THE DISSEMINATION OF DATA AMONG STATISTICAL OFFICES AND DATA ACCESS FOR RESEARCH PURPOSES. THE CASE OF ITALY

Luigi Biggeri, Italian National Statistical Institute, Rome
biggeri@istat.it

Key Words: Code of Ethics, Data Access, Safe Data, Safe Settings, Statistical Disclosure Control

Abstract The evolving technology and the growing request and availability of great mass of statistical information and elementary data are changing, provoking a corresponding evolution in the necessary procedures to test the confidentiality of data and in the dissemination regime implemented by NSI's. The paper discusses the problems and the challenges that the technological advancements give for the data access and dissemination among statistical offices and for research purposes assuring the protection of confidentiality. The solutions adopted in Italy, especially to solve the juridical and ethical issues, will be highlighted.

1 Introduction

After many scientific discussions and many researches that proposed a lot of adequate methods to solve the problems, the conflict between the protection of confidentiality and data access seems nearly ended.

However, today the huge opportunities provided by instruments available for the capture, storage, elaboration, and transmission of data, have modified and continue to modify both the demand and the supply for statistical information: on the demand side, users are requesting growing availability of great mass of statistical information and elementary data and they are much more prepared and exacting, and much more autonomous in terms of management and elaboration of data; while from the supply side, a substantial reorganisation of work within the National Statistical Institutes (NSI's) is being carried out. This led the NSI's to make a profound revision of their dissemination procedures, which now must aim at satisfying more flexible requirements than in the past, and this causes a corresponding evolution in the necessary procedures to test the confidentiality of data released.

This is particularly true when the NSI's have to face the demand of the elementary data by other statistical offices and agencies (at local, national and international levels) and by the research world.

In this context the paper, first of all, highlights the effects of the technological advancements on the protection of confidentiality and data access. Secondly, it discusses some critical issues that still remain to be faced to assess the disclosure risk, showing the possible methodological and organisational solutions of the strategies for data dissemination and access (safe data vs. safe settings; internet, databases and remote on-line analysis; role of the intermediaries). Thirdly, the paper focuses on the need for the development of sound juridical basis and regulations, in order to have an efficient protection of personal data, presenting the solutions adopted in Italy thorough a decree law on the protection of privacy for statistics and research, as well as a code of deontology for the Italian National Statistical System (Sistan).

2 Protection of confidentiality and data access: effects of the technological advancements

A simplified framework of the protection of confidentiality

The conflict of the protection of confidentiality *versus* data access is multifaceted but there is a general consensus that both the rights must and can be safeguarded and all the actions are devoted to reach this scope.

In any case, the NSIs have to define their strategies of protection of confidentiality and data access in a combination of measures and flexible methodological, organisational and legal framework, that includes at least the following steps and decisions (Biggeri and Zannella, 1991; Fienberg, 1997; Willemborg and de Waal, 1996):

- Definition of the disclosure risk (clear definition of data anonymous and of identifiable unit).
- Definition of the risk acceptable (threshold) taking into account a series of reasonable factors, e.g.:
  - characteristics of data (hierarchical or not, extent of detail in the survey, sample size, etc.);
  - characteristics of the respondent units;
  - condition of access to data (e.g. whether data access is in a safe setting or released to users);
  - different types of users;
  - behaviour and resources of the potential intruder.
- Definition of the legal and administrative framework.
- Definition of the dissemination strategy.
- Assessment of the risk for the specific data set.
- Release of data.

As far as disseminated data are concerned, the typologies of data available at present can be interpreted as logically derived from a single, large n-dimensional matrix that classifies elementary data derived from the survey as a function of n revealed variables. In this unified framework, the most analytic data (micro-data) are immediately obtainable from the original matrix (Biggeri and Zannella, 1991), while more synthetic data (tables or

other types of syntheses) are simply the marginal of the matrix obtained from the procedure of aggregation with respect to one or more variables. When one follows this framework, the risk of disclosure of a specific data set to be created can be evaluated in advance rather easily.

Effects of the technological advancements and changes

Referring to the previous framework, it is evident that the technological advancements are modifying the demand and the supply for statistical information and, as a consequence, also the situation of the factors that affect the decisions to be taken (Biggeri and Buzzigoli, 2001; Doyle et al., 2001).

Considering the *data users*, they have a quite marked diversification of needs and different levels of specialisation and autonomy in manipulation and management of the data. This leads to extremely varied requests for very differentiated output with different form of dissemination over varied types of medium.

As far as the local statistical offices and the research world, it easy to demonstrate the scientific importance for their access to micro-data; in fact, they need for original elementary data to produce detailed territorial information and to implement adequate researches on social and economic fields, but also to inform evidence based on policy and to monitor and evaluate the impact of socio-economic programmes (Trivellato, 2000). They also need for meta-data to replicate the studies and, usually, they have greater autonomy in data management.

In this context, the users have gained a more active interlocutory role with respect to official producers of statistics as it is revealed very clearly, for instance, by the increasing importance of the fact-finding surveys, etc. , aimed at evaluating the needs and the satisfaction of users (Doyle et al., 2001, chapter. 14).

From the point of view of the *data producers*, the technological advancements lead to new approaches and new methods of data collection, processing and dissemination and to a greater flexibility in the arrangements of those activities that should allow to satisfy the new user's exigencies. On the other side, in general, the official statistical offices are obliged to make more and best use of data produced with public resources in order to give a real contribution to the society. At this end, it is important to consider that the effective informational value of the disseminated data also includes the actual 'accessibility' to be understood not only and especially in a logical sense, as the ability to make use of an array of information and documentation that guarantee the usefulness of the wealth of information, is made available. Note also that the micro-data use is particularly challenging for the NSI's and can be of great help for them, because the usual series of checks and edits made by researchers in their data analysis can produce useful information for monitoring the data production process.

The previous discussion confirms that the technological changes are causing a corresponding evolution in the dissemination regime implemented by the NSIs.

The transition from a rigid and prearranged dissemination regime, in which the NSI's decide what will be carried outand in which form, to a more flexible regime that is more geared to the specific needs of its users, causes a corresponding evolution in the necessary procedures to test the confidentiality of data. This is due to the impossibility of carrying out ex-ante all the verifications necessary to guarantee that the disseminated data are sufficiently protected.

In any case it must be clear since now, that even if from one side the technology advancements increase the probability of a statistical disclosure, from the other side they help to reduce the disclosure risk.

3 Some open problems and challenges to assess the disclosure risk

The main critical issues for confidentiality still open and increasing are with respect to geographic referenced data, longitudinal data, administrative files and business data.

The large interest given to geographic information systems (GIS) presses the attention for those forms of output that allow for suitable integration of different informational sources.

GIS are becoming more and more popular as a tool to link administrative registers via spatial reference and, as a consequence, as a powerful mean to convey statistical information. On one side, geographical dimension is an essential information for analysing phenomena of different kinds (socio-economic, epidemiological, etc.), but, on the other side, it is well known that the geographical detail is maybe the most dangerous identification key for the variables in a survey. Therefore, small area statistics must necessarily be protected by means of appropriate disclosure control procedures. Recent results show that new micro-data protection methods, such as model-based methods, seem to outperform micro-aggregation for releasing geographically referenced data (Franconi *et al*. 2001).

Another challenge to disclosure control methods comes from longitudinal data. Longitudinal data deriving from panel surveys, which feed a wide range of research and behavioural modelling, usually imply a higher disclosure risk, especially when they have been linked to administrative records. An interesting debate on this subject has recently begun (Committee on National Statistics, 2000), which shows that longitudinal (as well as hierarchical) structures have not been adequately considered in the computing of disclosure risk. The definition of a per-record risk where the dependencies between variables or units are explicitly analysed could be a first step towards a more comprehensive approach (Benedetti and Franconi, 1998; Fienberg, 1997).

Another element to be considered is the use of administrative files as a statistical source. Sometimes the

availability of administrative records can be of invaluable importance for some research, but can seriously compromise data safety. In any case, it is necessary to make a conceptual and practical distinction between data which are compiled by linking administrative records and data obtained through censuses and surveys.
Finally, it is also important to point out that it is much more difficult case to give access to data with low risk of statistical disclosure when the units of collection are firms or business. Moreover, the identification of a firm could results in the release of sensitive information. Therefore, innovative methods to create anonymised micro-data files for business surveys is a potential very useful area of research (Franconi and Stander, 2001).
From a general point of view, the explosion of new technologies and electronic networks make people more aware of the uses of statistical data and of the possible abuses that can be made. A negative perception of the NSI's policy can have serious consequences on the collection of data and it is therefore useful to try to measure attitude towards confidentiality (Doyle *et al*, 2001).

4 Strategies for data dissemination and data access

Safe data vs safe settings
If different strategies of data storage and dissemination are possible, then NSI's need many different procedures to protect the confidentiality of data: the available alternatives must be evaluated in the light of the complex framework already presented.
The basic distinction among the various options available is between *safe data* and *safe settings*.
The traditional method used by the NSIs to protect the personal data is to limit or modify the informative content of the disseminated information with appropriate statistical procedures, called Statistical Disclosure Control (SDC methods): the data are partially cancelled or modified to avoid the spread of too highly detailed information (*safe data*). A vast literature is available on this argument (Willenborg and Waal, 1996).
The *Public Use Micro-data Files* (PUMF) are an example of micro-data files where identifying information has been removed to protect the confidentiality of the respondents. However, standard practices used for social public use files do not carry over to business micro-data, where the risk of identifiability of some units is particularly high: in this case it is advisable to apply the perturbative methods, consisting in modifying the informative content of the data, following adequate rules which guarantee the analytical validity of data with a minimum loss of information (Giessing and Hundepool, 2001). The model-based disclosure limitation methods seem to produce promising results (Franconi *et al*., 2001). Anyway, the effective utility of modified data depends on how well the perturbed data reproduce the characteristics of the original data.
Moreover, researchers are sometimes unsatisfied with these solutions, because they do not like to work with partly suppressed or manipulated data. These modifications make some specific studies difficult, because, for instance, small area characteristics cannot be analysed and model specification can be difficult.
In order to satisfy the research purposes, some NSIs (among them, the Italian National Statistical Institute, Istat) provide specific *Micro-data Files for Research* (MFR) with moderate level of acceptable risk, requesting to researchers to sign an agreement whereby they promise to not breach confidentiality, nor to try to link the data set with registers or other types of archives (Franconi and Seri, 2000).
However, to satisfy the needs of more sophisticated users an increasing number of NSI's give access to more detailed and - at the same time - less secure data, protecting them with very rigid protocols of use, involving technical, organisational and contractual aspects (*safe settings*).
In some cases specific locations are arranged (at the NSI's or at specific institutions, such as Universities) where *bona fide* researchers are provided with the necessary hardware, software and access to confidential data (e.g. in Statistics Netherlands, Statistics Canada, and Istat) sometimes under fellowship programs. In other cases statistical agencies issue licenses to researchers, who can analyse and elaborate restricted data in their own premises (Committee on National Statistics, 2000). In both cases severe penalties are provided for confidentiality violations.
Safe settings seem more adequate to research needs, but not all these kinds of solution are equally satisfactory. Bureaucratic protocols or organisational rigidities can reduce the effective usability of this kind of data release; moreover, restricted access modalities imply high costs (for the secure sites, for the personnel who have to help researchers as well as inspect their work, etc.).
The lively international debate shows that the concept of statistical disclosure control is sometimes substituted with that of 'statistical data protection' and considers not only the logical protection of the data, but also physical protection (the impossibility of violating the hardware and software protections of the data safeguard itself). This is becoming particularly important because new methods of data management (e.g. electronic data interchange, record matching techniques and data mining) are becoming more and more sophisticated and cause new threats for an adequate protection.

Internet, databases, and remote on-line analysis
Two fundamental aspects regarding the new opportunities in organising data access, are the role of Internet and the spread of database systems.
There is no doubt that as an instrument of data dissemination, the Internet presents clear advantages. First of all, it sets in motion unlimited possibilities for

dissemination, because it eliminates distances, and it facilitates a round-the-clock undifferentiated access. Secondly, it makes an enormous amount of very detailed information in time available and with low costs. Finally, it allows users to have an impressive degree of independence in managing their own access to data, even providing them, with data processing services.

For certain categories of data the Internet has an enormous dissemination potential, mostly thanks to the possibility of creating interactive geo-referential applications that allow users to access GIS applications at very low cost and with a minimum of computer expertise.

As significant examples of modern solutions in web based data dissemination with disclosure control see, for instance, the American Factfinder of the Census Bureau or the system adopted for the First National Agricultural Census of the People's Republic of China.

In any case, it is essential to evaluate the disclosure risk connected to the various dissemination protocols, in order to guarantee that the re-identification probability is sufficiently low.

However, management of dissemination over the Internet has not yet reached optimal levels, due to the presence of some important negative aspects.

In the first place, obviously, there are the technical problems: updating procedures are often critical, given the enormous number of web pages to check, and related to this, problems of managing the historical memory of the archives. An additional problem is to manage the *errata corrige*, which on the surface seems banal but is actually quite serious. While it is possible to correct errors quickly, it is not clear how to make the corrections visible to users. In addition, the advantages of this tool are effective only if as many steps as possible are automated in the chain of production of the statistical datum and if the connection with metadata allows an accurate interpretation of the disseminated data and an accurate disclosure protection of the disseminated data, both in term of limiting disclosure risk and in terms of access control.

A web-site is not a stationary product and the use of the Internet as the primary instrument of dissemination implies, therefore, a modernisation of working procedures within a NSI. It imposes a reorganisation and re-engineering of phases, protocols, and procedures with the goal of an integrated production of information oriented to the dissemination of the archives.

In database literature the problem of confidentiality becomes a problem of database security and means preventing illegal data access while maintaining the maximum data availability. For a brief review on the subject, refer to Brodsky *et al*. (2000).

The two aspects of remote access by web and database management are obviously connected and can interact to find an efficient solution to data access under strict confidentiality controls. Several other examples of Web-based dissemination systems which derive information from confidential databases can be found (Karr and Sanil, 2001; Fienberg, 2001), which should automate the application of disclosure control methods to statistical queries made by users to a web-based system.

The role of the intermediaries

The effective accessibility and usability of data is also boosted by the existence of auxiliary specialised structures (support) to satisfy the specific information needs of the researches, that integrate the dissemination activity of the producer institute.

We can distinguish the local branches of the NSI's from those other units which, instead, assume the role of actual intermediaries between statistical data producer(s) and user(s). The first one is the territorial support units both for operations of collection of surveys and for the dissemination of data produced. They assume a role particularly relevant in geographically larger states. The second one can be public or private entities that help the process of data dissemination or simply make them more visible, or add to their informative value by providing counselling activities and research.

Many of these structures turn to the academic world, or, more in general, to research institutions, for which they constitute preferential channels of access to data. The example of social science data archives is typical.

It is evident from the examples that supporting users is a full-time activity, not a secondary one.

The involvement of these intermediary figures has legal, financial and practical aspects. From a legal point of view it is necessary that the ownership of the data and the responsibility for its management be very clear; moreover, the set of legal penalties for misuse must be clearly defined. The financial aspect involves substantial decisions on rates policy and on the problem of finding financial support. The practical aspect calls for the predisposition of modern and efficient units often completely dedicated to the management of dissemination

The role of these intermediaries of dissemination seems, however, to assume particular significance in this society today, where, to complement an indiscriminate increase in the creation and exchange of information flows there often cannot be found an adequate culture of information in general, and particularly, that expressed as statistical data. The presence of a gap in information is addressed by the policy of dissemination to universities which should bridge it with their educational mission.

5 The need for a sound juridical framework: the solution adopted in Italy

The need for a juridical framework

Technical (i.e. organisational and methodological) solutions to data access for statistical and research purposes are not enough, if they cannot rely on sound

juridical basis, which should guarantee both the rights to data access and to privacy.
Without clear statements the judgements on the reasonable factors that affect the risk, and of the acceptable risk, of disclosure become too subjective. On the other hand, it is important to consider that small changes in the risk of disclosure can be translate into large changes in research and public benefit and, if the NSIs become risk-averse, many important researches will be hampered or prevented.
An articulated and complex legal architecture is consequently to be defined that operates at various levels (sovranational, national, etc.) and interacts with the laws concerning privacy. It should reflect the globalisation process which is modifying the socio-economic development cycles in our society and should be enough flexible to take into account the new technological advancements.
For instance, the legislation of European countries heavily relies on the conventions, regulations and recommendations developed by European institutions since 1981, of which the Rec. No.R(83) 10 and Rec. No.R(97) 18 are of greater interest to our field (for details see Biggeri and Buzzigoli, 2001).

The Italian juridical framework
The main rules for the protection of confidentiality in the Italian National Statistical System (Sistan) have been established for the first time in the decree law 1989/322 (Biggeri and Zannella 1999).
In 1996, the decree law n° 675/96 established a specific Authority for the Protection of Privacy and regulated very strictly all the treatment of personal data in order to safeguard the privacy of the citizens and juridical persons. This decree gave a definition of personal data ("it is considered any information on a personal identifiable or re-identifiable with any means") that is too narrow for statistical purpose preventing the dissemination of many statistical data. In fact, a contextual decree law (n° 676/96), published in the same date, asked for a specific regulation for personal data used for historic, research and statistical purposes.
The Italian Statistical Society, the Istat, the Committee for the guarantee of statistical information and many other scientific associations worked on the field in order to propose an adequate regulation. After a lot of meetings with the Authority for Privacy, it was decided to established an articulated legal architecture consisting in:

- a general decree law (referred to general rules for statistical data);
- specific codes of ethics and good conduct, for the different fields and subjects, to be subscribed by the interested bodies and scientific associations.

The general principles followed by the regulations
A specific decree law for the protection of personal data used for statistical and research purposes was promulgated in 1999 (n° 281/99). The preparation of codes is in progress: the first one has been promulgated one month ago.
Both them have been based on the mentioned recommendations developed by European institutions on: "the protection of personal data used for scientific research and statistics", n° R(83)10; and "the protection of personal data collected and processed for statistical purposes" n° R(97)/)18.
The general principles considered are the following:

(a) distinctive characteristic of the statistical purpose is the collective use of the micro-data;
(b) scientific research is considered equal to, and indistinguishable from, statistics for the use of micro-data;
(c) the notion of personal (or confidential) data is any information related to an identified or identifiable individual, but "an individual shall not be regarded as identifiable if the identification requires an unreasonable amount of time and manpower"; when an individual is not identifiable data are said to be "anonymous".

In term of disclosure risk, the definition recognises the principle of *reasonableness of the risk* (Franconi and Seri, 2000), accepting a low risk of identification in anonymous data sets (the provisions for data protection do not apply to anonymous data). This point of view is, in a sense, revolutionary because it allows to take into account the resources that an intruder can reasonably afford to employ for re-identification.
Also the following important principles have been considered:

(d) personal data collected and processed for statistical purposes shall serve only those purposes;
(e) personal data must be used only if they are strictly necessary (principle of parsimony);
(f) define a set of indications about measures to be taken to ensure the security of personal data and to disseminate only statistical results where the data subjects are no longer identifiable;
(g) statements for clear definitions of the rights and duties of official statisticians and researchers that must translate in a specific, personal responsibility of researchers who use confidential data.

The Codes of ethics and good conduct approved for the protection of personal data within the Sistan
The preparation of codes of ethics and good conduct is, in our opinion, a good choice for various reasons. Firstly, the codes leave enough degrees of flexibility, very useful for the future technology innovation. Secondly, the role of the statistical, scientific and professional bodies and associations is stressed both for the adoption of the codes and their functioning. Thirdly, the auto-discipline of the involved bodies and of their researchers is valorised.
In any case it is necessary to mention that some problems for 'sensitive' data (on health, sexual life, etc.) still exist

and for these data it is necessary to get the authorisation of the Authority for privacy.
At the moment various codes are in preparation, but the only code approved is the Code of ethics and good conduct for the use and dissemination of elementary data within the National Statistical System.
This code accepts all principles of the general decree law, before mentioned. Moreover it is important to mention that the code is in favour of the principle of the 'universal access to all elementary validated data by the statistical units within the Sistan', considering that those units are subject to the general rules of the protection of confidentiality. The code also includes:

- criteria of reasonableness of the risk of disclosure without reference to defined standard;
- criteria for assessing the risk of disclosure;
- criteria for the management, conservation and transfer of data within the Sistan;
- security for the data files and the rules of their conduct
- criteria for the use of sensitive data
- the different possible strategies for data dissemination and , above all, for the data access for bodies or researchers not belonging to Sistan.

Now there is the problem of the preparation of specific rules for the implementation of the code, but the objective to guarantee both the right to data access and the right to privacy of the individuals who provide the data, seems well achieved. Istat infrastructures should be able to do it (Franconi and Seri, 2000).

7. Concluding remarks

The work done in this field at national and international level seems satisfactory. But, it is necessary to verify the actual functioning of the established rules, at the light of the new changes in the society and in particular the technology advancements.
A broadly shared consensus towards codified standards could be a great help in ethics education and in favouring a positive public perception of research and research policy (Doyle *et al*., 2001).
Moreover, firstly, it is important to try to establish shared consensus towards codified standards in the field at international level. For the European countries very recently (May 2002) a new Commission Regulation (831/2002) concerning the access to confidential data for scientific purposes have been adopted. Secondly, we have to look for the best experiences and new solutions in order to solve the technical and organisational problems still existing for the protection of data and for the access to data. At this end, a CEIES (Eurostat) Seminar will be held in Lisbon (Portugal) the 26-27 September 2002 on "Innovative solutions in providing access to micro-data" (see the site: http://europa.eu.int/comm/eurostat).

References

Benedetti R. and L. Franconi (1998), *Applied Issues on Disclosure Avoidance Complex Micro-data Files*, Servizio Studi Metodologici, Istat, Roma.

Biggeri L and L. Buzzigoli (2001), Statistical Disclosure Control and Data Access for Research Purposes: Critical Issues and Possible Solutions, *Proceedings of the, 53rd ISI Session*, Seoul.

Biggeri L. and F. Zannella (1991), Release of Micro-data and Statistical Disclosure Control in the New National Statistical System of Ital, *Proceedings of the 48th ISI Session*, Cairo.

Brodsky A. and C. Farkas, D. Wijesekera and X.S. Wang (2000), *Contraints, Inference Channels and Secure Databases*, *Technical Report*, George Mason University.

Committee on National Statistics (2000), *Improving Access to and Confidentiality of Research Data.* National Academy Press, Washington, D.C..

Doyle *et al*. Editors (2001), *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies*, Elsevier Science.

Fienberg S.E. (1997), Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research, *Technical Report. 10/97*, Carnegie Mellon University.

Fienberg S.E. (2001), Statistical Perspectives on Confidentiality and Data Access in Public Health, *Statistics in Medicine*, 20:1347-1356.

Franconi L. and G. Seri (2000), *Micro-data Protection at the Italian National Statistical Institute*, Servizio Studi Metodologici, Istat, Roma.

Franconi L and J. Stander (2001), A Model-Based Disclosure Limitation Methods for Business Microdata, *Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Skopje.

Franconi L. et al. (2001), Experiences on Model-Based Disclosure Limitation, *Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Skopje.

Giessing S. and A. Hundepool (2001), The CASC Project: Integrating Best Practice Methods for Statistical Confidentiality, *NTTS & ETK 2001*, Crete, 18-22 June 2001.

Karr A.F. and A.P. Sanil (2001), Web Systems that Disseminate Information but Protect Confidential Data, *Proceedings of the, 53rd ISI Session*, Seoul.

Trivellato U. (2000), Data access versus privacy: an analytical user's perspective, CEIES Seminar 'Innovations in provision and production of statistics: the importance of new technologies', Eurostat.

Willenborg L., T. de Waal (1996) *Statistical Disclosure Control in Practice*, Lecture Notes in Statistics, 111, Springer-Verlag, New York.