AUSTRIAN JOURNAL OF STATISTICS
Volume 33 (2004), Number 1&2, 125-133

A Toolbox for Record Linkage

Rainer Schnell, Tobias Bachteler
University of Konstanz

Stefan Bender[1]
Institute for Employment Research (IAB)

Abstract: We developed a record-linkage toolbox in order to compare the performance of various string-similarity measures for German surnames. This ”Matching Tool-Box” (MTB) is made up by independent, highly portable JAVA-programs. MTB is currently used for prototyping pre-processing tools and the empirical comparison of string-similarity measures. Furthermore, MTB has been used successfully in sociological, economical and epidemiological research projects.

Zusammenfassung: Um die Verwendbarkeit der verschiedener Ähnlichkeitsmaße für fehlerbehaftete Namen auch für deutsche Namen vergleichen zu können, entwickelten wir eine eine ”Matching Tool-Box” (MTB). MTB besteht aus mehreren, transportablen JAVA-Programmen. MTB dient zur Entwicklung von Pre-processing-Werkzeugen und dem Vergleich von String-Ähnlichkeitsmaßen. MTB wurde erfolgreich in sozial- und wirtschaftswissenschaftlichen sowie epidemiologischen Forschungsprojekten eingesetzt.

Keywords: Record linkage, String Similarity.

1 Introduction

Statisticians are often faced with the problem of linking databases from other sources. Especially for human generated data like surveys, this can be a difficult task since respondent data are prone to various kinds of error (memory lapses, spelling and typographical errors, and so on). Matching these data to other data is problematic since even small errors prevent the use of exact-match algorithms. Since most database and statistical programs offer only exact-match routines, additional programs must be employed to perform the matches for data that contains errors. No such program seems to be public available. Two programs developed for use in official statistics are widely known: Matcher-2 of the US-Bureau of the Census (Porter/Winkler 1997) and GRLS of Statistics Canada (Fair, 1995). The most widely known program in medical research is OXLINK (Gill, 2001). Due to very special hard- and software requirements, OXLINK and GRLS are hardly portable (and in case of GRLS: very expensive). Matcher-2 runs on standard PCs, but don’t allow easy modifications of the algorithms used. On the commercial side, one of the few such programs was AUTOMATCH, which had been used widely in the medical research community. Since its discontinuation, there has been a dearth of commercial

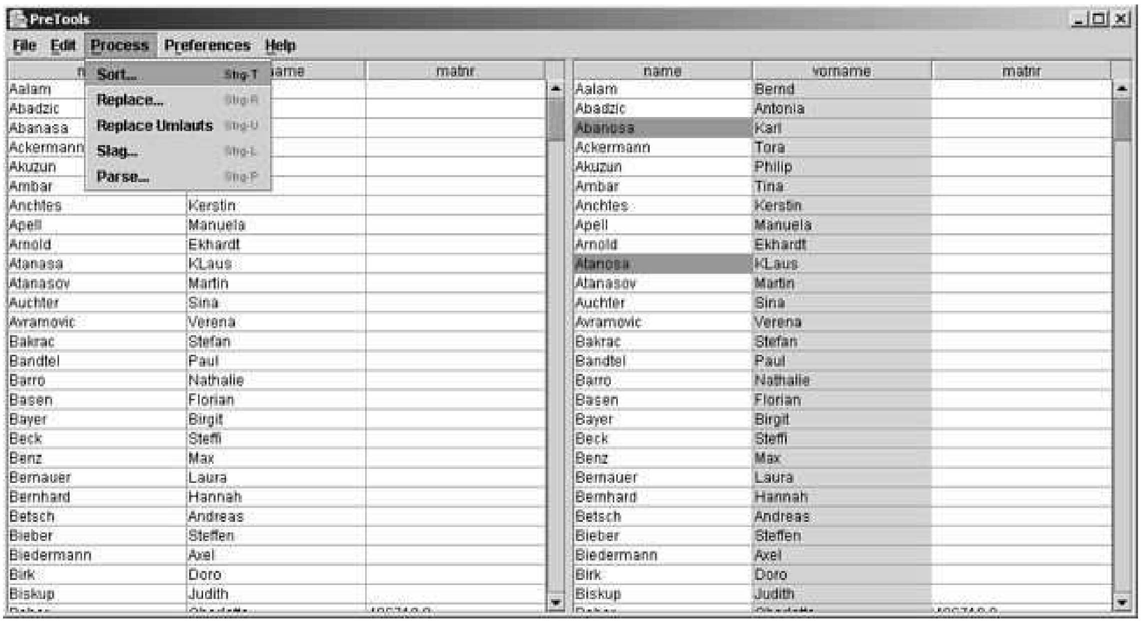
[1] This work was partially funded by the German Research Foundation and by the Institute for Employment Research (IAB). The survey was financed by the Max-Planck Institute for Human Development (MPI), the IAB and the European Social Fund. Perl and Java programming were done by Joerg Reiher.

products that fill this niche, at least of those with a price affordable for a German university or federal agency. Even if such a product were available, it would be difficult to know which string-similarity measure should be used, since only a few comparisons of string-similarity measures have ever been published in journals. Of course, even fewer comparisons have been published for German surnames. Due to the fact that all known comparisons using German surnames are based on artificially generated errors, not on actual human errors, we decided to conduct a study comparing various routines. Doing so required us to develop our own record linkage program.

2 Program Development

The program specification resulted in the assignment of subsets of the total task list to three programs: a pre-processing tool, a deterministic record-linkage module and a manual editing module. Since many of the program's subroutines were already available as modules in the "Comprehensive Perl Archive Network" (CPAN), the prototype programs were implemented in Perl. After one year of experimentation and program specification, we migrated to JAVA. All programs had been redesigned and are now much easier to modify than the original Perl versions. Furthermore, the JAVA implementation ensures an easier portability of the graphical user interface across many different platforms.

2.1 Pre-Processing Tool

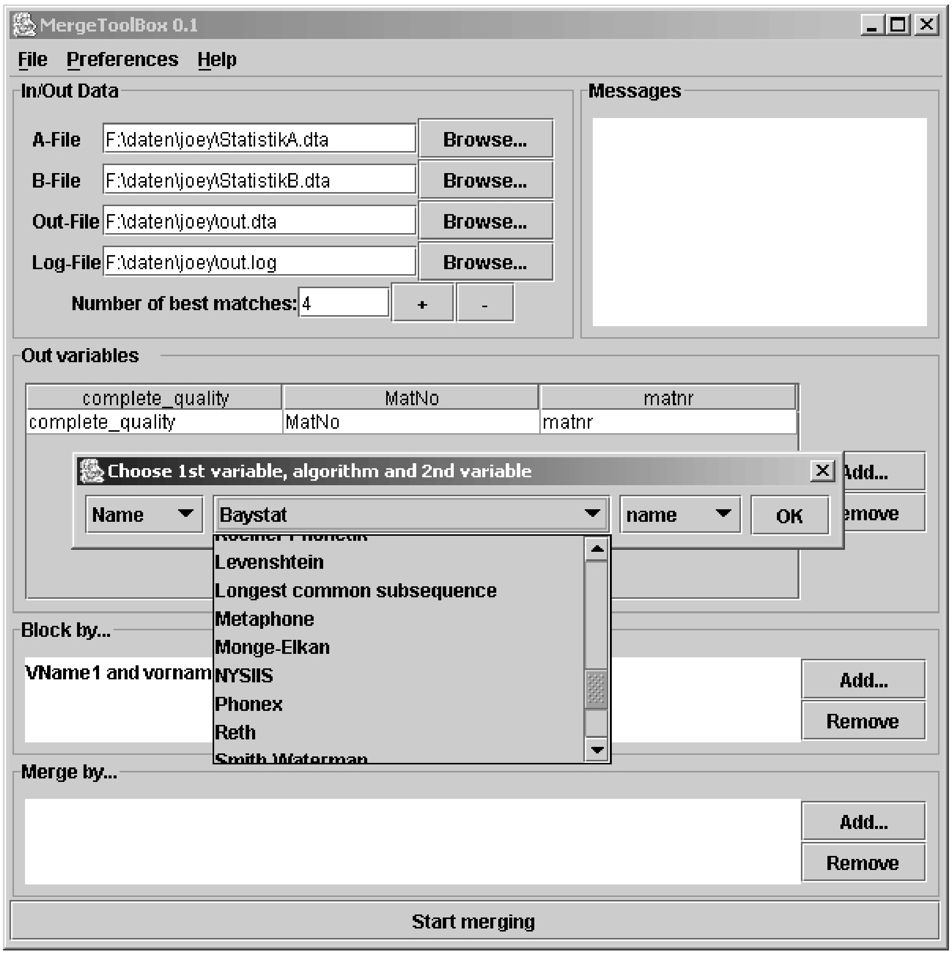


n	ame	matnr
Aalam		
Abadzic		
Abanasa		
Ackermann		
Akuzun		
Ambar		
Anchles	Kerstin	
Apell	Manuela	
Arnold	Ekhardt	
Atanasa	KLaus	
Atanasov	Martin	
Auchter	Sina	
Avramovic	Verena	
Bakrac	Stefan	
Bandtel	Paul	
Barro	Nathalie	
Basen	Florian	
Bayer	Birgit	
Beck	Steffi	
Benz	Max	
Bernauer	Laura	
Bernhard	Hannah	
Betsch	Andreas	
Bieber	Steffen	
Biedermann	Axel	
Birk	Doro	
Biskup	Judith	

name	vorname	matnr
Aalam	Bernd	
Abadzic	Antonia	
Abanosa	Karl	
Ackermann	Tora	
Akuzun	Philip	
Ambar	Tina	
Anchles	Kerstin	
Apell	Manuela	
Arnold	Ekhardt	
Atanosa	KLaus	
Atanasov	Martin	
Auchter	Sina	
Avramovic	Verena	
Bakrac	Stefan	
Bandtel	Paul	
Barro	Nathalie	
Basen	Florian	
Bayer	Birgit	
Beck	Steffi	
Benz	Max	
Bernauer	Laura	
Bernhard	Hannah	
Betsch	Andreas	
Bieber	Steffen	
Biedermann	Axel	
Birk	Doro	
Biskup	Judith	

Figure 1: Screenshot of the pre-processing tool

The purpose of this tool is to transform raw data entered using such programs as EpiData into a form suitable for record linkage. This tool is therefore able to read such common data-file formats as ASCII-CSV or Xbase. The tool then writes such data in a standard format: Since STATA is increasingly used in academic environments and its file format is openly published, we decided to use STATA (version 7) as the internal file format. A further function of the pre-processing tool is its ability to remove typical but unnecessary strings (academic titles, nobility or corporation prefixes, numbers, and so forth) and standardize the spelling of such prefixes as Mc or Mac. To this end, we have collected a number of different prefix lists (for companies, institutions and people). The program also allows for the replacement of German umlauts with usual ASCII codes < 127. A further issue is the treatment of women's married names. Revisions to German marriage laws mean that double and triple names are now quite common in Germany. The pre-processing tool therefore contains an option to parse double and triple names according to delimiters like dashes or blanks. The selection of this option creates separate records for the whole string and each sub-component. Finally, the program has an option to standardize different date formats. The graphical user interface displays two file windows, which can be scrolled separately. Commands are selectable by pull-down-menus. A special feature of the program is the ability to display (random) subsets of the files, so that the effect of a string transformation operation like prefix removal can be observed and judged immediately.



A-File F:\daten\joey\StatistikA.dta

B-File F:\daten\joey\StatistikB.dta

Out-File F:\daten\joey\out.dta

Log-File F:\daten\joey\out.log

Number of best matches: 4

complete_quality	MatNo	matnr
complete_quality	MatNo	matnr

Choose 1st variable, algorithm and 2nd variable

Name | Baystat | name | OK

Koelner Phonetik
Levenshtein
Longest common subsequence
Metaphone
Monge-Elkan
NYSIIS
Phonex
Reth
Smith Waterman

VName1 and vornam

Figure 2: Screenshot of the Deterministic Record-Linkage Module

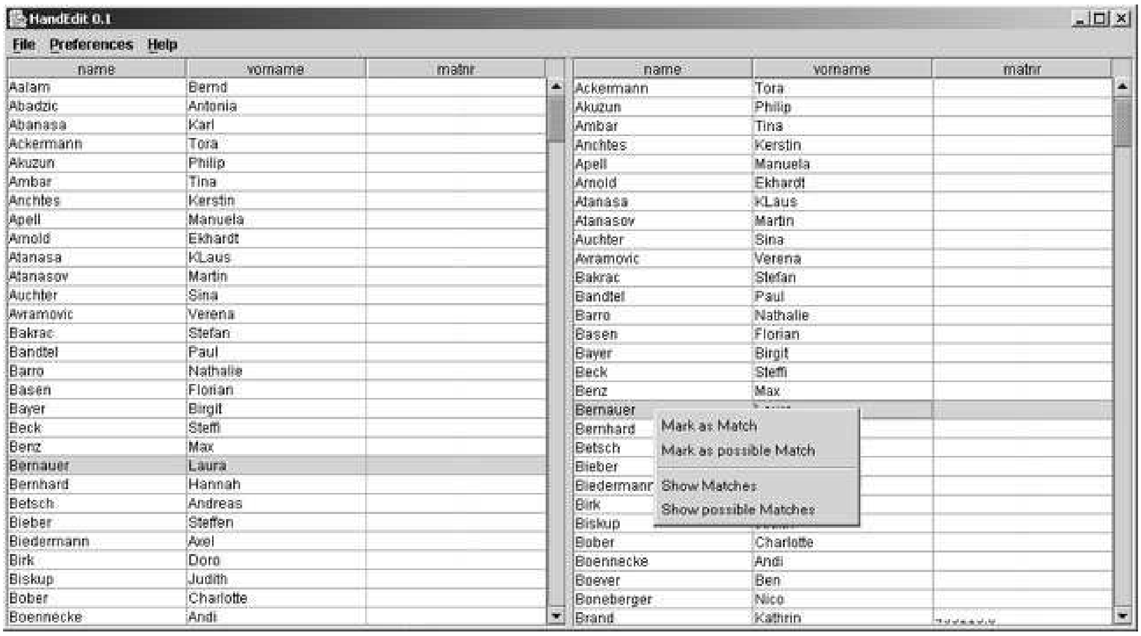
2.2 Deterministic Record-Linkage Module

This module operates on STATA files containing only preprocessed key variables and identification numbers. The main task of this module is the computation of a wide variety of string-similarity measures. Currently, the module computes string similarities based on the following measures:

- bi- and trigrams (Ukkonen, 1992)
- different versions of edit distances (Ukkonen, 1985)
  - LCS (Hirschberg, 1977)
  - Baystat (Fürnrohr, Rimmelspacher, & Roncador, 2002) [2]
  - Jaro's String Comparator and its variants (Porter & Winkler, 1997)
  - Monke-Elkan-algorithm (Monge & Elkan, 1996)
- Soundex (Knuth, 1998)
- German Soundex variants (Postel, 1969), (Reth & Schek, 1977)
- Metaphone (Philips, 1990)
- Double-Metaphone (Philips, 2000)
- Phonex (Lait & Randell, 1996)
- NYSIIS (Taft, 1970)
- Guth (Guth, 1976)
- Speedcop (Pollock & Zamora, 1984)
- Synoname (Borgman & Siegfried, 1992).

Within record subgroups formed according to several user selectable variables (called ”blocking”), string-similarities of all possible pairs are computed. The program output is a STATA data set containing both a selectable number of potential matches (pairs) for each identification variable and a similarity measure for each of those pairs. All program options can be selected by pull-down menus. A log-file of the selected options is saved for each run. This log-file can be used as command-file for batch-processing of this module. This feature is most useful for experiments with different similarity measures. The actual file-merge is not done within this module. This one-to-one-match of two different databases is done within STATA by using the output of this module as input file in STATA. For production runs, the whole process of pre-processing, blocking, similarity computation and file merge can be executed by an operating system shell script as a batch job.

[2] This measure had been developed by a Bavarian workgroup for the coming German census.



name	vorname	matnr
Aalam	Bernd	
Abadzic	Antonia	
Abanasa	Karl	
Ackermann	Tora	
Akuzun	Philip	
Ambar	Tina	
Anchtes	Kerstin	
Apell	Manuela	
Arnold	Ekhardt	
Atanasa	KLaus	
Atanasov	Martin	
Auchter	Sina	
Avramovic	Verena	
Bakrac	Stefan	
Bandtel	Paul	
Barro	Nathalie	
Basen	Florian	
Bayer	Birgit	
Beck	Steffi	
Benz	Max	
Bernauer	Laura	
Bernhard	Hannah	
Betsch	Andreas	
Bieber	Steffen	
Biedermann	Axel	
Birk	Doro	
Biskup	Judith	
Bober	Charlotte	
Boennecke	Andi	

name	vorname	matnr
Ackermann	Tora	
Akuzun	Philip	
Ambar	Tina	
Anchtes	Kerstin	
Apell	Manuela	
Arnold	Ekhardt	
Atanasa	KLaus	
Atanasov	Martin	
Auchter	Sina	
Avramovic	Verena	
Bakrac	Stefan	
Bandtel	Paul	
Barro	Nathalie	
Basen	Florian	
Bayer	Birgit	
Beck	Steffi	
Benz	Max	
Bernauer		
Bernhard		
Betsch		
Bieber		
Biedermanr		
Birk		
Biskup		
Bober	Charlotte	
Boennecke	Andi	
Boever	Ben	
Boneberger	Nico	
Brand	Kathrin	[illegible]

Figure 3: Screenshot Manual Editing Module

2.3 Manual Editing Module

The most labor-intensive part of a record-linkage project is the manual matching of cases not matched by the program. We therefore wrote a manual editing module intended to make this tedious task as easy as possible. The program displays two data sets in horizontally aligned data-browser windows. The program allows independent scrolling in both windows with a user-selectable view of different variables. Data sets can also be sorted according to variables of the user's choosing. The implementation of the AGREP routine means that searching by wild cards is also possible. The actual manual linkage is performed by pointing and clicking with a mouse. Optionally, cases already matched can be hidden from view.

3 A Comparison of String-Similarity Measures Using German Names

Having created a new tool for deterministic record-linkage that employed various string-similarity measures, we conducted an experiment to compare the performance of those measures. A string-similarity measure for the intended application should perform well on German names with human-generated errors. Furthermore, it should have an acceptable running time for data sets that contain even 2.2 million records. The test should be conducted with files with known true links and exactly one true link per key.

In order to test the algorithms, two studies were conducted. The results of an experimental comparison were reported in Schnell et al. (2003), here we will report a comparison based on an actual application. We tried to link a file of 5.092 persons with a second file of 29.542 observations (both files were membership lists of voluntary organizations).

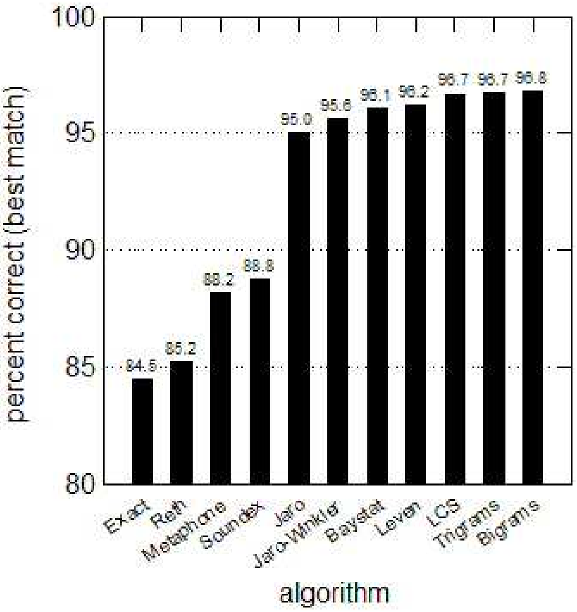


Figure 4: Correct links compared to a manual solution

Both files had a common hierarchical structure, so comparisons within 76 blocks of observations were possible. In total, each algorithm computed 2.390.142 pairwise comparisons using names and surnames. The computed best matches for each algorithm were compared with the results of a matching operation of an expert. Figure 4 shows 84.5% of exact matches in comparison of 85.2%-88.8% correct matches by using phonetic algorithms. In contrast to this result, the edit-distance-algorithms reached more than 95% of the manual solution. Interestingly, some of the phonetic algorithms need long computing times (compare figure 5). If speed and accuracy are equally important, we therefore would recommend Jaro, otherwise we would recommend Bigrams or LCS. One of the most interesting results of our comparisons is the performance of the phonetic algorithms. In all our studies, the phonetic algorithms form a distinct cluster. They perform much worse than the non-phonetic algorithms. The explanation of this at first surprising fact is their high specificity: If a phonetic algorithm indicate a match between two possible pairs, it most certainly is a true match. Furthermore, if one of the two files is generated by administrative actions, the keys in this file are probably formal correct but humans use short versions of the keys, for example by omitting titles or reversing the keys. In these cases, phonetic algorithms have no chance to correct this errors.

4 Application

We tried to link survey respondent data with company register data by using sex, month and year of birth, name of the employer and name of the workplace for 614 respondents. Names of the respondents were *not* used. In order to check the linkage, we used a known true-link file (Schnell, Bachteler, & Bender, 2003). We found 234 identical persons in both data sets (correct links) and 91 wrong links. Given response knowledge and a file of

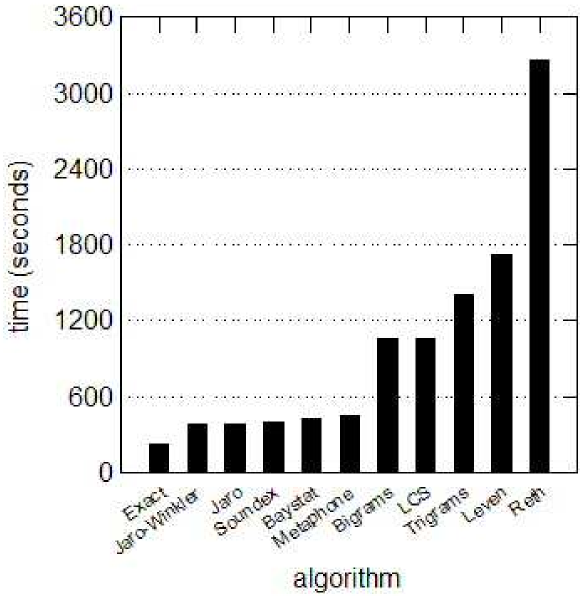


Figure 5: Running time of selected algorithms

the total population the keys used resulted in 38.1% correct links without using names of the respondents or unique identifiers like social security number.

In order to compare this result with a baseline, we conducted a record linkage with exact matching keys only. This yields 81 persons with at least one combination of employer number and company name. After removing statistical twins, 71 persons remained. 44 persons were correctly linked, 22 were wrong links. In sum, using Jaro-Winkler results in a 5-fold increase of linkage compared to exact matches in our data set (7.2% to 38.1%)[3]. Therefore, we feel confident that the record linkage toolbox is even in its current state a useful tool for survey research.

5 Future Work

Our results suggests, that work on pre-processing of potential keys seem to be more important than work on string comparators or similarity threshold selection. Therefore, we will try to develop conditional probabilities for errors in potential keys, which should be conditional on the assumed generating process of the keys. For example, we noted that respondents in work history surveys seem to use different cognitive heuristics to generate acronyms for work-places, company-names or organizations. We decided to implement a special acronym-generator for company-names in the preprocessing tool. The current version removes common company postfixes and builds a string of each initial letter after parsing the key at first along common delimiters (blanks, dashes, slashes) (type d-key) and then according to hyphenation rules (h-key). If a short string given by a respondent matches a d-key or a h-key, the key-pairs are considered as potential matches. Currently, we are experimenting with different weighting schemes for the similarity computing.

[3] This number (7.2%) had been previously reported incorrectly as 3.6% in Schnell et al. (2003, p.3716).

Nevertheless, our further work at the linkage module will concentrate on implementation of an EM-type probabilistic record linkage module, implementation of further string similarity measures and a comparison with neural nets for pattern recognition of names[4]. If the speedup of the program by algorithmic techniques will not result in acceptable running times (we try to accomplish the linkage to two national data bases within an hour), we will build a small simple LINUX-cluster, of which each cluster will be assigned to one linkage block.

Linked data-bases with human generated errors for which a true linkage is known are rare. Even the few known data-bases are not available due to data protection laws. Therefore, we are still in the process of locating such data-bases with German names as keys. We have now gained the permission to test our programs with two multi-million records federal data bases, which had been linked manually. We hope to publish the result of this comparison soon.

Issues of data protection are taken very serious in Germany. Record-Linkage without written permission of each respondent or patient in a data base must be approved by local and federal data protection agencies. To facilitate the approval process, a new linkage procedure is needed. The procedure should ensure privacy of the respondents by use of encrypted keys. Furthermore, the procedure should tolerate errors in linkage keys despite encryption. Recently, we started work on similarity-measures for strong cryptographic keys with errors in generating keys. Currently, we are experimenting with encryption functions based on augmentation of keys with randomly selected bigrams.

References

Borgman, C. L., & Siegfried, S. L. (1992). Getty's synoname and its cousins: A survey of applications of personal name-matching algorithms. *Journal for the American Society of Information Science, 43*(7), 459-476.

Fair, M. (1995). An overview of record linkage in canada. In *Proceedings of the social statistics section* (p. 25-33). American Statistical Association.

Fürnrohr, M., Rimmelspacher, B., & Roncador, T. von. (2002). Zusammenführung von Datenbeständen ohne numerische Identifikatoren: ein Verfahren im Rahmen der Testuntersuchungen zu einem registergestützten Zensus. *Bayern in Zahlen*(7), 308-321.

Gill, L. (2001). *Methods for automatic record matching and linkage and their use in national statistics*. Norwich: HMSO.

Guth, G. J. A. (1976). Surname spellings and computerized record linkage. *Historical Methods Newsletter, 10*(1), 10-19.

Hirschberg, D. S. (1977). Algorithms for the longest common subsequence problem. *Journal of the Association for Computing Machinery, 24*(4), 664-675.

Knuth, D. E. (1998). Sorting and searching. In *The art of computer programming* (2. ed., Vol. 3, p. 394-395). Reading/Mass.: Addison-Wesley.

[4] Details about program status and availability can be found on the project homepage: `http://www.uni-konstanz.de/FuF/Verwiss/Schnell/recordli.html`

Lait, A., & Randell, B. (1996). *An assessment of name matching algorithms* (Tech. Rep. No. 550). Department of Computing Science, University of Newcastle upon Tyne.

Monge, A. E., & Elkan, C. P. (1996). The field-matching problem. algorithms and applications. In E. Simoudis, J. Han, & U. Fayyad (Eds.), *Proceedings of the second international conference on knowledge discovery and data mining: Kdd-96* (p. 267-270). Menlo Park: AAAI Press.

Philips, L. (1990). Hanging on the metaphone. *Computer Language*, *7*(12), 39-43.

Philips, L. (2000). The double metaphone search algorithm. *C/C++ Users Journal*, *18*(6).

Pollock, J. J., & Zamora, A. (1984). Automatic spelling correction in scientific and scholarly text. *Communications of the Association of Computer Machinery*, *27*(4), 358-368.

Porter, E. H., & Winkler, W. E. (1997). Approximate string comparison and its effect on an advanced record linkage system. In W. Alvey & B. Jamerson (Eds.), *Record linkage techniques: Proceedings of an international workshop and exposition.* (p. 190-199). Arlington, VA.: Office of Management and Budget.

Postel, H. J. (1969). Die Kölner Phonetik. Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse. *IBM-Nachrichten*, *19*, 925-931.

Reth, H.-P. von, & Schek, H.-J. (1977). *Eine Zugriffsmethode für die phonetische Ähnlichkeitssuche* (technical report No. 77.03.002). Heidelberg: IBM Scientific Center.

Schnell, R., Bachteler, T., & Bender, S. (2003). Record linkage using error prone strings. In *Proceedings of the joint statistical meeting* (p. 3713-3717). American Statistical Association.

Taft, R. L. (1970). *Name searching techniques*. Albany, N.Y.: Bureau of Systems Development.

Ukkonen, E. (1985). Algorithms for approximate string matching. *Information and Control*, *64*(1-3), 100-118.

Ukkonen, E. (1992). Approximate string matching with q-grams and maximal matches. *Theoretical Computer Science*, *92*(1), 191-211.

Author's address:

Univ.-Prof. Dr. Rainer Schnell
Research Methods in Public Policy and Administration Science
University of Konstanz
P.O. Box 5560
78434 Konstanz Germany

Tel. +49 7531 3602
E-Mail: Rainer.Schnell@uni-konstanz.de
Homepage: `http://www.uni-konstanz.de/FuF/Verwiss/Schnell`