Austrian Journal of Statistics
Volume 33 (2004), Number 1&2, 55-67

Data Matching for the Maintenance of the Business Register of Statistics Austria

Alois Haslinger
Statistics Austria, Vienna

Abstract: The Business Register of Statistics Austria is the basic instrument for all surveys conducted in economic statistics. For the maintenance mainly four different administrative sources are used. Unfortunately, the units of the different registers do not agree exactly and there is no unique numerical key in the business register and the administrative registers. Each register uses its own key. The units of an administrative register belonging to a certain unit of the business register have to be found by comparing alphanumerical items like name and address. For that purpose we use the method of N-grams after some parsing and standardising of the texts. With that method above 90% of the profit-oriented units of the business register could be linked with a corresponding unit of the tax register (these linked units account for 99% of total turnover). 80% of the links were found fully automatically, the rest was checked manually.

Zusammenfassung: Das Unternehmensregister (UR) der Statistik Austria ist eine wesentliche Voraussetzung für alle wirtschaftsstatistischen Erhebungen. Für die Wartung dieses Registers werden hauptsächlich 4 administrative Datenquellen eingesetzt. Leider stimmen die Einheiten dieser externen Register nicht völlig mit den Einheiten des Unternehmensregisters überein und in jedem Register werden andere Schlüsselbegriffe verwendet. Die zusammengehörigen Einheiten des UR und eines anderen Registers müssen daher erst über einen Textabgleich von Namen und Adressen gefunden werden. Für den Abgleich der vorweg standardisierten und geparsten Textfelder setzt Statistik Austria die Methode der N-Gramme ein. Auf diese Art und Weise konnten über 90% der auf Gewinn ausgerichteten Unternehmen im Steuerregister gefunden und verknüpft werden (auf die verknüpften Einheiten entfällt an die 99% des Umsatzes). 80% der Verknüpfungen konnten vollautomatisch hergestellt werden, der Rest wurde manuell überprüft.

Keywords: data matching, record linkage, business register, bigram.

1 Introduction

It is a challenge for Statistics Austria in operating a Business Register (BR) to keep the units themselves and the data of all units up-to-date, because of various reasons:

- The BR is a basic instrument for all surveys conducted in economic statistics and even for some surveys in social statistics.

- Administrative registers (AR) are the main source of information for maintaining the BR because Statistics Austria is no longer allowed to perform an economic census. AR are the most important way to find births, deaths and changes of units of the BR.
- The requirements of the statistical BR have not been considered in the business processes of the AR, which are the sources of the BR (simply because the AR used for maintaining the BR started long before the BR was founded).
- The units of the AR do not conform exactly to the units in the BR, certain information in the AR is incomplete or wrong (e.g. the ÖNACE classification = Austrian version of the NACE classification of economic activities) or the timeliness in the AR is different from that in the BR.
- The greatest problem is the non-existence of a unique numerical identifier for the units in different registers.
- The matching of the register units has to be done by comparing mainly text fields like name and address of the company. These fields are not standardised and of different length in different registers.

This paper reviews how Statistics Austria solved some of the problems of maintaining the BR using administrative data. In the next section a rough overview of the structure and content of the BR is given. In section 3 some peculiarities of the four most important administrative sources for the maintenance of the BR are described. The method of N-grams and the standardising and parsing procedures for the linkage of different registers are presented in section 4. The paper ends with a report about the status of the linkage of the AR with the BR (section 5) and a short outlook (section 6).

2 The Business Register of Statistics Austria

The Business Register of Statistics Austria serves as an instrument for all surveys conducted in economic statistics. It has been designed according to the requirements of Council Regulation No. 2186/93 on business registers for statistical purposes within the EU and contains about 400.000 enterprises including their establishments and local units. All in all it has about 530.000 active and 120.000 inactive units (i.e. enterprises, establishments and local units) and has been in operation since mid-1995. Copies of historic files of the BR are kept also and accessible at any time.

The BR is a central register held in Statistics Austria for statistical purposes. It is designed to comply as far as possible with European requirements, but generally does not recognize any difference between a legal unit and an enterprise, like the registers of most EU member states.

The BR is used mainly for the following purposes:

- for the compilation of a directory of the statistical units included in surveys related to economic statistics for the mailing of survey forms,
- as a frame for the selection of the sampling units for a sample survey,

- as a basis for the grossing up of the results of sample surveys,
- for checking whether the survey forms have been returned by all relevant units,
- for improving the consistency of the results of different economic surveys,
- as a basis for enterprise demography: the register tracks changes of all units in time, i.e. their births, deaths, mergers/takeovers, break-ups and split-offs as well as changes of attributes observed on individual types of statistical units in the BR,
- as a basis for analyzing the economic units according to their regional or activity classification or some other attributes.

The BR covers three types of units:

- Enterprises: usually an enterprise corresponds to a unit registered in the register of corporations or to a physical person acting as an entrepreneur. Only in some cases, enterprises may consist of more than one legal unit.
- Establishments: come into being through the profiling of large important enterprises with heterogeneous activity. For practical reasons separate establishments within an enterprise are only registered if each subunit is a separate cost-accounting unit. The recording of establishments in the BR is not dictated by the EU regulation on business registers but is required for EU short-term economic surveys and for surveys of industrial output and structures. In addition, statistics are required at establishment level (in industry and services) for national and regional accounts.
- Local units: each statistical enterprise has at least one local unit (at the same location as the enterprise); other local units are recorded for enterprises in multiple locations.

The units of the BR have the following structure: Roughly 90% (360.000) of all active enterprises of the BR consist of only one establishment and one local unit. They are, in each case, shown as a single register unit, since enterprise, establishment and local unit are identical in all characteristics.

About 9% (38.000) of the active enterprises consist of a single establishment and more than one local unit (120.000 local units in total).

The BR has only around 1.000 multi-establishment enterprises (2‰ of the universe) with 2.500 establishments and 6.000 local units.

The BR covers all units of the market-oriented economy including the free-lance professions (e.g. physicians, lawyers, trustees, civil engineers) and also public and private non-profit organisations (e.g. clubs and societies) almost completely. Agricultural and forestry holdings are only included in the BR if they also perform another kind of activity in addition to one in ÖNACE sections A and B.

The number and kind of attributes maintained for each unit of the BR depends on the kind of unit. Generally, the items can be classified in 6 groups:

- identification items: identification number, type and status of unit, name of unit,
- address data: municipality, NUTS-3, postcode, address, reference number to the address in the register of buildings, which provides a standardised spelling of the addresses,

- mailing items: postal address, name of a contact person, telephone and fax number, identification number of a trustee, code for the scope of a survey,
- demographic items: date of birth and death, date of begin and end of economic activity,
- classification items: ÖNACE for the principal, secondary and ancillary activities of the units, code of activity according to the classification of the Austrian Federal Economic Chamber, number of employees, turnover, legal form or type of natural person, institutional sector,
- reference to administrative data sources: membership numbers of the Austrian Federal Economic Chamber, company number of the register of companies, identification codes of the social security register and the VAT payers register.

Technically, the BR was set up as a relational database in a DB/2 system. One of the key advantages of this system is that it can be extended at any time by the addition of (new) characteristics.

The multiplicity of characteristics per register unit and the complex structure of the units necessitate numerous plausibility checks to be run after editing a unit (currently around 90 criteria have to be checked). All individual register units with all the characteristics applicable to their type and their enterprise structure can be accessed online by staff members. In addition, the BR offers a programme for rapid search for units by the most important characteristics (direct search by index fields) such as identification number, unit type, name location, mailing address and ÖNACE code. Search conditions may also be defined individually by the logical connection of several characteristics.

3 Administrative Registers used for the Maintenance of the Business Register of Statistics Austria

3.1 Register of the Austrian Federal Economic Chamber

Before 2000 the most important administrative source of information for the updating of the BR were data on new members provided monthly by the Austrian Federal Economic Chamber. Every physical and legal person who wants to pursue a business has to apply for a trade licence. A separate licence is necessary for each different trade in which someone wishes to engage. The register of members of the Economic Chamber has about 350.000 entries.

The register of the Economic Chamber still is an indispensable source of information for maintaining the quality of the BR (local units, status of enterprises, at least also for specific tabulations needed by the Chamber of Commerce classified by the chamber-specific activity code). Monthly, Statistics Austria receives information on new, deleted or changed memberships and licenses as well as on the whole stock of businesses, whose owners are members of the Chamber.

Unfortunately, not every economic unit has to become a member of the Federal Economic Chamber, e.g. that holds for physicians, lawyers and civil engineers, which have their own chambers.

3.2 Register of companies

The register of companies is a public electronic register which is operated by special courts. It keeps record and informs about all facts, which have to be stored according to commercial law (name and address of the company, company number, legal form, enrolment of the submission of accounts, changes of the persons authorised to represent the company).

The register of companies contains only about 150.000 entries of corporations or merchants who have been entered as such in the register. Generally, such a person must have a net turnover above 400.000 € per year or above 600.000 € for food retailers or general stores.

Until recently, the information of the Register of Companies could be accessed only by separate online queries for each company. Since October 2002 the Register of Companies provides Statistics Austria with a copy of all births, deaths and changed units, monthly. The file contains the identification number, name and address and legal form. This access of the entire register in electronic form is a precondition for an automated matching of units of the BR with that of the register of companies.

3.3 Register of Employers in Social Security Institutions

Each Austrian employer has to register his employees in one of about 20 different social security insurance institutions. It depends on the region and the kind of contract of employment, which insurance institution is responsible for a certain employee. It is possible that the employees of an employer are registered at two or more different insurance institutions (e.g. if an employer has local units in more than one province). For each employment of a certain person by a certain employer, a data record is stored in the responsible social security insurance institution containing, among other things, the social security number of the person, an identification code of the employer, a code of the insurance institution, sex of the person and the kind of contract of employment.

Because of the federal organisation of the social insurance system most of the institutions are member of an umbrella organisation called Main Association of Austrian Social Security Institutions. The Main Association has access to the employment registers of its members and additionally maintains a register which holds one record for each combination of insurance institution and employer. This register of employer accounts (SSR) contains the name of the employer, postcode, address, place of the enterprise, NUTS-3 and ÖNACE codes, and contains about 350.000 units. The units of this register are not comparable with the units of the BR. Usually, one enterprise of the BR consists of 0 to n units of the Social Security register.

At present, Statistics Austria receives a copy of the register of employments and of employers monthly.

3.4 Register of Tax Payers

The most comprehensive administrative register used for the updating of the BR is the register of the tax authorities. It contains basic information like name and address, date of birth, sex and civil status (the last 3 primarily for persons), legal status and economic classification according to ÖNACE (primarily for enterprises) for about 5 million taxable units (persons, business partnerships, corporations, institutions, associations, …). The coverage of this basic tax file is much broader than that of the BR. To get a sub-file from the basic tax-file which is comparable with the BR it has to be merged with the turnover taxation file from the tax authorities containing about 600.000 units.

Statistics Austria receives both files of the tax register quarterly. Both files include a unique subject identification key which can be used for merging the two files. The turnover taxation file contains all units from the basic file which did a turnover tax return in at least one of the last 3 years. A problem is the lag between a fiscal year and the time, when all units have received their tax assessment. This lag is about 2-3 years. To get a realistic value of total turnover in 2000 you have to wait at least until mid 2003. The merged file of that date covers units which are no longer active at present. On the other hand, in the merged file units of the BR are lacking which are not liable for turnover taxation (e.g. turnover from medical activity). Nevertheless, most of the units of this merged file are in accordance with the enterprises of the BR.

From the start of 2003 on, the problem of the time lag of the turnover returns has been reduced, because now each enterprise with a turnover above 100.000 € in a year has to do a monthly turnover tax advance return beginning with January of the next year. Therefore, new enterprises are registered earlier than in the past in the basic tax file. Statistics Austria receives data on the turnover tax advance returns monthly.

3.5 Other Administrative and Statistical Sources

There are also other administrative or statistical sources which are used for the maintenance of the BR. Contrary to the above mentioned 4 main sources, these additional sources are used predominantly by the staff of the unit responsible for the manual updating of the BR. The main reasons that these sources are only used for interactive updating are that the information is not supplied in an electronic file which could be used as input for our text comparison program and/or that the information is only supplied on occasion but not on a regular basis. Another possibility is that information about a unit is received in the course of a statistical survey allowing an immediate online update of that unit in the BR.

Some examples of such sources are:

- online queries of the register of reliability of the borrowers (Kreditschutzverband),
- membership directories of the Medical Chamber, the Chamber of Lawyers, of Civil Engineers, Patent Agents, Notaries etc,
- respondents of the tourism statistics, freight transport statistics, structural and short-term business statistics and INTRASTAT,
- information from the notification of ÖNACE code: Each enterprise of the BR is informed about the ÖNACE code it has been classified by Statistics Austria. If the enterprise does not agree it can raise an objection.

4 Method used for record linkage of the BR with Administrative Registers

The AR contain name and address, legal form, number of employees, economic activity according to the ÖNACE classification and turnover of each enterprise. This information is very important not only for the maintenance of the BR itself but also as a substitute for some economic surveys. Before conducting a survey Statistics Austria is obliged by law to verify whether the whole survey or some items of it can be replaced by using information from an AR. Additionally the administrative data are now the most important way to find births and deaths of the BR.

There are however some problems in matching data from different registers: The units of the AR do not confirm exactly with the units in the BR, some information in the AR is incomplete or wrong (e.g. the ÖNACE classification) or the timeliness of some units in the AR is different from that in the BR. The greatest problem is the non-existence of a unique numerical identifier for the units in different registers. The matching of the register units has to be done by comparing mainly text fields like name of the company and address. These fields are not standardised and of different length in different registers. Fortunately, both the BR and the different AR store the postal and/or municipality code for each unit which diminishes the number of necessary comparisons highly.

Especially pronounced are the differences in the length of the name of the enterprise: the name of each unit of the BR consists of at most 50 characters and the street name of at most 40 characters. The name of the employer in the Social Security register is stored in two fields both of them can have up to 65 characters. The SSR thus has up to 130 characters for the name of the employer compared to 50 characters for the name of the enterprise in the BR. The situation in the tax register is quite converse: name and address of an enterprise are stored together in one field with up to 56 characters. Thus, the field length of comparable items in the tax register is usually much shorter than in the BR, whereas the text fields in the SSR are much longer than those of the BR.

The matching of units of different registers should be an automated process as far as possible, since a manual operation would be too time and/or person-consuming because

of the size of the different registers. To achieve satisfying results with an automated matching algorithm it is necessary that the compared text fields of the same unit in different registers are not written too differently. Before the text variables of two registers are compared for similarity, they must be standardised and parsed. This is essentially a statistical process. It is usually done by computing the frequency of all words in both texts. If the frequency of a string (like 'corp', 'inc', 'ltd', 'doctor') is very different in both registers, either one can delete that string in both registers, abbreviate it identically in both registers or replace it by a synonym at least in one register. An example for parsing is to write all parts of a name in the same order (e.g. family name, first name, title).

Before starting automated matching, the contents of each text field must be analysed and the following questions must be answered:

- Which language do most of the names in the text field come from?
- Are there combinations of letters that have a special phonetic value?
- Does the name field contain personal names, company names or both?
- Does the address field contain street names, city names or both?
- Are there any peculiarities in a text field, e.g. the string 'comp' precedes every company name?
- How was this register collected? Is it based on data collection by telephone (phonetic variations are likely to be found in the data) or by questionnaires (typographical variations and mistakes are probable)?

For standardising one has to choose between the following operations:

- deleting blanks at the begin or the end of a text field,
- converting lowercase characters in uppercase,
- deleting village name from the field for street name,
- converting special characters ('&' to 'U', 'Ä' to 'AE',…),
- converting Latin numerals in Arabic numerals,
- moving the most crucial word to the beginning of a text field,
- truncating of long words.

After accomplishment of the above steps, Statistics Austria uses the method of N-grams for matching a unit of an AR with a unit of the BR (this has the role of a dictionary). Matching takes place after the text fields have been standardised and parsed. Now a text field (or phrase) consists of words (or strings) separated by blanks and a word consists of some of the 26 capital letters of the alphabet. A unit of an AR is compared to some or all entries of the BR and the degree of similarity of selected text fields is computed. For this purpose compared text fields are disaggregated in all of their N-grams. An N-gram is a sequence of N successive characters of a string. For instance, the field 'MAYER KARL' can be split into 7 overlapping Bigrams 'MA', 'AY', 'YE', 'ER', 'KA', 'AR', 'RL' or into 5 Trigrams 'MAY', 'AYE', 'YER', 'KAR', 'ARL'.

If M(a) denotes the set of all different N-grams of a text field a and $|M(a)|$ its cardinality (number of different N-grams, several identical N-grams are counted only once) then the similarity S(a,b) of text fields a and b can be measured as

$$S(a,b) = 100 \frac{|M(a) \cap M(b)|}{\sqrt{|M(a)| * |M(b)|}} \quad (1)$$

As an example, when using bigrams, the similarity between a='MAYER KARL' and b='MAIER KARL' will be measured as

$$S(a,b) = 100 \frac{|\{'MA','ER','KA','AR','RL'\}|}{\sqrt{7*7}} = 100 \frac{5}{7} = 71.43 \quad (2)$$

A value of 0 stands for no common N-grams and a value of 100 signifies that the two compared phrases are identical. We have also experimented with other similarity measures but found no great differences between all of them. In the special situation when the text field in one register is generally shorter than the corresponding one in the other register it can be of advantage to divide by the minimum number of bigrams in the two strings instead by the geometric mean.

The N-gram method is useable for each language. It has also the ability to cope with spelling errors. For example, if 'VIENNA' was erroneously written as 'VIENA' then 4 of the 5 bigrams, but only 2 of the 4 trigrams, would still be in agreement. This example demonstrates that bigrams are better proof against spelling errors than trigrams. For the matching of registers we use therefore bigrams.

The N-gram method is also robust against permutations of the words in a phrase. The two phrases 'Technical advisor' and 'Advisor, technical' are very similar. The exact value of the similarity measure depends on the character set which is permitted for N-gram construction. If only capital letter N-grams are used (and no special characters like a blank or a comma) then the above-mentioned phrases bear a similarity of 100.

Fortunately, for N>1, the N-gram-method is not robust against permutations of the characters of a word. The meaning of a word is influenced essentially by the sequence of the characters. Usually, two words consisting of the same characters but in different order, also have a different meaning and generally should not be declared as similar. This characteristic can raise problems of misspellings of short words, e.g. one wrong character in a 3-digit word makes the similarity measure go down to zero. This weakness can be tempered by using blanks at the beginning and at the end of the phrase and by admitting a blank as a valid character for N-grams.

The idea to link the BR with an AR by using the bigram method came up already at the time when the BR went on line. It seemed useful to match the units of the business register of Statistics Austria with the corresponding units of the Federal Economic Chamber and to store the identification numbers of the Chamber-units on the corresponding records within the BR. With this linkage of the keys, it should be possible to take over the ÖNACE code or other information from the BR of Statistics Austria to the Chamber units or vice versa.

The two registers were matched by comparing the units' names and addresses for similarity. For each record of the BR, an attempt was made to find at least one corresponding record in the register of the Economic Chamber with a high similarity of name and address. The search for a given record of the BR was restricted to all Chamber-units belonging to the same municipality as the BR unit (blocking).

With the help of a batch program (see Müllauer, Statistics Austria 2003), the BR was divided into the following 4 subfiles or quality classes:

- File 1: All enterprises of the BR for which exactly one unit with a similarity measure of at least 80 was found in the AR of the Economic Chamber.
- File 2: All enterprises of the BR with at least two records of the Chamber register with a similarity above 80 or with at least one record with a similarity between 67 and 80.
- File 3: All enterprises of the BR with at least one record of the Chamber register with a similarity of name and address between 50 and 67.
- File 4: All remaining enterprises of the BR.

The matched records of file 1 were considered as identical without further manual control. For the matched records of file 2, specialists had to check these cases interactively with the help of an online application. On the screen, all the records at issue were presented with the suggested candidates for a match sorted by name and address together with additional characteristics from both registers. The person in front of the screen had to mark the records belonging together. File 3 was checked manually. In the few cases where identical records were found, the identification numbers of the BR and the Chamber register were written on a sheet and captured on a EDP file later. File 4 was regarded as containing all cases for which no counterpart could be found in the register of the Economic Chamber. Compared to a manual search the bigram method has accelerated the linkage of corresponding units by a factor of 30 (Haslinger 1997).

5 Status of Linkage of the Administrative Registers with the BR

Since the Federal Statistics Act 2000 has become effective, Statistics Austria also regularly receives extracts from the AR of Corporations, of Social Security and of the Tax authorities on EDP media. This information is not only used for the updating and maintenance of the BR but also as a partial or total surrogate for censuses and surveys. The idea is to reduce the response burden as much as possible. If any information which is needed for statistical purposes is already stored somewhere in the public administration, then Statistics Austria should get the information from the AR and not survey the enterprises or the citizens again.

This change in the basic principles on how statistics should be produced was not accompanied by integrating statistical requirements into the business processes of the external registers. The units of the AR do not exactly agree with the units of the BR and each register has its own system of identification keys for the units. Statistics Austria

decided that the most suitable kind of unit of the BR for linkage with all the external AR is the enterprise and that for each AR a table should be created in the DB/2-database of the BR where each enterprise gets assigned the identification keys of the corresponding units in external registers.

The first filling of these tables was done for each AR more or less similar to the filling of the table for the register of the Federal Economic Chamber. Due to insufficient manpower, quality class 3 could not be handled manually for each AR. The thresholds of the similarity measure defining which quality class a relation of two identification keys belongs to were determined empirically by checking a sample of automatically detected relations manually.

During the work for establishing links between the units of the BR and their corresponding units in the register of corporations, the register of the Social Security System and the tax register units were detected which were not present in the BR. For registering these missing units in the BR, the search and text comparison program had to be changed slightly. Instead of the AR, the BR was used as a dictionary: for each unit of an AR which had not yet been linked, a matching unit of the BR was looked for. At best the link with the highest similarity had a value below a certain threshold. In that case the external unit was considered as a new unit and was registered in the BR together with the relation to its matching unit in the AR. The value of the threshold was also determined empirically in this case. 'New units' of the tax register were considered as really new ones only if they were found also in one of the other AR.

Statistics Austria gets EDP-copies of the 4 most important AR monthly. Thus, the record linkage program for finding new relations and new units of the BR is also started monthly. The manual checking of relations and new units is mainly concentrated on the bigger units which are involved in the surveys for short term and/or structural business statistics.

6 Conclusions and Outlook

Table 1 demonstrates Statistics Austria's success story: In the last 3 years the number of active units of the BR has increased by one third from 300.000 to almost 400.000. This increase is partially caused by a better coverage of the public and non-profit-oriented sector and partially by a lower under-coverage of the profit-oriented sector. In any case, a higher coverage is also a sign of increased quality of the BR.

Table 1: Enterprises of the BR referenced to AR

Date	Active enterprises of the BR	thereof linked to register of			
		Corpora-tions	Economic Chamber	Social Security	Tax Authority
1.4.2001	305.233	77.424	226.978	0	23.941
1.4.2002	302.218	82.436	231.961	179.490	233.398
1.4.2003	389.913	107.704	293.408	223.404	332.467
1.4.2004	392.666	113.776	291.012	239.671	359.488

Not only the coverage has increased but also the number of enterprises which are linked to one or more AR. Now 92% of all active enterprises are linked to a unit in the tax register, 74% to a unit of the Federal Economic Chamber, 61% to at lest one employer account number of the Social Security System and 29% to a unit of the register of Corporations. According to some rough estimate 80% of all links were found fully automatically, the rest was checked manually.

Table 2 allows a more detailed assessment of the degree of linkage between units registered in the tax administration with a positive turnover in the year 2000 (when writing this paper this was the most recent year for which all liable units had done their taxes) and the BR seen from the tax side. When interpreting table 2 one has to bear in mind that table 2 was produced between April 2003 and April 2004. The first impression is that the tax register has many more units liable for tax on sales than there are units in the BR. This is indeed the case because 126.000 units, predominantly units with only income from rent and lease are liable for turnover tax but do not belong to the BR. The second impression is that only 311.000 units with a positive turnover in 2000 were linked to the BR, whereas nearly 360.000 units of the BR are linked to the tax register. The main reason for that difference lies in the fact, that not all linked units had a positive turnover in 2000 (that can happen for two reasons: either the enterprise was not active in 2000 or it was not liable for turnover tax).

Nevertheless, 93.000 units remain which are not linked to the BR, one half from the public and non-profit-oriented part of the economy (in that area missing links are not a problem, since these units have not been used as a sample frame for surveys until now) and the other half from the market-oriented area (which is important to be complete because it is used for surveys and the production of statistics).

Table 2: Status of linkage of the tax register with the business register

	Units	Turnover in billion €
Units of tax register with a positive turnover in 2000	532.000	483
Units linked to the BR	311.000	463
Units not in BR (thus not linkable)	126.000	13
Units in BR with ÖNACE<10 or >74, not yet linked	46.000	3
Units in BR with ÖNACE 10-74, not yet linked	47.000	4
thereof with a yearly turnover above 40.000 €	11.000	3
thereof with a yearly turnover above 20.000 €	9.000	0
thereof with a yearly turnover above 0 €	27.000	0

The 47.000 market-oriented units of the tax register which are not yet linked to the BR account for only 1% of total turnover, i.e. the problem concerns only very small units. One can imagine that a part of these small units is no longer active at the end of 2003 and therefore no longer an active unit of the BR. Nevertheless, the bigger part will consist of active units which should be incorporated into the BR in the future.

Other tasks for the future will be the abolishing of (the small but existing) under-coverage of the BR even in the non-market-oriented area, the locating and removing of doublets, the improving of structures of enterprises and the continuation of the work of

finding links of identical units in different AR on a monthly basis. Quality is a continuous process!

References

Council regulation (EEC) No 2186/93 of 22 July 1993 on Community coordination in drawing up business registers for statistical purposes, 1993.

Bundesstatistikgesetz 2000. Federal Statistics Act of 2000, BGBL I Nr.163/1999, idF BGBL I Nr.136/2001, Vienna, 2001.

A. Haslinger. Automatic Coding and Text Processing using N-grams. In *Conference of European Statisticians. Statistical Standards and Studies – No. 48. Statistical Data Editing, Volume No. 2, Methods and Techniques*, pages 199-209. UNO, New York and Geneva, 1997.

R. Müllauer. TST (MVS) 2.5X. Unpublished documentation about a collection of load-modules and program-skeletons for the purpose of matching two text fields used in Statistics Austria, Vienna, 2003.

Office of National Statistics. Methods for Automatic Record Matching and Linkage and their Use in National Statistics, National Statistics Methodological Series No. 25, London, 2001.

Author's address:

Dipl. Ing. Alois Haslinger
Registers, Classification and Methods Division
Statistics Austria
Guglgasse 13
A-1110 Vienna
Austria

Tel. +43 1 71128 / 7186
Fax +43 1 71128 / 7053
Elec. Mail: alois.haslinger@statistik.gv.at
http://www.statistik.at