BILETA
British & Irish Law, Education and Technology Association
2007 Annual Conference
Hertfordshire 16-17 April
UH

Data protection and database theory - Applying database design principles to personal data identification

Dr Boštjan Berčič
Institute for Legal Informatics,
Celovška 136, SI-1000 Ljubljana, Slovenia
Email: bostjan.bercic@ipri-zavod.si

Dr Carlisle George
School of Computing Science, Middlesex University,
The Burroughs, London, NW4 4BT, United Kingdom
Email: c.george@mdx.ac.uk

Abstract

The European Union (EU) directive on personal data and resulting data protection legislation of EU member states, require from data controllers, a notification of their activities to the appropriate supervisory authority. Included in this notification is also a description of the data or categories of data which are processed. Legislation in some EU member states (e.g. Slovenia) require that not only a description but also a concrete list of personal data attributes need to be included in this notification. In such cases, it is sometimes difficult to ascertain *in concreto* whether some collected attribute represents personal data (and should therefore be included in the list of attributes) or whether it is a non-personal attribute. Similarly, under the EU directive data, subjects have various rights, including the right to access their data, and data controllers are sometimes faced with the problem of determining whether various data items constitute personal data. Further, the impending case in the European Court of Human Rights, arising out of the decision of the UK case of *Durant v Financial Services Authority* (which narrowed the scope of personal data) has added some uncertainty as to the interpretation of the EU directive. In view of the legal uncertainty regarding what constitutes personal data, this paper examines whether relational database design principles can be applied to identifying personal data. Using this novel approach, the paper explores various parallels between personal data identification and principles of relational database design. The paper thus makes a novel contribution to the ongoing uncertainty in data protection law. The paper also discusses the wider issue of applying computing/scientific principles to interpreting the law, and comments on the success of the approach taken.

1. Introduction

The requirements needed for compliance with the EU data protection regime, mean that data controllers need to have a high degree of certainty regarding what constitutes personal data. The EU directive on data protection[1], and resulting data protection legislation of EU member states, require from data controllers, a notification of their activities to the appropriate supervisory authority[2]. Included in this notification is also a description of the data or categories of data which are processed[3]. Legislation in some EU member states (e.g. Slovenia) require that not only a description but also a concrete list of personal data attributes need to be included in this notification. In such cases, it is sometimes difficult to

[1] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data
[2] For example The UK Information Commissioner's Office
[3] Article 19 (1)(c) of Directive 95/46/EC

ascertain *in concreto* whether some collected attribute represents personal data (and should therefore be included in the list of attributes) or whether it is a non-personal attribute.

In the UK, the case of *Durant v Financial Services Authority (2003 EWCA Civ 1746)*, the Court of Appeal issued a landmark ruling narrowing the interpretation of what makes data 'personal' (within the meaning of personal data under the EU directive and UK Data Protection Act 1998). The Court ruled that personal data is information which: "*is biographical in a significant sense; has to have the individual as its focus; and has to affect an individual's privacy whether in his personal family life, business or professional activity"*. This ruling in effect, also narrows the right of subject assess under the EU directive and UK Data Protection Act 1998. The case is therefore currently being taken before the European Court of Human Rights as a breach of Article Eight of the European Convention of Human Rights[4]. The European Court's decision, however, will take several years, but in the meantime the uncertainty about 'personal data' remains.

In view of the legal uncertainty regarding what constitutes personal data, this paper examines whether relational database design principles can be applied to personal data identification. Using this novel approach, the paper explores various parallels between personal data identification and principles of database design. With the aid of examples, the paper discusses (and demonstrates) how knowledge of relational database design principles can greatly help to understand what is and what is not personal data. Consequently, data controllers can make use of this in their notification processes as required under EU data protection legislation. Data controllers can also use this approach to determine 'personal data' when complying with the right of subject access (subject to other provisions under the Directive such as the rights of third parties). The paper thus makes a novel contribution to the ongoing uncertainty in data protection law. Finally, the paper also discusses the wider issue of applying computing/scientific principles to interpreting the law, and comments on the success of the approach taken.

2. Definition of personal data

Following the definition given by Article 2 of the Directive 95/46/EC, "personal data" means:
" *any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity."*[5] The means of identifying should not cause excessive costs, effort and should not take much time.

The given definition is so broad, that when used, almost any information can qualify as personal data (the criteria are met if it applies to a concrete individual, for example: the mere fact that an individual is wearing a red shirt can constitute an item of personal data). On the other hand, this definition is semantically also rather vague. Even if we accept the fact that content-wise every item of information can be considered personal data provided it can be related to an individual, the directive's definition is still rather vague structurally (since it is not always clear what kind of internal structure every record has to have to be considered personal data). In relational database theory, a record structurally consists of two parts[6]: (i) a unique identifier (primary key) of the record viz. entity under consideration and (ii) one or several items of data related to it. The Directive's definition does not define personal data in this way, hence it is unclear, for example, whether a unique identifier of a person (such as the UK National

[4] Article Eight of the European Convention states that everyone has the right to respect to his private and family life, his home and his correspondence.
[5] Directive 95/46/EC can be found at: https://www.agpd.es/upload/directiva_95-46_ingles_pdf.pdf . Note that definition of personal data under the UK Data Protection Act 1998 is wider than the definition under the directive, This paper is based on the definition in the Directive since it will be too onerous to consider national legislation in each EU state.
[6] See: T. Connolly & C. Begg (2004), Database *Solutions: A step-by-step guide to building databases*, 2nd Edition, Pearson Education Ltd, UK, pp 21-35.

Insurance Number[7] referred to as NINO, or US Social Security Number[8] refereed to as SSN) already constitutes personal data, whether only items of data related to this unique identifier would be considered personal data (for example the fact that someone lives on Oxford street) or whether only a record that meets both criteria, inclusion of the unique identifier and data related to it would be considered personal data (for example the NINO of a person plus the fact that this person lives on Oxford street).

The question of what constitutes personal data is not as trivial as it seems, for another very important question hinges on it, the question of what constitutes a personal data filing system (which, following the Directive's definition, is: "*any structured set of personal data which are accessible according to specific criteria, whether centralized, decentralized or dispersed on a functional or geographical basis*."), and on this, cascading, another very important question, that of what constitutes processing of personal data (which, following the Directive's definition, is: "*any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction*."). The question of what constitutes processing of personal data is finally relevant when considering whether the provisions of the Directive are applicable to a case at hand (In article 1, the Directive states that: "*In accordance with this Directive, Member States shall protect the fundamental rights and freedoms of natural persons, and in particular their right to privacy, with respect to the processing of personal data.")*. So this structurally rather vague definition of personal data in the end determines whether the Directive will be used in a particular case.

This paper will attempt to more appropriately define the concept of personal data, applying concepts in relational database theory (viz. information systems theory). The paper argues that by using concepts from relational database theory, we can ascertain *in concreto* when a specific record contains (in the structural sense) personal data and when it does not.

3. Identifier and data related to it

In relational database theory a record consists of the record identifier (primary key) and data related to it[9]. The identifier is usually unique or full, which means that an individual is identified uniquely (e.g. name and surname, often only together with the added information of residence, which uniquely identify an individual; or the National Insurance Number - NINO). An identifier, that does not define an individual uniquely (for example, if only the name and surname are present, this can represent a rather large group of individuals with the same names and surnames) is not a real identifier, because it denotes a multitude of individuals and can therefore (based on the relational database theory) cannot be used as a real identifier. When dealing with questions of whether a specific record should be considered personal data, one often has to deal with cases where an individual is only partially identified, an important concept which for purposes of this paper will be called a "partial identifier".

The identifier can be explicit (name, surname, and residence if needed) or implicit (NINO or SSN). When using an explicit identifier, it is obvious which individual it relates to. To get to an individual's identity when using an implicit identifier one has to obtain additional information in a corresponding registry, which contains both the implicit identifier and the explicit one (for example, a National Identity Register linking national identity numbers with names and surnames). From the point of view of relational database theory there is not much difference between an explicit and implicit identifier, because they both identify an individual equally precisely and uniformly. On the other hand, from the point of view of human processors of information, it is only the explicit identifier that converts

[7] A national insurance number (NINO) is a number unique to an individual which is used to keep track of national insurance contributions and state benefits. The format of the NINO is two letters, six digits, and one optional letter, e.g. AB123456C.
[8] The Social Security Number (SSN) is a nine digit number which serves as a unique identifier for individuals within the United States. The number is divided into three parts and has the format "111-11-1111." The first three digits represent an area number (geographical region), the middle two are the group number and the last four are serial numbers.
[9] Connolly & Begg, *op.cit.*

information concerning abstract, unidentified individuals into data concerning concrete individuals. As far as human processors of information are concerned, implicit identifiers have a far lower information content, because humans are principally not good in processing data without intrinsic semantic content. In the information age, humans are just one kind of information processors with automatic processing of data just as, or even more important than, manual processing. Despite that, it is wise to uphold the distinction between explicit and implicit identifiers, because the end user of information is ultimately a person and for this person to understand any information, it ultimately has to be conveyed by way of an explicit identifier.

4. Explicit and implicit, full and partial identifier

There are four kinds of identifiers that can be used in real life. In Table 1, examples of all types of identifiers are shown across two dimensions (implicit/explicit and full/partial)

Table 1 Types of identifiers

Identifier	**Implicit**	**Explicit**
Full	Social Security Number (SSN)	Name, Surname, Residence
Partial	First few digits of SSN	Name, Surname

A full identifier (explicit or implicit) defines an individual uniquely. An example of an implicit full identifier is Social Security Number (also UK National Insurance Number). An example of an explicit full identifier is Name, surname and date of birth taken together (i.e. presuming that name and surname alone do not suffice to identify an individual uniquely, while in combination with date of birth they do). There is usually more than one implicit full identifier for an individual (for example in addition to NINO a person has a National Health Service Number). There are also many possible explicit full identifiers (for example, name, surname and residence, or, if this does not suffice to identify an individual, name, surname, residence and date of birth). A full identifier can also be over determined; for example, if a name, surname, and date of birth are sufficient to uniquely identify an individual, then a name, surname, date of birth and residence is still a unique identifier, but it is no longer minimal (cf. the notion of minimal keys in relational database theory).

A partial identifier, however, does not identify an individual uniquely. For unique identification, a partial identifier has to be supplemented with other (full or partial) identifiers, so that it becomes a full identifier. A partial identifier can also be explicit or implicit. The first three digits of SSN, which uniquely define the geographical region, but not an individual person is an example of the partial implicit identifier. A partial identifier can be changed to a full identifier, for example if one adds other ciphers of SSN. A name and a surname, which is an explicit partial identifier, for example, changes to a full identifier if one adds to it date of birth or residence or both.

In order for the legal system to be predictable, it is important that there are rules to decide whether a record with an identifier similar to one of the four categories discussed above should be considered ‘personal data’. In the following paragraphs, records with identifiers similar to the four categories are discussed in the order of likeness that they should be considered personal data.

5. Records with explicit full identifier

Explicit full identifiers (for example name, surname, date of birth) most certainly define an individual. Everything communicated about this individual (for example where he lives, who he is married to, what kind of a car he drives) can most certainly be considered personal data.

6. Records with implicit full identifier

An implicit full identifier (SSN for example) does identify an individual uniquely, but the identification is aggravated nonetheless. Despite the fact, that the record relates to a uniquely identified individual, one still needs, in order to gain full identification, to use an appropriate register (for example a central

register of citizens) which links explicit identifiers with individuals. At this point, context comes into play: Something that is personal data for somebody is not necessarily personal data for somebody else. Whether something can be considered personal data or not depends in this case on the data processor's access to such a register, because he can only identify an individual through such means. There are many registers that link individuals' implicit identifiers with their explicit identifiers and they are not limited to the public sector. Employers in the private sector for example process social security numbers of their employees in accordance with the labour legislation requirements.

The issue at hand is whether records identified with full but implicit identifiers, such that one needs to obtain additional data in order to fully identify an individual and data possibly being kept in non-publicly accessible registers, should be deemed to fall in the category of personal data according to the Directive.

This paper argues that they should. There is a multitude of reasons for that. The most pragmatic one is that processing of such data (for example publishing in the media) is still processing of data which are undisputedly personal (identified with explicit identifier, for example, name and surname) at least for a certain group of processors of data; those that have access to registers that identify individuals[10]. Because it is impossible to exclude in advance when processing such data the possibility that the data would reach a processor who could link implicit identifiers with names and surnames, it makes sense to consider all such data personal data and protect them in accordance with the Directive[11]. Besides this pragmatic reason it is easy to come up with other types of arguments. Even if data identified with an explicit identifier comes into possession of individuals who themselves don't have access to register needed for identification, they can still pass them on to those that have such access. One simply cannot ignore the broader social danger of identifying an individual down the road in one of the next steps of data processing, possibly even in extended form. That is another reason why data processing on the basis of the same unique identifier (e.g. NINO or SSN), should be seen as processing of personal data. Finally, there is a rather more metaphysical argument for considering all data identified with implicit identifiers as personal data. As explained above, the difference between explicit and implicit identifier, especially if they are full identifiers, is more or less arbitrary. It is relevant if data are being processed by humans, who process mental images of other individuals usually with respect to their names and surnames, never with respect to their SSNs. Such processing is nevertheless only one type of processing of personal data, the other one being automatic, by means of information systems that do not distinguish between explicit and implicit identifiers. Data processing systems for example do not need to know names and surnames of individuals they process (even for example, if they process their rights and obligations). They also do not need to know names and surnames to link existing data with new data as this integration can be done on the basis of implicit identifiers alone. The distinction between implicit and explicit identifiers is therefore merely a cognitive human concept - processing of data identified with one or other type of identifier does not differ much when one thinks about privacy concerns.

7. Records with explicit partial identifier

Next, perhaps an even more uncertain question is whether data identified with an explicit, but partial identifier, for example name and surname (where personal data attached to this identifier would for example be individual's salary) can qualify as personal data. There can be many individuals with the same name and surname, so partial identifiers identify merely a set of (similar) individuals and no single individual. These then are not proper personal data because no specific individual is taken out of the crowd. Rather, such records state something about some multitude of individuals. On the other hand it

[10] For example in addition to a central register of citizens and a multitude of employee registers that can contain links between SSNs and individuals names, there are also public phonebooks that link telephone numbers with names, surnames and other data and other such registers.

[11] It is interesting to note the wide definition of personal data under the UK Data Protection Act 1998. Part I states that personal data is "*data which relate to a living individual who can be identified (i) from those data or (ii) from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller*". As noted earlier however this definition has been narrowed by the Durant Case which states that data is personal if it "*is biographical in a significant sense; has to have the individual as its focus; and has to affect an individual's privacy whether in his personal family life, business or professional activity*".

is also true that this kind of data still retains some informative value: if we know that one of the individuals defined with name and surname has a certain salary, then a partial explicit identifier (name and surname) combined with concrete data lead us closer to a full explicit identifier, although it is yet to be fully defined. In this example, the additional information, which relates to the partial identifier, not only informs us about something (salary of someone with a certain name and surname), but also affects the further specification of the identifier. Even if by use of this triple of name, surname and salary an individual person is not yet fully defined (but for example he might be, if the salary was extremely high and someone with that name and surname would be known as a successful businessman, and it would be highly unlikely that two individuals with same name and surname were so successful), he is still much closer to full identification than he was before when he was described only his name and surname. Partial identifiers do denote a multitude of individuals, but every specification thereof narrows this multitude, until one last specification reduces the number of possible individuals down to one, thus obtaining a full identifier. The question here is, should records with partial explicit identifiers (name, surname), possibly together with some data related to them (for example persons' salary) which specify the identifier, be considered personal data or not?

Here there are many possibilities. If an individual becomes uniquely identified with the added information (lets presume that name, surname and salary do identify him uniquely), then it surely must be considered personal data. On one hand there is a unique full identifier (name, surname and salary), and on the other hand the identifier itself contains data about this uniquely identified individual (salary), with a unique identifier of the person together with at least one piece of data relating to it exactly corresponding to the structural definition of the personal data set forth in this paper.

Another possibility, where it is also certain that the data are personal, is a rather hypothetical example of a record which contains a combination of partial identifier and data, related to it , but with the important difference, that this data is not valid just for some individual from this multitude, but for all of them. The partial identifier can be for example John Smith (which, as explained above, denotes a set of individuals with this name; also, for the sake of this example, let us assume that all of them live in London) and data related to this identifier (hence data related to all the individuals that this identifier refers to) is the fact that whoever goes by that name lives in London. Individuals are only partially identified, but because the data is valid for the whole set, it is not possible to argue, that it is not known to whom specifically it relates to. So if data is valid for everyone, one cannot but agree that it is again personal data.

In the end the most interesting question remains: Should one consider partial explicit identifier (for example name and surname that together denote a multitude of individuals and that considerably narrows down the set of all individuals to those with the same name and surname but that still doesn't define one individual uniquely) together with data related to it (for example individual's salary) personal data or not?

Since records with such partial identifiers do not define individuals uniquely, we believe that they should not be considered personal data. Despite the fact that every added item of data not only communicates something which was not known before, but also helps with further identification of an individual, therefore focusing identification of an individual to a progressively narrower selection, the person in question is still not completely identified and the record therefore still not a personal data record. The fact that an individual could be uniquely identified in a few further steps arguably should not mislead one into believing, that data related to this as yet unidentified individual should qualify as personal data, because it is ultimately true that in a finite number of steps it is possible to convert any record about a set of individuals into a record of fully identified individual. If this was not so, then, for example the whole concept of anonymisation of data would be irrelevant since the anonymisation procedure refers exactly to the removal of unique identifiers from records about individuals in such a way, that only data relating to unidentified individuals remain.

When concluding that certain data is not personal data, one must be very cautious, because every partial identifier can be converted into data that uniquely identifies a person by the addition of additional data. Name and surname, for example, can constitute a unique identifier, if there is only one individual with this name and surname. If next to name and surname, there are further items of data

that can additionally specify an individual (e.g. residence, age), then the possibility that an individual gets individuated (identified) in this way is already quite large. Despite the fact that in theory data that cannot be attributed to fully specified individuals are not considered personal data, sometimes it will be very hard to argue that some records are not personal data even if they are only partially specified. Firstly, the majority of records about individuals will be based on name and surname, which do not require much, if any, new information, to become personal data. Secondly because data about individuals only have value for the data processors if they know who they relate to, such records will usually have to be fully specified and will fall under the Directive's definition of a "personal data filing system".

Such criteria for the delineation of records into records with personal data and those without also make sense with respect to Directive's notion of anonymisation. Anonymisation refers to the procedure by which unique identifiers of records get deleted, therefore leaving one with data that are still individual but that cannot be traced back to any specific individual. Anonymisation would make no sense if data about only partially identified persons were already considered personal data, since then there would be hardly anything more to do in order to further de-individualise such records. In the process of anonymisation, database administrators will of course also have to be very careful and will have to make sure that data is anonymised in such a way that it will not be possible to identify any individual person from it in any way whatsoever.

8. Records with implicit partial identifier

The last category are partial and implicit identifiers (for example the first few digits of SSN). The question is whether data about individuals identified with such implicit partial identifiers should be considered as personal data. The answer to the question is similar to the one given above for explicit partial identifiers: if such a partial and implicit identifier together with data related to it identifies an individual uniquely, then it should be considered personal data, otherwise not. These kinds of questions will of course have to be answered *in concreto,* in every case separately.

In summary, personal data are all data, that relate to individuals that are fully identified (explicitly or implicitly), but not data that relates to individuals that are identified only partially.

9. Attempt at a more specific definition

From the foregoing discussions, a more concise definition of 'personal data' in the Directive would be:

> Personal data is any data that relate to a fully identified individual, whereby identification is achieved either by means of an explicit identification (which requires no further linking of data) or by means of an implicit identification (which requires further linking of implicit identifiers with their more readable explicit counterparts but which does not require further collection of data).

This definition is, as a matter of fact, quite similar to the Directive's, that *'personal data 'shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity*. For similar definitions in EU members' data protection laws and a discussion about what they mean, see for example: Carey (2004)[12], page14; Lloyd (2004)[13], page 86; or Reed & Angel (2003)[14], page 431.

This paper however argues that the problem with the Directive's definition (and with jurisprudence) is that from the definition alone, one cannot infer that personal data are data with full identifiers,

[12] Peter Carey (2004) *Data Protection: A practical guide to UK and EU Law*. 2nd Edition, Oxford University Press,
[13] Ian Lloyd (2004). *Information Technology Law*, 4th Edition, Oxford University Press, UK
[14] Chris Reed and John Angel (2003), *Computer Law: The Law and Regulation of Information Technology*, 5th Edition, Oxford University Press, UK.

irrespective of whether they are explicit or implicit and not data with explicit identifiers, irrespective of whether they are full or partial. A shorten version of the Directive's definition can also be:

> Personal data is any combination of a unique (full) identifier of an individual and data related to it, whether this identifier is explicit or implicit.

10. Conclusion

This paper discussed how principles of relational database theory can be applied to determining personal data. The relational database concept of a record having a unique identifier (primary key) and related data was successfully applied to analysing data collected and protected under the EU data protection regime. The paper argued for the classification of data identifiers into full versus partial identifiers, and explicit versus implicit identifiers. The paper then proceeded to examine whether each of the four combinations of the two categories of data qualified as personal data, i.e. full and explicit identifiers, full and implicit identifiers, partial and explicit identifiers and partial and implicit identifiers. It was argued that a full identifier (whether explicit or implicit) qualified as personal data, however, a partial identifier (whether explicit or implicit) only qualified as personal data under certain circumstances, (i.e. where additional data lead to the identification of an individual or a set of individuals).

The paper was an attempt at applying computing/scientific principles to interpreting the law. An interesting issue which arose was the semantic interpretation of data and hence the addition of meaning when processed by the human mind. For example the distinction between implicit and explicit identifiers is a cognitive concept, and also an important aspect of the foregoing discussions. Human processing and cognition is different to machine processing especially in ascribing meaning to data. This use of meaning and contextualisation of language presents a difficulty when applying computing/scientific principles to law. The essence of law is extrinsically linked to the semantic interpretation of rules/laws, and the analysis (to find intention and meaning) of legal policy. Computing/Scientific principles such as relational database theory on the other hand, does not place any importance on the semantics of data, and hence, neglects and important aspect of law and jurisprudence. Nonetheless, the work done in this paper, proved that computing principles can help illuminate legal issues.

This paper represented the first part of a wider body of research into using relational database principles in the area of data protection. The approach taken in this paper was successful in placing data into four distinct categories, and subjecting such categories to further analysis and determination as to whether or not each category qualified as personal data. While some categories could not be definitively specified in all cases, further analysis of these categories represents an interesting challenge for future work.