BILETA
2008 Conference
GLASGOW CALEDONIAN UNIVERSITY
Law Shaping Technology; Technology Shaping the Law

Data, Information and Knowledge Under EU Data Protection Directive

Dr Boštjan Berčič
Institute for Economics, Law and Informatics, Ljubljana, Slovenia
Email: bostjan.bercic@iepri.si

Abstract

Information science (and knowledge management) distinguish between data, information and knowledge (some authors, e.g. Ackoff 1989, include understanding and wisdom as well). Data is raw piece of data by itself (e.g. attribute of a data record). Information is data brought in the context (e.g. data related to its primary key). Knowledge is many pieces of information considered together (e.g. a database).

In realm of information law (containing such legal institutes as personal data protection, copyrights, rights to privacy and personal integrity etc.) where the protected matter consists of intangibles (e.g. personal data, literary works, words of defamation) it is interesting to determine for relevant legal institutes what has to be there in order to warrant legal protection: bare data, information (data acquiring meaning in some context) or knowledge. In most cases, law protects information (e.g. personal data), but it sometimes protects bare data (e.g. anonymized personal data which do not relate to any known individual) as well as knowledge (complex information, for example works protected by copyright). The cases where the developed knowledge (such as in copyrighted works) is protected by law are the most complex.

This research into the connection between various forms of semantic intagibles and their legal protection could as well be developed in the other direction. One could set the stage by starting from semantic forms and then ask oneself how the law protects:

- complex knowledge (whereby procedural and structural knowledge can be protected by patents, literary works by copyright, personal integrity by actions agains defamation etc.)
- information (whereby personal information is for example protected by personal data legislation, public information is governed by freedom of information statutes, confidential information by non-disclosure agreements etc.)
- naked (pieces of) data itself.

This paper will try to map selective legal notions from the information law area (such as identifiers, personal data - which is in fact personal information, defamation etc.) to their underlying semantic forms (notions from areas of information science and knowledge management) in the form of a twodimensional correlation matrix. By way of this author hopes to obtain some new insights into connection between law and information sicence/knowledge management.

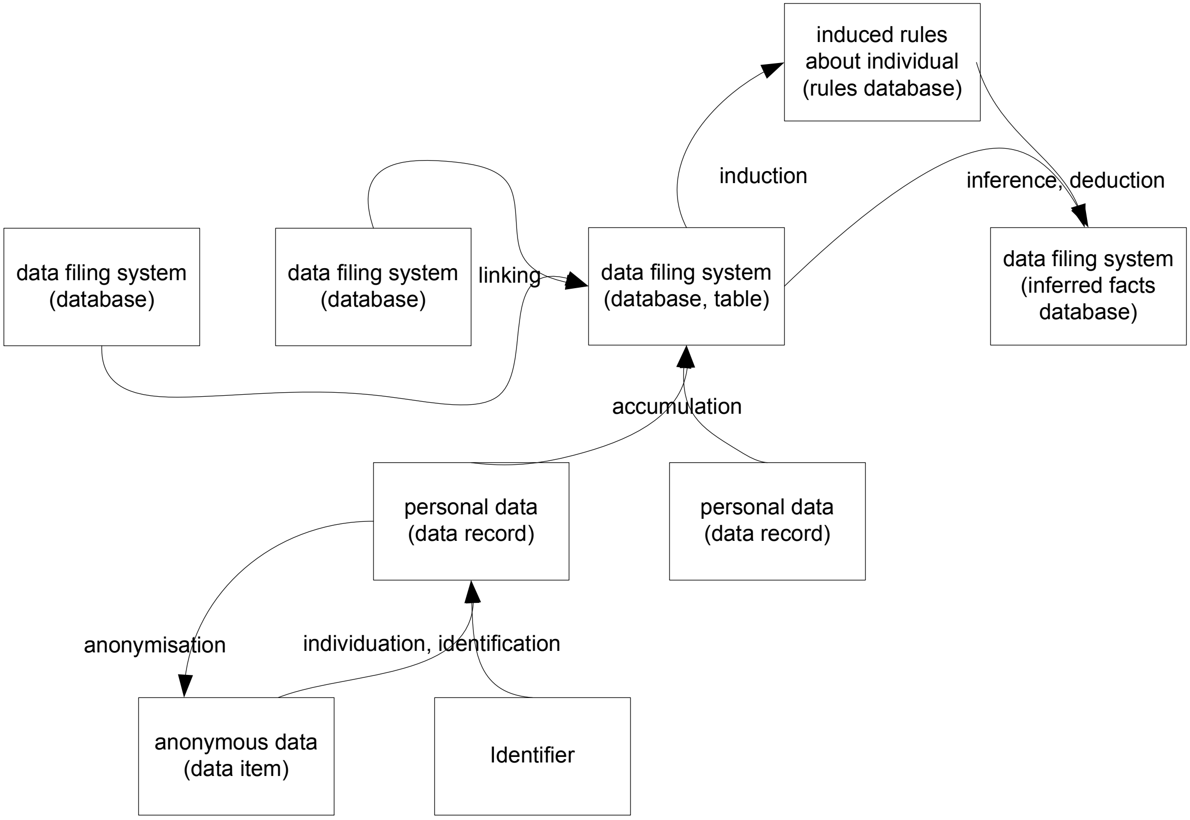
Data Information Knowledge Wisdom Hierarchy and Interpretation of it With Respect to Data Protection

According to the Data Information Knowledge Wisdom (DIKW) hierarchy (see for example Ackoff, R. L., "From Data to Wisdom", Journal of Applies Systems Analysis, Volume 16, 1989 p 3-9) **data** is raw data which does not have any meaning of itself. When connected to other data (e.g. the name of the entity it refers to), that is by way of a relational connection, it becomes **information,** data with meaning obtained from the context. **Knowledge** is the collection, amassing of information. **Understanding** is cognitive and analytical. It is the process by which new knowledge is synthesized from the previously held knowledge. The difference between understanding and knowledge is the difference between "learning" and "memorizing".

In the data protection context, data refers to anonymous information, that is data about an unidentified individual or anonymised data. Information refers to personal data *stricto sensu*. Knowledge refers to personal data filing system according to letter c of the Article 2 of the Directive. Deduced knowledge forms part of yet another data filing system. Induced knowledge (rules database) does not have a specific correlative in the data protection parlance. Interpretation of the Data, Information, Knowledge, Wisdom Hierarchy in the terms of data protection (hereinafter referred to as Data, Information, Knowledge, Rules Hierarchy or DIKR) is depicted in figure 1.

Typical operations between levels are also presented in figure 1. In some way, moving up the hierarchy means adding additional information and extracting new knowledge from the exisiting one. There are several ways of adding information to existing information. One is individuation, the change of data into information by way of identification. Another is collection of further pieces of information about an individual. The third is linking of such data and thus creation of a database. The fourth is linking of disparate databases. The fifth is application of common knowledge and methods (rules, categorisation schemes, statistical methods) to information about individual thus obtaining new facts about individual. The most advanced one is furthering hypotheses (induction) about an individual on the basis of knowledge contained in the database about him. These are all depicted in figure 1.

Figure 1 Interpretation of the DIKR hierarchy in terms of data protection


induced rules about individual (rules database)
induction
inference, deduction
data filing system (database)
data filing system (database)
linking
data filing system (database, table)
data filing system (inferred facts database)
accumulation
personal data (data record)
personal data (data record)
anonymisation
individuation, identification
anonymous data (data item)
Identifier

Terms from the definition of data processing from letter b) of Article 2 of the Directive such as collection (of data records form existing sources), recording of data records (obtained for example from data subjetcs), alignment (linking of several data records in a data table) or combination (linking of various data records) can be interpreted in terms from figure 1.

Data

Data is raw data which does not have any meaning of itself. In terms of data protection, data would refer to unidentified or anonymised personal data. Unidentified personal data would be information relating to an unidentified natural person. Anonymised personal data (according to the Article 26 of the Recitals of the directive) would be personal data rendered anonymous in such a way that the data subject is no longer identifiable and retained in a form in which identification of the data subject is no longer possible.

But there is another possibility, namely that of data which is not identified, but is identifiable. According to DIKR hirarchy, this would sit on the data level because the data is not identified, it is merely identifiable. Directive turns in other direction, though, and protects identifiable data (data) as personal data (information), that is, it lends in this case to data the same protection it usually lends to information! According to the wording of the directive, 'personal data 'shall mean any information relating to an identified or identifiable natural person ('data subject') whereas 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.

An example of an unidentified data would be anonymised DNA analysis (I'll build on this example later on and show how other semantic forms such as knowledge and rules can be understood in terms of thi example). Data is there, but it is unrelated to a specific individual by way of anonymisation.

Information

When data is connected to other data (e.g. the name of the entity it refers to) it becomes **information,** data with meaning stemming from the context. Data protection directive talks about personal data. Personal data is in fact (personal) information because it is one of the necessary prerequisites for the presence of personal data that there be an identification of whom the data refers to (according to the wording of the directive, letter a of Article 2 'personal data 'shall mean any information relating to an identified or identifiable natural person ('data subject')).

Typical operation connecting the previous and this level (data and information) is identification, that is individuation by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity, which turn data into information.

Another typical operation between the first and the second level of DIKR hierarchy is anonymisation, a process in which personal data (information) is rendered anonymous (data) in such a way that the data subject is no longer identified or identifiable and retained in a form in which identification of the data subject is no longer possible.

An example of personal data in the DNA analysis context is the DNA analysis of a known individual (or better even, analysis of DNA fragments if we reserve DNA analysis in its entirety for the example of a knowledge).

A typical operation that occurs between the second (information) and the third (knowledge) level is accumulation (adding) of information.

Knowledge

Knowledge is the collection, amassing of information. In the data protection context, knowledge might refer to a database of personal data, that is to 'personal data filing system' which according to letter c of Article 2 of the Directive shall mean any structured set of personal data which are accessible according to specific criteria, whether centralized, decentralized or dispersed on a functional or geographical basis.

A very important article of the Directive is Article 3. It says that the directive is applicable 'to the processing of personal data wholly or partly by automatic means, and to the processing otherwise than by automatic means of personal data which form part of a filing system or are intended to form part of a filing system'. This is a very important definition since it separates protected from the unprotected matter under the directive. While a particular record or data might very well be personal data, it si not neccessarily protected under the directive.

One typical operation that occur on the third level of the DIKR hierarchy is linking of existing databases.

An example of knowledge in the DNA analysis context is the DNA analysis of a known individual in its entirety. It's about a known individual and it is his genetic profile.

Deduced Knowledge

Deduced knowledge is knowledge extracted from knowledge database according to some predetermined procedure (e.g. by application of rules from rules database or by application of statistical procedures). Diretive refers to deduced knowledge in Articles 15 (Automated individual decisions) and 12 (Right of access). Article 15 talks about automated decisions (which are kind of deduction because rules are automatically applied to the facts database) producing legal effects for individuals and based solely on automated processing of data with the intention to evaluate certain personal aspects relating to individual, such as performance at work, creditworthiness, reliability, conduct, etc. Article 12 confers subjects that have been subjects of such procedures the right to obtain from the controller knowledge of the logic involved in any automatic processing of data concerning him. This logic is, in fact, rules database.

Typical operations connected to deduced knowledge are statistical inference (using knowledge database and statistical procedures) and deduction from a rules database (using knowledge and rules database).

An example of deduced knowledge in the DNA analysis example would be an application of know diagnostic tests (rules) to individual's DNA profile (knowledge). It surely represents personal data processing and falls under the Directive's definition of data processing (see letter b) of Article 2 of the Directive) but it is not individuated as a special kind of processing. I believe that application of known general knowledge to individual's profile merits special attention and should be either mentioned separately or understood as a special kind of processing. I believe that it should merit aproximately the same protection as the automated processing of data intended to evaluate aspects relating to individual such as his performance at work, creditworthiness, reliability, conduct accorinig to Article 15 of the Directive, or better protection !

Nothing as such is mentioned in the Directive presently. It has to be said, however, that personal data processing is limited by the directive in a general way, above all by Article 6, which requires that 'personal data must be collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes.' This means that if an application of general knowledge to individual's profile at hand is to be carried out, it should definitely fall within the scope of the purposes for which the profile was erected. If this does not involve diagnostics (individual might have, for example, paid the institution to decipher his DNA which he will store for his own use but has thereby not given the institution the permission to also run diagnostic on it), it should not be done.

Induced Knowledge

Induced knowledge are general laws and rules, usually in the form of a knowledge database, about the domain from the description of which they were extracted using induction (formation of rules). Induced knowledge (rules database) does not have a specific correlative in the data protection parlance. It is mentioned in the Article 6 of the Directive according to which personal data must be processed only in a way compatible with purposes for which they were collected. However, 'further processing of data for historical, statistical or scientific purposes shall not be considered as incompatible provided that Member States provide appropriate safeguards' and 'member States shall lay down appropriate safeguards for personal data stored for longer periods for historical, statistical or scientific use'. If not used in the historical, statistical or scientific use context, data should otherwise be 'kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed'. Article 11 also states that the information to be given to the data subject where the data have not been obtained from the data subject should not be as extensive as normaly, 'in particular where, for processing for statistical purposes or for the purposes of historical or scientific research, the provision of such information proves impossible or would involve a disproportionate effort '.

Typical operations that occur between the third (knowledge) and the fourth level (rules) are induction (scientific discovery of rules on the basis of present knowledge) and deduction (infering new facts from rules and present knowledge).

An example of induced knowledge in the DNA analysis context would be testing individual responses to drugs (pharmacogenomics) or, in the context of e-marketing, inducing buyers' individual buying habits on basis of their buying history.

This kind of processing, as is the case with deduction from individual's profiles and common knowledge, is not specifically mentioned in the directive. In my opinion, it should merit special protection too (one could try to interpret it in terms of automated processing of data, if in fact it was processed in an automatic way, but this does not exhaust all possibilities). If anything, such processing could according to the present wording of the Directive warrant a lesser degree of protection than ordinary processing of personal data, because:

- it shall not be considered as incompatible with the Directive provided that Member States provide appropriate safeguards (letter b, paragraph 1, Article 6),
- the results of such processing could be stored for longer provided that member States lay down appropriate safegurads and (letter e, paragraph 1, Article 6),
- the information to be given to the data subject where the data have not been obtained from the data subject should not be as extensive as it would ordinarilly be (paragraph 2, Article 11).

Recapitulation

Above facts and mappings from semantic forms (notions from areas of information science and knowledge management) to legal notions from the information law area (such as identifiers, personal data - which is in fact personal information, defamation etc.) are summarized in the table 1 bellow.

Table 1

semantic concept	...is incarnated in...	...and contains ...	..and is examplified by...
data	Unidentified data record	fact about unidentified individual	anonymous DNA analysis
information	identified data record	fact about identified individual	analysis of DNA fragments of a known individual
knowledge	collection of personal data records (database)	many facts about identified individual;individual's profile, linking databases	DNA analysis of a known individual
deduced knowledge	collection of inferred facts from the facts and rules database (deducing new facts with rules form the theory or expert system)	inferred facts from the facts and rules databases	deducing possible health problems from DNA analysis
induced knowledge	rules database (inducing new theories about individual)	rules about individual	testing individual responses to drugs (pharmacogenomics); inducing buyers' buying habits on basis of their buying history

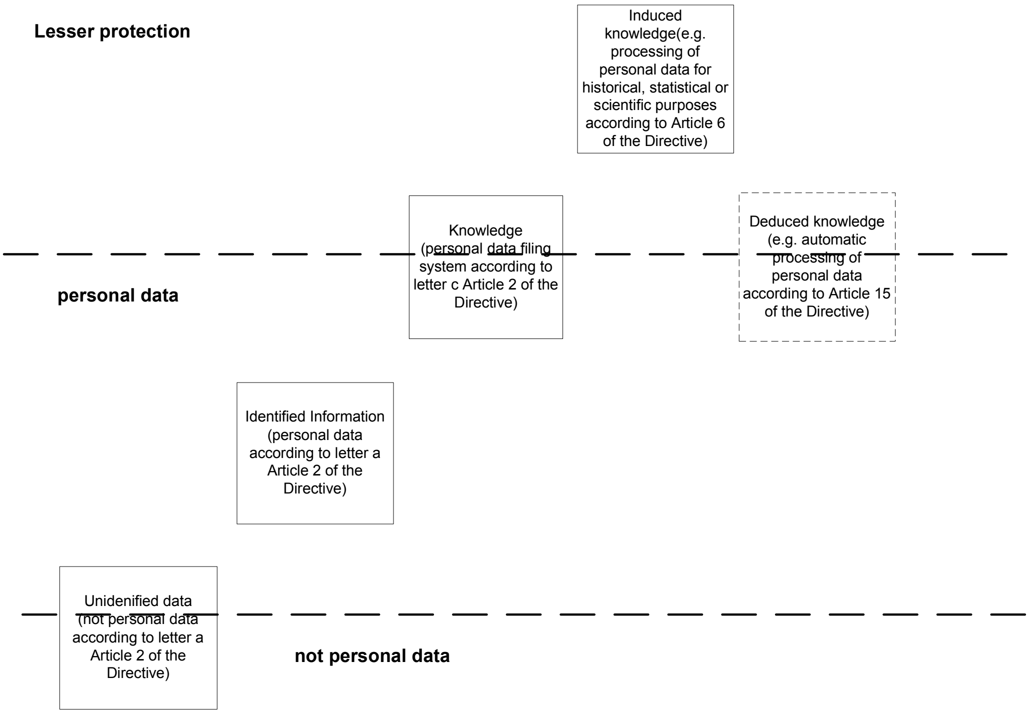
semantic concept	...and mentioned in the Directive	typical data operation
data	anonymised personal data (Article 26 of the Recitals)	anonymization
information	personal data (Article 2 of the Directive)	individuation
knowledge	data filing system (Article 2 of the Directive)	accumulation and linking of facts
deduced knowledge	a) automated processing of data intended to evaluate certain personal aspects relating to individual, such as his performance at work, creditworthiness, reliability, conduct (Article 15 of Data Directive, b) processing for statistical purposes (paragraph 2 of the Article 11 of the Directive)	inference, deduction, application of statistical and other scientific metods
induced knowledge	processing for the purposes of historical or scientific research (Article 29 of the Recital, Article 6 of the Directive, paragraph 2 of the Article 11 of the Directive)	induction, theory formation, hypothesis testing, learning rules

Protected Concepts in the Processing of Personal Data

In the context of personal data, unidentified data is not personal data according to the Directive and is therefore not protected. As soon as the data refers to identified or identifiable individual (it is therefore information) it becomes personal data according to the Directive and is warranted protection (with caveat that it has to either form part of a filing system or is intended to form part of a filing system, Article 3 of the Directive). In accordance with the same definition, knowledge about individuals in the form of a filing system is also protected and so is deduced knowledge about an individual (Directive mentions automatic processing of personal data intended to evaluate certain personal aspects which in fact is application of rules to data about individual, that is deduction). Induced knowledge might be another category. Article 6 of the Directive states that »further processing of data for historical, statistical or scientific purposes shall not be considered as incompatible provided that Member States provide appropriate safeguards«. That is, personal data which normally has to be « collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes« might be used for these other purposes (historical, statistical or scientific) for which it was not collected. Article 11 also states that the controller's duty to provide information to the data subjects as regards the processing of personal data related to him does not apply » where, in particular for processing for statistical purposes or for the purposes of historical or scientific research, the provision of such information proves impossible or would involve a disproportionate effort or if recording or disclosure is expressly laid down by law.«. There are therefore important exceptions form the normal data protection regime when data is used for historical (knowledge), statistical (deduced knowledge) or scientific (induced knowledge) purposes. It is tempting to say that the induced knowledge is to certain extent excluded from the data protection whereas knowledge and deduced knowledge about individuals is always fully protected, but this generalisation is probably wrong. There are historical purposes and statistical purposes for which (any) personal data can be used, without being originally collected for it, which means that general knowledge (e.g. historical data) and deduced knowledge (e.g. statistical data) can also be to certain extent exempt from the normal data protection regime. It is clear, however, that induced knowledge (if we equate induced knowledge by scientific method referred to in the Article 29 of the Recital) is somehow less protected than other types of knowledge.

A probably more accurate description of what gets protected under EU data directive is depicted in figure 2 where some data are included (identifiable but as yet unidentified data) and where some knowledge (historical and statistical use) is excluded to certain extent.

Figure 2



Lesser protection
Induced knowledge(e.g. processing of personal data for historical, statistical or scientific purposes according to Article 6 of the Directive)
Knowledge (personal data filing system according to letter c Article 2 of the Directive)
Deduced knowledge (e.g. automatic processing of personal data according to Article 15 of the Directive)
personal data
Identified Information (personal data according to letter a Article 2 of the Directive)
Unidenified data (not personal data according to letter a Article 2 of the Directive)
not personal data

References

Ackoff, R. L., "From Data to Wisdom", Journal of Applies Systems Analysis, Volume 16, 1989, p 3-9

(Data, Information, Knowledge, and Wisdom by Gene Bellinger, Durval Castro, Anthony Mills, http://www.systems-thinking.org/dikw/dikw.htm)

BERČIČ, Boštjan, GEORGE, Carlisle. Identifying personal data using relational database design principles. *Int. j. law inf. technol.*, 2008, vol. 16