Austrian Journal of Statistics
Volume 32 (2003), Number 4, 305-321

# Data Integration and Record Matching: An Austrian Contribution to Research in Official Statistics

Michaela Denk[1] and Peter Hackl[2]
[1] ec3 – Electronic Commerce Competence Center, Vienna
[2] University of Economics and Business Administration, Vienna

**Abstract:** Data integration techniques are one of the core elements of DIECOFIS, an EU-funded international research project that aims at developing a methodology for the construction of a system of indicators on competitiveness and fiscal impact on enterprise performance. Data integration is also of major interest for official statistics agencies as a means of using available information more efficiently and improving the quality of the agency's products. The Austrian member of the project consortium comprises university departments, representatives from the Bundesanstalt Statistik Austria, from the Statistical Department of the Austrian Economic Chamber, and from ec3, a non-profit research corporation. This paper gives a short report on DIECOFIS in general and on the Austrian contribution to the project, mainly dealing with data integration methodology. Various papers that have been read at the DIECOFIS workshop last November in Vienna, will be published as a Special Issue of the Austrian Journal of Statistics.

**Zusammenfassung:** Techniken der Integration von Datenbeständen sind eines der Kernthemen von DIECOFIS, ein EU-finanziertes, internationales Forschungsprojekt mit der Zielsetzung, Methoden für die Konstruktion eines Systems von Indikatoren der Wettbewerbsfähigkeit und der steuerlichen Auswirkungen auf die Unternehmensleistung zu entwickeln. Die Integration von Datenbeständen ist heute von größtem Interesse für die Amtliche Statistik, da diese Techniken es ermöglichen, die verfügbaren Daten effizienter zu nützen und die Qualität der statistischen Produkte zu verbessern. Das österreichische Mitglied des Projektkonsortiums umfasst Universitätsinstitute, Repräsentanten der Bundesanstalt Statistik Österreich, der statistischen Abteilung der Wirtschaftskammer Österreich und des ec3, eines gemeinnützigen Forschungsinstituts. Dieser Artikel berichtet allgemein über DIECOFIS und über den österreichischen Beitrag zum Projekt, der vor allem mit der Integration von Datenbeständen befasst ist. In einem kommenden "Special Issue" der Österreichischen Zeitschrift für Statistik werden die Beiträge publiziert werden, die Mitte November beim DIECOFIS Workshop in Wien vorgetragen wurden.

**Keywords:** DIECOFIS, EU-funded Research, Data Integration, Record Matching, Multi-source Database, Meta-information, Statistical Indicators, Official Statistics.

# 1 Introduction

DIECOFIS (Development of a System of Indicators on Competitiveness and Fiscal Impact on Enterprises Performance, cf. DIECOFIS, 2003) is an EU-funded international research project, coordinated by the Italian national statistical agency ISTAT. The main goal is to foster the development of "best" policy impact and evaluation techniques in the field of taxation, to further the Lisbon objectives and EU governance.

The choice between tax policy convergence, co-ordination, or harmonisation remains a thorny issue in the EU policy agenda. Member countries do not always move synchronically. Yet, as economic and financial integration spreads globally, the unavoidable conclusion is that some form of co-operation is a necessity, no longer an option. One drawback for this policy area is that "facts" on the impact of taxation are charted with a high degree of approximation, in spite of extensive discussions, experts' and working groups' meetings and a crowd of reports. Tax indicators have well-known pitfalls. Understanding how taxes affect economic performance is central to endow the EU with a set of efficient and fair tax policies.

DIECOFIS takes up this challenge of developing a system of micro-founded indicators. It aims at (i) assembling a wide ranging system of statistical information including data from economic, tax and social insurance sources into an *integrated multi-source enterprise database*, and (ii) creating *micro-simulation models* for enterprise taxation in two European countries, Italy and the UK, with a view to eventually producing an "EU demonstrator" as a foundation for the development of similar models in the whole EU. For the creation of such a multi-source database of enterprise data as a basis of micro-simulations of effects of fiscal policy measures on enterprise competitiveness and performance, *data integration*, mainly record matching, is a core issue of the project. Actually, the project shows the importance of data integration as a means of generating comprehensive statistical databases as a sound foundation for deliberate decision making.

From an official statistics' point of view, data integration is of major interest as a means of using available information more *efficiently* and improving the *quality* of a statistical agency's products. By using integration methods, the *value added* that can be extracted from the existing stock of information is greatly augmented. The *Bundesanstalt Statistik Austria*, being an early user of data integration methods, contributed actively in the project by collaborating in an empirical study on measuring multi-source dataset quality and as co organizer of an international workshop.

# 2 DIECOFIS: The Project

Aiming at the development of an appropriate methodology for the construction of a system of indicators on competitiveness and fiscal impact on enterprise performance and the application of the developed methods, the DIECOFIS project is structured into three steps, viz. (i) the development of an integrated multi-source database which can be the basis for a broad range of micro-founded statistical indicators; (ii) the creation of enterprise datasets for micro-simulation purposes, specifically to simulate and monitor the impact of public policy on enterprise performance, that are flexibly modulated ac-

cording to whether one wishes to simulate national or other EU member or EU-wide policies; and (iii) the construction of national policy models that can be integrated into country tax algorithms in order to see which effects one country's rules would produce if applied in another country, or to investigate the impact of EU policies across member countries taking into account their specialization and socio-economic structures.

These three steps were organized in the following way:

- *multi-source, integrated database*: this step was organized in three work packages and covered statistical issues of the integration of cross-section and longitudinal micro data from surveys and/or administrative registers, the conceptualization and development of software for the creation of a multi-source database, and the development of methodology and software for measuring the quality of multi-source, integrated databases.
- *database for micro-simulation purposes*: this step provided data related to a set of virtual companies that are the basis for simulation and monitoring the impact of public policy on enterprise performance. Issues are data validation and incompleteness, methods of updating and projection, general data quality and methods for sensitivity analysis. Within this task, also tax compliance by firms and analysis of responses to tax-rates and enforcement are discussed.
- *system of indicators*: this step included the conceptualisation and development of indicators and of the general framework for micro-simulations that allow for different types of national databases and fiscal rules.

The project was started in November 2001 and will end in February 2004. Consortium members include the Italian national statistical agency, ISTAT, a UK government department, Inland Revenue, academic institutions (the London School of Economics, the Microsimulation Unit from the University of Cambridge, the University of Rome Tor Vergata, the University of Florence, and Wirtschaftsuniversität), other research institutions (CERES and the EU Joint Research Centre, both located in Italy) and an Information technology company, INFORMER SA Computer System & Management Consulting from Greece.

The Austrian member of the consortium is represented by the Division of Business Statistics from Wirtschaftsuniversität (University of Economics and Business Administration) in Vienna. Other involved institutions and individuals are the Bundesanstalt Statistik Austria (the Austrian national statistical organization, short ST.AT) and the Statistical Department of the Austrian Economic Chamber (Wirtschaftskammer Österreich), both important agencies for the Austrian databases on business and industry; members from the Department of Statistics and Decision Support Systems of the University of Vienna, consortium member of an earlier EU-funded project ISMIS (Design of an integrated Statistical Metainformation System) and coordinator of the EU-funded project IDARESA (Integrated Documentation and Retrieval Environment for Statistical Aggregates); and members from ec3 (Electronic Commerce Competence Center), a non-profit research corporation with both business enterprises and university departments as its members. In an advisory council individuals from all partners are represented.

The Austrian member of the consortium mainly was engaged in the first step of the project concerning database integration. Work package 1 was intended to survey available methods of data integration, to provide a critical assessment of different data integration methods with a focus primarily on statistical issues and to provide an overview of assessment criteria for multi-source databases from a theoretical perspective, in particular statistical indicators of multi-source database quality. All these activities have been seen in view of the concrete application within DIECOFIS. Contributions have been produced to all three deliverables of Work Package 1. In particular, the survey of available methods of data integration (Denk and Oropallo, 2002) has been provided as well as a discussion of relative merits of the various methods in the context of databases to be encountered in the national statistics context (Denk, Inglese, and Calza, 2003) and on quality aspects of multi-source databases and related quality indicators (Denk, Inglese, and Oropallo, 2003). In addition, an empirical study has been designed that compares the applicability of various integration procedures in the context of the Austrian business register (a comprehensive register of Austrian companies) and that demonstrates the use of quality indicators for multi-source databases. The results of the study will be added as a supplement to Deliverable 1.3.

On November 13-14, the "DIECOFIS Workshop on Data Integration and Record Matching" took place in Vienna. Hosted by Statistik Austria, consortium members from DIECOFIS were brought together with practitioners, researchers, and developers from other institutions for an exchange of ideas and experiences. DIECOFIS partners from national statistical offices and research institutions have gained experience both in methodological questions concerning database integration and in the application of such techniques. Theoretical aspects primarily include statistical issues, such as the assessment of different approaches and statistical indicators of multi-source database quality. Various national statistical organizations apply corresponding techniques and actively work on their enhancement; some have developed their own software tools. Academics and statistical journals publish on theoretical issues of dataset integration methods such as probabilistic record matching, statistical matching or data fusion. About 40 participants – from eight countries including Canada – heard lectures on the following range of topics: Database Integration Methodology; Applications to Health and Census Data; Applications within DIECOFIS; Other Applications to Business Data; Integration Software. As a proceedings volume a Special Issue of the Austrian Journal of Statistics is in work.

# 3 Data Integration

Data integration is a broad field of research and can be viewed from various perspectives. In DIECOFIS, main emphasis was on *statistical* data integration *methodology* and *quality indicators* for the assessment of different approaches and applications. In particular, *record-based* or *micro* integration strategies bringing together records representing the same (or at least a similar) real-world entity in different micro datasets were investigated. However, also some *technical considerations* with respect to multi-source database characteristics were made since technical problems must be overcome first when integrating data from different databases. More generally speaking, not only the

analysis of technical but rather of *semantic discrepancies* and similarities of data sources is a precondition for actual application of integration procedures (or, even more general, for the application of any statistical method to data from different sources): *data source integration* as a prerequisite of *dataset integration*. A *metadata* oriented approach for the detection and formalised representation of semantic heterogeneities following the ideas and concepts of IDARESA was proposed.

## 3.1 The Database Point of View

According to Sheth and Larson (1990), in particular, three characteristics of multi-source database systems have to be accounted for, viz. *distribution*, *heterogeneity*, and *autonomy*.

In DIECOFIS, distribution and autonomy were not the main problems. The data required for the micro-simulations originated from multiple distributed autonomous source database systems from different organizations that have collected the data for different reasons. Yet, the source data were all made available at ISTAT's database system – so, there were different datasets, but all located on one site, using the same database system.

With regard to heterogeneity, basically, three different types may be discerned, viz. heterogeneity due to *technological* differences, heterogeneity due to *structural* diversity (cf. Chatterjee and Segev 1992) and heterogeneity due to differences in the *semantics* of the data. *Technological* heterogeneity encompasses differences in hardware, operating systems, and in database management systems. *Structural* heterogeneity essentially covers the problem of different data models, subsuming differences concerning which real-world entities are represented, which characteristics of these entities are included in the data model, as well as differences in data types, data formats, measurement units or granularity of corresponding attributes.

*Semantic* heterogeneity occurs when there is a disagreement about meaning, interpretation, or usage of the same or related data. Equally named attributes may refer to different characteristics, and, vice versa, attributes referring to the same characteristic may be named differently. Similarly, on the entity level, the same entity may be identified by different identifiers in different databases (*synonyms*), or different entities may be identified by the same identifier (*homonyms*). Moreover, meaning of codes are often local to databases, as are the representation and meaning of missing or incomplete information. The temporal validity of data may also be heterogeneous.

Ventrone and Heiler (1991) point out the problem of *domain evolution* referring to the changes in the meanings of the real-world counterparts of domain values as a source of semantic heterogeneity, which might occur even in a single database. One example of domain evolution is termed *heterogeneous instances*: over time, different occurrences of the same value may have different meanings. For instance, enterprises may be split up, merged, bought up by other enterprises etc. If (one of) the "new" enterprise(s) has one of the "old" identifiers, it is not sure if it makes sense to regard those two enterprises to be the same – they are heterogeneous instances (see also Pu, 1991). This is one of the most serious problems when integrating enterprise data as necessary in DIECOFIS – even if there are identifiers for individual records that allow deterministic matching of data, one cannot be sure that the matched records really refer to identical real-world

entities. Further examples of domain evolution are encoding changes which may occur when switching from a nomenclature to a new version of that nomenclature, or time and unit differences.

Usually, structural heterogeneity and most types of technological heterogeneity are quite straightforward to resolve, since they can be tackled on a rather general level – for instance, a strategy how to translate between data models of different organizational types, such as relational and object-oriented, can be developed independently of a particular application. Obviously, things become a lot more complicated if semantic heterogeneity is involved. Typically, database management system schemas do not provide enough semantics to interpret data consistently. Only an increased integration of metadata (as information on actual data) into information systems enables consistent handling and interpretation of data and detection of semantic heterogeneity.

## 3.2 Meta-Information – a Prerequisite

Ventrone and Heiler (1991) discuss the need for metadata usage to facilitate the solution of the semantic heterogeneity problem in more detail. They argue that (i) metadata should be used to capture the semantics of domains, (ii) comparison operators for these metadata should be provided to enable detection of differences in domain semantics, (iii) the life cycle of data and metadata should be synchronized, for which the creation of semantic information for derived data must also be supported, and (iv) that some kind of relevance evaluation is required, allowing to determine whether particular semantic differences affect the results of an application.

Froeschl (1999ab, 2004) and Froeschl and Grossmann (2000) draw similar conclusions. They emphasize the representation of semantic overlaps and discrepancies in a joint data context and the explication of processing levels of datasets through a universal process model. That is to say, standardized meta-information providing data documentation and setting up a unified data context is required to enable the joint usage of data sources (*data source integration*), and thus, also the application of statistical data integration methods to datasets from different sources (*dataset integration*).

In IDARESA (cf. for instance IDARESA 1997 and 1998ab), a coherent formal metadata framework for the joint statistical usage of data from different sources was developed, essentially based on previous work of Froeschl (1997). Continuing IDARESA research, Denk and Froeschl (2000) propose a data mediation approach (cf. Wiederhold 1992, Wiederhold and Genesereth 1997) to overcome semantic heterogeneities and integrate sources, and Denk (2002) includes actual statistical dataset integration methodology into the IDARESA metadata framework.

In DIECOFIS, integration of data sources has not been formalized in a metadata framework; only minor semantic differences have been detected, and then been resolved and documented manually.

## 3.3 Statistical Data Integration Methodology

Once integration of data sources has been realized, statistical datasets can be integrated. According to D'Orazio, Di Zio and Scanu (2001), two broad classes of integration pro-

cedures can be distinguished, viz. (i) *micro procedures* integrating datasets at record level by combining records representing the same (or a similar) real-world entity in different datasets, and (ii) *macro procedures* where the main interest is on aggregates of the integrated data.

Several different terms are used for micro data integration: the most common seem to be *object* or *instance identification* (e.g., Neiling, 1998, Neiling and Lenz, 1999 & 2000, Wang and Madnick, 1989), *record matching* (e.g., Fellegi and Sunter, 1969, Winkler, 1995, Fair and Whitridge, 1997, FCSM, 1980, or Alvey and Jamerson, 1997), and *data fusion* (e.g., Raessler, 2002). *Exact* and *statistical matching* procedures as well as *imputation* methods fall into this category, while the macro category encompasses all kinds of *weighting* procedures (e.g. adapting estimates resulting from surveys in order to comply to population structures or parameters) and procedures for combining summary level data into one single table, as for instance Malvestuto's *Universal Table Model* (e.g., Malvestuto 1989, 1991, 1993).

Exact matching is used when datasets with substantial overlap (with regard to observed entities as well as variables) have to be integrated, and matching of records belonging to identical entities is aimed at. If this is not possible (or not even necessary), e.g., because of different survey samples that rarely overlap, statistical matching (as an approximation of exact matching) can be used. (For a discussion of exact and statistical matching see, for instance, FCSM 1980). Imputation and statistical matching are also closely related: imputation replaces missing or obviously erroneous values in a dataset, while statistical matching inserts values for variables not originally included in the survey.

For the creation of the DIECOFIS integrated and systematised enterprise statistical information system needed for micro-simulation purposes, exact matching was used to combine administrative data (from the business register, commercial accounts, tax returns and foreign trade) and survey data. Statistical matching was relevant to integrate different ISTAT surveys (like structural business statistics and industrial production) that do not contain the same enterprises in order to reduce responder burden. Imputation was applied to complete still missing data. In the Austrian empirical study where ST.AT's business register was integrated with tax authority data, only exact matching was used.

### 3.3.1 Exact Matching

Obviously, in case of availability of identifiers valid in all datasets to be combined, integration simply amounts to a natural database join on the basis of these identifiers. Yet, this ideal situation is rather unlikely. Even if datasets contain identifiers, their equivalence across datasets of different data sources is not necessarily provided. So usually, other identifying characteristics (such as names or addresses of persons or enterprises) have to be taken into account which, in general, does not allow unique identification of identical units. Basically, exact matching methods classify all record pairs that can be built from source datasets into *non-links*, *possible* (i.e. *indeterminate*) *links*, and *links*. Possible links are then clerically reviewed, and in most cases, linked pairs are checked to obtain a 1:1-assignment of records. In practice, in order to reduce the number of pairs that have to be investigated by the matching procedure, the set of all record pairs is decomposed into (i) *blocks* containing candidate pairs that agree on selected blocking

variables which are then further analysed, and (ii) a residual set of determinate non-linked pairs that do not satisfy blocking criteria.
The following subsections briefly introduce exact matching methods. However, in most real-world applications, a combination of available methods seems to work best. A quite common pragmatic approach is to use deterministic linkage, followed by probabilistic linkage (including string comparators, if necessary), followed by clerical review (Gill, 2001).

### 3.3.1.1 Quality Classes

In the quality class approach record pairs are assigned to different *compliance* or *quality classes* of record pairs based on their extent of agreement or disagreement on specified matching variables. By this means, a hierarchy of compliance classes is established. Record pairs in classes with high compliance ("*high quality match*") are linked, those in classes with low compliance are designated as non-links. Pairs in between are sent to clerical review.

Usually, selection of variables as well as definition of classes is based on experience. Otherwise, the method is quite of an ad-hoc nature which makes the results hard to interpret. It is quite easy to implement and easy to use; yet, a disadvantage is that the clerical review region might be large. There is no statistical model underlying. Anyhow, there is danger of overfitting, since there are many parameters to be set (selection and combination of variables, setting of thresholds, and designation as link/non-link). Matching systems working with compliance classes might have to be adapted very often. If training samples with true matching status are available, a justification of used class definitions might be achieved by statistical classification algorithms, such as classification trees (cf. Breiman et al., 1984).

.

### 3.3.1.2 String Comparator Metrics

When comparing values of string variables like names or addresses, it usually does not make sense to just discern total agreement and disagreement. Typographical error may lead to many incorrect disagreements. Several methods for dealing with this problem have been developed: string comparators are mappings from a pair of strings to the interval [0, 1] measuring the degree of compliance of the compared strings (Winkler, 1990). String comparators may be used in combination with other exact matching methods, for instance, as input to probabilistic linkage, discriminant analysis or logistic regression. The simplest way of using string comparators for exact matching is to define compliance classes based on the values of the string comparator.

In order to make reasonable comparisons of string variables, adequate pre-processing by *standardizing* (i.e., replacing words of little distinguishing power with consistent abbreviations) and *parsing* (decomposing a string variable into a set of string components which are then individually compared) the strings is essential (cf. Winkler, 1995). This holds, in particular, when matching business data, since inconsistencies of name and address information are typically even greater for this kind of data (Winkler, 1999). Problems with addresses are due to different types of addresses that might be used by an enterprise in different situations, such as the mailing address, the physical address, or

the address of the lawyer. The only method that will work even if the order of different components of a string variable is not fixed for all records is the bigram method.

A common string comparison method consists in comparing the *bigrams* that two strings have in common. A *bigram* is two consecutive letters of a string. The return value of the bigram function is the total number of common bigrams in the two strings divided by the average number of bigrams in the two strings (Porter and Winkler, 1997). Other bigram variants use a different denominator: instead of the average number of bigrams the number of bigrams in the first (or in the second) string is used. Bigrams are known to be a very effective, simply programmed means of dealing with minor typographical errors. They are widely used by computer scientists working in information retrieval (Frakes and Baeza-Yates, 1992). Porter and Winkler (1997) have shown empirically that bigrams work well, and ST.AT has also made positive experience with the bigram method which is applied in the update process of the business register.

An early string comparator is the *Damerau-Levenstein (D-L) Metric* (Damerau, 1964, Levenstein, 1966), which is in fact only one particular comparator metric from the class of *edit distance metrics*. Its basic idea is the fact that any string can be transformed into another string through a sequence of changes via substitutions, deletions, insertions, and possibly reversals. The smallest number of such operations required to change one string into another divided by the maximum length of the two compared strings is a measure of the difference between them which is easily converted to a string comparator rating the degree of agreement of the two strings. For a discussion of several enhancements of the D-L metric see Hall and Dowling (1980).

*Jaro* (see for instance Winkler, 1985, 1990) introduced a string comparator more straightforward to implement and maybe more closely related to the type of human decisions in comparing strings than the D-L metric. Basically, it accounts for the proportion of common characters in both strings and the number of transpositions that have to be made to create the sequence of common characters of one string from the sequence of common characters of the other string. Several enhancements to the Jaro comparator have been developed, in particular by Winkler (e.g. Porter and Winkler, 1997).

Many more string comparators are presented in Gill (2001) and Cohen, Ravikumar and Fienberg (2003).

### 3.3.1.3 Probabilistic Record Linkage

In probabilistic record linkage (cf. Fellegi and Sunter, 1969, Kilss and Alvey, 1985, Alvey and Jamerson, 1997), conditional probabilities of observing agreement (disagreement) on a matching variable given a pair is actually a match (or a non-match, respectively) are used to define *matching weights* measuring the evidence that a pair is a match or not. Usually, the dual logarithm of the likelihood ratio of these conditional probabilities is used as weight, with the probability given a true match in the enumerator. Each matching variable is associated an *agreement* and a *disagreement weight*. The individual variable weights are assembled to a composite matching weight for each record pair. *Weight thresholds* are then determined for the classification of record pairs into links, possible links and non-links based on fixed error levels.

This kind of linkage rule defined by Fellegi and Sunter (1969) is optimal in the sense that the number of possible links is minimised for fixed error levels. It is also intuitively appealing. If a particular comparison outcome consists primarily of agreements, then it is more likely to occur among matches than non-matches and the corre-

sponding weight will be large. On the other hand, if the comparison outcome consists mainly of disagreements, the matching weight will be small.

In practice, matching weights are computed using some variant of the EM algorithm (Dempster, Laird and Rubin, 1977, Wu, 1983, Meng and Rubin, 1993).

#### 3.3.1.4 Classification Methods

Micro data integration can also be viewed as a well-known statistical problem, viz. a *classification problem*. Record pairs have to be assigned to the class of matches or the class of non-matches, respectively. However, there is one problem: a training sample must be available to enable estimation of classification rules.

A classical methodological choice is discriminant analysis. One approach based on discriminant analysis is the Belin-Rubin method (Belin and Rubin, 1995) which tries to predict class membership conditional on the matching weight assigned to record pairs. Usage of discriminant analysis based on original values of identifying characteristics or comparison outcomes instead of matching weights is also conceivable. Non-parametric methods whose applicability is independent of distribution assumptions, such as nearest neighbour approaches or classification trees, are often used (cf. Neiling 1998).

Another classification method that might be used is logistic regression. Again, comparison outcomes or matching weights may serve as input variables. Chatterjee and Segev (1992, 1994) suggest fitting a logistic regression model to estimate matching weights.

### 3.3.2 Statistical Matching

In statistical matching the linkage of data for the same real-world entity either is not sought or is not essential to the procedure (FCSM, 1980). Usually, datasets have very few (or no) entities in common. Thus, the linkage of data for similar entities rather than for the same entity is acceptable and expected. Actually, except in rare cases, linked records do not represent real-world entities, but rather what is referred to as a synthetic entity (Rodgers and DeVol, 1981), as opposed to exact matches, where, apart from erroneous assignments, linked records refer to identical entities.

Statistical matching originated in the field of economics, initially primarily targeting at the combination of income data and data on tax returns (e.g., Okner, 1972, 1974, Radner, 1978, Radner and Muller, 1977). Statistically matched datasets have been used extensively in micro-simulation modelling (e.g., Cohen, 1991) to examine the impact of policy changes on population subgroups, and, hence, allowed the expectation of reasonable results for DIECOFIS tax simulation studies.

Among statistical matching methods, there are(i) techniques separating datasets into equivalence classes and then selecting records to be linked randomly, (ii) distance measures for the selection of most similar records, and (iii) regression-based techniques (see Kadane, 1978, Moriarity and Scheuren, 2001, Rodgers, 1984, or Raessler, 2002). Imputation techniques are very closely related to statistical matching (e.g., Kovar, Whitridge 1995). Essentially, statistical matching differs from imputation only with regard to its purpose: in a statistical match two different datasets are matched and (in almost all cases) the purpose is the addition of variables not present for any entity in the

base dataset, whereas in imputation often only one dataset is used and values missing for several entities are completed.

### 3.3.3 Imputation

Imputation is used to reconstruct values missing for a record (item non-response, partial missing answers). If a full unit non-response (total missing answers) occurred (i.e., there is no record in the dataset for a sampled unit) usually macro integration procedures (such as weighting) are utilized. A broad introduction to imputation and other types of missing data analysis is given in Little and Rubin (1987).

The most simple imputation approach is *deterministic* imputation, where all missing values of a variable are replaced with the same value, such as the mean, median or mode of the variable. If a large portion of a dataset has to be imputed, this method yields extremely unrealistic distributions with high peaks at the imputed values.

*Model-based* methods hypothesize a probabilistic relation between the variable with missing values and the matching variables. An auto-regression model is often used, so that the variable itself (taken from previous surveys) supplies the information. The probabilities for the occurrence of observed values of a variable are estimated. The imputation value is then randomly drawn from this probability distribution.

In donor-based approaches like *hot-deck* or *nearest-neighbour* imputation, the imputation value is taken from a so-called *donor*, which is a complete and correct record that is similar to the incomplete record. The similitude between donor and receiving record is determined via matching variables. Several donors might be available for the same record – then, one of them is chosen randomly.

Multiple imputation (Rubin 1987) is a simulation-based approach to the statistical analysis of incomplete data. Each missing value is replaced by $m>1$ simulated values. The resulting $m$ versions of the complete data are then analysed by standard complete data methods, and the results combined to produce inferential statements (e.g. interval estimates or p-values) that incorporate missing data uncertainty. So, actually, multiple imputation is not one particular imputation algorithm, but rather a means of evaluation of imputation results.

## 3.4 Quality Assessment

No matter what the objective of data integration actually is, an evaluation of the procedures carried out and the resulting multi-source database is indispensable. Apart from the assessment of the quality of source data, which plays an important role in the integration process, the description of the methods applied and the variables used for integration, measures on the variability or reliability of the results as well as method-specific or application-specific measures are required to evaluate the quality of the integrated database.

The quality of matching variables is crucial to any of the integration procedures presented. For its assessment, a precise definition of the concept captured, the amount of missing data, the discriminating power and reliability should be reported (e.g., Hassard, 1986). For a reasonable overview of the usability of particular personal characteristics

as matching variables see Jabine and Scheuren (1986) or Gill (2001). For enterprises, Winkler (2001) provides some empiric evidence.

In exact matching, particularly misclassification rates and the size of the grey zone of possible matches is of interest. Depending on the aim of matching, gross or net error rates may be considered. The accuracy of the estimation of error rates mainly depends on the availability of training data with known matching status. Moreover, quality indicators for individual processing stages (like blocking or 1:1-assignment) are available (see for instance Baxter, Christen and Churches, 2003).

In statistical matching, where the linkage of records belonging to similar entities is sought for, error rates are not defined, since there is no "true matching status". Rather, distributions of distances of linked records or the number of times individual records are used in linkage (in case that multiple linkage is enabled) are used as quality indicators.

As a matter of the close affinity of imputation (or at least, particular imputation methods, like donor-based imputation) and statistical matching procedures, imputation results may be judged by the same or at least similar criteria as statistical matching results.

# 4 Concluding Remarks

DIECOFIS is an EU-funded international research project, coordinated by ISTAT. The objectives are the development of an appropriate methodology for the construction of such a system of indicators and the illustrating application of the developed methods. Data integration, mainly record matching, and the generation of multi-source databases that are to be used as a basis of micro simulations are a core issue of the project. The Austrian member of the consortium mainly was engaged in the issues of database integration. Contributions concern the surveying of available methods, a critical assessment of different data integration methods with a focus primarily on statistical issues, and an overview of assessment criteria for multi-source databases from a theoretical perspective, in particular statistical indicators of multi-source database quality, all these activities having in view of the application within DIECOFIS. An empirical study has been designed that compares the applicability of various integration procedures in the context of the Austrian business register and that demonstrates the use of quality indicators for multi-source databases. Also a two-day workshop on "DIECOFIS Workshop on Data Integration and Record Matching" was organized by the Austrian group. Hosted by Statistics Austria, consortium members from DIECOFIS exchanged ideas and experiences with practitioners, researchers, and developers from other institutions. Many of the participants came from official statistical organizations, for which integration of data files is of major interest as a means of using available information more efficiently and improving the quality of their products.

# References

W. Alvey and B. Jamerson, editors. *Record Linkage Techniques*. Federal Committee on Statistical Methodology (FCSM), Washington, DC, 1997.

R. Baxter, P. Christen, and T. Churches. A Comparison of Fast Blocking Methods for Record Linkage. To appear in *Proc. First Workshop on Data Cleaning, Record Linkage, and Object Consolidation, 9th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, Washington, DC, 2003.

T.R. Belin and D.B. Rubin. A Method for Calibrating False-Match Rates in Record Linkage. *JASA*. 90(430):694–707, 1995.

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth, Monterey, 1984.

A. Chatterjee and A. Segev. Resolving Data Heterogeneity in Scientific Statistical Databases. In H. Hinterberger and J.C. French, editors, *Proc. 6th Int. Conf. on Scientific and Statistical Database Management*, pages 145-159. ETH Zürich, 1992.

A. Chatterjee and A. Segev. Supporting Statistics in Extensible Databases: A Case Study. In H. Hinterberger and J.C. French, editors, *Proc. 7th Int. Conf. on Scientific and Statistical Database Management*, pages 54-63. IEEE Computer Society, 1994.

S. Cohen. Micro-simulation of Firm Investment. Presented at the *Symposium on Economic Modelling*, London University, 1991.

W.W. Cohen, P. Ravikumar, and S.E. Fienberg. A Comparison of String Distance Metrics for Name-Matching Tasks. Submitted to *18th International Joint Conference Workshop on Information Integration on the Web*, 2003. Also available at *http://www.niss.org/dg/technicalreports.html*.

F.J. Damerau. A Technique for Computer Detection and Correction of Spelling Errors. *Communications of the ACM*. 7(3):171-176, 1964.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *JRSS B* 39:1-38, 1977.

M. Denk *Statistical Data Combination: A Metadata Framework for Record Linkage Procedures*. Dissertation, Department of Statistics and Decision Support Systems, University of Vienna, 2002.

M. Denk and K.A. Froeschl. The IDARESA Data Mediation Architecture for Statistical Aggregates. *Research in Official Statistics*. 3(1):7-38, 2000.

M. Denk and F. Oropallo. *Overview of the Issues in Multi-Source Databases*. DIECOFIS Deliverable 1.1, ISTAT, Rome, 2002.

M. Denk, F. Inglese, and M.G. Calza. *Assessment of Different Approaches for the Integration of Sample Surveys*. DIECOFIS Deliverable 1.2, ISTAT, Rome, 2003.

M. Denk, F. Inglese, and F. Oropallo. *Report on Statistical Indicators for the Assessment of Multi-source Databases*. DIECOFIS Deliverable 1.3, ISTAT, Rome, 2003.

DIECOFIS. DIECOFIS Web Site, *http://petra1.istat.it/diecofis/index.html*, 2003.

M. D'Orazio, M. Di Zio, and M. Scanu. Statistical Matching: a tool for integrating data in National Statistical Institutes. In *Proc. of the Joint ETK and NTTS Conference for Official Statistics*, Crete, 2001.

M.E. Fair and P. Whitridge. Tutorial on Record Linkage. In W. Alvey and B. Jamerson, editors, *Record Linkage Techniques*, pages 457-479. FCSM, Washington, DC, 1997.

FCSM – Federal Committee on Statistical Methodology. *Report on Exact and Statistical Matching Techniques*. Statistical Policy Working Paper 5, U.S. Department of Commerce ,Washington, DC, 1980.

I.P. Fellegi and A.B. Sunter. A Theory for Record Linkage. *JASA*. 64:1183-1210, 1969.

W.B. Frakes and R. Baeza-Yates, editors. *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, Upper Saddle River, NJ, 1992.

K.A. Froeschl. *Metadata Management in Statistical Information Processing*. Springer, Wien, Berlin, 1997.

K.A. Froeschl. On Standards of Formal Communication in Statistics. Working Paper No. 16, UN-ECE/METIS, *Work Session on Statistical Metadata*, 1999a.

K.A. Froeschl. Metadata Management in Official Statistics - An IT-based Methodology Approach. *Austrian Journal of Statistics*. 28(2):49-79, 1999b.

K.A. Froeschl. A Sketch of Statistical Meta-Computing as a Data Integration Framework. To appear in *Austrian Journal of Statistics, Special Issue on Data Integration and Record Matching*. 2004.

K.A. Froeschl and W. Grossmann. The Role of Metadata in Using Administrative Sources. *Research in Official Statistics*. 3(1):65-82, 2000.

L.E. Gill. *Methods for automatic record matching and linking in their use in National Statistics*. GSS Methodology Series, NSMS25. Office for National Statistics, UK, 2001.

P.A.V. Hall and G.R. Dowling. Approximate String Matching. *ACM Computing Surveys*. 12(4):381-402, 1980.

T.H. Hassard. Writing the Book of Life: Medical Record Linkage. In Brook, et al., editors, *The Fascination of Statistics*, pages 25-46. Dekker, New York, 1986.

IDARESA. *The Data Model – Final Version*, Deliverable 3.4.2, Dept. of Statistics, University of Vienna, 1997.

IDARESA. *IDARESA Tandem Structures*, TPR–*viu*–3.4.2/3, Dept. of Statistics, University of Vienna, 1998a.

IDARESA. *The IDARESA **info**-Net*, TPR–*viu*–3.2.1, Dept. of Statistics, University of Vienna, 1998b.

T.B. Jabine and F.J. Scheuren. Record Linkages for Statistical Purposes: Methodological Issues. *Journal of Official Statistics*. 2(3):255-277, 1986.

J.B. Kadane. Some Statistical Problems in Merging Data Files. In *1978 Compendium of Tax Research*, pages 159–171. US Dept. of the Treasury, 1978. (Reprinted in *Journal of Official Statistics*. 17(3):423-433, 2001.)

B. Kilss and W. Alvey, editors. *Record Linkage Techniques*. FCSM, Washington, DC, 1985.

J.G. Kovar and P.J. Whitridge. Imputation of Business Survey Data. In B. Cox et al., editors, *Business Survey Methods*. John Wiley, New York, 1995.

V.I. Levenstein. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Sov. Phys. Dokl.* 10:707-710, 1966.

R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, 1987.

F.M. Malvestuto. A Universal Table Model for Categorical Databases. *Information Sciences*. 49:203-223, 1989.

F.M. Malvestuto. Data Integration in Statistical Databases. In Z. Michalewicz, editor, *Statistical and Scientific Databases*, pages 201-232. Ellis Horwood, Chichester, 1991.

F.M. Malvestuto. A Universal-Scheme Approach to Statistical Databases Containing Homogeneous Summary Tables. *ACM Transactions on Database Systems*. 18:678-708, 1993.

X.L. Meng and D.B. Rubin. Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*. 80(2):267-278, 1993.

C. Moriarity and F. Scheuren. Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure. *Journal of Official Statistics*. 17(3):407-422, 2001.

M. Neiling. Data Fusion with Record Linkage. Presented at the *3rd Workshop* "*Föderierte Datenbanken*", Magdeburg, 1998. Also available at *http://wwwiti.cs.uni-magdeburg.de/fdb98/online-proc/.*

M. Neiling and H.J. Lenz. The Creation of the Register Based Census for Germany in 2001: An Application of Data Integration. In *Betriebswirtschaftliche Reihe: Diskussionsbeiträge des Fachbereichs Wirtschaftswissenschaft der FU Berlin* 34. Freie Universität Berlin, 1999.

M. Neiling and H.J. Lenz. Data Fusion and Object Identification. Presented at *SSGRR* 2000 (*Advances in Infrastructure for Electronic Business, Science and Education on the Internet*). Available at *http://www.ssgrr.it/en/ssgrr2000/proceedings.htm.*

B.A. Okner. Constructing a New Data Base from Existing Microdata Sets: the 1966 Merge File. *Annals of Economic and Social Measurement*. 1:325-342, 1972.

B.A. Okner. Data Matching and Merging: An Overview. *Annals of Economic and Social Measurement*. 3(2):347-352, 1974.

E. Porter and W. Winkler. *Approximate String Comparison and its Effect on an Advanced Record Linkage System*, RR97-02, U.S. Bureau of the Census, 1997. Available at *http://www.census.gov/srd/www/byyear.html*.

K. Pu. Key Equivalence in Heterogeneous Databases. In *Proc. 1st Int. Workshop on Interoperability in Multidatabase Systems*, Kyoto, Japan, pages 314-316. IEEE Comp. Soc. Press, 1991.

D.B. Radner. The Development of Statistical Matching in Economics. In *Proc. Social Statistics Section*, pages 503-508. American Statistical Association, 1978.

D.B. Radner and H.J. Muller. Alternative Types of Record Matching: Costs and Benefits. In *Proc. Social Statistics Section*, pages 756-761. American Statistical Association, 1977.

S. Raessler. *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. Springer, New York, 2002.

W.L. Rodgers. An Evaluation of Statistical Matching. *Journal of Business and Economic Statistics*. 2:91-102, 1984.

W.L. Rodgers and E.B. DeVol. An Evaluation of Statistical Matching. In *Proc. of the Survey Research Methods Section*, pages 128-132. American Statistical Association, 1981.

D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*.: John Wiley & Sons, New York, 1987.

A.P. Sheth and J.A. Larson. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. *ACM Computing Surveys*. 22(3):183-236, 1990

V. Ventrone and S. Heiler. Semantic Heterogeneity as a Result of Domain Evolution. *ACM SIGMOD record*. 20(4):16-20, 1991.

Y.R. Wang and S.E. Madnick. The Inter-Database Instance Identification Problem in Integrating Autonomous Systems. In *Proc. of the 6th International Conference on Data Engineering*, Los Angeles, pages 46-55. IEEE, 1989.

G. Wiederhold. Mediators in the Architecture of Future Information Systems. *IEEE Computer*. 25(3):38-49, 1992.

G. Wiederhold and M. Genesereth. The Conceptual Basis for Mediation Services. *IEEE Expert*. 12(5):38-47, 1997.

W. Winkler. Preprocessing of Lists and String Comparison. In B. Kilss, W. Alvey, editors, *Record Linkage Techniques*, pages 181-187. FCSM, Washington, DC, 1985.

W. Winkler. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In *Proc. Section on Survey Research Methods*, pages 354-359. American Statistical Association, 1990.

W. Winkler. Matching and Record Linkage. In B. Cox et al., editors, *Business Survey Methods*, pages 355-384. J. Wiley, New York, 1995.

W. Winkler. *The State of Record Linkage and Current Research Problems*, RR99-04, U.S. Bureau of the Census, 1999. See *http://www.census.gov/srd/www/byyear.html*.

W. Winkler. *Quality of Very Large Databases*, RR2001/04, U.S. Bureau of the Census, 2001.

C.F.J. Wu. On the Convergence Properties of the EM-Algorithm. *Annals of Statistics*. 11(1):95-103, 1983.

Authors' addresses:

Dr. Michaela Denk
ec3 – Electronic Commerce Competence Center
Donau-City-Straße 1
A-1220 Vienna
Austria

Tel. +43 1 522 71 71 / 19
Fax +43 1 522 71 71 / 71
Elec. Mail: michaela.denk@ec3.at
http://www.ec3.at/

Univ.-Prof. Dr. Peter Hackl
Department of Statistics
Vienna University of Economics and Business Administration
Augasse 2-6
A-1090 Vienna
Austria

Tel. +43 1 31336 / 4751
Fax +43 1 31336 /
Elec. Mail: peter.hackl@wu-wien.ac.at
http://www.statistik.tuwien.ac.at/oezstat/