Austrian Journal of Statistics
Volume 33 (2004), Number 1&2, 173-194

A Sketch of Statistical Meta-Computing as a Data Integration Framework

Karl A. Froeschl
ec3 Electronic Commerce Competence Center, Vienna

Abstract: Statistics defines itself as a methodological discipline providing a rigorous, formal framework for scientific empirism based on a mapping of contingent observable phenomena to (real) numbers that can be dealt with, or analysed, computationally. Application of the statistical methodology of data reduction, in turn, requires some representation of the problem context. Most of the time, this amounts to encoding (a part of) the problem context of observation data into another layer of data – called *metadata*. Based on metadata, procedures of data analysis might be enhanced to encompass also the analysis and transformation of metadata alongside the accompanied data itself. The paper sketches the outline of a systematic approach to statistical "meta-computing" as a dual-mode proposal of statistical data processing.

Zusammenfassung: Statistik kann als eine methodische Disziplin definiert werden, die eine formal rigorose Grundlage der wissenschaftlichen Empirie anstrebt, indem sie kontingente beobachtbare Phänomene auf (reelle) Zahlen abbildet, die sodann im rechnenden Sinne weiterverarbeitet bzw. analysiert werden. Die Anwendung der statistischen Methodik zur Datenreduktion erfordert ihrerseits eine Repräsentation des jeweiligen Anwendungsproblems. Praktisch läuft dies dann in der Regel meist darauf hinaus, den die Daten beschreibenden Beobachtungskontext zumindest teilweise selber in Form von Daten – sogenannten *Metadaten* – abzubilden. Auf diesen Metadaten aufbauend lassen sich nun die Verfahren der Datenanalyse so erweitern, dass sie auch die Analyse und Transformation der Metadaten – zusammen mit den zugrundeliegenden Daten selbst – einbeziehen. Dieser Beitrag skizziert die Umrisse eines systematischen Ansatzes zu einem statistischen „Meta-Computing“ in Form eines Vorschlags zu einer solchen dualen statistischen Daten/Metadaten-Verarbeitung.

Keywords: statistical metadata, metadata management, statistical data processing, data and metadata integration.

1 Introduction and Background

In general, 'data integration' is considered a preparatory procedure to merge two or more datasets into one larger, or augmented, dataset undergoing further analysis or processing (Wiederhold and Genesereth, 1997). Integration of data, though, is based on some underlying pairing or matching logic justifying the putting together of data of different record sets so that the resulting, “integrated” dataset, in a sense, carries the joint information of its integrated predecessors. In the statistical context, accordingly, very often techniques useful in this respect are dubbed “record matching”, “data linking”, and the like (Winkler, 1995). To be legitimate, such kind of statistical data linkage typically presupposes that observation records strung together contain or pool variables (data elements) referring to some shared or, at least, compliant statistical unit and population, alongside possibly further criteria to be met. Naturally, the combination conditions depend on the specific subject-matter or information processing context whence, as a matter of fact, even in statistics the range of data integration modes happens to be variegated (Denk, 2002). The point here is that both feasibility and mode of data integration depend on information *external* to the data to become so integrated, as this is highlighted, for instance, in the DIECOFIS project (Inglese and Oropallo, 2004). Thus, in addition to record matching, there is in fact another legitimate reason for data integration quite different in its intention and purpose from the one stated above: any reasonable processing of statistical data (viewed as symbol-coded statements about empirical reality) refers to information *about* this data, a good deal of which can itself be encoded as data. Generally, this kind of second-order data is termed *metadata* for some thirty years now (the term being used reportedly the first time by Sundgren as early as 1973).

Given that metadata is indeed amenable to formal processing – based on structured representations of metadata – the notion of computing with metadata makes perfect sense. Quite naturally, this type of computing could be called “meta-computing”, as the objects of symbol transformation are not statistical observations any more but coded information *about* such statistical observations. An immanent feature of all higher-order information, informed processing of lower-order information relies on processing upper-order information first. So, except for trivial cases, “ordinary” data integration generally requires the *preceding* integration of respective metadata to enable meta-computed decisions about subsequent data linkage. Yet, the real virtues of meta-computing depend on keeping data and metadata *always* integrated, calling for specific computational architectures tightly linking data and data descriptions (metadata) within all statistical transformations to make metadata available alongside the referred-to data throughout (Bethlehem et al., 1999).

This contribution seeks to define statistical meta-computing mainly as a task of metadata integration and outlines a couple of design considerations towards the development of data/metadata-integrated statistical transformation systems. In so doing, it draws heavily on previous work, mostly carried out in a series of research projects with the participation of the Data Analysis and Computing unit (headed by Prof. Wilfried Grossmann) of the Dept. of Statistics, University of Vienna, from the late-1980ies onwards. Thus, both methodology and proposed design elements of the proposed framework emanate from some 15 years of research work carried on, most of the time, in collaboration with research partners and (national) statistical institutes from all over Europe, including – among others – Paul Darius and Michel Boucneau (University of Leuven, Belgium), Gerda van den Berg (University of Leiden, Netherlands), Dennis Conniffe (Trinity College, Dublin, Ireland), Sally McClean and David Bell (University of Ulster), Haralambos Papageorgiou (University of Athens), Eric Schulte Nordholt (CBS, Netherlands), Joanne Lamb (University of Edinburgh), Hans-Joachim Lenz (FU Berlin, Germany), Jana Meliskova (UN/ECE, Geneva, Switzerland), and ranging over projects such as “Automated Generation of Statistical Tables” (Austrian Office of the Chancellor; Grossmann and Froeschl, 1994), “Modelling Metadata” (IST FP3/DOSIS; Darius et al., 1993), “Integrated Documentation and Retrieval Environment for Statistical Aggregates (IDARESA)” (IST FP4/DOSES; Denk and Froeschl, 2000), a book on “Metadata Management in Statistical Information Processing” (Froeschl, 1997), a pilot re-implementation for the UNIDO Industrial Statistics Database (Froeschl et al., 2002), and the “MetaNet” Network of Excellence (IST FP5; Froeschl et al., 2003).

In this paper, the exposition of ideas on statistical meta-computing is organised as follows. Section 2 juxtaposes the concepts of statistical computing and meta-computing, elucidates the virtues of metadata modelling, and presents a (preliminary) definition of meta-computing based on the analysis of an appropriate modelling abstraction. Next, Section 3 develops the operational framework for meta-computing with specifically designed object models integrating both data and metadata components, exemplifying the proposed operand structures in a general weighting context. For reasons of limited space, however, the presentation gives only a sketchy account of application. Section 4 indicates how the outlined structure of statistical meta-computing could be extended to more comprehensive meta-information structures, giving a flavour of the meta-information designs needed. Finally, Section 5 touches briefly the option of goal-driven statistical computing by means of metadata, and tentatively evaluates the potential of statistical meta-computing, given the present state of affairs.

2 Definitions and Terminology

Statistical computing, while primarily dealing algorithmically with numbers encoding empirical observation data, always relies on and is justified in terms of information about this observed (part of the) reality – as it is construed in the eyes of the observer or according to some theoretical consideration. This information about data – henceforth termed *meta-information* – comprises varied knowledge such as about

- the coding of observations in so-called codebooks recording the mapping from observed phenomena to number scales or code systems;
- the concepts and processes underlying or determining the actual measuring and data capturing activities (that is, how the mapping of observations to data takes place, for example, through a questionnaire);
- the design of the data collecting schema, or sampling structure (that is, the choice and configuration of the entities actually observed) as well as
- various subject-matter considerations providing the motivation for actually carrying out the whole empirical process, its reason and conceptual framing.

Correspondingly, any sound empirical reasoning has to rely on such kinds of “situation-dependent” statistical meta-information. Moreover, statistical data carry statistical information only so long as they can be interpreted in the light of this surrounding meta-information; hence, the ultimate purpose of meta-information is to provide, or maintain, the semantics of the *context* the data originates from.

In addition to this situation-dependent context information, statistics as a methodology proceeds in a structured way to analyse statistical data. Statistics, as a practical method of empirical reasoning, rests on various foundations including

- first of all, statistical theory, and, corresponding to this, a specific terminology comprising the salient notions and discernments supporting theory development;
- data management methodology – for organising the storage of both observation data and descriptive data context – and (numerical) computational principles and algorithms – for algebraic data transformation;
- familiarity with the subject-matter issues the formal methodology of statistics becomes applied to (that is, about what is actually encoded in statistical data);
- a toolbox of modelling techniques – so-called "applied statistics" – helping to fit formal explanation structures to the data captured.

From a theoretical perspective, statistical methodology seeks to abstract from individual empirical investigations in order to sift out, depending on stereotypical constellations, generalised schemata of empirical reasoning (Box, 1976). For such constellations (for example, linear models, or time series models), a canonical methodology is developed grounded on the fundamental concepts of the discipline (such as axiomatic probability theory, maximum likelihood principle, distributional assumptions, etc.), and equipped with normative status. In practice, of course, these recipes happen to be softened, not in their formal conduct but as to their theoretical conditions of admission, to remain applicable in a wider range of subject domains.

Clearly, a sufficient formalisation of methodology is an indispensable precondition for any serious attempt at statistical meta-computing. However, traditional methodology of statistical estimation and inference predominantly focused on what might be called "intra-inferential aspects of methodology" in that, in general at least, it dealt with one (statistical) function applied to a dataset at a time, and scrutinised the mathematico-statistical properties of this function in a rather abstract way. Typically, this amounts to prescribing the various conditions to hold for the function's arguments, viz. the dataset (or "sample"), to make the claimed statements about the function – such as unbiasedness, sufficiency, … – valid. Of course, also numerical issues (accuracy, stability, …) are extensively dealt with in this respect.

Compared to that, in mathematical statistics seemingly less emphasis has been laid on the development of "coherent macro-strategies", that is, the methodology of deriving consistent – or, at least, *coherent* in some meaningful way – sets of summaries or conclusions from a body of related observation data. In this respect, attitudes began to change in the (late) 1970ies, notably propelled through the immensely influential work of Tukey (1977) but certainly also supported by the improved capabilities of computing machinery placing more and more computing power and data management facilities on a statistician's desk. A further stimulus was triggered by the then rising popularity of Artificial Intelligence-based so-called expert systems – in the 1980ies several such statistical expert systems aiming to encode the analytical tactics of applied statisticians were proposed (Haag, 1994). Although, meanwhile, the interest in expert system technology has declined for several reasons (Streitberg, 1988), this is nevertheless to say that, more recently, attention seems to have shifted towards more processing-oriented aspects of statistics, or "extra-inferential aspects of methodology" – whether particular statistical functions could be applied legitimately under varying practical conditions (with, correspondingly, a tangible focus on robust and non-parametric methods as well as the management of errors of the *third* kind; cf. Kimball, 1957).

In addition to the indicated changes in focus on methodology in statistics, also an emerging preponderance of data is noticeable from the 1980ies onwards – the former method-centric approach now became somewhat overshadowed by a focus on data. This data-centric turn benefited from a lasting change in data provenance, with a growing abundance of available data, compared to the traditional data scarcity most of empirical work used to struggle with. Thanks to decreasing storage costs, the incredibly swelling waves of electronically captured or even "digitally borne" data (much of which non-observational, by the way) are gathered over long time periods in huge statistical repositories; with the advent of the Internet data has additionally gained an unprecedented mobility and become easily accessible even on a global scale (Ryssevik, 2002). At the same time, data structures are becoming more and more complex. Little surprise, then, that data turns into a new resource of its own, seeing data-centric statistics (also known as "data mining" techniques; cf. Han and Kamber, 2001, or Fayyad et al., 1996) receiving tremendous importance, and the role of data documentation getting sharply elevated, as data is increasingly available for use outside its originating context.

A strategic response to these changes and developments – based on the heritage of data description approaches and requirements that evolved from the former attempts towards building statistical expert systems – consists in advancing the methodology of data documentation. Without any pretensions to truly codify any (procedural, heuristic, etc.) knowledge of statistical analysis, it is reasonable to assume that a standardised set of universally relevant data-descriptive elements can be inferred by way of abstraction and theoretical argument. This approach towards formal documentation aims at the systematic capture of (non-observational) higher-order data, called *metadata*, and seeks to establish a sustained linkage of data with the metadata describing it. Clearly, "metadata" is a term with relative meaning: by definition, metadata is data as well, and so it is always conceivable to establish metadata for data of whatever order (thus, the "metadata" of some data's metadata is its "meta-metadata", and so on). Thinking of metadata in terms of (otherwise ordinary) data, however, highlights an often overlooked feature: as data, metadata is amenable to a schema-compliant representation (that is, it meets the structural requirements such as those of database schemata) and, furthermore, metadata become well-defined symbolic objects to formally operate with – although, apparently, meaning and structure of computation is different from computing with usual statistics data. At any rate, though, considering data and related metadata as a tightly integrated "tandem" structure (Darius et al., 1993) is paving the way towards a metadata-controlled mode of data transformation bringing into reach benefits such as

- the preservation of documentation integrity through always concordantly co-transforming both data and data description;
- the automated verification of the (formal) preconditions of an operation to be applied to statistical data by matching these preconditions against operand metadata;

- the assurance of self-consistent statistical processing sequences as, even in "out-of-context" processing, at least a minimum of origin context is preserved by metadata.

Actually, the idea of metadata-controlled computing is by no means radically new: most familiar statistical packages use internal data descriptions to organise "local" data processing but – and this makes a significant difference – metadata is generally not accessible from outside. In order to fully exploit the potential of metadata, statistical data processing environments have to be thoroughly re-engineered such that metadata stays seamlessly linked to the data proper, regardless of all the transformation it undergoes, throughout its whole lifetime, and beyond any single system's boundaries. This way, conversely, metadata modelling becomes a prerequisite to effective interoperable – and, hence, "metadata-mediated" – statistical data processing systems utilising devolved, distributed, or federated data resources (Denk and Froeschl, 2000).

In drawing together, in the present context, metadata means the formalised share of data context mapping data semantics to (syntactic) information structures; correspondingly, meta-computing refers to the application of algorithmic (symbol) transformations to metadata, in complete analogy to computing referring to the application of symbol transformations to (numerical, in general) data. In particular, *statistical* meta-computing denotes that kind of meta-computing using dual, self-describing computation structures encompassing both statistical data and metadata components such that all data is "wrapped up" in metadata.

Now, demanding that all data must be wrapped up in its own metadata sounds like begging the question: where to stop appropriately with all the wrapping before cycling infinitely?

2.1 Meta-Information Modelling

Any modelling (of information) starts with the definition of a language to express ideas. Given the ambition of universally describing empirical discourses – the formal negotiation of models of an outer reality – by means of a (semi-) formal language, the task of determining a suitable language amounts to abstract bottom-up from "discourse instances" until a stable, self-sufficient set of *universal empirical discourse notions* is attained the first time. This search process can be visualised very well using Del Vecchio's model pyramid (cf. Froeschl et al., 2003) comprising altogether four modelling levels on top of a ground "reality" level:

- L4: methods that define methods (meta-metamodels)
- L3: methods for the making of definitions (meta-models)
- L2: definitions (models of the data)
- L1: extensions (data)
- L0: "reality"

In this hierarchy, L1 denotes the level of direct symbol representation (terms, codes) of "real" contingencies whereas L2 signifies the level regulating the semantics of L1-designations. Apparently, in so doing, L2 structures conform to some language convention, viz. an L3 schema, or "meta-model" providing these language definitions. As a typical example, the ISO/IEC 11179 ("specification and standardization of data elements") standard might be considered: while the standard provides a description methodology at L3 in terms of so-called 'data elements', data element descriptions is L2 data. In turn, the meta-model language is expressed and communicated using a yet higher-level specification language, at L4. Fortunately, L4 happens to be self-encompassing (provided a very powerful language is used, as assumed here), so the hierarchy happily remains finite.

Given the generality of L4 modelling and the domain specificity of L2 models, L3 turns out to be the appropriate level of formalising models of data description, that is, providing the information structures and concepts *invariant* with respect to individual applications. In other words, L3 notions and relational structures provide the appropriate means of expression for general *statistical* domain modelling and, thus, provides the terminology and formal constructs of statistical meta-computing. Dropping L0 und using a more familiar statistical terminology, Tab. 1 (simplified from Froeschl et al., 2003) summarises the proposed names, intended content and envisaged "carrier" information structures for the respective model levels.

Table 1: Statistical Meta-Information Modelling Hierarchy

Level	**Name**	**Content**	**Carrier**
1	instance	data	individual
2	domain	data schema	collective
3	application model	domain schema	category
4	meta-model	model schema	class

L3 modelling, then, consists in singling out a parsimonious, clear-cut set of fundamental statistical discourse notions (or types), for which the designation "metadata categories" is proposed, simply because it is the modelling hypothesis that any discernible entity occurring in an empirical discourse belongs to exactly one of these categories. In other words, L3 language is based on a taxonomy of discourse objects, classifying objects into categories, and determining the internal structures (type description for each category) and relationships between categories, or, more specifically, instances of categories. Moreover, L3 is also the layer for defining a set of elementary operations for transforming (representations of) category instances according to the semantics of statistical operations.

As an example, Fig. 1 (reproduced from Froeschl et al., 2003) illustrates an excerpt of L3 structures, focusing on the metadata category of 'dataset' and its immediate model neighbourhood. Note that the graphic presentation of the model excerpt uses, though without explicit mention, L4 language. Note further that both shown metadata categories and relationships between them hold generically and inherit automatically to all individual category instances.

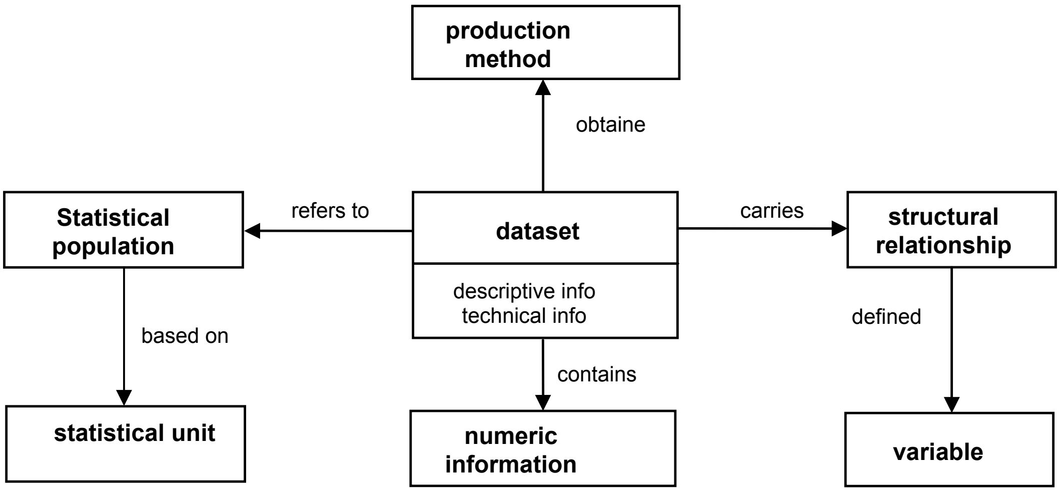

Figure 1: Sample L3 Structure Model

Using L3 structure elements, application (L2) models can be built such that problem context of empirical discourses is captured in terms of metadata categories and category interrelations independent of any particular application context.

2.2 Statistical Meta-Information Structures

Quite obviously, L3 formalisation still leaves open various choices. The main guiding principle adopted here consists in a formalisation supporting both structural context description and transformation dynamics. Aiming at a formal representation of the meta-empirical (L3) discourse comprising well-defined computable operands with neat algebraic properties throughout, the proposal cuts the set of core constructs – called meta-information *dimensions* – to only three of them, viz.

- *population* structures (P) comprising the statistical units and collectives undergoing observation/measurement;
- *measurement* (observation) structures (V) encompassing the observed phenomena ("variables"), and
- *value* structures ($P \times V$) capturing the observational design (incl. sampling structure).

Not by mere incidence, these dimensions mirror the elementary set-up of probabilistic experiments with a random variate (V), its domain component (P), and – typically for some random sample drawn – a materialised outcome of the experiment, or value, within an event space induced by $P \times V$. From a meta-information perspective, however, there are different implementation layers for metadata discernible, at least indicating a distinction between a 'conceptual' and an 'operational' layer as shown in Tab. 2 (modified from Froeschl et al., 2003). This "3-by-3 meta-information breakdown" highlights a marked asymmetry between data and metadata: while meta-information gives rise to metadata on *all* implementation layers, data is present usually only on the material layer (notably, in terms of datasets) and, partly, on the operational layer (for instance, extensional representations of: population registers, code systems of hierarchical classification systems, questionnaires). Clearly, as the conceptual layer concerns data intensions (mostly in terms of non-formal definitions) only, this layer covers metadata exclusively. To explain briefly, 'statistical unit' refers to a type of carrier of 'statistical characteristics' (that is, an observable feature of an observed individual) whereas a 'data source' frames the observational setting (including the sampling scheme, factorial design, etc.).

Table 2: Statistical 3-by-3 Meta-Information Breakdown

	Implementation layer		
Dimension	*conceptual*	*operational*	*material*
P	statistical unit	statistical population	case(s)
V	statistical characteristic	codomain (value range)	observation(s)
$P \times V$	data source	event space	dataset

As to transformation dynamics, all P, V, and P×V elements qualify equally as operands (Denk et al., 2002). This establishes two views on statistical meta-computing, viz.

- a *structural* (representational) view, focusing on information structures and information structure transformations, and
- an *operational* (state-transitional) view, comprising inputs (sources), algebraic/ numerical operation(s), and outputs (for presentation/storage/transfer).

Structurally, operands are composed, in general, of **(i)** statistics data proper (that is, "first order" observation data), **(ii)** companion data (such as paradata and peridata; cf. Froeschl et al., 2003: 'paradata' – a notion introduced by Fritz Scheuren (2000) – is a special category of still by nature observational data accounting for salient features of taking observations/measurements; 'peridata' – an artificial notion indeed – refers to structural and numerical non-observational quantitative information describing the observation set-up in terms of sampling fractions, non-response rates, etc.), **(iii)** context linkage (reference data embedding operands into a broader transformation context; cf. Section 4), and **(iv)** content data (providing structural operand self-description; cf. Section 3). Transformation structure separates into operand derivation (how it is obtained) and operand lineage (where it comes from). In the operational view, operands are divided into several types (algebraic sorts) comprising, among others, statistical datasets, matrices, tables, time series, etc. each sort bearing, of course, specific operations. These, in turn, may affect either the data itself or transform data structures (schema transformations). Finally, operators divide into elementary (primitive) operators and compound operators (that is, expressions of nested operations).

The main technical implication of always keeping together data and its description is a compound operand representation stacking, if need be, several data/metadata-level pairs as shown in Fig. 2a (modified from Denk, 1999). In particular, because part of an operand's metadata is the schema for the data contained, transformation management generally amounts to a stepwise level unfolding procedure: before a data component of an operand can be written, its structure (schema) must be established first by another operation at the metadata level. In fact, there is no cogent reason to transform data and metadata levels synchronously; rather, it is often even advantageous to string together metadata transformations τ'_i on metadata operands d'j and determine their compound effect ($\tau'_1 \circ \tau'_2(d'_1,d'_2) = \tau'_2(\tau'_1(d'_1,d'_2))$ in example Fig. 2b) before data components are filled in.

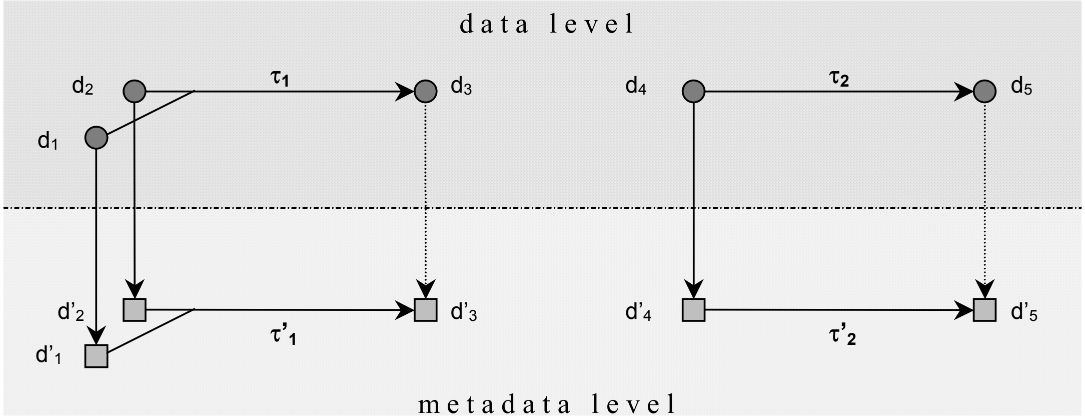

Figure 2a: Stacked Operand Levels

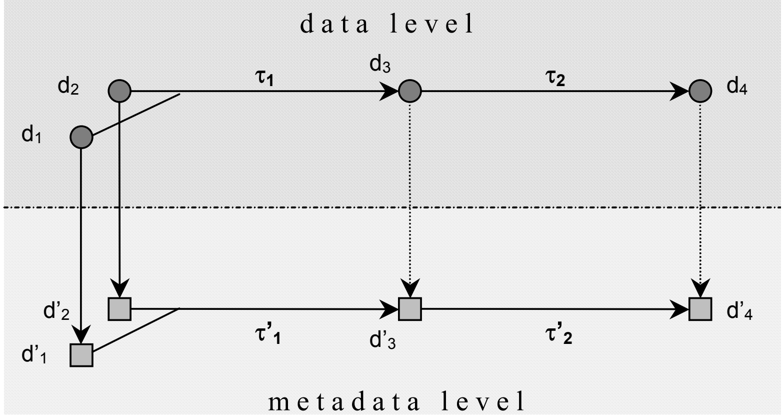

Figure 2b: Level-locked Operators Lined Up

Apparently, the management of compound operands introduces a considerable overhead in structure administration, mainly because (even) elementary transformations affect, in general, several schema/data-level pairs. As a consequence, most (statistical) operations become complex event-condition-action rules for changing both operand state and structure (schema). However, this is compensated by conducting all transformations "in-context", that is, by maintaining all explicit context linkages even for operands changed or newly created as output of an operation.

3 A Design for Operand Structures

This section presents a proposal for a transformation-invariant meta-computing operand structure such that **(i)** all operations can be stated as operators of a calculus over a many-sorted algebra, and **(ii)** all statistical computing can be expressed as a sequence of stepwise schema/data updates preserving structural consistency of the operand and context network. To focus the presentation and for limits of space, however, only datasets (cf. Tab. 2) are considered as operand specimen. The ensuing foundational structure is first explicated structurally and then exemplified in a simple weighting context (cf. Denk et al., 2002).

3.1 Statistical Composites

Statistical composites as a data structure are built out of a few basic elements as shown in Fig. 3. Essentially, each statistical composite (SCo, henceforth) consists of a *container directory* as its "top level" registry listing all SCo components (except the container directory itself), an *attribute directory* gathering all "attributes" used in the SCo, and a variable number of so-called *bucket schemata* and *buckets*, respectively. In spite of this flexibility, the basic building blocks may still conform to conventional relational (or tabular) structures. Dashed arrows in Fig. 3 indicate the metadata/data dependencies established: attributes describe content used in bucket schemata; bucket schemata describe bucket structures.

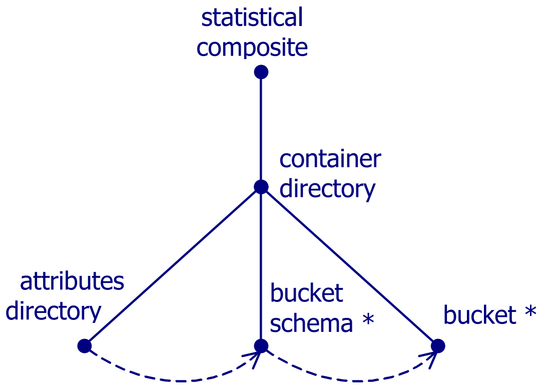

Figure 3: Skeleton SCo Structure (from Denk et al., 2002)

Buckets, as a generic tabular container structure, come in *two formats* and subdivide into different *classes*. First, there is a ‘case’ format good for storing standard observation (that is, “case-by-variate”) matrices, and a ‘summary’ format representing generalised, multi-dimensional statistical tables. Regardless of the format, the following bucket classes are distinguished:

- *data*: buckets for storing statistics data as well as paradata (Scheuren, 2000) where the latter refers to a special kind of microdata assuming a bivalent role of either a regular observable or a variable encoding mode information about the observation (such as the number of trials to successfully contact a respondent in a telephone interview);
- *sampling* and *weighting*: buckets storing information about the sampling scheme and data grossing-up, respectively;
- *method*: buckets to store estimates or model parameters typically obtained as output of statistical procedures;
- *annotation*: buckets for gathering comments and footnotes (Silver, 1993; Froeschl et al. 2002).

Buckets are composed of attributes that might be defined either for use within a specific bucket (‘bucket’ level attributes), for use within an individual SCo (‘composite’ level attributes), or used to link a SCo to its outside context (‘context’ level attributes). Moreover, (bucket) attributes are classified into ‘data’ attributes (C=categorical, Q=quantitative, K=key, …), ‘summary’ (S), ‘weight’ (W), and ‘method’ (M) attributes.

Fig. 4 illustrates the type definition of SCo using a language of semi-structured data (Abiteboul et al., 2000); the question mark denotes an optional attribute, reserved keywords are typeset in italics, an ampersand (‘&’) prefix signals a reference to an instance of the indicated type. A (fictitious) instance of this defined type of SCo is ‘SCO-2’ shown in Fig. 5; this composite encodes a case-level micro-dataset in use within some transformation stage. ‘&PCO-4’, ‘&PCO-6’, and ‘&SRC-1’, respectively, refer to external instances of types ‘POPULATIONCOMPOSITE’ and ‘SOURCE’.

```
type STATISTICALCOMPOSITE =
   {
         (Label : string) ?,
         (Description : string) ?,
         Origin : source | derived,
         Context : input | transformation | output,
         Format : b_format,
         ProcessingLevel : raw | micro | macro | adjusted | …,
         Components : CONTAINERDIRECTORY,
         StatisticalPopulation : &POPULATION | &POPULATIONCOMPOSITE,
         SamplingPopulation : &POPULATIONCOMPOSITE,
         GeneratedBy : &SOURCE | &TRANSFORMATIONSTEP
   }
```

Figure 4: SCo Type Definition

{
Label : “Example composite”,
Origin : *source*,
Context : *transformation*,
Format : *case*,
ProcessingLevel : *micro*,
Components : CDIR2,
StatisticalPopulation : &PCO-4,
SamplingPopulation : &PCO-6,
GeneratedBy : &SRC-1
}

Figure 5: SCo Instance

The container directory ‘CDIR2’ is actually a nested sub-structure (dependent component) of SCo instance ‘SCO-2’ separated from the SCo here for presentation purposes, and might look like shown in Fig. 6.

{
Attributes : ADIR2,
Contains :

Class	Schema	Bucket
data	&SCH-1	&B-3
sampling	&SCH-3	&B-7

}

Figure 6: Nested SCo Container Directory

{
Contains :

ID	Class	Role	CorrespTo
SCOATT1	C	strat	&ATT-17
SCOATT2	Q	obs	&ATT-8
M	M	M	M
SCOATT8	U	sel_prob	&ATT-12
SCOATT9	Q	strat_size	&ATT-15

}

Figure 7: Nested SCo Attributes Directory (Excerpt)

The attributes directory ‘ADIR2’ of ‘SCO-2’ is again a nested sub-structure listing all the attributes used in the SCo. The running example is continued in Fig. 7. The ‘role’ column specifies the formal meaning of the respective attribute (that is, the variable stated in the ‘CorrespTo’ column) within the SCo; for example, ‘strat’ declares ‘SCOATT1’ as a stratification variable whereas ‘sel_prob’ indicates that ‘SCOATT8’ provides the case selection probabilities used for random sampling.

{

Format : *case*,
Class : *data*,
Contains :

ID	CorrespTo
BATT1	SCOATT1
BATT2	SCOATT2
M	M

}

Figure 8: Example Data Bucket Schema (Excerpt)

{

Schema : &SCH-1,
Contains :

BATT1	BATT2	…
F	35210	…
F	14700	…
M	53890	…
M	M	M

}

Figure 9: Example Data Bucket (Excerpt)

Schemata for both container and attributes directories are, by definition, constant. For each bucket, however, its schema has to be declared separately by listing all bucket attributes and identifying these with the respective composite attributes, as shown in Fig. 8 for ‘&SCH-1’.

Using the defined schema ‘&SCH-1’, bucket ‘B-3’ eventually provides the content of the data matrix as indicated in Fig. 9. Note that observations are actually stratified according to variable ‘ATT-17’, and ‘ATT-8’ is one of the dataset’s observables accessible through ‘BATT2’ (both defined elsewhere outside composite ‘SCO-2’).

3.2 A Simple Weighting Example

For the sake of illustration, this subsection sketches the practical use of SCo structures in a simple weighting application (Ofner, 2001). Generally, the calculation of weights for a statistical data sample depends on **(i)** the structure of the dataset (whether its format is ‘case’ or ‘summary’), **(ii)** the type of additional information available (for example, sample sizes, stratification variables, etc.), and **(iii)** the method of weighting (for example, using base weights, or calibration weights to compensate for non-sampling errors, etc.). Clearly, the analytical target of weighting guides the choice of a weighting method, usually selected by the analyst. If the target is adjustment of data to the sampling process, base weights will be used, whereas for adjustments to the population structure a kind of calibration applies. The feasibility of a particular weighting method as well as details of the computational procedure are determined by both, the input SCo (including its component data) and additional user data (such as method parameters), if any. The ensuing transformation procedure consists of basically three stages, viz. **(i)** checking of method feasibility (accomplished mainly by inference on metadata), **(ii)** the computation of (new) component data, and **(iii)** the creation of a well-defined output SCo.

The example scenario assumes the application of a standard Horvitz-Thompson estimator with weight multipliers 'weight_base' computed as 1 over 'sel_prob' (cf. Fig. 7) using a 'case'-format 'data' bucket (cf. Fig. 9) combined with a 'summary'-format 'sampling' bucket 'B-3' shown in Fig. 11; the bucket schema 'SCH-3' for this sampling bucket is exhibited in Fig. 10.

{

Format : *summary*,
Class : *sampling*,
Contains :

ID	**CorrespTo**
BATT1	SCOATT1
BATT2	SCOATT8
BATT3	SCOATT9

}

Figure 10: Example Sampling Bucket Schema

{

Schema : &SCH-3,
Contains :

BATT1	**BATT2**	**BATT3**
F	0.2	600
M	0.5	800

}

Figure 11: Example Sampling Bucket

Linking back to the attributes directory, the columns of the sampling bucket 'B-3' refer to, in this ordering, the stratification variable 'SCOATT1' (sex, in this case), 'SCOATT8' (sel_prob), and 'SCOATT9' (strat_size). Obviously, the Horvitz-Thompson estimator based weighting method (using inclusion instead of selection probabilities for grossing-up) applies because the sampling bucket provides the required weighting factors appropriately stratified (namely, by the same variable, 'SCOATT1'). Hence, computation of base weights is fairly easy. A consistent update of the description network amounts to generating an output SCo structure through the following steps:

- define a container directory ('CDIR15'; cf. Fig. 12) by adapting the container directory of the input SCo through adding both a weighting bucket entry and its schema entry;

- accordingly, also the attributes directory (‘ADIR13’; cf. Fig. 13) is adapted by adding another composite attribute (‘SCOATT10’);
- creating the new weighting bucket schema (‘SCH-11’; cf. Fig. 14), and
- creating the new weighting bucket (‘B-14’; cf. Fig. 15) with the computed weights (‘BATT2’) inserted.

In all figures below, the grey-shaded parts denote added or newly created entries. The resulting output SCo (‘SCO-3’) is exhibited in Fig. 16.

{

Attributes : ADIR13,
Contains :

Class	Schema	Bucket
Data	&SCH-1	&B-3
Sampling	&SCH-3	&B-7
Weighting	&SCH-11	&B-14

}

Figure 12: Adapted Container Directory

{

Contains :

ID	Class	Role	CorrespTo
SCOATT1	C	strat	&ATT-17
M	M	M	M
SCOATT8	U	sel_prob	&ATT-12
SCOATT9	Q	strat_size	&ATT-15
SCOATT10	Q	weight_base	&ATT-23

}

Figure 13: Adapted Attributes Directory (Excerpt)

While highlighting, to some degree, basic mechanisms underlying the sketched meta-computing principles, the example is incomplete as, in several places, references to objects outside of the considered SCos occur or are introduced during the computation (‘&ATT-23’, ‘&WT-1’) that, of course, all have to consistently refer to outside context structures establishing cross-SCo meaning assignments.

{

Format : *summary*,
Class : *weighting*,
Contains :

ID	CorrespTo
BATT1	SCOATT1
BATT2	SCOATT10

}

Figure 14: Example Weighting Bucket Schema

{

Schema : &SCH-11,
Contains :

BATT1	BATT2
F	5
M	2

}

Figure 15: Example Weighting Bucket

{

Label : “Example weighting composite”,
Origin : *derived*,
Context : *transformation*,
Format : *case*,
ProcessingLevel : *micro*,
Components : CDIR15,
StatisticalPopulation : &PCO-4,
SamplingPopulation : &PCO-6,
GeneratedBy : &WT-1

}

Figure 16: Output SCo Instance

4 Context Linkage

By design, every SCo is embedded into a “surrounding” meta-information structure depending on the role and purpose of the respective SCo. As outlined in Subsection 2.2, this context is composed of (typed) structural elements interwoven in a referential net providing the denotations encoding (only part of) the semantics of embedded SCo components. Thus, for example, prior to any data processing the surrounding meta-information structure of any SCo is a ‘source context’ defining, among other things, the statistical population the dataset of the SCo relates to, and the meaning of the variables contained in the dataset (cf. Tab. 2). A source context, in turn, will be enclosed in a still broader information structure, the 'domain context' providing descriptions of a whole subject matter area (such as, say, employment or education statistics) which itself, again, may be embedded in a yet more general 'institutional context' comprising definitions or prescriptions of fundamental discourse elements such as standardised nomenclatures, statistical unit delineations, registries, and so on.

Structurally, a SCo is but a sub-net deeply entangled into its context, raising the question of what to include in SCos and what to place outside in the surrounding context for reference by SCos. Generally speaking, SCo "interior" is determined pragmatically, that is, by computational considerations and requirements: whenever structural relationships of composition or association are – typically – affected by transformations within an otherwise unchanged surrounding network structure, these sets of "related bits and pieces" undergoing joint transformational reshaping legitimise SCo entities. Analogously, of course, the very same argument of structural organisation applies to the recursive formation of sub-contexts nested within (super-) contexts, as contexts by themselves are legitimate operands of (context) processing.

For the sake of SCo processing, the model assumes that one or more (input) SCos are first moved from source contexts to a 'transformation context' maintaining a structural description of all entities relevant in the conduct of a specific data transformation. With respect to the weighting example of Section 3.2, Fig. 17 sketches how input SCo 'SCO-2' and output SCo 'SCO-3' might be formally related in such a transformation context.

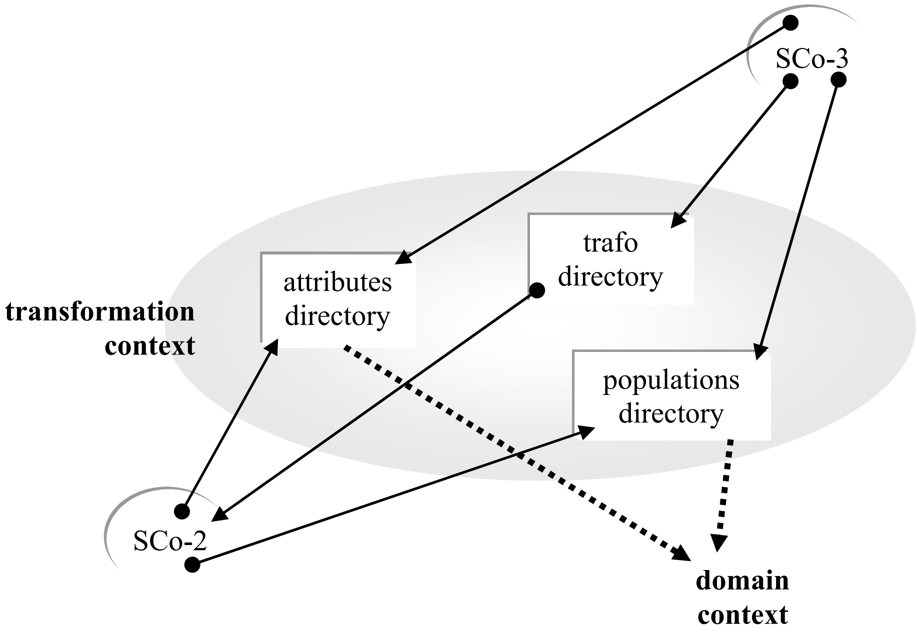

Figure 17: Weighting Transformation Context (Excerpt)

Basically, in Fig. 17, input and output SCo are linked by references to (i) the same attributes (‘&ATT-*x*’) and populations (‘&PCO-*x*’), recorded in the respective directories within the transformation context, and (ii) another directory gathering the actual transformation(s) applied, such as object ‘&WT-1’ (not shown in Fig. 17) establishing the functional relationship between ‘&SCO-2’ and ‘&SCO-3’ by memorising the applied weighting operation and its parameters. Making extensive use of object referencing in structure composition, by the way, helps to save space in that most of the time only references are copied (“shallow copies”). So, while ‘&SCO-3’ still contains the components of ‘&SCO-2’, virtually no sub-structures are duplicated.

Like within SCos, the directories in a transformation context accomplish the local identification of sub-structure components and, thus, need to be linked to “non-local” meta-information structures for external reference. For brevity, assume that all these links point directly to a ‘domain context’ although, in practice, contexts might be nested recursively several times, yet always following the same linkage principles. This holds particularly for the creation of federations of (otherwise autonomous) statistical databases implying, in fact, the arrangement of yet another shared context linking the sub-structures of the participating databases according to semantic co-incidences (Denk and Froeschl, 2000).

From an application point of view, the definitional and terminological entirety describing the content of statistical information systems as large as national statistical institutes, or even supra-national statistical agencies or federations, constitute the essential ‘institutional context’, at least to the extent it ever becomes explicated through formalised entities and relationships between those. Correspondingly, meta-computing depends crucially on a far-going resolution of context semantics into higher-order data structures amenable to algorithmic transformation. While there are many initiatives and projects (both in academia and statistical offices) driving the formalisation of statistical context, proposed meta-information models so far are converging rather slowly in specific areas only, such as classification systems, data element documentation, or multidimensional (table) data descriptions (cf. *http://www.epros.ed.ac.uk/metanet/* for an overview of recent activities in this respect).

5 Summary and Outlook

This paper has tried to point out possible contributions of meta-computing to the integration of statistical data and metadata. First and foremost, *integrated* statistical information systems can be seen as repositories purposively interlacing means of access through description with means of access through self-description: capturing a good deal of statistical semantics of data collections in terms of formal relations provides a convenient way to explore and analyse data bodies on condition that both descriptive and self-descriptive structures are co-transformed with what is so described (that is, the statistics data proper). Amongst the major benefits of this approach might be stated:

- the *reduction of documentation effort* because, once the basic descriptive structures are set up, documentation is maintained (semi-) automatically whenever content is updated, transformed, or distributed;

- *documentation integrity* – in terms of both completeness and consistency – is assured due to the algebraic properties of the operators co-applied to both data and metadata;
- formal documentation is amenable to algorithmic processing, enabling more powerful modes of information access and transformations easier to describe and accomplish.

As to the latter, a particularly interesting application of meta-computing principles is goal-driven information processing (Froeschl, 1997): given that the content of a statistical database is "marked-up" using context structures as outlined, the formal metadata specification of a retrieval target (say, a statistical table) can be used as (4GL) query statement. Meta-computing commences then **(i)** to investigate whether the requested aggregate is derivable from the database, and if so, **(ii)** to generate a transformation plan for turning candidate source data into the target structure (or something quite similar to it). In other words, meta-computing facilitates an inferential calculus for the derivation of output structures from a database using purely formal output *descriptions*.

At present, the sketched statistical meta-computing is barely more than a theoretical proposal, although some partial pilot implementations have been accomplished. While the fundamental issues of meta-computing appear more or less clarified, many practical and technical underpinnings are still lacking. More research is needed mainly with respect to canonical computable meta-information structures for statistical documentation, and the development of efficient implementation models.

While still in its infancy, statistical meta-computing clearly addresses process optimisation and workflow improvement (as to both cost and speed) of statistical information processing, particularly focussing on the requirements of online databases providing "table-on-demand" services: responding adequately to a practically non-finite set of custom-tailored information requests necessitates advanced modes of metadata management built on meta-computing principles as proposed.

References

S. Abiteboul, P. Buneman, D. Suciu. *Data on the Web: From Relations to Semi-structured Data and XML*. Morgan Kaufmann, San Francisco et al., 2000.

J.G. Bethlehem, J.-P. Kent, Ad Willeboordse, W. Ypma. On the Use of Metadata in Statistical Data Processing. UN/ECE Work Session on Statistical Metadata Report, Working Paper No.23, Geneva, September 22-24, 1999, 11pp., 1999.

G.E.P. Box. Science and Statistics. JASA 71: 791-799 (Applications Section), 1976.

P. Darius, M. Boucneau, P. de Greef, E. de Feber, K. Froeschl. Modelling Metadata. *Statistical Journal of the United Nations Economic Commission for Europe* 10(2): 171–180, 1993.

M. Denk. *Metadata Driven Production of Statistical Aggregates*. Diploma Thesis, Dept. of Statistics and Decision Support Systems, University of Vienna, 1999.

M. Denk. *Statistical Data Combination: A Metadata Framework for Record Linkage Procedures*. Doctoral Thesis, Dept. 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, K.A. Froeschl, W. Grossmann. Statistical Composites: A Transformation-bound Representation of Statistical Datasets. In J. Kennedy (editor). Proc. 14th Int. Conf. Scientific and Statistical Database Management (Edinburgh, UK), pages 217-226. IEEE Computer Society Press, Los Alamitos, Ca., 2002.

U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (editors). *Advances in Knowledge Discovery and Data Mining*. AAAI Press / The MIT Press, Menlo Park, Ca. et al., 1996.

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

K.A. Froeschl, T. Yamada, R. Kudrna. Industrial Statistics Revisited: From Footnotes to Meta-Information Management. *Austrian Journal of Statistics* 31(1): 9–34, 2002.

K.A. Froeschl, W. Grossmann, V. Del Vecchio The *Concept of Statistical Metadata*. Project Deliverable D5, MetaNet (IST-1999-29093) Workgroup 2 (Harmonization of Metadata: Structure and Definitions), vi+134 pages, 2003.

W. Grossmann and K.A. Froeschl. *Automated Table Generation Through Metadata* (in German). Project Report, Dept. of Statistics, Univ. of Vienna, 160 pages, 1994.

U. Haag. Knowledge-Based Systems in Statistics: A Tutorial Overview with Examples. In P. Dirschedl, R. Ostermann (editors). *Computational Statistics*, pages 211-236. Physica (Springer), Heidelberg, 1994.

J. Han and M. Kamber. *Data Mining—Concepts and Techniques*. Morgan Kaufmann, San Francisco et al., 2001.

F. Inglese and F. Oropallo. The Development of an Integrated and Systematized Information System for Economic and Policy Impact Analysis. *Austrian Journal of Statistics*, **this issue**.

A.W. Kimball. Errors of the Third Kind in Statistical Consulting. *J. Amer. Statistical Association* 52: 133-142, 1957.

P. Ofner. *Embedding of Weighting Algorithms into Metadata Structures*. Dissertation Thesis, Dept. of Statistics and Decision Support Systems, University of Vienna, 2001.

J. Ryssevik. Metadata for Traveling Statistics—The World of Statistics Meets the Semantic Web. Invited Talk at the 14th Int. Conf. On Scientific and Statistical Database Management (Edinburgh, UK), 2002.

F.J. Scheuren. Macro and Micro Paradata for Survey Assessment. Paper presented in a satellite meeting to the UN/ECE Work Session on Statistical Metadata (Washington D.C., November 2000), U.S. Bureau of Labor Statistics. 16 pages, 2000.

M. Silver. The Role of Footnotes in a Statistical Metainformation System. *Statistical Journal of the United Nations Economic Commission for Europe* 10(2): 153-170, 1993.

B. Streitberg. On the Non-Existence of Expert Systems—Critical Remarks on Artificial Intelligence in Statistics. *Stat. Software Newsletter* 14(2): 55-62, 1988.

B. Sundgren. *An Infological Approach to Data Bases*. Report, Statistics Sweden, 1973.

J.W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Ma., 1977.

W.E. Winkler. Matching and Record Linkage. In Cox B.G. (editor). *Business Survey Methods*, pages 355-384. Wiley, New York, 1995.

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

Author's address:

Dr. Karl A. Froeschl
Electronic Commerce Competence Center (ec3)
Donau City-Straße 1
A-1220 Vienna DC,
Austria

Tel. +43 1 522 71 / 71 0
Fax +43 1 522 71 / 71 71
Elec. Mail: Karl.Froeschl@ec3.at
http//www.ec3.at