest between the administrative institutions and scientific institutions to which the statistical institutes belong. Sometimes, the statement of these differences can be provocative. "No further strengthening of the legal provisions regarding confidentiality but rather an active campaign to inform the public." "Technical precautions are less important because there is already an overkill in limitation information imposed in the name of confidentiality." Although these impressions are

confirmed by the number of measures which are taken, we should avoid creating an either or situation in the discussion. Both are necessary and information on these measures should be part of a total campaign of 'going public'. We should inform the public of the ways of maintaining confidentiality of individually identifiable information, including the use of legal barriers to disclosure and physical security procedures, which are intended to minimize the intrusion of privacy.

**The distinction must be clear between the use of data for administrative and statistical purposes (functional separation). This means that even use of statistical data by the police or intelligence services must be excluded in all Statistics Acts of the Member States of the European Union.**

47. Experience demonstrates that one of the most important reasons for respondents' distrust is the fear that data collected for statistical purposes will be used for administrative purposes. Therefore in most countries of the European Union it is emphasised that administrative data may be used for statistical purposes, but statistical data may not be used for administrative purposes.

48. However in the Statistics Acts of some countries, provision is made for use of emergency data for undefined purposes such as judicial review and national security. As there are no concrete examples of the necessity of this use all possibilities of release for the purposes should be excluded by law.

#### **Future developments<sup>42</sup>**

49. It is clear that public distrust is substantially influenced by technological developments which continue apace. For that reason Nobel<sup>43</sup> states that data access and data confidentiality do not cease to be ethical problems once practical and legal solutions have been found for the issues of the 1980s. We also need to find answers for new questions about the use of data for statistical research such as:

- \* using and scanning aerial pictures for statistical purposes, which can lead to unwanted administrative and penal consequences;
- \* chipcard information on shopping and spending patterns, in particular in combination with other information e.g. mobility;
- \* for a worthwhile description of European society we need to be able to collect and publish information on sensitive issues such as illegal immigration;
- \* knowledge about the operations of supranational companies is essential for a proper description of the global economy.

50. For Thygesen<sup>44</sup> the growth of data communication is the most promising development. In particular he mentions the Local Area Networks (LANs) which facilitate sharing data and messages between workstations within an organisation. Other increasingly important developments are the Electronic Data Interchange (EDI) for companies to exchange (confidential) business information and the use of EDI techniques for statistical institutes to exchange data amongst themselves and also supply information also to other agencies such as Eurostat. These developments in the field of data

---

<sup>42</sup> For recent technological developments in the field of dissemination of statistical data, see Euro-stat, New Techniques and Technologies for Statistics II. Proceedings of the Second Bonn Seminar. Amsterdam: IOS Press, 1997.

<sup>43</sup> Nobel, Joris, Data Confidentiality and Data Access - Practical and Legal Issues in the Netherlands, in Eurostat (1994), 207-214.

<sup>44</sup> Thygesen Lars, Technological Aspects of Confidentiality: New Technology - Threat or Greater Protection?, in Eurostat (1994), 299-305.

communication require specific measures like call back procedures, identification of communication equipment, cryptographic techniques and trusted third parties to certify the keys. He concludes that these technologies will face new threats to confidentiality. Two years later the same author<sup>45</sup> gave an overview of new collection methods, such as the computer assisted data capture Computer Assisted Telephone Interviewing (CATI), Computer Assisted Personal Interview (CAPI), Computer Assisted Direct Input (CADI) and telephone based techniques: Touch- tone data entry (TDI) and Voice Recognition self response (VR).

51. These developments converge in what is known as the electronic highway: the definitive solution to all problems of collecting, storing and disseminating information. It is quite possible that this electronic highway raises new problems, many of which are related to data protection, confidentiality and disclosure.

## VI. CONCLUSION

52. In the course of time, the position of statistical institutes fundamentally has changed. From an institute that collected data and disseminated macrodata it became an institute with a heavy demand for microdata. Universities, research centres and governmental authorities wanted in an increasing way more and more detailed data for making their own (re)analysis. At the same time with the dissemination of this data the problem of reidentification and disclosure got renewed attention. This attention was stimulated by the growing use of the computers for matching of data, by the introduction of EDI and by the lasting privacy discussion. As a consequence all institutes took more and more legal, administrative and organisational measures to diminish the risk of unwanted disclosure of data.

53. Taking this as a starting point, based on the collection of material in this study, some interesting observations can be made. Firstly, the (statistical) probability of disclosure is low and there is little motivation for potential intruders to re-identify and disclose personal data. The main threat of invasion comes from individuals and organisations wishing to show that confidentiality can be - breached. Secondly, on community, country and (statistical) institutional levels, many (legal, ethical, organisational, administrative and technical) measures have already been taken to prevent re-identification and disclosure. Thirdly, public disquiet is unrelated either to a deficiency of protective measures or to the motives of intruders. Often these are related to data collection or to disclosure of data. However this applies mainly to statistical data and information. This leads to a fourth observation that it is the statistical data and statistical institutes themselves which form the weak link in a vulnerable system.

54. Several authors have correctly concluded that statistical institutes are increasingly becoming the target for criticism of information technology and the information society. In some instances, the institute has also become a symbol in the fight for protection of privacy. Information technology is becoming increasingly opaque and individuals who are becoming alienated from their own data may react by attacking the statistical institutes and their surveys and censuses. On the one hand, they symbolize a government which needs ever more detailed information about individual citizens, and on the other hand, the collection of data is seen as one of the real possibilities to say 'no' to a governmental - and in particular to an administrative - appetite for data which seems insatiable. Moreover there is a general sense of distrust arising from the belief that data collected for one purpose (statistics) will be used for another (administration). It is questionable whether simply increasing the legal, organisational, administrative and technical measures is the solution to this complex issue of powerlessness and distrust. These measures are necessary but insufficient.

---

<sup>45</sup> Thygesen, Lars, The Influence of the Technological Development on Data Collection Methods in Survey. How is the Protection of Personal Data Affected, in Eurostat (1995), 197-199.