COUNCIL OF EUROPE CONSEIL DE L'EUROPE

Strasbourg, 11 January 2008 T-PD(2008)01

**CONSULTATIVE COMMITTEE OF THE CONVENTION
FOR THE PROTECTION OF INDIVIDUALS
WITH REGARD TO AUTOMATIC PROCESSING OF PERSONAL DATA**

(T-PD)

24th meeting
13-14 March 2008
Strasbourg, G01

Application of Convention 108 to the profiling mechanism

Some ideas for the future work of the consultative committee (T-PD)

By Jean-Marc Dinant, Christophe Lazaro, Yves Poullet,
Nathalie Lefever , Antoinette Rouvroy

Final version

The experts signing this report
are here expressing their personal opinion
which does not necessarily reflect
that of the Council of Europe

Secretariat document prepared by
the Directorate General of Human Rights and Legal Affairs

INDEX

1. Attempt to define profiling

This report refers to profiling in the abstract based on identifying information, making predictions and, finally, inference.

Profiles are sometimes based on the collection and analysis of information about specific individuals, with no inference or prediction based on external sources. This second type of profiling is often referred to as "personal"[1] or "specific"[2] profiling.

In practice, abstract profiles are often partly based on specific/personal profiles and vice versa. We believe that specific profiling comes within the scope of Convention 108 and that specific or individual profiles constitute personal data in connection with which those concerned have the rights specified in that Convention.

Our work focuses on abstract profiling because certain crucial operations in the profiling process use data on non-identifiable and non-identified individuals as their raw material.

In any (abstract) profiling operation, therefore, three stages may be identified.

1. The first, that of "*observation*", is a stage in which personal or anonymous data are collated. If the data refer to an identifiable or identified individual, they will generally be anonymised during this stage. In the following we shall assume that the outcome of this first stage is an anonymous data set describing certain aspects of the personality of an unidentifiable individual. We shall call this stage **data warehousing**. The data may be of internal or external origin. For example, a bank might draw up an anonymous list of its customers who are bad payers, together with their characteristics. A marketing firm might acquire a list of the major supermarket chains' "shopping baskets" without the shoppers being identified.

2. This first stage is followed by a second set of operations which are carried out by statistical methods and whose purpose is to establish, with a certain margin of error, *correlations* between certain observable variables. For instance, a bank might establish a statistical link between a long stay abroad and one or more missed loan repayments. We shall call this stage **data mining**. The concrete outcome of this stage is a **mechanism** whereby individuals are categorised on the basis of some of their observable characteristics in order to infer, with a certain margin of error, others that are not observable.

3. The third and last stage, known as "*inference*", consists in applying the mechanism described above in order to be able to infer, on the basis of data relating to an identified or identifiable person, new data which are in fact those of the category to which he or she belongs. Very often, only this last operation is referred to as "profiling". We feel it is essential, however, to see this final stage as part of a process. Even if it is at this point that the effects of profiling can be felt in relation to a specific person, the profiling mechanism begins with

---

[1] Clarke Roger, Customer profiling and privacy implications for the finance industry, May 1997; Future of Identity in the Information Society (FIDIS) Deliverable 7.2: Descriptive analysis and inventory of profiling practices, no 2.3; 2.6 and 3.3

[2] Bygrave Lee Andrew, Data protection law: Approaching its rationale, logic and limits, 2002, p. 303

data warehousing, or even earlier, when the individual is observed with the use of information and communication technologies.

By way of illustration, we shall cite examples of profiling throughout this report:

1. The ATS (Automated Targeting System) which has been developed in the USA in order to evaluate the probability of a given individual being a terrorist.

2. Cable digital TV. This is a revolution which should provide programme distributors with precise information regarding channel selection and channel hopping by viewers who receive television channels via the telephone cable by means of DSL technology.

3. Profiling of taxpayers by governments in order to detect tax evaders. The system in the process of being deployed should allow governments to identify potential evaders so as to carry out better targeted checks and adapt the legal safeguards against such evasion. On a more positive note, the system can be used to inform the persons concerned of tax aids available.

4. The on-line advertising system put in place by Google. On most frequently visited sites there is a box called “Ad by Google”. What many consumers do not know is that the commercial links appearing in this box are generated by Google on a case-by-case basis and in real time on the basis of the referrer information supplied by the browser. Google can therefore monitor step by step each Internet user’s surfing through the pages of frequently visited sites (eBay, on-line newspapers, search engines, stock market sites, real estate sites, etc).

5. In a shop, tubes of lipstick are freely available for testing. Anyone can try out the lipsticks in front of a mirror provided for this purpose. Each tube of lipstick incorporates an RFID tag enabling a system equipped with a camera to film the customers’ lips and detect the characteristics of each lipstick used. The pictures are kept and analysed, but they show only the lips and the person filmed is completely unidentifiable.

1.1. History of profiling and recent developments

Historically, the term “profiling” first came to prominence in connection with the training of crime profilers in the USA. In theory, these people are supposed to be capable of determining a criminal’s personality type by analysing traces left at the scene of the crime. One of the most famous profilers was the psychiatrist James A Brussel who managed to work out a detailed profile of the so-called “mad bomber” in the 1950s, thanks to which the FBI was able to track down this criminal.

According to a French Wikipedia article, in our modern societies the term “profiling” is also used as a synonym of “behavioural analysis”.

*“**Behavioural analysis of an individual** consists in observing a person’s behaviour in response to stimuli (behaviourism), rather than investigating his/her thoughts, with the following aims:*

- *normally to help him/her, through counselling, but also through work on stimuli (eg reward/punishment), to avoid harmful behaviour,*

- *sometimes also to condition him/her to act in a way that does not match his/her interests,*

- *or to render him/her harmless (criminal profiling).*

This method is sometimes criticised as being a form of "training" rather than helping to resolve deep-seated psychological problems."

For our purposes, we shall define profiling as a computerised method involving data mining from data warehouses, which makes it possible, or should make it possible, to place individuals, with a certain degree of probability, and hence with a certain induced error rate, in a particular category in order to take individual decisions relating to them.

This concept of profiling differs from criminal profiling, where the aim is to get inside and understand the criminal's mind, but is similar to behavioural analysis since the aim is not to understand the motives which lead or might lead an individual to adopt a given behaviour, but to establish a strong mathematical correlation between certain characteristics that the individual shares with other "similar" individuals and a given behaviour which one wishes to predict or influence. As this approach does not depend on human intelligence, but on statistical analysis of masses of figures relating to observations converted to digital form, it can be practised by means of a computer with minimum human intervention.

1.2. Brief etymology

The word "profile" (*profil* in French) was originally used in the artistic field. It denoted the "outlines and features of a face seen from one side" (1621) or, more broadly, the "portrayal of an object seen from one side only"[3].

By extension, this term eventually took on the figurative sense of "all the characteristic features of a thing, a situation or a category of people"[4]. Where people are concerned, this term thus refers to "all the characteristic features exhibited by a person or category of persons" (1925). In this connection, reference can be made to the concept of "psychological profile" or, more specifically, to the "set of characteristics which a person must exhibit in order to hold or be recruited to a post" (1967)[5].

In the context of our report, attention should be focused on the original use of this term in the fine arts field. A profile is merely an image of a person based on different features. In the artist's profile, features are sketched; in profiling, data are correlated. In neither case can the profile be equated with the person him or herself. As we shall see, one of the most acute dangers of profiling is the fact that it tends to reduce the

---

[3] See *Le Trésor de la langue française informatisé* (TLFI), http://atilf.atilf.fr/dendien/scripts/tlfiv5/advanced.exe?8;s=3843709350.

[4] *Centre National de Ressources Textuelles et Lexicales* (CNRTL), http://www.cnrtl.fr/lexicographie/profil.

[5] E. CLAPARÈDE, *Arch. de psychol.*, vol. 19, p. 267 in QUEM. *DDL* vol. 29), quoted by *Centre National de Ressources Textuelles et Lexicales* (CNRTL), http://www.cnrtl.fr/etymologie/profil.

person to the profile generated by automated processes which are liable to be used as a basis for decision-making.

Overview of the issues (→ some ideas at random)

Leaving aside the preliminary stage of data collection and storage (pre-profiling), a definition of profiling can be proposed which hinges on two aspects: the creation of a profile and its application to a given situation. L A Bygrave proposes the following definition: "*Generally speaking, profiling is the process of inferring a set of characteristics (typically behavioural) about an individual person or collective entity and then treating that person/entity (or other persons/entities) in the light of these characteristics. As such, the profiling process has two main components : (i) profile generation – the process of inferring a profile ; (ii) profile application – the process of treating persons/entities in light of this profile*"[6]. He draws no distinction between data mining and data warehousing. We feel it is useful to differentiate between these two distinct stages.

As suggested above, three stages are therefore necessary in order to perform what we call profiling:

- large quantities of digitised data from **observation** of the behaviour and characteristics of individuals,

- determination of the probability relations (**correlations**) between certain behaviours/characteristics and other behaviours or characteristics,

- **inference**, based on certain behavioural variables or observable characteristics of an individual identified in general terms, of new characteristics or past, present or future behavioural variables.

Each of these three stages is detailed below.

2. Warehousing observation data

Businesses collect increasingly detailed information about the behaviour of their customers or staff and keep it sometimes for long periods. In practice, data are rarely destroyed, despite declarations made to the data protection authority regarding the storage period. In contrast to Convention 108, under which the keeping of data (which is a form of processing) is lawful only if an explicit and specified purpose can be shown, there is a temptation in practice not to destroy data "which may come in useful". All these data, typically financial transactions, geographical data, sale or purchase data, but also, where government departments are concerned, data relating to medical care and refunds, employment, marriage, private property, wealth, moveable and immoveable property, income, savings etc, can be stored in huge data warehouses taking three technically distinct forms:

---

[6] L. A. BYGRAVE, « Minding the machine : Article 15 of the EC Data Protection Directive and Automated Profiling », *Computer Law & Security Report*, 2001, vol. 17, pp. 17-24, available on-line at http://www.austlii.edu.au/au/journals/PLPR/2000/40.html.

1. Nominative data or data identified by a personal number (eg customer number, social security number, national number etc) or by a pseudonymous identifier specific to the data controller.

2. Coded data: a third party possessing a decoding “key” has access to the person’s identity.

3. Anonymous data: there is no identifier. When two data sets relating to the same person are examined, it is impossible, using state-of-the-art resources, to be reasonably certain that the two data sets concern the same person.

To illustrate this point, let us consider the example of a sales receipt issued in a department store. If the receipt shows a customer number or the number of a personal loyalty card, what we have here are personal data.

If the receipt does not show a customer number but the number of the bank account used for the purchases, we have a coded data item. A third party (in this instance the bank) is capable of identifying the person who made the purchases.

If the receipt gives only a list of the goods sold, the date and a fairly broad time frame, the third case applies.

It should also be noted that the data chosen may be from one or more sources. When profiling is performed in connection with the risk management of a ministry such as the Ministry of Finance, for example, the data are obtained from different sources within the department and from external sources (other ministries). The term “distinct sources” is used when referring to databases compiled for distinct purposes.

As regards the profiling performed by Google, this involves collection of data from all the visitors to the major on-line sites displaying an “Ad by Google”. Although its aim is to provide an advertising service on the Internet, Google can nevertheless monitor a particular individual’s clickstream, at least on the frequently visited sites where it is present via invisible hyperlinks. Generally, this enables it to measure and/or modify each individual’s exposure to advertising. For this purpose, Google processes the IP address which, under the mandatory rules on preservation of traffic data, enables the Internet access provider to identify the holder of the IP address at any given time. This is therefore a coded data item. Furthermore, if Google, like DoubleClick, places a persistent identifier cookie on the hard disk of the user’s terminal, that may be regarded as identification by a personal number, this globally unique number being assigned to a particular individual’s clickstream and potentially revealing his/her social, economic or cultural identity.

In the case of the lipstick, one might imagine that the data are purely anonymous since the system only films a person’s lips and, generally speaking, it is reasonably impossible to identify a person positively from a picture of his/her lips. However, it would be possible to teach the system, automatically or possibly with some human intervention, to distinguish a man from a woman or a person with white skin from a person with black skin. Since the monitoring system in question is coupled to an RFID tag reader and, as required by law, this RFID tag contains a single global identifier (serial number), it would be possible, when the lipstick tube equipped with its RFID tag is put through the check-out, to link a particular customer with a

particular video sequence and thus be able to infer the customer's ethnic origin, gender and even sexual preferences, if the lipstick tester is identified as a man.

In the case of digital television, the programme is technically transmitted by the Internet access provider. The decoder serves to receive the flow of IP packets from the telephone exchange via a DSL connection, to reassemble these packets and to encode them to supply a conventional video signal (eg PAL, SECAM, VGA, DVI etc). When a user switches channels, the decoder, at the user's request, transmits IP packets to the exchange via the DSL connection (which is bi-directional) in order to request the sending of another channel's IP packets. In traditional terrestrial broadcasting methods (aerial or satellite and dish), which are one-directional, broadcasters cannot know which programme is being watched. They cannot even know whether the television is switched on or not. In digital TV, service providers can know whether the decoder is switched on and what programme is selected. They can monitor and memorise each change of channel. They can thus create and keep a perfectly accurate viewing profile for each user. It therefore becomes technically possible to tailor advertisements to the user's profile. Notwithstanding the fact that the service provider knows its subscriber's identity (in the sense of such particulars as their surname, first name and address), the IP address of the television set constitutes personal data item. Indeed, the set's IP address enables it to be differentiated from all other television sets (see below the work of Pfeizmann). More generally, we feel that there can be no anonymity (and therefore no non-personal data) when a telecommunications device possesses a single address, since it is this single address which makes it possible to transmit individual information to a single device distinct from all other telecommunications devices.

One could imagine a digital television operator selling its viewer profiles identified by their IP address, but without any nominative data, to a third party, who could in turn purchase "anonymous" clickstream data from cybermarketers (such as DoubleClick or Google[7]). It would then be possible to intersect the behaviours of TV viewers and web surfers on the basis of their IP addresses and the date and time of connection in order to draw up a particularly precise profile of everything a family consults on the web or watches on TV.

In the context of data warehousing, the data chosen can be subjected to various anonymisation or coding operations, if appropriate by a third party, before they are put into the data warehouse, thus making them available for data mining.

3. Data mining

In this study devoted to profiling, we cannot dispense with a discussion of the developments in information technology which today make profiling activities increasingly easy and sophisticated, particularly by significantly improving the traditional approaches to analysis of data and statistics. These various techniques based on recent innovations in artificial intelligence (decision trees, rules of association, neural networks, score grids etc) have fostered the emergence of data mining.

[7] Google bought DoubleClick in May 2007.

3.1. Concepts

Data mining can be defined as "the application of statistical, data analysis and artificial intelligence techniques to the exploration and analysis with no preconceived ideas of (often large) computer data bases in order to extract fresh information that may be of use to the holder of these data"[8]. This concept therefore covers all the new techniques and methods used to exhaustively explore and bring to the surface complex relationships in very large data sets which may be from different sources and databases[9]. In other words, the value of data mining is that it is an IT tool which can "make the data talk".

Generally speaking, the methods on which data mining is based can be divided into two categories: some are *descriptive* and others *predictive*, depending on whether the aim is to explain or predict a "target" variable.

Descriptive methods are used to bring out information that is present but hidden within the mass of data, while predictive methods are used to "exploit a set of observed and documented events in order to try and predict the development of an activity by drawing projection curves. This method can be applied to management of customer relations in order to predict a customer's behaviour. The aim is for example to determine the profile of individuals with a high purchasing probability or to predict when a customer will become disloyal"[10].

3.2 Applications of data mining

The possibilities offered by data mining in terms of profiling are numerous and cover different areas of application.

3.2.1 Customer relationship management and marketing

Generally speaking, data mining is an extremely valuable tool in the area of marketing and customer management. It is one means of moving from mass marketing to genuinely personalised marketing. Data mining thus reflects the concern for personalisation seen in contemporary trends such as marketing one-to-one and customer relationship management[11]. The fact is that businesses want an increasingly detailed knowledge of their customers' habits, tastes and purchasing behaviour in order to personalise their offerings in the form of targeted solicitation.

---

[8] S. TUFFÉRY, *Data mining et statistique décisionnelle. L'intelligence dans les bases de données*, Technip, Paris, 2005, p. VII.

[9] B. MOXTON, « Defining Data Mining », DBMS Data Warehouse Supplement, 1996, quoted by S. OMARJEE, *Le data mining: aspects juridiques de l'intelligence artificielle au regard de la protection des données personnelles*, Université de Montpellier I, ERCIM, 2001-2002, available on-line at http//www.droit-ntic.com.

[10] H. BENALI, « Analyses décisionnelles et data mining », *SUPINFO Projects*, Ecole Supérieure d'Informatique, Paris, 20 August 2006, available on-line at http://www.supinfo-projects.com/fr/2006/decisionnel%5Fdatamining/.

[11] PH. LEMOINE, "Commerce électronique, marketing et liberté", in Groupe d'études Société d'information et vie privée (P. Tabatoni sous dir. de), *La protection de la vie privée dans la société d'information*, vol. II, 2000, available on-line at http://www.asmp.fr/travaux/gpw/internetvieprivee/rapport2/chapitr7.pdf.

With this in mind, they have therefore gradually expanded their commercial databases or have developed data warehouses. Effective exploitation of these databases calls for special tools to extract relevant information from them.

Where strategic marketing is concerned, the uses of data mining include the following: aid to the creation of packages and special offers; aid to the design of new products; customer loyalty policy; adjustment of marketing communication or the prices of products and services (in a "dynamic pricing" perspective) to each customer segment, etc. So, for example, large-scale distributors make use of data mining methods to manage large databases stemming from the development of store loyalty cards and supplied with behavioural information from sales receipts. Analysis of the product associations on sales receipts enables stores to establish a customer profile and, in so doing, improve product selection and adapt its merchandising strategy more subtly.

Customer relationship management stands out as another key application of data mining. The customer relationship covers a wide range of activities which can all benefit from data mining inputs. The following are some of the advantages[12]: identification of likely prospective customers; better response rate in marketing campaigns; personalisation of pages on the company's website according to each user's profile; choice of the best distribution channel or determination of the best locations for bank branches or chain stores; analysis of letters of complaint from customers[13], etc.

Researchers at MIT are working on the development of sophisticated software tools known as recommendations systems which are not limited to a single area of application and to certain specific data like those used on e-commerce sites such as Amazon or eBay. The aim pursued in this research is to go beyond data relating to the *user* viewed in the context of a single given application and to give priority to profiling of the whole *person*: "we must start modeling the person rather than the user"[14]. The technique, known as "social data mining", is designed to create richer profiles, particularly by analysing data collected on the web-based social networks that have recently sprung up on the Internet, such as Friendster, MySpace or Facebook. These popular digital forums are invaluable sources of information about individuals and the socio-cultural communities with which they identify. Indeed, as well as talking about their friends and acquaintances on these sites, individuals describe themselves and keep a detailed record of their activities, interests and hobbies (reading, music, television, shows, films, sport, food, etc)[15].

3.2.2 Risk management

It is in the field of *risk management*, which tends to take on a considerable importance in contemporary societies, that data mining is coming to be regarded as an indispensable tool. In this field it can be useful, inter alia, for determining the

---

[12] The examples given are partly taken from the book by S. TUFFÉRY, *Data mining et statistique décisionnelle. L'intelligence dans les bases de données*, Technip, Paris, 2005.

[13] This is based on analysis of textual data, forming part of what is known as text mining.

[14] H. LIU & P. MAES, "InterestMap: Harvesting Social Network Profiles for Recommendations", *Workshop: Beyond Personalization 2005 IUI'05*, 9 January, 2005, San Diego, available on-line at http://ambient.media.mit.edu/assets/_pubs/BP2005-hugo-interestmap.pdf.

[15] These "self-introductions" are to some extent akin to profiles pre-created by the individuals themselves.

characteristics of high-risk customers. Aims include: adjustment of insurance premiums; prevention of arrears; aid to payment decisions where current account overdrafts exceed the authorised limit in the banking sector; use of a risk "score" in order to offer individual customers the most appropriate loan or refuse a loan depending on his/her probability of honouring repayment deadlines and the terms of the contract, etc.

One unusual application of data mining has recently been developed in connection with judicial risk in England and Wales. The OASys (Offender Assessment System) project, as this application is known, is designed to carry out systematic evaluation of offenders and to define their profile in order, in particular, to assess the risk of re-offending[16]. The variables on which the scheme is based cover various offending related factors such as education and training, income management, accommodation, lifestyle and associates, analysis of offences, etc. Broadly speaking, the scheme is intended to provide practitioners and managers with a tool for improving the quality and consistency of their assessment practice, planning resources and tailoring individual interventions.

4. Application of profiling rules to a specific individual

This stage illustrates the distinction between profiling and statistical processing. The purpose of the latter is to generate results leading to the description or understanding of a situation or to the taking of abstract private or public decisions which, once taken, will have effects on people. In other words, what distinguishes statistical processing is that its purpose is to assist not individual but overall decision-making. Although profiling includes statistical operations, we consider that it differs from statistical processing insofar as its aim is to aid a decision modifying a course of action while ensuring that choices already made are applied automatically and more effectively. For example, when Google defines appropriate profiles according to the specific characteristics of a product or service which it wishes to advertise on a one-to-one basis, the idea is not to determine the state of the market or to take a strategic decision regarding the development of its own products (something which statistical operations can help with): once a decision has already been taken, profiling simply makes it possible to ensure that it is as effective as possible in its individual application by identifying the most appropriate criteria and the richest correlations.

Clearly, profiling allows for immediate application of its results. For example, when a person reacts in a particular way in front of his or her interactive television set and can be characterised by a particular choice of programmes, the correspondence observed in real time between that person's choices and profile X will permit the immediate display of a particular banner or commercial.

In other cases, profiling gives rise to subsequent individual application. In the example of the Ministry of Finance, it is perfectly conceivable that, before the tax

---

[16] R MOORE, "The Offender Assessment System (OASys) in England and Wales", *Probation in Europe,* Bulletin of the Conférence Permanente Européenne de la Probation, June 2006, pp 12-13, available on-line at http://www.cep-probation.org/bulletin/june06-E.pdf. Further aims of the scheme include identifying and classifying offending related needs, assessing the risk of serious harm, linking the assessment to the sentence plan, indicating the need for further specialist assessments, measuring change during the sentence, etc.

evader profile is applied to individual cases, there should be prior identification of citizens matching that profile.

Decisions based on automatic processing

Strangely, Convention 108 does not contain any provision prohibiting the taking of an administrative or private decision involving an assessment of human behaviour solely on the basis of automatic processing of information defining the profile or personality of the individual concerned.

In this respect it differs from Directive 95/46/EC, Article 15 of which deals explicitly with “automated individual decisions”. Article 15 undoubtedly constitutes an unusual provision in the body of the directive as it concerns a type of decision and not data processing as such. It is also unusual for being the only provision in the directive to deal with certain aspects of profiling activities[17].

In the context of this study, it is worth dwelling briefly on the background and limits to this provision. Article 15 of Directive 95/46/EC provides as follows:

1. Member States shall grant the right to every person not to be subject to a decision which produces legal effects concerning him or significantly affects him and which is based solely on automated processing of data intended to evaluate certain personal aspects relating to him, such as his performance at work, creditworthiness, reliability, conduct, etc.

2. Subject to the other articles of this Directive, Member States shall provide that a person may be subjected to a decision of the kind referred to in paragraph 1 if that decision:

(a) is taken in the course of the entering into or performance of a contract, provided the request for the entering into or the performance of the contract, lodged by the data subject, has been satisfied or that there are suitable measures to safeguard his legitimate interests, such as arrangements allowing him to put his point of view, or

(b) is authorized by a law which also lays down measures to safeguard the data subject's legitimate interests.

It needs to be specified here that in the particular case of an individual about whom a decision is taken solely on the basis of automated processing, the right of access to data includes "*knowledge of the logic involved in any automatic processing of data concerning him*" (Article 12 of Directive 95/46).

In his study on Article 15 of the Directive, L A Bygrave shows that this provision stems from several concerns on the part of European lawmakers[18].

---

[17] Account should also be taken of Article 12, a), third sub-paragraph, of the directive, which, with reference to the right of access, states that every data subject has the right to obtain from the controller “*knowledge of the logic involved in any automatic processing of data concerning him at least in the case of the automated decisions referred to in Article 15(1)*”.

[18] L. A. BYGRAVE, « Minding the machine : Article 15 of the EC Data Protection Directive and Automated Profiling », *Computer Law & Security Report*, 2001, vol. 17, pp. 17-24, available on-line at http://www.austlii.edu.au/au/journals/PLPR/2000/40.html.

The major concern is the growing automation of decision-making processes in respect of individuals. As emerges from the preparatory documents, European lawmakers came to worry about this automation because of the extent to which it reduces the role played by individuals in decision-making processes: "*This provision is designed to protect interest of the data subject in participating in the making of decisions which are of importance to him. The use of extensive data profiles of individuals by powerful public and private institution deprives the individual the capacity to influence decision-making processes within those institutions, should decisions be taken on the sole basis of his 'data shadow'*"[19].

Another concern relates to the fact that rampant automation of decision-making processes engenders virtually automatic acceptance of the validity and relevance of those decisions and, as a corollary, a lack of involvement and abdication of responsibility on the part of "human" decision-makers. In this connection the Commission notes that "*the results produced by the machine, using more and more sophisticated software, and even expert system, has an apparently objective and incontrovertible character to which a human decision-maker may attach too much weight, thus abdicating his own responsibilities*"[20].

Lastly, the preparatory documents also refer to the risk that the registered data images of persons (their "data shadows") might eventually usurp the constitutive authority of the physical self despite their relatively attenuated and often misleading nature. This concern, no doubt of a more general nature, refers to the risk of alienation and the threat to human dignity[21].

Despite being designed to cover risks in respect of automated individual decisions, Article 15 contains several ambiguities that could make it difficult to apply. Without undertaking an exhaustive analysis of the way in which this provision is applied, we may nevertheless note certain difficulties.

Article 15 concerns a decision "*significantly*" affecting the person concerned[22]. What is to be understood by "significant"? Does this term have any objective meaning independent of the person's own perceptions? Must the effect of the decision be purely material or financial? Must the decision necessarily have an effect contrary to the interests of the person concerned? This latter condition is no doubt necessary but not sufficient to consider a decision as having a significant effect. It would seem that the sending of a brochure to a list of people selected on the basis of automated processing cannot be considered as significantly affecting the person within the meaning of Article 15. However, other types of advertising used in cybermarketing seem more problematical, particularly where they involve unfair discrimination based

---

19 COM(90) 314 final – SYN 287, 13 September 1990, p. 29.

20 COM(92) 422 final – SYN 287, 15 October 1992, p. 26. See in this connection the edifying remark by S TUFFERY (*op cit,* p 1): "Data mining makes it possible to *limit human subjectivity* in decision-making processes and also, thanks to the increasing power of IT tools, to process large numbers of files increasingly rapidly".

21 L A BYGRAVE, *op cit,* p 4. See recital 2 of the directive: "Whereas data-processing systems are designed to serve man; whereas they must, whatever the nationality or residence of natural persons, respect their fundamental rights and freedoms, notably the right to privacy, and contribute to (…) the well-being of individuals".

22 Article 1 of recommendation no R (97) 18 refers more succintly to "*decisions or measures concerning a particular individual*".

on analysis of clickstream data (for example, a person visiting a website who is offered goods or services at a higher price than others, or a person who is refused the opportunity to purchase goods or services that are available to others).

Another difficulty relates to the nature of the processing referred to in Article 15. What is to be understood by "*automated processing of data intended to evaluate certain personal aspects*"? Is the purpose of such processing precisely to develop and use *profiles* to aid decision-making? It will be noted in this connection that in this provision the data to be processed are not described explicitly as "personal". This being so, the processing referred to in this provision could cover the development of a profile derived from data which are not necessarily and directly personal within the meaning of the relevant legislation, which is frequently the case in the various profiling activities to which we have already referred. That being said, the reference in this provision to "certain" personal aspects of an individual is also ambiguous. Does this mean that not all personal aspects are relevant for the application of this provision? But then where and how should the limit be set, despite the few examples mentioned in Article 15, such as performance at work, creditworthiness, reliability or conduct?

While at first sight Article 15 of the Directive seems to stand out as a valuable counterweight to the risks connected with the automated processing used in profiling activities, it should be noted however that its application is rendered difficult by the ambiguities of its wording. Added to this is the fact that its applicability depends on several cumulative conditions being met:

- a decision must have been taken;

- this decision must have legal effects in respect of a person or affect him/her significantly;

- this decision must have been taken solely on the basis of automated data processing;

- the data processed must be designed to evaluate certain personal aspects of the individual affected by the decision.

If any of these conditions is missing, the right enshrined in Article 15 will not be recognised.

5. Analysis of the Swiss law

The Swiss law is interesting in that to our knowledge it is the only national legislation to deal explicitly with profiling.

The law even sets out a definition of the "profiling" concept: Article 3 d defines a "personality profile" as *an assemblage of data facilitating appraisal of the essential features of an individual's personality.* This definition raises a number of questions as to its scope. A profile is an assemblage of data whose aim is to collect various data which are not necessarily correlated to the extent that assembling such data is statistically linked to a specific feature of the individual or of a group of individuals (eg a family, all the students in a given class, or the residents of a particular

neighbourhood) which enables the person producing or using the profile to act in respect of individuals or groups of individuals.

To quote a simple example, it might be ascertained that 80% of persons of male gender who pay with a credit card, do their shopping on Saturdays from 4 pm onwards and have dietary products and bottles of wine within such-and-such a price range in their shopping baskets are ideal targets for short breaks in luxury hotels on paradise islands, and that 85% of the professionals among them dodge direct taxation.

This example brings us to the second part of the definition: *the assemblage of data facilitates appraisal of the essential features of an individual's personality.*

The concept of "essential features" is problematical. Are the specific tastes in the area of travel arrangements of a group thus profiled or the tax evasion aspect to be deemed the essential features of the personalities of members of this group?

This "essential features" concept should no doubt be put into perspective and viewed from the angle of the operation which the person using the profile is intending to conduct. For instance, for an official in charge of checking compliance with tax regulations, the important thing is to access the various data assemblages characterising the tax evader, and similarly for tour operators the main thing is to find data enabling them to pinpoint the potential market for their products as reliably as possible.

The law in question gives the "profile"" datum a status similar to that of sensitive data, for example in Articles 12, 17 and 18.

The following section is an analysis of the three provisions in question:

-Article 13 mentions "justification" for a finding of infringement of personality. It states that in addition to the data subject's consent, infringement of personality can be justified by an overriding interest on the part of the data controller.

The provision points out that evaluation of an individual's credit is valid justification provided that it is not based on sensitive data and does not constitute a personality profile. This provision is surprising in that it apparently prohibits the creation of credit-rating systems based on profiling of the various customer types.

Articles 17, 17 a and 18 relate to public data processing. Article 17 requires processing of "profiles" to be strictly governed by a law in the formal sense, or to be permitted exceptionally where the data subject consents or "has made his/her data accessible to all and sundry", in which case the authorisation of the Federal Council is required on the basis of a finding of absence of threat to the rights of the data subjects in question, and lastly where the profiling is vital for the accomplishment of a task clearly defined by law. This latter case would preclude the possibility of the tax authorities using profiles to check on tax payments because this legal task can be conducted by other supervisory means than profiling. On the other hand, the law does provide for using modern methods of identifying persons suspected of terrorism for the purposes of airport security.

Article 17 a, which has been introduced since the adoption of the law on 24 March 2006, enables the Federal Council, on the basis of a prior opinion from the data protection authority, to authorise profiling under pilot experiments before the adoption of a law in the formal sense. One example might be trial runs of administrative activities to be covered by future legislation, using persons recruited on the basis of defined profiles.

Lastly, Article 18 states that where public data processing is concerned, the collection of vital data or personality profiles must be effected in a manner recognisable to the individual concerned. The provision accordingly stresses the obligation to inform the data controller when such data are involved.

6. Anonymous data

There is some controversy about the concept of personal data and anonymity on Internet. The Globally Unique Identifier concept has been widely used in this context, particularly for the drafting of Privacy Policies via P3P. For instance, Microsoft's privacy policy[23] specifies that "the Windows operating system generates a Globally Unique Identifier that is stored on your computer to uniquely identify it. The GUID does not contain any personal information and is not used to identify you…". DoubleClick, a Cybermarketing company recently bought up by Google, declares "*DoubleClick does not collect any personally identifiable information about you, such as your name, address, phone number or e-mail address*."[24].

So it would seem that the main players in the NICT industry do not accept such data as a number associated with a telecommunications terminal (eg a serial number associated with a software or hardware component (or both) or an item inserted by a specific operator (eg a persistent cookie) as personal data items. The same applies to the IP address, whose classification as a personal data item has been challenged despite the extremely clear position adopted by Group 29 on the basis of recital 26 of Directive 95/46[25][26].

In order to shed some light on this debate, we might usefully consider the position adopted by technicians on the functional concepts behind the concept of personal data, outlining the issues at stake. In this field we must refer to ISO Standard 15408, "Common Criteria for Information Technology Security Evaluation". These common criteria facilitate evaluation of security functions on the basis of eleven functional classes and guarantee requirements: security audit, communication, cryptographic support, user data protection, identification and authentication, management of security functions, privacy, protection of security functions, use of resources, access to components and trust channels. Each of these eleven functional classes is divided into 66 families, each comprising component criteria[27][28]. It is interesting to note that under this standard privacy is to some extent an aspect of NICT security.

---

[23] Seen on http://v4.windowsupdate.microsoft.com/fr/default.asp in May 2004.

[24] Seen on http://www.doubleclick.net/company_info/about_doubleclick/privacy

[25] http://www.cnil.fr/index.php?id=2244&news[uid]=484&cHash=f2a66a27ee

[26] See also on this subject the conclusions of Advocate General Juliane Kokott in the case C-275/06 *Productores de Música de España (Promusicae)* v. *Telefónica de España SAU* on line http://eur-lex.europa.eu/LexUriServ/.do?uri=CELEX:62006C0275:FR:HTML

[27] http://www.cases.public.lu/documentation/normalisation/ISO_15408/index.html

[28] http://www.commoncriteriaportal.org/

One vital contribution of ISO Standard 15408 is that the "privacy" section sets out four different levels of protection[29], which we might summarise as follows:

- Unobservability: at this level it is impossible to distinguish a human behaviour from the surrounding noise.

- Anonymity: this level is reached if, in a given context, a set of data on several individuals cannot be linked to one individual in particular.

- Pseudonymity: at this level, the individual may be identified by a third person in a given context. An individual may have one or more pseudonyms in a given context. This may also be referred to as coded data.

- Chainability: this level of protection enables a connection to be made within the data available on one individual, ie between two sets of data relating to him/her.

In the same context we should mention the work of Andreas Pfitzmann of Dresden University, Germany. The main value of the work being done by Pfitzmann and his team is the goal of devising terminology which might ultimately be approved by the W3C and used as a basis for the vocabulary of the technical regulations (standards) issued by the World Wide Web Consortium. A further aim is to create a community made up of consumers, representatives of the industry and experts mandated to assess the level of "privacy" or "privacidity" of the ambient technologies.

On the legal front, considering that one "essential requirement" of telecommunications terminals is *incorporation of measures to guarantee the protection of the personal data and privacy of users and subscribers* (in accordance with Article 3.3c of Directive 99/5), it is technically possible for the European Commission to use the option provided under Article 5.2[30] of Directive 99/5 to impose technical rules to protect privacy rather than personal data in the strict sense.

It should also be noted that like Directive 95/46, even as duly complemented by Directive 2002/58, Convention 108 cannot claim to be the sole defender of the right to respect for privacy[31]. Other international legal standards that have been transposed into the national legislations of member States are also geared to protecting privacy. One example is the Council of Europe Convention on Cybercrime, Articles 2 and 3 of which prohibit illegal access to or interception of computer data, whether or not they are of a personal nature. At the EU level, Article

---

[29] See on this subject Andreas Pfitzmann: Anonymity, Unobservability and Pseudonymity – a Proposal for Terminology, available on line at http://dud.inf.tu-dresden.de/Anon_Terminology.shtml

30 "Where a Member State or the Commission considers that conformity with a harmonised standard does not ensure compliance with the essential requirements referred to in Article 3 which the said standard is intended to cover, the Commission or the Member State concerned shall bring the matter before the committee", or Article 15 of Directive 2002/58 which states "2. Where provisions of this Directive can be implemented only by requiring specific technical features in electronic communications networks, Member States shall inform the Commission …3. Where required, measures may be adopted to ensure that terminal equipment is constructed in a way that is compatible with the right of users to protect and control the use of their personal data, in accordance with Directive 1999/5/EC …"

[31] We should remind the reader that the very title of Convention 108 indicates that it strives to protect individuals with regard to automatic processing of personal data, with Directive 95/46 setting out to protect individuals with regard to the processing of personal data.

8.4c of Directive 2002/21[32] requires national regulatory authorities to contribute to "*ensuring a high level of protection of personal data and privacy".*

Furthermore, it is interesting to note that email addresses have always been considered as personal data by both the industry and the data protection agencies. However, we cannot overlook the fact that many email addresses are anonymous (of the type toto234321@yahoo.com), thus merely allowing individuals to be contacted rather than identifying them.

In concluding this section on personal data,

- people's privacy must be protected, regardless of whether the data identifying them or facilitating their identification is actually used. This principle is in fact used in many other European Union and Council of Europe prescriptive texts. Observing and regularly recording the behaviour of anonymous individuals in any case constitutes invasion of their privacy;

- Council of Europe Convention 108 and the European Directives on data protection are not adequate legal instruments for protecting people's privacy in the NICT world;

- the technical standardisation work conducted by the ISO defines different levels of data security protection (unobservability, anonymity, pseudonymity and non-traceability) applicable to information on individuals, whether identifiable or not. Technically speaking, anonymity is impossible where any component of a set of data on an individual enables the latter to be connected with another set of data on the same individual. Anonymity presupposes non-traceability, or more concretely, the inability to conduct an automatic matching of two sets of data, or, even more precisely, the absence of an identifier, whether coded or not, that is reasonably persistent and is regularly present in the various sets of data concerning the same individual or his or her family;

- it cannot reasonably be contested that the third stage of the process of profiling an individual, which concerns an identifiable or identified person, is clearly an instance of personal data processing, even if this mechanism was originally designed on the basis of completely anonymous data.

7. "Statistical purposes" and "profiling

7.1 General comments

Recommendation (97) 18 relates to the collection and processing of personal data for statistical purposes[33]. The scope of the Recommendation covers not only strictly statistical activities but also those based on statistical procedures comprising

[32] Directive 2002/21 of the European Parliament and of the Council on a common regulatory framework for electronic telecommunications networks and services ("Framework Directive"), OJEC 24 April 2002, L108/33.

[33] Council of Europe Recommendation No. R (97) 18 concerning the protection of personal data collected and processed for statistical purposes, adopted by the Committee of Ministers on 30 September 1997.

personal data collecting and processing operations, such as opinion polls and market research[34].

In order to characterise operations involving data processing for statistical purposes, the Recommendation sets out the following definitions (Art. 1):

- "*"**For statistical purposes**" refers to any operation of collection and processing of personal data necessary for statistical surveys or for the production of statistical results. Such operations exclude any use of the information obtained for decisions or measures concerning a particular individual*".

- ***"Statistical results"** means information which has been obtained by processing personal data in order to characterise a collective phenomenon in a considered population.*

Broadly speaking, statistics is a scientific discipline and an activity geared to highlighting the collective characteristics of a specific population or group by extracting the essential facts from individual sets of information.

As the explanatory memorandum to the Recommendation puts it, "*statistics aim at analysing mass phenomena. Statistics allow, by means of a condensation process, the drawing of a general affirmation from a series of systematic individual observations. The results of this process often become available as information in figures, on a given phenomenon or a target population. In this way, although statistics are based on individual observations, their objective is not to acquire knowledge of the individuals as such, but to produce synthetic and representative information on the state of a population or of a mass phenomenon. Statistical activities can be distinguished from other activities, in particular by the fact that they are not directed at taking decisions or individual measures, but rather at gathering knowledge of large entities - such as economic cycles, the living conditions of a social group or the structure of a commercial market - as well as at the analysis of phenomena - such as epidemics, opinion trends, fertility or consumer behaviour of households - and therefore arriving at collective judgments or decisions*"[35].

his is how the "statistical purposes" mentioned in Article 1 of Recommendation No. R (97) 18 should be understood. So all statistical activities are geared to characterising a collective phenomenon[36]. While statistics are based on data relating to various individuals comprising the population or group under consideration, **the statistical result actually detaches the information from the individual.**

Nevertheless, if we were to confine ourselves to pure "statistical purposes" and to the fact that they produce only anonymised results, we would be missing the close link between this discipline and the issue of privacy and personal data protection, because statistics is inherently based on the possibility of collecting and processing a series of micro-data, some of which may be personal in nature.

Two types of risk can be pinpointed in the statistical field:

---

[34] See para. 61 of the explanatory memorandum to the Recommendation (hereafter "explanatory memorandum).

[35] Ibid., para. 2.

[36] Ibid., para. 8.

- Diversion from the original purpose: "*there is a risk, therefore, of the data in question being diverted from the purpose for which they were collected and thus being used for purposes relating to individuals. This may be the case, for example, when statistics come into contact with the administration and the police and where one could be tempted to use these data collected for statistical purposes for individual judgments and decisions*"[37];

- Cross-checking of data: "*moreover, in spite of their anonymous and aggregate nature, statistical results sometimes can be such that their analysis or recomposition enable the individuals upon whose personal data the results are based to be identified"*[38]. This is why principle 2.1 includes statistical results within the scope of the Recommendation.

In order to gauge the impact of statistical activities on privacy, therefore, we must consider the whole process of producing and disseminating statistical information.

In general terms, moreover, we should stress the fact that statistical knowledge is not an end in itself. It can be a means to a variety of ends which are usually broken down into three categories:

- general information;

- assistance for planning and decision-making;

- scientific purposes.

So regard must be had to the "mediate purposes" underlying this type of activity, while bearing in mind that the statistical information provided for such mediate purposes always relates to mass phenomena and cannot, therefore, entail direct or individualised consequences for individuals[39]. Therefore, where the purpose of assisting planning and decision-making processes is concerned, personal data collected and processed for statistical purposes can under no circumstances serve as the basis for individual decision-making. This concerns decisions of an administrative, judicial, fiscal or financial nature, as well as decisions by public authorities and measures affecting individuals in their workplace and in their community or corporate activities[40] (similarly to decisions on admission or exclusion, taxation, benefit, medical treatment, etc[41]).

The statistical activities covered by the Recommendation are therefore supposed to involve minimum interference in the affairs of the individuals providing the basic information; at the very most they may lead to decisions which are general in scope (laws, scales of assessment, vaccination campaigns, organisation of transport, design of blueprints, entry into production, etc), which, despite their favourable or unfavourable impact on specific individuals, do not involve any personalised effect on the individuals in question.

---

37 Ibid., para. 3.
38 Ibid., paras. 3 and 27d.
39 Ibid., para. 12.
40 Ibid., para. 68a.
41 Ibid., para. 13.

7.2 The principles

It is customary in the statistical field to differentiate between two modes of data collection, namely the primary collection, which is directly conducted among the individuals concerned, and the secondary one carried out among private or public bodies possessing data on individuals.

From the methodological angle, the primary collection process generally goes as follows: the collection is preceded by a phase designed to determine the scope of the collection, ie the set of individuals who are deemed relevant for the type of statistics planned. This phase is followed by the actual data collection procedure, which uses a variety of techniques, the most familiar being the questionnaire-based survey. The collection is then followed by a control phase geared to ensuring data quality and relevance. Lastly, statistical processing in the strict sense can then take place, producing the statistical results. It will, however, be noted that this methodological process is not always quite so linear. For instance, anomalies sometimes emerge in the statistical results, forcing those responsible to implement upstream controls, or else the collection process is staggered over time, so that some results are established before the process is completed.

However, statistical methods are not confined to processing data collected by means of surveys. Frequent recourse is had to so-called "secondary collection" procedures. This second method of acquiring data has a number of undeniable advantages in terms of time, cost and data reliability, the latter having, in theory, already been checked by the first data collector.

We shall now analyse the principles conducive to protecting personal data collected and processed for statistical purposes in the light of the above methodological framework.

7.3 Statistical purposes

As we have pointed out, although data are only used to produce statistical results, statistical activities can lead to certain risks affecting privacy and personal data. Furthermore, there are two main criteria for maintaining a certain level of protection: exploiting data exclusively for statistical purposes and producing impersonal results[42].

"Statistical purposes" are strictly defined in Recommendation No. R (97) 18, which requires data not to be used in a manner incompatible with the purpose of their collection[43]. Principle 4.1 of the Recommendation specifies that "*Personal data collected and processed for statistical purposes shall serve only those purposes. They shall not be used to take a decision or measure in respect of the data subject,*

---

42 Ibid., para. 27.

43 It will be noted that this provision echoes the definition of the expression "for statistical purposes" set out in Principle 1 of the Recommendation: ""For statistical purposes" refers to any operation of collection and processing of personal data necessary for statistical surveys or for the production of statistical results. Such operations exclude any use of the information obtained for decisions or measures concerning a particular individual" (my underlining). As if to stress its fundamental importance, the Recommendation thus mentions, on two separate occasions, the rule that statistical results cannot be used for taking decisions on a specific individual.

nor to supplement or correct files containing personal data which are processed for non-statistical purposes".

This principle is a specific transposition of Article 2 of Convention 108, which stipulates that data collected for a specific purpose must not be used for other purposes incompatible with the latter[44].

As the explanatory memorandum to the Recommendation points out, this provision has three consequences[45]:

1. where personal data have been collected and processed for statistical purposes, they must not in any event be used to take decisions or measures in respect of the data subjects;

2. data collected and processed for statistical purposes must not be used to supplement or correct data files used for non-statistical purposes. If the initial statistical purpose were to be distorted in this way, there would be no guarantee that the data concerned could not be used to take decisions or measures in respect of the data subjects;

3. controllers may possibly carry out multi-purpose collections (for statistical purposes on the one hand and non-statistical ones on the other). In case of multi-purpose collection, Principle 4.1 provides for separation of the processing operations in the light of their respective purposes.

Article 4.2 concerns the case of secondary collection. Specifically, it authorises processing for statistical purposes of personal data collected for non-statistical purposes, as such an operation is not incompatible with the purpose(s) for which the data were initially collected. On the other hand, processing for non-statistical purposes of data collected for statistical purposes is deemed completely unlawful.

7.4 Anonymisation

In order to prevent risks of infringement of privacy, Principle 3.3 of Recommendation No. R (97) 18 calls for the anonymisation of data: "*Personal data collected and processed for statistical purposes shall be made anonymous as soon as they are no longer necessary in an identifiable form*"[46].

The issue here is the risk inherent in holding identifiable data, especially "identification data".

At the time of collection "identification data" may be gathered. The "identification data" concept is defined in the Recommendation as covering "*those personal data that allow direct identification of the data subject, and which are needed for the collection, checking and matching of the data, but are not subsequently used for*

---

[44] In particular, Article 5 b) of the Convention provides that "*Personal data undergoing automatic processing shall be stored for specified and legitimate purposes and not used in a way incompatible with those purposes*".

[45] Explanatory memorandum , para. 68.

[46] Specific provisions in this field are set out in Principles 8.1 and 10.1 of the Recommendation.

drawing up statistical results"[47]. More specifically, these data identify individuals for the purposes of the collection and control (eg, date of birth or place of residence); they may be collected with a view to carrying out repeat surveys, supervising the work of the market research interviewers or checking doubtful data with the interviewees.

In this context, anonymisation is required as an elementary protective measure. We might note in passing that the Recommendation does not define this operation, although the explanatory memorandum sheds some light on the matter: "*anonymisation consists in erasing identification data so that individual data may no longer be attributed to the various data subjects by name*"[48]. More specifically, in the statistical field anonymisation is usually effected by withdrawing identification data as soon as they are no longer required for the processing operations.

In connection with identification data, Principle 10.1 of the Recommendation explicitly provides that "*when identification data are collected and processed for statistical purposes, they shall be separated and conserved separately from other personal data, unless it is manifestly unreasonable or impracticable to do so*"[49].

However, anonymisation is not an infallible protective measure because of the risk of re-identification of data which are anonymous in principle[50]. As stipulated in the explanatory memorandum, the data will nonetheless be deemed anonymous "*if identification requires an unreasonable amount of manpower, that is excessively complicated, lengthy and costly operations (see paragraph 28). Conditions for anonymity are relative, especially in relation to the technical means available for identifying data and taking away their anonymity. In this way, in view of the rapid progress in technological and methodological developments, the time and manpower required to identify a person, which would today be considered "unreasonable", might no longer be so in the future*". According to the explanatory memorandum, however, the present wording is sufficiently flexible to cover such developments[51].

7.5 Lawfulness

Recommendation No. R (97) 18 lays down conditions for lawfulness vis-à-vis the collection of personal data for statistical purposes. This is a specific application of the lawfulness principle enshrined in Article 5 a) of Convention 108 requiring personal data to be obtained fairly and lawfully.

---

47 See Principle 1, on definitions, of the Recommendation.

48 Explanatory memorandum, para. 53b.

49 Principle 11.1 indent 2 of the Recommendation on conservation of data should be read in parallel: "*In particular, identification data shall be destroyed or erased as soon as they are no longer necessary: for the collection, checking and matching of the data, to ensure the representativeness of the survey, or to repeat the survey with the same persons*".

50 Furthermore, even though statistical results cannot be considered as personal data, by means of cross-checking operations they can be used to recover, at least approximately, specific data on certain individuals and establish their identities.

51 Explanatory memorandum, para. 52d. See Principle 1 on definitions, of the Recommendation: "*"Personal data" means any information relating to an identified or identifiable individual ("data subject"). An individual shall not be regarded as "identifiable" if the identification requires an unreasonable amount of time and manpower. Where an individual is not identifiable, data are said to be anonymous*".

Principle 4.3 of the Recommendation envisages various hypotheses. Data collection and processing are considered lawful:

- in cases where collection and processing of personal data for statistical purposes are provided for by *law*. This hypothesis covers statistical activities conducted under a public-interest assignment involving compulsory information to be provided by citizens;

- in cases where the law so permits[52]:

  - if the data subject or his/her legal representative has given his/her *consent.* This hypothesis covers situations where the data subject is directly interviewed during the statistical survey;

  - if the data subject has been informed of the collection or processing of his/her data and has not *opposed* it. This latter hypothesis concerns special cases where the data subject is entitled to oppose data processing in the context of a secondary collection[53].

    We should note that in the case of this type of collection – which we have already identified as a valuable methodological instrument – the information cannot be provided to the data subjects in the same way and by the same means as during a primary collection effected directly from these subjects. As the explanatory memorandum puts it, "*it would be very expensive, and may even be impossible, to trace all the data subjects, who, in addition, might be surprised, or even needlessly worried, at receiving the information*"[54].

    In this connection, the compulsory information to be given by the controller must be guaranteed, under principle 5.4 of the Recommendation, by means of "suitable publicity". In the case of company statistics appropriate publicity may be notification either sent routinely to clients and suppliers or appearing on invoices and order forms. In the case of public statistics, suitable publicity may consist in official publication of the statute or other measure of internal law authorising secondary collection[55].

Principle 4.4 of the Recommendation sets out two additional conditions for the lawfulness of secondary collection of data initially obtained for non-statistical purposes. As Principle 4.4 explicitly states, the aim of this provision is to avoid collection of the same data again. Processing such data for statistical purposes will be considered lawful where that is necessary[56]:

- for the performance of a task carried out in the public interest or in the exercise of official authority;

---

[52] Collection and processing are considered "permitted" in the sense that there is no legal hindrance to them.

[53] For such cases the Recommendation lays down three specific conditions: the data subject must be appropriately informed, has not opposed the processing of his/her data, and the processing must not concern sensitive data.

[54] Explanatory memorandum, para. 81.

[55] Ibid., para. 81a.

[56] Under the same conditions, data collected for a statistical purpose can also be processed for other statistical purposes.

- for the purposes of the legitimate interests pursued by the controller except where such interests are overridden by the rights and fundamental freedoms of the data subject.

As the explanatory memorandum points out, "*such interests may be of various kinds: scientific (secondary collection may sometimes yield data of better quality than are obtained by direct collection), technical (secondary collection is simpler to organise), financial (the costs are lower), or courtesy to the data subject (the data subject is not caused the disturbance of having someone seek information he/she has supplied before). One or more of these considerations may prompt a decision to use secondary collection*"[57].

7.6 Proportionality

Here again, the Recommendation draws on one of the basic requirements of Convention 108: data collected must be adequate, relevant and not excessive in relation to the purposes established by the controller (Article 5c). In the statistical field, Principle 4.7 of the Recommendation provides that "*only those personal data shall be collected and processed which are necessary for the statistical purposes to be achieved*". Moreover, the Recommendation pays particular attention to so-called "identification data", which must only be collected and processed if this is necessary for example in order to conduct technical controls or create statistical files[58].

The concept of "data necessary for the statistical purposes to be achieved" must, however, be understood in the specific context of statistical activity. This involves considering the methodology used in organising, collecting and processing data for statistical purposes. From this angle, the Recommendation does not impose any criteria regarding relevance or necessity, because that would mean undermining the scientific independence of statisticians in defining and combining the variables required for characterising a collective phenomenon. Practice shows that the statistician's work primarily concerns creating a set of variables, only dealing with the data to be collected at a much later stage. But the fact is that in some surveys it might prove difficult to define exactly which data are needed to established the desired set of variables[59].

As the explanatory memorandum points out, "the drafters thus acknowledged that the statistical purpose of a collection could not always be defined in terms of an expected statistical outcome or a specifiable number of statistical results. They therefore agreed that the proportionality principle, while remaining a general reference criterion in the design of questionnaires and surveys, must be applied in such a way as firstly not to interfere with the statistician's scientific freedom to decide the methodological

---

[57] Explanatory memorandum, para. 72.

[58] See above, p. …

[59] Explanatory memorandum, para. 75b: "For instance, although it is easy to decide what data are necessary to establish school examination pass rates or produce statistics on household income and expenditure, it is much less so when it comes to statistics on, and indicators of, sex inequality or a social group's standard of living. In addition, some collection of personal data, such as population censuses, is not aimed solely at generating a predefined set of statistical results but also at amassing general-purpose data that can be used for a whole range of specific statistical purposes over a lengthy period (ten years, say)".

approach and secondly to make allowance for how specific the objective of the particular collection was"[60].

8. The purpose of profiling

As part of this study, it is important to identify and consider the different ways in which profiles are drawn up and used, since these do not all pose the same threat to privacy. Profiles may be used for widely differing purposes. We therefore propose the following typology[61]:

Use of the profile for:

a) establishing a general rule:
   i. in science
   ii. as a business and marketing resource
   iii. used by the organisation to take general organisational decisions
   iv. used by the organisation to take specific organisational decisions
b) application to an individual and practical case
   i. to decide whether or not to offer something to a group
      1. when the individuals are not yet really interested
      2. when the individuals want it
   ii. to decide whether or not to offer something to an individual
      1. as such
      2. with a reduction
      3. at a higher price
   iii. to decide whether or not to accede to an individual request
      1. which is not vital for the individual
      2. which is vital for the individual

The typology shows, for example, that the demarcation line between traditional statistics[62], data mining techniques and profiling is not always easy to identify.

The descriptive techniques of data mining seem to be closely related to traditional statistics. One example is the analysis of household shopping baskets in supermarkets, from which it is possible to determine which products are bought at the same time and organise shelf layout and promotions accordingly. Descriptive

---

[60] Ibid.

[61] Based on the reply of B.-J. KOOPS in M. HILDEBRANDT & S. GURWITZ (eds.), *Implications of profiling practices on democracy and rule of law*, Future of Identity in the Information Society (FIDIS), Report D7.4, 5 September 2005, pp. 66-67.

[62] However, the methodological basis of data mining differs from that of traditional statistics. The former is concerned with more extensive and less precise data that are often based on limited samples in which there are sometimes gaps, and that were initially collected for other purposes than statistical analysis. In such cases, the issue of data quality is particularly marked.

techniques also include automatic classification operations known in marketing as client segmentation.

Modelling based on data mining is generally compatible with the objectives authorised by Recommendation (97) 18, since it contributes to our understanding of groupings of persons, which can form the basis for judgments or decisions of a collective nature.

More specifically, traditional statistics are concerned with highlighting what are termed "aggregates", that is totals, averages, percentages and breakdowns, such as the identification of groups characterised by certain common situations or behaviour patterns, geographical location or collective professional classification[63].

However, while firms cannot be criticised for seeking to classify their customers and using relevant variables to guide their strategic and other business decisions, it needs to be recognised that segmentation techniques, among others, may enable them to draw up profiles, that is homogeneous categories of customers based on observed behaviour identified from a number of variables. So-called behavioural segmentation involves the use of information deduced from observation of behaviour to establish socio-economic, or even psychological, profiles of individuals, who will then be allocated to a particular "segment". For example, as part of their marketing strategies firms have developed methods and tools that enable them to segment their customers in order to select the ones most likely to be interested in the products and services on offer.

Segments as such cannot readily be equated with personal data[64], but their use may raise problems when they are associated with identified or indirectly identifiable persons and are an integral part of automatic data processing. The French CNIL considers that in so far as they derive from statistical processing, segments cannot be classified as basic information collected from the persons concerned and do not therefore, in themselves, constitute personal data. However, they do become personal data when they are associated with an identified or indirectly identifiable person and are an integral part of automatic processing[65].

---

[63] Explanatory report of Recommendation (97) 18, No 9.

[64] Similarly, statistical results as such are not personal data because they cannot be linked to identified or identifiable individuals. However, cross-tabulation may be used to identify individual data from statistical results and establish a link between the latter and the persons concerned. To ensure that such results do not lead to breaches of privacy, principle 14 of Recommendation (97) 18 makes their dissemination and publication subject to a number of requirements.

[65] CNIL - national data commission, report 93-032 on the inspection of the Dordogne regional agricultural bank carried out on 2 October 1992. The report found that in the case in question, the information must be consistent with, relevant to and not exceed the purposes for which it was recorded. In particular, the purpose for which the information was collected should be quite clear, and segmentation should not be based on information whose collection was prohibited or which bore no relation to the undertaking's activities, particularly concerning aspects of individuals' private lives that it was not entitled to know.

Another risk associated with segmentation is that it is based on a very large number of variables. In the banking sector, for example, such practices may be incompatible with the principles of legitimacy and proportionality enshrined in Convention 108. Examples include data on the number and size of transfers to competing credit institutions, the net balance of the household rather than that of the account holder, insurance policies with outside institutions, the number of credit card payments abroad and the level of a customer's life insurance[66].

The predictive techniques of data mining in connection with profiling require even more careful consideration. Particular attention should be paid to the process of scoring, which is one of the most widely used applications in sectors such as banking, insurance and telephone services.

Scoring is targeted more specifically at individuals. It may set out to identify:

- *risk*, such as the risk of a credit applicant defaulting;

- *partiality*, that is the likelihood that a customer will purchase a product, which is particularly useful for focusing mail shots on customers most likely to respond favourably;

- *attrition*, that is the loss or transfer of a customer to a competitor. By extension, this helps to assess undertakings' ability to retain customers and their loyalty. It helps them to define the customer defection rate.

The scoring system therefore makes it possible to identify and draw up customer typologies, by giving each of them a score, which is commercially available and can be used by businesses to decide whether to offer a product or service and the risk of default[67]. In the case of credit, for example, the system works as follows. Credit scoring packages apply specific weightings derived from statistical data and probabilities to certain personal information on applicants for credit[68], and those whose score exceeds a certain number of points will be granted a loan automatically. If, on the other hand, the computerised system gives a customer a low score, the bank is very likely to refuse the loan.

In such cases, if profiles are associated with identified or identifiable persons and are produced automatically, the process must be compatible with existing legislation on personal data protection.

8.1 Aims

Data mining techniques are distinguished by the fact that they permit alternative uses of data collected. In other words, data collected for one particular purpose may

---

66 CNIL, guidance note on the consequences for the amended "data and liberties legislation" of bank customer ratings (McDonough ratio - Basle II), 3 March 2005, p. 5.

67 *Ibidem*, p. 4.

68 The CNIL identifies three categories of personal information: "administrative" information on customers, such as age, socio-economic status and the length of custom with the bank, behavioural information, such as the number and type of banking products subscribed, average balance of the account over the last twelve months, regular income and financial capacity, and banking "events", such as any payments stopped.

subsequently be processed for other uses. For example, transaction information collected for the purposes of credit card payments may later be used in other ways, such as data mining operations. By its very nature, data mining always entails subsequent processing that can modify the original object of the exercise and thus pose a threat to privacy.
For some, one of the main difficulties in terms of data protection legislation is that the final objective of an effective data mining programme cannot be specified in advance, since the aim is to extract unknown information from a vast range of available data. As Ann Cavoukian has stated, "the data miner does not know, cannot know, at the outset, what personal data will be value or what relationships will emerge. Therefore, identifying a primary purpose at the beginning of the process, and then restricting one's use of the data to that purpose are the antithesis of a data mining exercise"[69].

8.2 Interconnection

The question of the purpose of data processing must be considered in conjunction with that of interconnection (and associated notions such as correlation and data matching).

- According to principle 4.6 of Recommendation (97) 18: "*Personal data or sets of personal data may be matched or interconnected for statistical purposes if domestic law offers appropriate safeguards to prevent their being processed and communicated for non-statistical purposes*."
- Section 2 of the Luxembourg personal data protection legislation of 13 August 2002 offers the following definition[70]:

  j. "interconnection": any form of processing that involves the correlation of data processed for a particular purpose with data processed for identical or linked purposes by one or more data controllers.

- Section 25 of the French Act 78-17 of 6 January 1978 on computerisation, files and liberties specifies certain activities that require the authorisation of the national data commission and are not covered by sections 26 and 27. These include

  5. automatic processing for the purpose of:

  - the interconnection of files of one or more legal persons managing a public service whose goals correspond to different public interests;

  - the interconnection of the files of other persons/bodies whose main purposes differ.

The CNIL appears to be moving towards a broad interpretation of interconnection[71], which it defines it as any automatic processing by one or more data controllers that consists of relating/correlating data with a particular purpose with other data with the same or another purpose. This process may involve the transfer of a file to

---

[69] A. CAVOUKIAN, *Data Mining : Staking a Claim on Your Privacy*, Information and Privacy Commissioner Ontario, January 1998, p. 13, available on line at http://www.ipc.on.ca/images/Resources/up-datamine.pdf

[70] Official journal, 13 August 2002, A – N° 91, p. 1836.

[71] http://www.cnil.fr/index.php?id=1735.

supplement or be merged with another file, or relating several files that are normally managed separately for a specific purpose, for example by establishing one of these files as a call file, which is used to interrogate the other files, the results of which are then used to supplement the call file. Information from several files may also be brought together in the same data base, such as a so-called data warehouse, to which various forms of software may then be applied for data mining purposes. Alternatively, a technical link may be established between several personal data bases, so that they can be consulted simultaneously. An example of the latter is portal sites, from which other bases can be accessed via hypertext links.

The commission has also adopted a broad interpretation of the interconnection of files belonging to different persons or bodies and whose main purpose also differs. This might include interconnections between the files of different private sector bodies, those of a private and a public sector body or ones produced by the same private sector body, but with differing main purposes.

8.3 Self-regulation

The question of self-regulation needs to be considered, for example from the standpoint of rules on:

- internal procedures, which firms need to draw up concerning the use of customer scoring systems;
- individual customers' score or segment, which should be not permanently accessible on their computerised file, to avoid any stigmatisation;
- the identification of segments, which should not include pejorative, unfavourable or subjective descriptions of groups of customers, such as liking for material goods.

9. Conclusions and recommendations

9.1 Is profiling a form of personal data processing?

The explanatory report to Convention 108 states that the object of this convention is to strengthen data protection, i.e. the legal protection of individuals with regard to automatic processing of personal information relating to them.

It therefore needs to be asked whether profiling constitutes a form of personal data processing. Traditionally, Council of Europe Convention 108 and European Directive 95/46 have offered protection against infringements of individual freedoms and privacy, but only in the case of inappropriate uses of personal data, that is data on identified or identifiable individuals. What distinguishes profiling is that it involves the processing of data that is anonymous, anonymised or coded in the first two stages or of personal data, in the strict sense of the term, to which the rules governing the profiling of identified or identified persons have been applied.

There must be complete agreement on what fully anonymised data signifies. We would refer here to Pfitzmann's functional definition[72]: **"Anonymity of a subject means that the subject is not identifiable[73] within a set of subjects, the anonymity set"[74].** In other words, data that contain an individual identifier are not anonymous if other personal data contain or might contain the same individual identifier.
The processing of data that are purely anonymous at the outset, such as pictures of unrecognisable individuals derived from video surveillance or a basket of household goods purchased by cash, fall outside the scope of Convention 108, but this does not mean that such operations are lawful and legitimate from the standpoint of other rules of law, in particular Article 8 of the ECHR.

This last point, which to those involved in data protection is self-evident, must be strongly emphasised. Both the general public and the information and communication technologies industry seem to believe that there are no rules or regulations governing the processing of the personal data of non-identified and non-identifiable individuals. This is certainly not the case, partly because an increasing number of rules of law, particularly European directives, use the dual term "privacy and protection of personal data" and partly because the right to respect for privacy remains a fundamental right embodied in numerous national constitutions and Article 8 of the ECHR. Individuals are entitled to a degree of privacy that is largely or totally incompatible with anonymous surveillance and the systematic observation of their tiniest deeds and actions, however routine.

9.2 Statistical purposes and the purposes of profiling

Even though it uses statistical methods, profiling based on current data warehousing and data mining techniques does not correspond to the processing for statistical purposes for which specific rules like those in Recommendation (97) 18 were drawn up.

In principle, data processing for statistical profiling purposes poses a greater threat to data protection, particularly as it may lead to discrimination since the aim is not simply to establish the facts or influence policy but to apply the findings, directly or indirectly, to individual situations. Profiling can help to make activities such as advertising, enforcing regulations or granting credit more effective.

The obvious implication is that the restrictions placed on statistical processing must, as far as possible, apply equally to processing for the purposes of profiling, while at the same time the latter should be subject to additional measures and certain rules to encourage statistical processing should, from the outset, be declared inapplicable to profiling.

---

[72] Anon Terminology, available on line at http://dud.inf.tu-dresden.de/Anon_Terminology.shtml

[73] "not identifiable within" means "not uniquely characterised within".

[74] From [ISO99]: "[Anonymity] ensures that a user may use a resource or service without disclosing the user's identity. The requirements for anonymity provide protection of the user identity. Anonymity is not intended to protect the subject identity. [...] Anonymity requires that other users or subjects are unable to determine the identity of a user bound to a subject or operation." Compared with this explanation, our definition is more general as it is not restricted to identifying users, but any subjects.

One example is the principle that "processing for statistical purposes of personal data collected for non-statistical purposes is not incompatible with the purpose(s) for which the data were initially collected if appropriate safeguards are provided for, in particular to prevent the use of data for supporting decisions or measures in respect of the data subject" (principle 4.2 of the Recommendation, which also appears in article 6 of Directive 95/46/EC).

If this condition is deemed not to have been met, it then becomes necessary to seek other traditional legal bases to justify data collection for profiling purposes. It would be difficult to argue that the secondary use of data for profiling was necessary "for the performance of a task carried out in the public interest or in the exercise of official authority", in accordance with principle 4.4 of the recommendation, with regard to statistical processing by public authorities. Such processing would probably require direct legal authorisation, in the material sense of the term. Moreover, can profiling in the private sector be justified simply as being necessary "for the purposes of the legitimate interests pursued by the controller", as also stated in this principle? Admittedly, this part of the recommendation then adds "except where such interests are overridden by the rights and fundamental freedoms of the data subject", but is this sufficient and should not processing for profiling purposes at least be subject to a procedure for establishing whether the controller's legitimate interests clearly outweigh those of the individuals concerned?

9.3 Profiling as a goal and the purposes of profiling

As noted earlier, profiling is not a goal in itself but a technical means of achieving a particular result. In practice, the aim is always (see the earlier taxonomy):

- to refuse to supply goods or services or to offer them at a different price – higher or lower;
- to contact individuals for marketing purposes;
- to take other specific decisions likely to have varying degrees of impact on individuals.

Clearly, profiling may well have much more significant consequences than simple statistical processing, since it may exclude individuals from access to employment (for example, the United States government appears always to check the ATS score of all potential federal employees), credit (based on a poor credit rating), housing and so on. There are other, less dramatic, examples, such as marketing and whether to carry out individual tax inspections. Individuals are thus affected to varying extents by the consequences of profiling, which may be aggravated by the associated decision-making process. In practice, profiling entails a certain error rate. This is deemed to be minimal, but certain individuals are nevertheless adversely affected by decisions taken in their case without any justification. This is particularly irksome in the case of individuals forced to deal with a machine, for example via the Internet, which is incapable of applying common sense. Following the example of Article 15 of Directive 95/46, we believe that individuals who are the subject of automatic profiling decisions should have a right of redress via a non-automated channel, particularly when these decisions affect the exercise of a fundamental right.

9.4 Lawfulness, transparency and proportionality

These three pillars of personal data processing must be applied at an early stage of any profiling operation, even if the human data being processed are not personal. We believe that the threat of unlawful and disproportionate processing lacking in transparency arises early on, at the warehousing and data mining stages. A risk prevention policy, based on the precautionary principle, must therefore be applied before profiles are applied to specific individuals. The security measures in Article 7 of the Convention may even be applicable to the collection of totally anonymised data. For example, we consider it good security practice to ban the warehousing of sensitive data and data mining for correlations between sensitive data such as race or sexual deviance and consumer profiles. Such data mining might end up by identifying logical correlations between particular baskets of goods and particular deviances or races. It would then be possible to infer or deduce the values of these two variables from particular patterns of purchases, which in principle should be unlawful.

9.5 The need for special protection against profiling operations

We have drawn attention to the particular arrangements specified in articles 12 to 15 of Directive 95/46, which offer individuals legal protection against decisions that have legal effects concerning or significantly affecting them and that are taken solely on the basis of automated processing designed to assess certain aspects of their personality.

The directive also allows individuals to understand the "logic" that underlies the automated processing of data, at least in the case cited above.

Convention 108 does not specify any particular arrangements to provide information on, access to or the right to challenge automated processing for the purposes of assessing certain aspects of individuals' personality.

Technically, such protective arrangements could be added to Convention 108 in the form of an additional protocol.

Such protection would be justified by the particular nature of the fresh and increasing risks attached to the exponential growth of profiling. These risks are linked to the three factors that make data mining operations increasingly powerful: the growing number of types of data stored, with the automatic and systematic storage of routine operations, the increasing scale of data warehousing and the opportunities for interconnection, and the growing sophistication of data mining techniques, with their associated computing power, opacity and complexity.

An increasing number of commercial undertakings have used or will use such techniques to grant or refuse access to their goods and services or to adjust their prices.

More generally, individuals who are caught up in profiling operations of whose complexity, logic or even existence they are unaware are in danger of finding themselves in a Kafkaesque situation. They then become quite incapable of

exercising any control over their computer image or even understanding by what means this image has been created.

Without such protective arrangements there is a significant risk that commercial undertakings will make increasing and systematic use of rapid and inexpensive profiling of their customers. Such profiling will inevitably result in certain individuals being excluded from particular goods or services or having to pay a higher price for them. Individuals therefore face the risk of a prediction of their future behaviour that is nothing to do with them and is no more than a forecast based on the previous behaviour of other individuals whom they do not know.

We believe that in the current and reasonably predictable economic and technological context the additional protocol must go further than the protection offered in Directive 95/46, and in particular that the protection should not be limited to decisions that have legal effects that concern or significantly affect individuals. The simple fact of individuals' being subjected to automated profiling to assess certain aspects of their personality should be sufficient by itself to entitle them to be informed of this profiling and of its underlying logic, and to challenge it, at least in certain cases of automated processing deemed to be capable of making such assessments.

9.6 ***Proposed recommendation***

In conclusion, we believe that the Council of Europe should prepare a recommendation setting out rules for profiling activities. There are several reasons for such a recommendation:

a) The Council has already produced an extremely judicious recommendation on statistics. We consider that in terms of its purpose, or technical approach, profiling is quite distinct from statistics, notwithstanding any confusion between the two in the minds of the general public. The increasing use of profiling could have serious consequences for a growing number of individuals by assigning them profiles that do not necessarily correspond to their particular circumstances or impinge excessively on their privacy. Hence the particular need for a recommendation on profiling.

b) Profiling has expanded enormously in recent years. The new information and communication technologies have almost completely overcome the problems of data overload, cost, speed and reliability that characterised them some ten years ago. At the dawn of the 21st century, most day-to-day and banal human activities – buying and selling, undertaking searches, moving around, reading newspapers, consulting books, sending and receiving mail, changing television channels or testing a tube of lipstick in a shop - can be observed and stored more easily, rapidly, cheaply and invisibly than ever before;

c) There is currently considerable confusion as to what constitutes personal data and many of those concerned may well consider Convention 108 and European directives 95/46 and 2002/58 as the only legal rules protecting individual privacy. In cases which fell outside the material scope of these instruments, because the individualised data concerned did not constitute personal data, certain data controllers might be led to conclude – wrongly – that processing such data was fully compatible with respect for privacy, as

defined in Article 8 of the European Convention on Human Rights. A recommendation on profiling could serve to draw their attention to the requirements of dignity and privacy underlying Article 8 of the ECHR, which of course apply to all individuals, whether or not they are identifiable.