Record Linkage of Anonymous Data by Control Numbers

W. Thoben, H.-J. Appelrath, S. Sauer

OFFIS, Westerstr. 10-12, D-26121 Oldenburg

Summary: The processing of personal data in information systems is of decisive importance in regard to data protection. Several applications exist where the personal identifying data make no difference but are needed for the record linkage with other data sets. This paper describes a concept which enables to compare anonymous data and to fulfil the complex requirements of data protection in the processing.

1. Introduction

The right of access to personal files means that everyone decides, in principle, for himself on the revelation and use of his personal data. This right is one of the basic rights protected by constitution and should protect a person against the abuse of his personal data while the information technology is expanding more and more.

However there are applications in which this right competes with the need of a population-based data aggregation with the possibility to identify individuals and where no personal identifying data is needed for the processing. In this case a data record can be coded, i.e. to encrypte the personal identifying data or to separate it from the other data and to use only the anonymous data.

A decisive problem which follows from such a proceeding is to join different anonymous data records. The primary criteria for searching in personal identifying data are generally identifying features. But as well as they are of no account for the processing their use is not allowed in regard to data protection, so it is unalterable to develop a concept which does not use these features but nevertheless permits a record linkage.

2. Cancer Registry Lower-Saxony

An example for an application in which the personal identifying data makes no difference is an epidemiological cancer registry. The different cancer reg-

www.manaraa.com

istration models are described in (Schrage (1991), Gruner et al. (1989)). In cooperation with the institute OFFIS (Oldenburger Forschungs- und Entwicklungsinstitut für Informatik-Werkzeuge und -Systeme) the concept of Prof. Michaelis (Michaelis and Krtschil (1992), Schmidtmann et al. (1993)) is proved in the years 1993/94 in Lower-Saxony (Appelrath et al. (1993), Brand et al. (1993)). This concept, which is taken into account in the actual bill of an Act for cancer registries (KRG (1994)), contains a registration of cancer divided in a position of trust and a register. Personal reports to the cancer registry are collected in the position of trust, where the personal identifying data are separated from the epidemiological data. Applying a public-key cryptosystem (using two keys - public key for encryption and secret key for decryption (Rivest et al. (1978)) the personal identifying data is encrypted. These anonymous reports are transfered to the register and condensed to a population-based epidemiological cancer registry.

Because each cancer case is repeatedly reported to the cancer registry and to increase the quality and quantity of the registry by integration of external data sets - e.g. information on the death certificates, reports from pathologists - a record linkage of the anonymous data records has to be realized in the register. At this point the problem is to recognize several anonymous data records, which have to be attached to the same respectively different persons as the same respectively different reports.

3. Record linkage system

3.1 Robustness of the record linkage system

A decisive aspect for the record linkage of data records is to guarantee the robustness of the record linkage system, if there should be errors in the plaintext. In this case several data records, which describe the same person, but include some different identifying data, are matched to the same person. This so-called fault tolerance is opposed to the aim of a high discriminating power, i.e. to attach data records of different persons even with negligible differences in their identifying data to different persons. Typical causes for differences of natural language character strings, especially in personal identifying data, which result from errors during the registration or transfer of the personal identifying data or incomplete or incorrect information about the person, are (Ultsch (1987), Gruner et al. (1989)):

- **Typing error:** Writing or typing errors are mainly independent of an application and can be subdivided in four elementary error categories (missing, additional, incorrect character or transposition of characters).

www.manaraa.com

- **Phonetic misunderstandings and ambiguouses:** Phonetic misunderstandings can result during verbal transfer of information, especially during the registration of personal data records from other than written sources. A similar problem is represented by different styles of equally sounding names (e.g. Meier or Meyer), where confusions conduct to ambiguouses.

- **Differing information:** Another error cause is differing information in reports, which describe the same person. E.g. in one data record a person with double name is registrated completely while in another only one name is available. A similar problem are different addresses when a person is migrated or has several places of residence.

- **Missing information:** In this case the question is to recognize the similarity of data records and furthermore, which and how much missing information in a data record is tolerable.

3.2 Control numbers

In principle there is the possibility of a record linkage through the ciphertexts of single reports whereby already little errors in the plaintext cannot be detected in the ciphertexts. Using a nondeterministic cryptosystem (adding a random number to the encryption to prevent chosen plaintext attacks), which is realized in the Cancer Registry Lower-Saxony, several identical data records of one person are enciphered to different ciphertexts and, thus, they cannot be matched to one data record.

One approach for a record linkage of data records is the use of control numbers (Michaelis and Krtschil (1992), Schmidtmann et al. (1993), Appelrath et al. (1993)), a deterministic encryption of characters of personal data records, e.g. name, birthname, address. The control numbers are generated in parallel to the encryption of the personal identifying data by using a one-way function (deterministic function for which the inverse function cannot be practically developed). The identity of a person cannot be inferred from these control numbers. Decisive for the work with control numbers is the fact that the one-way function is deterministic, because several reports of the same person must be represented by the same control numbers. Two types of errors occur in record linkage (Gruner et al. (1989)):

- **Homonyms:** Reports of different cancer cases are matched to the same person, i.e. the control number (variable) of different persons is assigned with the same value.

www.manaraa.com

- **Synonyms:** Different reports of one cancer case are represented to different persons, i.e. one control number (variable) is assigned with different values, in spite of the equality of the person.

Causes for these errors are the typing errors, phonetic misunderstandings and ambiguouses represented in section 3.1, but also the conception of the control numbers - extracts of personal data records - can generate homonyms itself.

The aim of the record linkage system is to find attributes or parts of them for the specification of control numbers, which produce minimal synonyms and homonyms. Therefore the similarity of data records must be defined in form of metrics respectively the development of suitable heuristics has to be forced (Newcombe (1985)).

4. Examinations

4.1 Examinated control numbers

During the first examinations with the prototype of the record linkage system in the course of the Cancer Registry Lower-Saxony 16 control numbers have been generated (Appelrath et al. (1993)), which can be - if required - adapted respectively reduced or completed. The following 6 control numbers are presented exemplary:

(1) surname, first name, date of birth, sex
(2) surname[1], first name[1], date of birth, sex
(3) soundex(surname), soundex(first name), date of birth, sex
(4) ASCII-$\sum$(surname) + LEN(surname), ASCII-$\sum$ (first name) + LEN(first name), date of birth, sex
(5) surname[3], first name[3], month/year of birth, sex
(6) surname[3], first name[3], year of birth, sex

The digit in the parentheses behind the attributes shows the number of viewed places from the beginning of a word and the functions ASCII-$\sum$ and LEN are definded as follows:

- ASCII-$\sum$(string): sum of the ASCII-encoding of the characters,
- LEN(string): number of characters of the string.

On the one hand the control numbers are generated directly from the attributes of the data records (see (1), (2)), on the other hand by using functions (see (3), (4)), which try to recognize errors like a transposition of characters

www.manaraa.com

(in this case the results of the functions SUM and ASCII-$\sum$ are equal) or phonetic misunderstandings with phonetic codes like soundex (Mresse (1984)).

4.2 Rates

Suitable measures to judge the quality of the control numbers have to be determined. In this case that are the rate of homonyms and the rate of synonyms, which are defined as follows:

Given are n reports without synonyms, i.e. all n reports are from different cases. A control number is generated for each of these cases and then the number of differently generated control numbers k is determined. The rate of homonyms H results from $H = \frac{n-k}{n}$, the ratio of equally generated to the total number of all control numbers.

To determine the rate of synonyms of a control number there are given n different cases with 2 reports for each case. A control number is generated for each of these $2n$ reports and the number of correct matches l is determined, i.e. in l of n cases the control numbers of both reports of a case process the same value. The rate of synonyms S of a control number results from $S = \frac{n-l}{n}$, the ratio of all falsely matched to the total number of different cases.

4.3 Examinated data sets and examination process

In the first step of the project of Cancer Registry Lower-Saxony 3 data sets are examinated (Appelrath et al. (1993)):

- Nachsorgeleitstelle Oldenburg (NLS): 13.269 data records,
- death certificates from 1990-92 of the city Oldenburg (DC): 4.587 data records,
- Ma.Ca.-death certificates from 1990-92 of the city Oldenburg and the district Ammerland (Ma.Ca.): 88 and 47 data records.

For all data records of the 3 data sets 16 control numbers are generated and are saved with the data records in a database. The control numbers of every data record are compared with those of all other data records and in the case of identity they are saved in a special database. Homonyms are calculated by the comparison of each data record of a data set with every other data record of the same data set. Therefore, it is important that there are no synonyms in these data sets. To determine the rates of synonyms by using the double registrated reports the 3 data sets are compared with each other.

For the evaluation of the examinated control numbers all equal entities of a

www.manaraa.com

control number are visually controlled by using the plaintext of the data sets. Thereby it is checked if both data records really describe the same person. In the following it can be determined for each control number,

- how many similarities have been correctly recognized,
- how many data records have been detected as similar but do not belong to the same person,
- how many data records have not been detected as similar although they should describe the same person.

5. Results

5.1 General survey

We have to distinguish between two kinds of record matching process: internal record linkage (see tab. 1) to determine the rates of homonyms and external record linkage (see tab. 2) to determine the rates of synonyms. Under the section matches there is noted how many reports are related to at least another one and with how many control numbers this correspondence has been ascertained.

	NLS	DC	Ma.Ca.
matches			
data records	387	74	0
control numbers	596	123	0

Table 1: internal matches

	NLS-DC	NLS-Ma.Ca.	DC-Ma.Ca.
matches			
data records	519	73	90
control numbers	4.512	854	1.120
errors			
data records	181	3	2
control numbers	291	5	4

Table 2: external matches

In the table of internal matches every match is of course a fault. The total number of examined data records produces the denominator for calculating

www.manaraa.com

the rate of homonyms for each control number (e.g. NLS: 13.269).

The table of external matches shows also the absolute numbers of matches. These cases have been visually controlled and the discovered wrong matches (synonyms) have been determined. The difference between matches and faults (e.g. NLS-DC at tab. 2: 519-181 = 338) is the denominator for calculating the rate of synonyms for each control number because these are the data records for which exactly two reports exist.

5.2 Rates of homonyms and synonyms

On the base of the denominator from tab. 1 and tab. 2 and with further detailed examinations of each control number the rates of homonyms and synonyms have been determinated by using the definitions of chapter 4.2.

control number	NLS	DC	Ma.Ca.
1	0	0	0
2	0,12	0,13	0
3	0	0	0
4	0	0	0
5	0,17	0,07	0
6	1,31	0,7	0

Table 3: rates of homonyms

control number	NLS-DC	NLS-Ma.Ca.	DC-Ma.Ca.
1	30,77 (20,41)	28,57	28,41 (11,36)
2	7,69	4,29	0
3	18,34 (11,24)	8,57	19,32 (5,68)
4	30,18 (19,82)	28,57	28,41 (11,36)
5	10,36	7,14	5,68
6	8,88	5,71	5,68

Table 4: rates of synonyms

Tab. 3 and tab. 4 show the percentages for the control numbers given in chapter 4.1. The numbers in parentheses within the table of synonyms mark the part of faults that are caused by a non-uniform registration of first names in the different data sets if more than one is entered in the original report.

In principle the rates of synonyms are much worse than the rates of homonyms.

www.manaraa.com

One reason is that these rates are a result of a comparison of two different data sets and therefore different guidelines for registration. This fact shows especially the faults that are exclusively caused by the non-uniform registration of first names.

Besides, you have to recognize that the rates influence each other. Control numbers with a low rate of homonyms (e.g. control number (1), (3), (4)) have a considerably higher number of synonyms or vice versa. This is affected by the fact that a control number with a slight rate of homonyms is characterized by a high discriminating power while it generates more synonyms at the same time. In contrast a control number with a low rate of synonyms has a high fault tolerance but it neglects the discriminating power. Therefore, the aim has to be to find a suitable mixture of control numbers that combine both qualities.

6. Prospect

The present examinations of anonymous data records have been exclusively made on the base of isolated control numbers. In fact the approach by means of control numbers seems to be promising but the results in chapter 5.2 show that isolated control numbers cannot ensure reliable record linkage with minimal rates of homonyms and synonyms. So in a second step combinations (conjunctions of single control numbers) will be examined which guarantee both the aspect of discriminating power and a sufficiently high fault tolerance. The third step will then be the construction of rules (disjunctions) out of single control numbers and combinations of them. As a result there will be a system of rules that ensures minimal rates of homonyms and synonyms.

Furthermore, we have to consider other data sources to improve the validity of previous results. Especially the examination of synonyms should be forced, i.e. to find data sources with double registration of individuals and apply the methods of record linkage on them.

Another aspect of further efforts is improving the quality of registration. The results show remarkably the problems that arise by recording personal data and so they form the basis for an improvement of the quality of data.

References

APPELRATH, H.-J., THOBEN, W., RETTIG, J., and SAUER, S. (1993): CARLOS (Cancer Registry Lower-Saxony): Tätigkeitsbericht für den Zeitraum 1.4.-1.11.1993. Oldenburg.

www.manaraa.com

BRAND, H., REICHLING, I., APPELRATH, H.-J., ILLIGER, H.-J., UNGER, I., and WINDUS, G. (1993): CARLOS (Cancer Registry Lower-Saxony) - Pilotstudie für ein bevölkerungsbezogenes Krebsregister in Niedersachsen. In: S.J. Pöppl, H.-G. Lipinski and T.Mansky (eds.): *Medizinische Informatik - Ein integrierender Teil arztunterstützender Technologien*. MMV Medizien Verlag, München, 404-406.

GRUNER, G., HARTMANN, S., MEISNER, C., PIETSCH-BREITFELD, B., and SELBMANN, H.K. (1989): Forschungsvorhaben "Epidemiologisches Krebsregister Baden-Württemberg". Bericht Nr. 6/1989, Institut für Medizinische Informationsverarbeitung, Universität Tübingen.

KRG (1994): Entwurf eines Gesetzes über Krebsregister (Krebsregistergesetz KRG). Stand: 26.5.1994, Drucksache 12/6478, Bonn.

MICHAELIS, J., and KRTSCHIL, A (1992): Aufbau des bevölkerungsbezogenen Krebsregisters für Rheinland-Pfalz. *Ärzteblatt Rheinland-Pfalz*, 45, 434-438.

MRESSE, M. (1984): Information Retrieval - eine Einführung. Leitfäden der angewandten Informatik, Teubner, Stuttgart.

NEWCOMBE, H.B. (1985): Handbook of Record Linkage - Methods for health and statistical studies, administration, and business. Oxford University Press, Oxford.

RIVEST, R.L., SHAMIR, A., and ADLEMAN, A. (1978): A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*, Vol. 21, No. 2, 120-126.

SCHMIDTMANN, I., POMMERENING, K., and MICHAELIS, J. (1993): Pilotstudie zum Aufbau eines bevölkerungsbezogenen Krebsregisters in Rheinland-Pfalz, In: S.J. Pöppl, H.-G. Lipinski and T.Mansky (eds.): *Medizinische Informatik - Ein integrierender Teil arztunterstützender Technologien*. MMV Medizin Verlag, München, 399-403.

SCHRAGE, R. (1991): Zur Krebsregisterfrage - modifiziertes Melderechtsmodell zur Verbesserung des Datenschutzes. *Öffentliches Gesundheits-Wesen*, 53, Georg Thieme Verlag, Stuttgart, 746-752.

ULTSCH, A.G.H. (1987): Control for knowledge-based information retrival. Dissertation, no. 3, institute of computer science, ETH Zürich.

www.manaraa.com