Schnell, Rainer

Working Paper

Record-Linkage from a Technical Point of View

RatSWD Working Paper, No. 124

Provided in Cooperation with:
German Data Forum (RatSWD)

Suggested Citation: Schnell, Rainer (2009) : Record-Linkage from a Technical Point of View, RatSWD Working Paper, No. 124, Rat für Sozial- und Wirtschaftsdaten (RatSWD), Berlin

This Version is available at:
https://hdl.handle.net/10419/186229

German Council for Social
and Economic Data (RatSWD)

www.ratswd.de

RatSWD
Working Paper Series

Working Paper No. 124

Record-Linkage from a Technical Point of View

Rainer Schnell

August 2009

Federal Ministry
of Education
and Research

Working Paper Series of the Council for Social and Economic Data (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

Starting in 2009, some of the empirical research papers that originally appeared in the *RatSWD Working Papers* series will be published in the series *RatSWD Research Notes*.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD.

Record-Linkage from a Technical Point of View

Rainer Schnell

University of Duisburg-Essen (rainer.schnell[at]uni-due.de)

Abstract

Record linkage is used for preparing sampling frames, deduplication of lists and combining information on the same object from two different databases. If the identifiers of the same objects in two different databases have error free unique common identifiers like personal identification numbers (PID), record linkage is a simple file merge operation. If the identifiers contain errors, record linkage is a challenging task. In many applications, the files have widely different numbers of observations, for example a few thousand records of a sample survey and a few million records of an administrative database of social security numbers. Available software, privacy issues and future research topics are discussed.

Keywords: Record-Linkage, Data-mining, Privacy preserving protocols

1. Introduction

Record linkage tries to identify the same objects in two different databases using a set of common identifiers.[1] If the files have error free unique common identifiers like personal identification numbers (PID), record linkage is a simple file merge operation. If the identifiers contain errors, record linkage is a challenging task. In many applications, the files have widely different numbers of observations, for example a few thousand records of a sample survey and a few million records of an administrative database of social security numbers. Most research applications of record linkage use the linking process for preparing sampling frames, deduplication of lists and combining information on the same object from two different databases.[2]

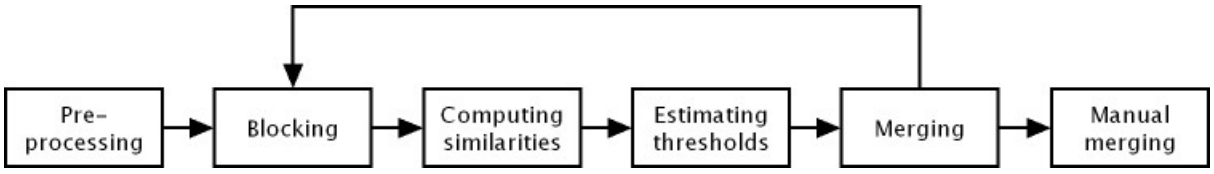
2. Current Applications

Searching for the keyword „record linkage“ will currently yield a few thousand papers on applications in medicine (foremost in epidemiology), but only a few dozen papers in social sciences. Nevertheless, record linkage is often used by social science research companies as part of the fieldwork contracted to them; in many such cases the record linkage process is unknown by the client. Constructing sampling frames in practice often implies joining information from different databases on objects like names, addresses, birthdays, phone numbers and geo-data by using record linkage.[3] Record-Linkage is often used to combine information based on a survey with information from a database. Very often such linkages have been done for business surveys, where information on performance, business size and business type have been added by record linkage to business survey data.[4] Record Linkage may be used to build panels after data collection, for example by using historical data as in the „Victorian Panel Study“ (VPS). The VPS is intended as longitudinal dataset based on the British censuses 1851-1901 (see Crockett et al. 2006). Such linkages are possible in many cases even without the use of unique personal identifiers. One such application is the „Statistical Longitudinal Census Data Set“ (SLCD). The Australian Bureau of Statistics

1 The label “record linkage” is most often used by statisticians. In computer science, many different labels are common, for example ”deduplication”, "reconciliation" or "merge/purge processing".
2 Record linkage tries to identify the same objects in two databases. Do not confuse record linkage with statistical matching: Statistical matching (or data fusion) tries to find records of very similar values of different objects; thereby deliberately joining data files with no common objects. For applications of statistical matching, see D'Orazio et al. (2006).
3 Some examples for German surveys may be found in Schnell (2008).
4 Details on such application can be found in a paper by Winkler (1995).

(ABS) will build the SLCD by linking a 5% sample of people from the 2006 population census to subsequent censuses. In order to minimize privacy problems, ABS will use record linkage without the use of name and address (Bishop and Khoo 2006). Furthermore, record linkage is an essential tool for conducting any census in general and the most important tool for a registry based census like the German census 2011. After taking the census, record linkage is necessary for the estimation of coverage rates.[5] As a final example, in nonresponse research linking data of nonrespondents to administrative data files is one of the few methods to assess nonresponse bias with empirical data.

Figure 1: The linking process



3. Record linkage process

Record linkage is the process of linking two files which have data on the same objects using common identifiers. This process follows a standard sequence (see figure 1). Usually, the identifiers must be standardized, which is called „pre-processing“. Since the number of comparisons is in general too high to be computed directly, the computations are split up between disjunct subsets of observations (called „blocks“) and repeated for different blocking criteria.[6] The similarity of records within a block is computed using similarity functions, most often today either with a edit-distance or the Jaro-Winkler-String-similarity function.[7] Then a decision on thresholds of similarity has to be made: Records above a threshold are considered as a link, records below the threshold are considered as a non-link. Records between the thresholds are usually submitted to clerical review. linkprocess.pdf A record linkage process The statistically most interesting part of the process is the decision which pairs of the elements of the two datafiles should be considered as true links. This decision can be based on different computational models, for example classification trees (CART), support vector classifiers (SVM) or statistical decision rules.[8] Most record linkage programs today use a

5 There is a rich literature on using record linkage for census undercount estimates, starting with Winkler/Thibaudeau (1991) and Ding/Feinberg (1996).
6 For example, in a cancer registry, persons living within an area with a common postcode are treated as a block.
7 Details on the computation and performance of string similarity functions can be found in Herzog et al. (2007) and Schnell et al. (2003).
8 Detail on SVMs and CART can be found in any textbook on statistical learning, for example Bishop (2006).

probabilistic decision rule due a model suggested by Fellegi/Sunter (1969). The parameters of the model are usually estimated by some variants of an EM-algorithm (Herzog et al. 2007). Special situations (for example: a known one-to-one correspondence between the two files) require modifications of the decision rules.

4. Available software

There are many record linkage systems available. Most of the systems are special purpose programs for use in official statistics or cancer registries.[9] Furthermore, there are a couple of commercial programs for office applications. Of course, there are some academic proof-of-concept-implementations of special algorithms. The historically most important program and three contemporary programs in the public domain will be described in some detail.

4.1 Automatch

The most widely known probabilistic record linkage program is „Automatch“. The last version (4.2) has been released in 1992. Automatch is now a part of a large collection of programs (IBM's „WebSphere QualityStage“) and can not be licensed or bought as a stand-alone program. The cost of the IBM Web-Sphere is far beyond the scope of research groups, therefore Automatch is no more used in research contexts. Only a few cancer registries use the old DOS-version of Automatch with a special permission of IBM. Automatch is often used for validation of other programs. It should be noted, that the limitations of an old DOS programs had been evaded by some clever programming shortcuts; therefore Automatch is not a perfect baseline for comparisons.

4.2 Link Plus

Link Plus is primarily a probabilistic record linkage program for cancer registries. The program has been developed for the „National Program of Cancer Registries“ (NPCR) of the Center for Disease Control and Prevention. It is a windows based program for detecting duplicates and linking cancer registry files with external files.[10] The program offers different similarity functions and phonetic encodings. Furthermore, it handles missing data and special cases like middle initials.[11]

---

9 A highly selective review from an official statistics point of view can be found in Herzog et al. (2007). There is also a list of criteria which should be used in evaluations of record-linkage software.

10 Since the development team want to include the Microsoft .NET framework and Access-databases, the binding of Link Plus to windows will be even closer in the future.

11 The program is available at no charge under www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm

4.3 Link King

„Link King“ is an SAS-based probabilistic record linkage program developed by Kevin M. Campbell. The program requires a base SAS license. The program can work with SAS-files, SPSS portable files, and CSV-files. The most interesting features are nickname matching, gender imputation for 20.000 (American) names and the calculation of distances between (American) zip codes.[12]

4.4 The Merge-Toolbox: MTB

A project group of the author (funded by a research grant of the German research foundation) has developed a „merge toolbox“ (MTB) for probabilistic record linkage (Schnell et al. 2005). MTB is written in JAVA and therefore highly portable to any modern computer system. The program consists of a preprocessing module, a linkage module and a manual editing module. The program can read and write STATA and CSV-files, computes nearly all known string similarity functions and can perform deterministic and probabilistic record linkage. MTB is being used by cancer registries and research groups in epidemiology, sociology and economics in Germany.[13]

4.5 Empirical comparisons of programs

Since most record-linkage programs for probabilistic linkage use the same algorithms for making link decisions, the programs should yield very similar results, given the same input. Since the programs differ in pre-processing, some studies compare different parts of the linkage process. Only identically preprocessed data files should be used for linking; but this is often of no practical relevance. So for practical applications, the complete linkage-process between optimally tuned programs should be compared: This is no trivial task and therefore rarely such studies have been published (Campbell et al. 2008). From a theoretical point of view, comparing different programs using different decision rules (for example, CART, SVM and Fellegi-Sunter) on non-preprocessed data and identically pre-processed data would be more interesting. Systematic studies are lacking up to now. However, working on an optimized combination or sequence of decision rules after extensive standardization and preprocessing seem to be more promising than naive empirical comparisons.

12 The program is available at no charge under www.the-link-king.com

13 A restricted version of the program is available at no charge under www.uni-konstanz.de/FuF/Verwiss/Schnell/mtb. For scientific purposes, the full program is available at no charge by writing to the author.

5. Privacy Issues

Record linkage may be misused for de-anonymization of scientific research files. This possibility of misuse is simply due to the fact that the programs try to minimize distances between objects in a high-dimensional space. Therefore, de-anonymization by minimizing distances can be done by every program for cluster analysis.[14] So this misuse is not specific to record-linkage programs.

The result of a successful record-linkage is a data set C with more known characteristics of the objects than in the original data files A and B. Using this enhanced data file C by comparing these characteristics with another data file D makes a identification of objects in D much more likely than identification by using A or B alone, since the number of observations with a given combination of characteristics is declining with every added variable.[15] The risk of disclosure is therefore higher after the record-linkage. It might be necessary to use additional standard risk disclosures measures for the enhanced datafile C.[16]

6. Research perspectives

From a statistical perspective, the theoretical problems of record linkage are well defined and some interesting solutions have been found. Many applied researchers consider record-linkage as a trivial task. In practice, it is not. It is remarkable, that the actual performance of record-linkage programs in practice is often disappointing for the layman.[17] The main cause of the lack of performance is usually the quality of the input data: If many identifiers are missing or poorly standardized, any automatic method will fail. Therefore, we need more work on preprocessing of identifiers. Since preprocessing depends on language and country specific details, programs and algorithms must be fine tuned with local datasets and expert systems. Therefore, experts from statistics and computer science need to use real data from actual data generating processes.

14 For an application, see Torra et al. (2006).

15 This can be seen as a direct consequence of the definition of k-anonymity: In a k-anonymized dataset, each record is indistinguishable from at least (k-1) other records.

16 Examples of such techniques can be found in Willenborg/de Waal (1996) and Domingo-Ferrer (2002); for record linkage and privacy issues in general, see United States General Accounting Office (2001).

17 For example, Gomatam et al. (2002) note higher sensitivity and a higher match rate but a lower positive predicted value of Automatch in comparison to a stepwise deterministic strategy. These results could be changed easily by a change of matching parameters and the preprocessing.

6.1 Real-world test data sets

Interestingly, a standard data set for comparing record linkage procedures has not been published. Instead some research groups build data generators with specified error generating mechanisms. Since such error structures may be different from those of real-world applications a collection of test data sets based on real world data would be highly desirable. Since the details of name conventions, addresses, post codes etc. differ between countries and data bases, a German reference data base is needed.

6.2 Expert systems and key standardization

Database fields contain many different ways of storing information of key values used for record linkage. This fields must be standardized by using expert knowledge on the distinctive features of German addresses, phone numbers (land-line and mobile), name conventions (for example, historical rules for name changes after marriage), academic titles, historical hereditary titles, legal forms of companies etc. Compiling such lists and generating transformation rules is a tedious and labor intensive task. Currently, the required amount of work to generate such exhaustive lists and standardization rules have been expended solely by private companies.[18] Of course, the cumulated commercial knowledge bases are not available for academic use. Therefore, German official statistics will have to buy such standardization services for large scale operations like the Census 2011 on the commercial market with obvious consequences. In the long run, statistical offices, cancer registries and other public funded research organizations need a common knowledge bases for key standardization.

6.3 Reference data bases

For practical record linkage, several reference data bases are needed, which are currently not public available for research purposes. For example, simple lists of all German municipalities with old and new German zip codes, correspondence lists of zip codes and phone numbers, regional identifiers like city codes („Gemeindekennziffer“), Gauss-Krüger-coordinates and street addresses are not available for public use. Every record linkage group has to compile rough versions of these reference lists. Since some of these list are quite expensive, there should be a scientific license for this data gathered by public money.[19] Furthermore, frequency

18 The unit on "`Postal Automation"' of Siemens I&S (Konstanz) employs more mathematicians and computer scientists for producing such expert systems than all German cancer registries in total. Given the published lists of customers of other companies in the same sector in Germany (for example, "`Fuzzy Informatik"', a spin-off of Daimler) it is safe to assume that currently more than 50 experts in Germany work on such standardization tasks.

19 For example, the list of all geo-coordinates of all German buildings, which would be useful for many research purposes in record linkage and epidemiology, is a considerable expense at about the costs of a research assistant per year.

tables of names and surnames conditioned on gender, nationality and year of birth would be highly useful for imputing gender, nationality and age given a name. Other data bases can be used for the same purpose, for example for certain ICD- or ISCO-codes gender can be imputed. This imputed information can be used for record linkage with incomplete keys.

6.4 Candidate generation

One interesting idea, which has not been studied in detail so far, is the generation of candidates for matching based on an search string. The candidates can be generated by introducing random errors or according to pre-specified rules (Arusu et al. 2008). The resulting candidates will be compared to the existing identifiers. This step should follow unsuccessful standard linkage attempts.

6.5 Blocking

Data files for record linkage are usually quite large. In many applications, we have a small file (for example, a survey) with about 1000 observations and an administrative data base with, for example, 10 million records. This would result in 10^{10} comparisons, taking 278 hours at 10.000 comparisons per second. Using standard hardware and standard programs, this is unacceptable. Therefore, the computation time is usually reduced by using a simple idea: Compute the similarity matrix only within subgroups. These subgroups are called „blocks“ and the strategy is called „blocking“. For example, we don't compare every company name in Germany with each other; instead we compare only all pairs of company names within each city. Using a suitable blocking variable reduce the computing time of one typical record linkage run (10.000 observations linked to a five million record data base) to less than a hour. Of course, this speed comes with a price. The variable used for blocking must be considered as a perfect classification variable: Exhaustive, disjunct and error free. Since blocking variables are in many cases proxy variables of geographical identifiers like dial prefixes, post codes or administrative units, there is no guarantee for error free perfect classification of units. Currently, there is a lot of research activity in computer science in modifications of blocking algorithms in order to improve on simple blocking schemes (for example, „adaptive blocking“, Bilenko et al. 2006). These new blocking techniques still have to be implemented in production software for record linkage.

6.6 Algorithms for large similarity matrices

As an alternative to blocking, algorithms for computing approximate similarity matrices could be used. Such algorithms have been proposed in the technical literature, for example „Sparsemap“ (Hristescu and Farach-Colton 1999), „Boostmap“ (Athitsos et al. 2004) and „WEBSOM“ (Lagus et al. 2004). Another interesting approximation has been recently suggested by Brandes/Pich (2007). None of these techniques has been systematically used for record-linkage up to now. Special data structures or algorithms used for high-dimensional indexing (Yu 2002) have rarely been applied for large scale record-linkage projects.

6.7 Special hardware

Since the blocking of data sets reduce the task of computing a *n*n* similarity matrix to the independent computation of *k* matrices of size *m*m*, the computation can be done by several independent machines or processors. This is a very simple version of a parallel computing process, which requires only a trivial modification of existing programs. Of course, parallel searching of similarity index structures by special algorithms (Zezula et al 2006, chapter 5) or the separate standardization of each record may also be done with such hardware. However, the resulting program can be run of the shelf hardware like standard PC boards. Since such a system should be portable, a compact server rack can be used. Currently available server boards house 4 processors with 4 cores each, so a special machine with 64 cores can be build by using only 4 server boards. In order to reduce power consumption, smaller mobile processor boards may be used instead, requiring 8 boards with 2 quad-core mobile processors. Such a system will drain less than 1000 Watt in total, so it do not require special cooling or power supply. The machine should be equipped with at least 1 Gbyte RAM for each processor. In order to minimize the risk of data leaking, the machine can be build as a diskless server: The machine need no hard-disk at all, since the operating system can be booted from a memory stick and the data to be processed may reside on removable memory sticks.[20] The sticks should be destroyed after reading; the linked data file should be written to an empty new stick. In slightly less security demanding computing environments, the input files may be copied to the machine by using VPN. Such a portable secure special purpose record-linkage machine can be build at the price of three small enterprise servers. It would be highly desirable to have at least one such machine within a trusted computing center with restricted access, for example within one the research data centers.

20 Even a data file with 30 million records and 100 bytes of ID-information per record fits on a 10 Euro 4-Gbyte USB-stick.

6.8 Privacy preserving record linkage

In most practical applications record linkage has to be done with the standard keys name, surname, gender, date of birth, place of birth. Since people hesitate to use of such identifiers, in many applications encrypted keys have to be used. Since the input data for encryption is prone to errors, a slight deviation between the keys of a true link pair is probable. Such slight deviations result in keys which can not be matched, since similarity distances between encrypted keys are pointless. Therefore, privacy preserving record linkage requires special algorithms. Starting with the publication by Churches/Christen (2004) some protocols for record linkage with encrypted alphanumeric keys with errors have been suggested (Pang and Hansen 2006; Scannapieco et al. 2007). Independent comparisons of these protocols have not been published and are badly needed. All protocols seem to be awkward to implement with mistrustful database owners. To overcome this problems, we have developed a new protocol, which seems to be very fast and reliable (Schnell et al. 2007). Currently, we test the protocol on different simulated datasets. A complete record linkage solution for encrypted keys must include a protocol for computing distances between encrypted metric data. One very interesting protocol has been proposed by Inan et al. (2006). A really secure record linkage program for error prone numeric and alphanumeric keys will need a few years of testing and programming. This seems to be the most important research task before record linkage can be used widely given the increasing privacy concerns in western populations.

7. Three recommendations

7.1 Training data sets and reference data sets

In order to improve the performance of record-linkage programs and algorithms, large training and reference data sets should be produced. This should be real-life datasets, containing only linkage variables. The links have to be established by a common error free key or careful clerical work. Simulated data sets are no substitutes for such data sets. Therefore, privacy concerns must be take care off by standard procedures of statistical disclosure control.

7.2 Research program on pre-processing and privacy preserving record linkage

We need a european research program on pre-processing keys for privacy preserving record linkage. Such a research program should be multi-national, since the ethnic composition of european countries differ and therefore the distribution of ethnic surnames. Furthermore, the

legal situation on record-linkage differs widely within Europe. Therefore, a multi-national and multi-disciplinary research group of computer scientists, lawyers, linguists, historians and social scientists is needed to solve the problems of privacy-preserving record linkage using standard identifiers like names and surnames.

7.3 National Record Linkage Center

Currently, we don't have research centers for record linkage in Germany. We just have the cancer registries, which do a very limited kind of record linkage for a single purpose. Every research team in criminology, sociology, medicine or economy must organize its own record linkage infrastructure. In many cases, the cost of doing so exceeds the available research funds. Therefore, at least one National Record Linkage Center is needed. The center should have special machines (massive parallel processors), a team trained in record linkage and the data protection facilities necessary to act as a data trustee for large scale projects.

References:

Arasu, A./Chaudhuri, S. and Kaushik, R. (2008): Transformation-based Framework for Record Matching, International Conference on Data Engineering.

Athitsos, V./Alon, J./Sclaroff, S. and Kollios, G. (2004): BoostMap: a method for efficient approximate similarity rankings. CVPR, 268-275.

Bilenko, M./Kamath, B. and Mooney, R.J. (2006): Adaptive Blocking: Learning to Scale Up Record Linkage, ICDM06: Sixth International Conference on Data Mining, 87-96.

Bishop, C.M. (2006): Pattern Recognition and Machine Learning, Berlin (Springer).

Bishop, G. and Khoo, J. (2006): Methodology of Evaluating the Quality of Probabilistic Linking, Proceedings of Statistics Canada Symposium 2006, www.statcan.ca/english/freepub/11-522-XIE/2006001/article/10401-en.pdf

Brandes, U. and Pich, C. (2007): Eigensolver Methods for Progressive Multidimensional Scaling of Large Data. Proc. 14th Intl. Symp. Graph Drawing (GD '06). LNCS 4372, 42-53.

Campbell, K.M./Deck, D. and Krupski, A. (2008): Record linkage software in the public domain: a comparison of Link Plus, The Link King, and a 'basic' deterministic algorithm. Health Informatics Journal, 14, 1-15.

Churches, T. and Christen, P. (2004): Some methods for blindfolded record linkage. BMC Medical Informatics and Decision Making, 4, 1-17.

Crockett, A./Jones, C.E. and Schürer, K. (2006): The Victorian Panel Study, unpublished manuscript, University of Essex.

Ding, Y. and Feinberg, S.E. (1996): Multiple sample estimation of population and census undercount in the presence of matching errors. Survey Methodology, 22, 55-64.

Domingo-Ferrer, J. (Ed.) (2002): Inference control in statistical databases: from theory to practice. Berlin (Springer), Lecture notes in computer science Vol. 2316.

D'Orazio, M./DiZio, M. and Scanu, M. (2006): Statistical Matching: Theory and Practice, New York.

Elmagarmid, I.A./Ipeirotis, P.G. and Verykios, V. (2006): Duplicate Record Detection: A Survey, IEEE Transactions on Knowledge and Data Engineering, 2006, (19), 1-16.

Fellegi, I.P. and Sunter, A.B. (1969): A Theory for Record Linkage. Journal of the American Statistical Association, 64.

Gomatam, S./Carter, R./Ariet, M. and Mitchell, G. (2002): An empirical comparison of record linkage procedures. Statistics in Medicine 21, (10), 1485-1496.

CMIS Technical Report No. 03/83, CSIRO Mathematical and Information Sciences.

Herzog, T.N./Scheuren, J.J. and Winkler, W.E. (2007): Data Quality and Record Linkage Techniques, New York/Berlin.

Hristescu, G. and Farach-Colton, M. (1999): Cluster-preserving embedding of proteins. Technical report, Rutgers University.

Inan, A./Saygin, Y./Savas, E./Hintoglu, A.A. and Levi, A. (2006): Privacy Preserving Clustering on Horizontally Partitioned Data. Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06).

Lagus, K./Kaski, S. and Kohonen, T. (2004): Mining massive document collections by the WEBSOM method Information Sciences, Volume 163, Issues 1-3, 135-156.

Pang, C. and Hansen, D. (2006): Improved record linkage for encrypted identifying data. HIC 2006 and HINZ 2006 Proceedings, Brunswick (Health Informatics Society of Australia), 164-168.

Scannapieco, M./Figotin, I./Bertino, E. and Elmagarmid, A.K. (2007): Privacy preserving schema and data matching, Proceedings of the 2007 ACM SIGMOD international conference on Management of data, 653-664.

Schnell, R./Bachteler, T. and Bender, S. (2003): Record linkage using error prone strings. American Statistical Association, Proceedings of the Joint Statistical Meetings, 3713-3717.

Schnell, R./Bachteler, T. and Reiher, J. (2005): MTB: Ein Record-Linkage-Programm für die empirische Sozialforschung: ZA-Information 56, 93-103.

Schnell, R./Bachteler, T. and Reiher, J. (2007): Die sichere Berechnung von Stringähnlichkeiten mit Bloomfiltern, Discussion paper, University of Konstanz, September 2007.

Schnell, R. (2008): Avoiding Problems of Traditional Sampling Strategies for Household Surveys in Germany: Some New Suggestions, Discussion paper for the GSOEP, www.diw.de/documents/publikationen/73/86107/diw_datadoc_ 2008-033.pdf

Torra, V./Abowd, J.M. and Domingo-Ferrer, J. (2006): Using Mahalanobis Distance-Based Record Linkage for Disclosure Risk Assessment; in: PSD 2006, Privacy in Statistical Databases, Berlin, 233-242.

United States General Accounting Office (GAO) (2001): Record Linkage and Privacy. Issues in Creating New Federal Research and Statistical Information, April 2001, GAO-01-126SP, Washington.

Willenborg, L. and de Waal, T. (1996): Statistical Disclosure Control in Practice, New York.

Winkler, W.E. (1995): Matching and Record Linkage; in: Cox,B.(et al. eds.): Business Survey Methods, New York, 355-384

Winkler, W.E. and Thibaudeau, Y. (1991): An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. decennial census technical report, US Bureau of the Census, www.census.gov/srd/papers/pdf/rr91-9.pdf.

Yu, C. (2002): High-Dimensional Indexing. Transformational Approaches to High-Dimensional Range and Similarity Searches, Berlin.

Zezula, P./Amato, G./Dohnal, V. and Batko, M. (2006): Similarity Search. The Metric Space Approach, New York.